UNIVERSITE DE REIMS
CHAMPAGNE-ARDENNE

Application de techniques de séparation de sources à la spectroscopie Raman et à la spectroscopie de fluorescence

THÈSE

présentée et soutenue publiquement le 27 mars 2006

pour l'obtention du titre de

Docteur de l'Université de Reims Champagne-Ardenne

Spécialité Génie Informatique, Automatique et Traitement du Signal

par

Cyril GOBINET

Composition du jury

Rapporteurs :	Monsieur Yannick DEVILLE
	Monsieur Jérôme MARS
Examinateurs :	Monsieur Didier COQUIN
	Monsieur Christophe HADJUR
	Madame Danielle NUZILLARD
Directeur de thèse :	Monsieur Michel MANFAIT
Co-directeur de thèse :	Monsieur Régis HUEZ

Mis en page avec la classe thloria.

Remerciements

Je tiens en premier lieu à remercier Monsieur Michel Manfait de m'avoir choisi comme doctorant sur un sujet en interface entre la spectroscopie moléculaire et le traitement du signal, pour m'avoir fait confiance tout au long de cette thèse, pour m'avoir permis de présenter mon travail dans des congrès internationaux. Mais ma reconnaissance va bien plus loin puisque sans son influence et sa générosité, mon travail de thèse n'aurait pas obtenu les qualificatifs qui lui ont été attribués lors de la soutenance, et je n'aurais pas eu l'honneur de pouvoir poursuivre mon travail par une collaboration avec L'Oréal. Je le remercie aussi d'avoir pensé à moi pour un contrat de post-doctorat et de m'accueillir dans son unité MéDIAN dans le cadre de ce post-doctorat. Je lui présente donc ici l'expression la plus sincère de ma reconnaissance.

Je suis également reconnaissant à Monsieur Janan Zaytoon de m'avoir accueilli au sein du CReSTIC.

Je remercie également Monsieur Régis Huez d'avoir cru en mes capacités scientifiques pour achever ce travail de thèse, pour m'avoir guidé tout au long de ces trois ans et demi de labeur, de m'avoir soutenu lors d'un gros passage à vide, et de m'avoir permis de partir être l'ambassadeur de notre équipe lors de congrès internationaux.

Je remercie Madame Danielle Nuzillard d'avoir présidé ce jury et de m'avoir accueilli au sein de l'équipe de traitement du signal du CReSTIC.

Je tiens à exprimer l'honneur que m'ont fait Messieurs Jérôme Mars et Yannick Deville en acceptant d'être rapporteurs de mon travail de thèse. Leurs questions, suggestions, corrections et commentaires ont apporté des améliorations indéniables et précieuses au mémoire, et ont ouvert d'autres perspectives à ce travail. De plus, leur renommée scientifique et l'intérêt qu'ils ont manifesté pour mon travail m'assurent de sa qualité. Leurs commentaires enthousiastes m'ont réconforté et m'ont permis d'aborder la date fatidique de la soutenance avec plus de sérénité et d'oublier certains commentaires et actions déstabilisants de tierces personnes.

Je remercie évidemment Messieurs Didier Coquin et Christophe Hadjur d'avoir examiné ce travail de thèse. Leurs questions pertinentes et constructives m'ont permis d'avoir une autre vision axée sur le traitement d'image de mon travail et de constater qu'un large champ d'exploration reste ouvert dans l'application des techniques du traitement du signal à la spectroscopie vibrationnelle. Je suis reconnaissant à Monsieur Hadjur de la confiance qu'il m'accorde actuellement dans le cadre d'une collaboration avec L'Oréal et de l'intérêt qu'il exprime pour mon travail. J'espère que cette collaboration naissante va grandir durant de nombreuses années.

Je suis très redevable à Valeriu Vrabie qui a su donner un coup de fouet à mon travail depuis son arrivée au CReSTIC il y a maintenant un an et demi. Sa vivacité d'esprit, ses connaissances et sa rigueur scientifiques ont permis de hausser considérablement le niveau scientifique de ma thèse. Merci d'avoir

su être autoritaire tout en restant diplomate et de me tenir tête lorsque je faisais un caprice par peur d'aller présenter mon travail lors d'une session orale d'un congrès international. Merci aussi de toujours m'avoir remonté le moral lors de mes périodes de doutes. Et un grand merci pour m'avoir aidé, en priorité psychologiquement, pendant l'heure ayant précédé la soutenance de thèse.

Je remercie aussi Olivier Piot, Ali Tfayli, Ganesh Sockalingum et Isabelle Adt pour leur aide précieuse sur l'interprétation biophysique de mes résultats, interprétation sans laquelle ce travail n'aurait aucune valeur. Et merci également de m'avoir intégré à votre équipe de l'unité MéDIAN dans le cadre d'un post-doctorat, et merci de votre sympathie et bonne humeur face aux questions d'un novice en biophysique.

Je suis reconnaissant à Fabien Belloir d'avoir essayé de m'empêcher de faire une boulette pendant une période creuse de ma thèse et de m'avoir aider lors de mes répétitions, à Éric Perrin pour m'avoir orienté vers l'utilisation des méthodes de traitement du signal les plus pertinentes et adaptées à mon problème pendant mon DEA et ma thèse, et pour sa bonne humeur et ses blagues salaces. Merci également à Nicolas Pétrochilos, fournisseur officiel du journal *Docteurs & Co*, pour avoir effacé certains de mes doutes lors d'une soirée de rédaction au LAM.

Je dis un grand merci à un sous-groupe clandestin du CReSTIC, les LAMpins, composé de Dieu, de Doc, de Jeff et de Valoche, sans qui jamais je n'aurais su faire la vaisselle, écrire des cartes postales, jouer dans un clip musical, passer un jour de l'an inoubliable, avoir des fêtes surprises d'anniversaire arrosées de cadeaux très utiles ou encore passer de longues soirées de discussions existentielles au labo, en particulier avec Valoche. Pour toutes ces heures de bonne humeur, merci.

Je remercie aussi tous les membres (Damien, François et Benoît entre autres) du Cercle des Joueurs Amateurs de Tarot du LAM (le CJATLAM) pour avoir tenté en vain de faire de moi un joueur confirmé, concentré et réfléchi. Désolé d'avoir failli à ce niveau.

Merci aussi à tous les autres membres du LAM qui m'ont supporté pendant ces années de galère, en particulier Pascale pour les bons petits gâteaux qu'elle prépare régulièrement pour les doctorants du LAM, Kévin pour le resto offert gracieusement à Prague, Linda pour son "bonjour" souriant du matin, Lanto pour m'avoir supporté dans son bureau pendant aussi longtemps, Stéphane pour ses mails bourrés de blagues et pour avoir donné un grand coup de main à ma famille lors du pot de thèse, et Sylvie pour les pauses détente à n'importe quel moment de la journée.
Je suis reconnaissant aux membres du département EEA (David et Olivier en particulier) pour m'avoir confié des enseignements pendant trois ans et pour avoir cru en mon sérieux, contrairement aux remarques infondées et hors propos de certaines personnes.

J'ai gardé pour la fin les remerciements les plus importants à mon cœur.
Merci, merci, merci et encore merci Anne de m'avoir soutenu chaque jour, d'avoir facilité mon quotidien pendant la longue période de rédaction et de préparation de la soutenance en assurant toutes les tâches liées à notre vie de couple, merci de m'avoir montré et exprimé toute ta confiance en mes capacités, merci

d'avoir supporté de ne nous croiser que le matin et le soir pendant trois longs mois, merci de m'avoir aimé chaque jour plus fort.
Je crie un grand merci à ma Maman pour s'être sacrifiée pour que je puisse toujours étudier dans des conditions optimales, entouré de sa présence et de son amour à n'importe quelle heure du jour ou de la nuit, pour ses encouragements quotidiennement répétés, pour ses Tupperwares toujours remplis de bonne choses et prêts à être réchauffés au micro-onde, pour son amour de toujours. T'as vu Maman, on a réussi à l'avoir ce doctorat ! ! ! Et ne t'inquiète pas, je n'oublierai pas de couper le diplôme en morceaux pour te donner la part qui te revient de droit.
Merci beaucoup beaucoup beaucoup à mon joyeux frère Johan pour sa bonne humeur quotidienne sauf les jours où il est de mauvais poil, pour ses coups de gueule réguliers mais si marrants, pour ces bonnes années passées tous les deux à Reims, pour les vacances passées ensemble, notamment en Italie, et pour tous ses cadeaux si personnels. Et pis mon frère, ce que pensent les autres on s'en fout.
Je remercie évidemment tout le reste de ma famille, ma belle famille et mes amis qui ont contribué de près ou de loin à l'achèvement de mon doctorat.

À tous merci.

À Anne, ma mère et mon frère

Table des matières

Table des figures

Liste des tableaux

Introduction

Depuis un demi siècle, la recherche en biologie connaît un essor considérable grâce aux développements d'outils d'analyse toujours plus performants pour percer les secrets du vivant. Des techniques d'imagerie médicale telles que l'échographie, l'Imagerie par Résonance Magnétique fonctionnelle (IRMf), la Tomographie par Émission de Positons (TEP), sont utilisées par la médecine pour le diagnostic et le traitement de nombreuses pathologies. Des études physico-chimiques complexes forment les standards pour le contrôle de la qualité des matériaux agroalimentaires.

Depuis 20 ans, les techniques de spectroscopie optique se sont rendues indispensables dans des domaines aussi variés que la microélectronique, la géologie, l'environnement et le contrôle des matériaux. Ce succès provient de la puissance de ces méthodes à analyser la structure et la composition moléculaire des échantillons étudiés. Le biomédical a su tirer avantage des spectroscopies optiques pour faciliter le diagnostic de nombreuses pathologies humaines. Des appareillages spécifiques sont développés depuis quelques années pour assurer l'analyse de tissus *in vivo*. Le bien-être du patient et la rapidité de l'analyse en sont des avantages indéniables. De même, l'industrie agroalimentaire s'appuie sur ces technologies d'investigation simple et rapide de la constitution chimique de matériaux pour le contrôle de leur qualité.

La complexité d'analyse d'un spectre acquis grâce aux méthodes de spectroscopies optiques est proportionnelle à la complexité chimique de l'échantillon analysé. Pour compenser les limites cognitives de l'homme, la chimiométrie s'est développée et a offert quelques algorithmes de séparation des spectres optiques. Souvent développés pour une application particulière, ils n'exploitent pas toute l'information contenue dans les spectres. Aucune méthode chimiométrique de séparation des spectres ne s'est réellement imposée de manière générale.

Le traitement du signal a pour vocation de développer des méthodes et des algorithmes d'extraction des informations utiles pour répondre à une problématique donnée. Les réseaux de neurones, dont les cartes de Kohonen, l'analyse temps-fréquence et temps-échelle, et l'Analyse en Composantes Principales sont quelques exemples de la richesse et la puissance du traitement du signal. La séparation de sources s'est imposée depuis 20 ans dans de nombreux domaines de la recherche comme un outil performant de traitement du signal. Cette technique a pour but de réorganiser sous forme de sources simples les informations mélangées de manière complexe au sein de signaux enregistrés par des capteurs.

Dans ce contexte, l'objectif principal de ce travail est d'étudier les possibilités d'application des techniques de séparation de sources aux spectres issus des spectroscopies de fluorescence et Raman. L'étude générale des propriétés physiques des mécanismes régissant ces spectroscopies et des propriétés statistiques des spectres enregistrés par ces méthodes va permettre de désigner la Factorisation en Matrices Non-négatives et l'Analyse en Composantes Indépendantes comme des méthodes de séparation de sources aptes à séparer les composantes chimiques de tout type d'échantillon biologique.

Ce mémoire est organisé de la manière suivante :

Le premier chapitre est consacré à l'introduction des spectroscopies optiques. Les principes physiques sont exposés pour faciliter la compréhension des propriétés physiques de ces méthodes. L'instrumentation qui assure la mesure de ces phénomènes physiques est expliquée au travers de la description de la chaîne d'acquisition. La notion de spectre, objet numérique disponible à la sortie de la chaîne, est introduite grâce à des exemples. La polyvalence des techniques de spectroscopies optiques est suggérée par la description de nombreuses applications dans de nombreux domaines scientifiques différents.

Dans le deuxième chapitre, les caractéristiques et les propriétés fondamentales communes aux spectroscopies Raman et de fluorescence sont analysées. Les dissimilitudes de leurs propriétés statistiques sont exposées. Elles conduisent au développement de méthodes de séparation des spectres totalement différentes. L'étude des propriétés des spectres de fluorescence et des méthodes dédiées à leur analyse numérique est alors menée indépendamment de celle des spectres Raman. L'absence de traitements adaptés, généraux et efficaces de ces spectres ressort de cette étude.

Le troisième chapitre se propose de montrer que la Factorisation en Matrices Non-négatives est une théorie candidate pour la résolution du problème général de séparation des spectres de fluorescence. Les fondements de cette théorie seront expliqués. Ses hypothèses de base sont en parfait accord avec les propriétés des spectres et des mécanismes de la spectroscopie de fluorescence. Deux applications sur des grains de blé et d'orge seront développées et illustreront le caractère général de la Factorisation en Matrices Non-négatives à extraire les spectres des espèces chimiques pures de l'échantillon analysé.

Le quatrième chapitre démontre le potentiel de l'Analyse en Composantes Indépendantes à traiter les spectres Raman. Dans un premier temps, le problème général de la séparation de sources est présenté. L'Analyse en Composantes Indépendantes est décrite comme une méthode possible de résolution de la séparation de sources. Sa définition, ses hypothèses de travail, ses indéterminations, ses prétraitements nécessaires à la simplification de la résolution, ses mesures d'indépendance et ses algorithmes y seront décrits. Son efficacité lui a valu une grande popularité qui se traduit par des applications nombreuses et diverses. Une nouvelle application sera proposé en spectroscopie Raman dont les spectres conviennent parfaitement à la modélisation imposée par l'Analyse en Composantes Indépendantes. Une nouvelle technique de déparaffinage des échantillons biologiques sera proposée sur l'association de la spectroscopie Raman et de l'Analyse en Composantes Indépendantes pour éliminer numériquement la paraffine de tissus biologiques. Cette méthodologie sera appliquée avec succès sur des échantillons d'épiderme de peau.

Une nouvelle modélisation du spectre de la paraffine sera également proposée sur l'analyse de ces résultats. La puissance de la méthode sera prouvée par la discrimination entre un mélanome et un naevus, deux tumeurs respectivement maligne et bénigne de la peau, à partir de spectres Raman enregistrés sur des échantillons d'épidermes paraffinés.

Chapitre 1

Les techniques de spectroscopie optique

1.1 Introduction

La curiosité pousse l'homme à chercher à comprendre comment fonctionne le monde qui l'entoure. Au fil des siècles, sa démarche timide et hasardeuse l'a conduit à avoir une vision grossière mais indispensable des phénomènes physiques et biologiques visibles à l'œil nu. L'accumulation de son expérience et la démocratisation de la science ont entraîné le développement d'une démarche scientifique rigoureuse pour l'analyse et la résolution des problèmes posés aux chercheurs. Cependant, au fur et à mesure de l'accroissement des connaissances de l'homme, le besoin s'est fait sentir de ne plus se cantonner aux phénomènes visibles à l'œil nu mais de s'engager dans la compréhension de phénomènes liés à l'infiniment petit. L'étude structurelle des matériaux s'est longtemps faite grâce à la chimie et à la microscopie. Afin d'améliorer l'analyse des matériaux inertes et vivants, les techniques d'analyse doivent être les plus informatives, rapides et non-traumatiques.

L'analyse et la synthèse de la lumière blanche par Isaac Newton, l'introduction de la notion du quantum par Max Planck dans l'interprétation du spectre d'émission du corps noir incandescent suivie de sa confirmation par Albert Einstein et enfin l'étude du spectre de l'atome d'hydrogène par Niels Bohr ont permis d'expliquer les innombrables phénomènes qui se manifestent dans les interactions entre atomes ou molécules et ondes électromagnétiques. Couplées à ces découvertes en physique théorique, les avancés technologiques dans les sources laser et les dispositifs de détection ont mené au développement de la spectrométrie optique. Les techniques de spectroscopie optique infrarouge, de fluorescence et Raman sont des méthodes d'analyse de la composition chimique d'un échantillon par mesure des phénomènes d'interactions entre onde et matière. Ces méthodes ont initialement servi à élucider la structure moléculaire de nombreux composés issus de la chimie organique. Les révolutions techniques qu'ont constitué le couplage spectromètre-micro-ordinateur, l'interférométrie, l'amélioration des optiques et des détecteurs ont engendré des applications des spectroscopies optiques à l'étude de molécules de plus en plus complexes telles que

des macromolécules d'origine biologique. Un nouveau pas a été franchi avec l'étude des cellules et tissus entiers, échantillons d'une complexité encore plus élevée. La formidable quantité d'informations contenues dans un spectre optique fait de la spectrométrie optique une technique aux applications nombreuses.

Ce chapitre est une brève introduction aux techniques de spectroscopie optique. Les trois grands types d'interaction onde-matière : l'absorption, la diffusion et l'émission, qui sont respectivement à la base des spectroscopies optiques de type infra-rouge, Raman et de fluorescence, seront décrits dans la section 1.2. Les rayonnements résultants de ces interactions sont mesurés par une chaîne d'acquisition constituée d'une source laser, d'une optique de collection, d'un système d'analyse spectrale et d'un détecteur de rayonnement. L'étude de cette chaîne, quasiment identique pour ces trois types de spectroscopies, fera l'objet de la section 1.3. L'information moléculaire de l'échantillon analysé est obtenue, à la sortie de cette chaîne, sous forme d'un spectre. Cette notion de spectre, différente selon la spectroscopie considérée, sera étudiée à la section 1.4. Sa définition, ses propriétés générales, quelques exemples et les différents types de mesures spectrales, qui conditionnent la représentation matricielle des spectres enregistrés, y seront décrits. Les spectroscopies optiques, de par leur instrumentation et leurs principes physiques, possèdent des propriétés qui leur sont propres et qui leur assurent des applications nombreuses dans de nombreux domaines hétérogènes. Quelques applications sont succinctement décrites à la section 1.5 afin d'illustrer la puissance d'analyse des spectroscopies optiques.

1.2 Principes physiques des spectroscopies optiques

Un rayonnement électromagnétique désigne le transfert d'énergie entre deux sources sous la forme d'ondes électromagnétiques. Lorsqu'une onde électromagnétique percute de la matière, de l'énergie peut être fournie aux atomes ou molécules de cette matière. Ce changement d'énergie peut se manifester sous forme d'absorption, de diffusion ou d'émission d'ondes électromagnétiques. L'analyse de ces différents types d'interactions entre des ondes et de la matière a donné naissance à différents types de spectroscopies optiques que nous allons étudier dans cette section.

1.2.1 Onde et matière

Les spectroscopies optiques sont basées sur l'interaction de la lumière (*onde*) avec le nuage électronique des liaisons chimiques (*matière*).

1.2.1.1 Onde électromagnétique

La lumière désigne les ondes électromagnétiques visibles par l'œil humain. Une onde électromagnétique est une variation périodique des champs électrique et magnétique associés à un flux continu de particules

appelées *photons*. L'énergie E d'un photon de fréquence ν est donnée par la relation :

$$E = h\nu = h\frac{c}{\lambda}$$

où h est la constante de Planck ($h = 6.626 \times 10^{-34}\ Js$), c est la vitesse de la lumière dans le vide ($c = 3 \times 10^{8}\ ms^{-1}$) et λ est la longueur d'onde en mètre.

Le spectre électromagnétique s'étend des ondes radio aux rayons cosmiques. En fonction de la fréquence de l'onde qui va interagir avec un échantillon, différentes formes de transitions vont être plus ou moins privilégiées [104]. La figure 1.1 présente les transitions provoquées par les différentes grandes catégories d'ondes électromagnétiques.

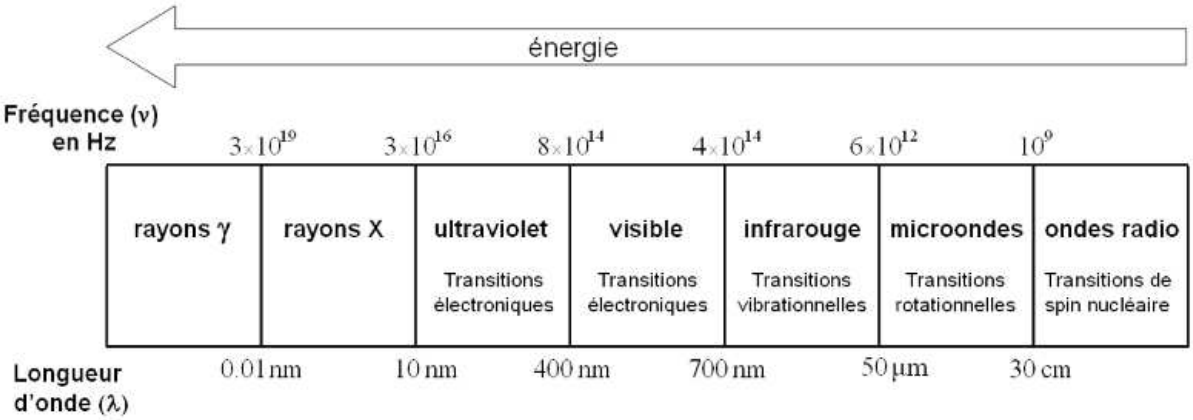


FIG. 1.1 – Les divers domaines spectraux du rayonnement électromagnétique

1.2.1.2 Matière

Une molécule peut, schématiquement, être considérée comme formée d'atomes dont les électrons assurent la liaison chimique ou liaison covalente. L'approximation de Born-Oppenheimer permet de découpler le mouvement des noyaux de celui des électrons, bien plus légers, et donc à découpler leurs énergies respectives [28, 7]. Ainsi, en première approximation, l'énergie E_T d'une molécule peut s'écrire comme la somme de trois énergies :

$$E_T = E_e + E_v + E_r \tag{1.1}$$

avec $E_e \gg E_v \gg E_r$. Le terme électronique E_e est dû à l'énergie des électrons (cette énergie dépend du nombre d'électrons et de la forme de la molécule). Le terme E_v est dû à l'énergie vibrationnelle des noyaux (cette énergie dépend de la masse des atomes et de leur arrangement), c'est-à-dire au déplacement des noyaux les uns par rapport aux autres. Le terme rotationnel E_r est dû à la rotation de la molécule autour de directions privilégiée.

Les énergies E_e, E_v et E_r sont quantifiées. Les transitions ont donc lieu entre des valeurs discrètes des grandeurs E_e, E_v et E_r. Nous sommes ainsi amenés à considérer les phénomènes radiatifs d'absorption

et d'émission comme résultant de transitions entre les différents niveaux d'énergie E_T de la molécule.

Une description plus détaillée peut faire apparaître des couplages entre ces diverses formes d'énergie.

Énergies électroniques : La spectroscopie de fluorescence dépend des énergies électroniques d'une molécule [91]. Un bref rappel en est donné ci-dessous.

L'énergie E_T d'une molécule est quantifiée et dépend de nombres entiers qui sont nommés nombres quantiques. Pour des atomes, cette énergie est quantifiable par une formule dépendante du nombre quantique. Pour des molécules, cette énergie n'est pas quantifiable directement par une équation et dépend de la géométrie nucléaire de la molécule. Le niveau de plus faible énergie d'une molécule correspond à l'état le plus stable de la molécule. Un apport extérieur d'énergie sous forme de photon amène la molécule dans un état excité plus instable. La molécule va chercher à retrouver sa stabilité en restituant de l'énergie. Cette perte d'énergie s'accompagne de la relaxation de la molécule qui retrouve la stabilité de son état fondamental. Les niveaux électroniques d'une molécule étant quantifiés, l'énergie apportée ou cédée par la molécule est elle même quantifiée. Un apport d'énergie égal à l'énergie nécessaire à la transition de la molécule de son état fondamental à un état excité est autorisé et mènera la molécule à son état excité.

Énergies vibrationnelles : Les spectroscopies infrarouge et Raman mesurent les énergies vibrationnelles des molécules [28, 7]. Une brève introduction aux modes de vibrations est donc nécessaire.

▷ ***Molécules diatomiques :*** Dans le cas d'une molécule diatomique, un seul type de vibration est possible. En supposant l'approximation du modèle du vibrateur harmonique vérifiée, l'énergie des transitions entre les niveaux de vibrations peut prendre les valeurs :

$$E_v = \frac{h}{2\pi}\sqrt{\frac{k}{\mu}}(v + \frac{1}{2})$$

où h est la constance de Planck définie à la page 7, v est le nombre quantique de vibration, k est la constante de raideur de la liaison[1] et μ est la masse réduite définie par :

$$\mu = \frac{m_1 m_2}{m_1 + m_2}$$

avec m_1 et m_2 les masses respectives des deux atomes composant la liaison. Les règles de sélection quantiques imposent que la variation du nombre quantique v appartienne à l'ensemble $\{-1, 0, 1\}$. Une molécule diatomique possède ainsi trois modes de vibration.

▷ ***Molécules polyatomiques :*** Dans le cas de molécules polyatomiques, le nombre de liaisons augmente et la géométrie des liaisons se complexifie. Si nous considérons n atomes isolés, chacun possède 3 degrés de liberté de translation, 3 degrés de liberté de vibration et 3 degrés de liberté de rotation

[1] typiquement, pour une liaison simple $k \approx 500\ N.m^{-1}$, pour une liaison double $k \approx 1000\ N.m^{-1}$, et pour une liaison triple $k \approx 1500\ N.m^{-1}$

dans un espace à 3 dimensions. Lorsque ces atomes sont reliés entre eux par des liaisons d'angles et de longueurs variables, la molécule possède $3n$ degrés de liberté dont 6 peuvent être attribués à la translation et à la rotation de la molécule entière. Il reste donc $3n-6$ degrés de liberté pour les modes de vibration des liaisons. Ce nombre est ramené à $3n-5$ pour les molécules linéaires du fait de la rotation selon l'axe de la molécule.
Ces modes de vibration donnent naissance aux 3 principaux types de bandes d'absorption :

- les bandes de valence : elles sont caractéristiques d'un des modes de vibration fondamentaux d'une liaison ;
- les bandes harmoniques : leur fréquence est un multiple entier de la fréquence d'une vibration fondamentale ;
- les bandes de combinaison : elles résultent de la combinaison (addition ou soustraction) des vibrations de plusieurs liaisons.

Il existe également les bandes nées de la résonance de Fermi qui sont une combinaison d'un mode fondamental et d'un mode harmonique.

▷ ***Groupes de vibrations :*** Les fréquences de vibration[2] dépendent des masses des atomes et des forces des liaisons de covalence. De ces considérations se dégagent deux classes de vibrations moléculaires :

- les vibrations de valence ou d'élongation (symétriques ou antisymétriques) qui font intervenir une ou des variations de longueurs de liaisons, les angles formant ces liaisons restant constants ;
- les modes de déformation pour lesquels, au contraire, les liaisons gardent leur longueur, mais les angles qu'elles forment varient.

Afin de mieux visualiser ces différents modes de vibrations, la figure 1.2 présente l'exemple du groupement CH_2 d'une molécule.

La considération des propriétés de symétrie laissant la structure de la molécule invariante permet de simplifier l'analyse de ses vibrations. Toute molécule appartient à l'un des 32 groupes ponctuels de symétrie [7]. La connaissance du groupe d'appartenance d'une molécule et des calculs simples assurent de prévoir les modes de vibrations de cette molécule.

La structure d'une molécule conditionne ses fréquences de vibrations qui se révèlent être une véritable signature de la molécule étudiée. L'interaction entre un rayonnement lumineux, de fréquence judicieusement choisie, et une molécule met en jeu des phénomènes radiatifs capables de sonder ces fréquences de vibrations. C'est ce que nous allons exposer dans la section suivante.

[2] Les nombres d'onde de vibrations sont donnés par l'équation de Hooke : $\bar{\nu} = \frac{1}{2\pi c}\sqrt{\frac{k}{\mu}}$

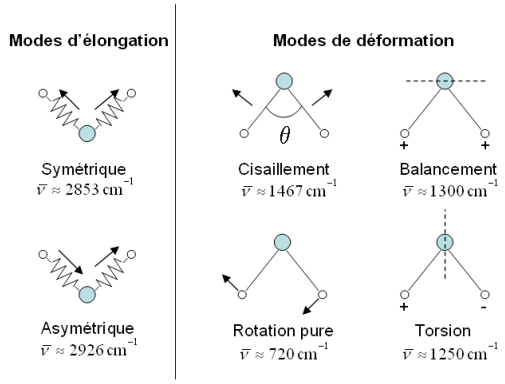


FIG. 1.2 – Exemple des vibrations localisées du groupement CH_2 d'une molécule

1.2.2 Phénomènes radiatifs et spectroscopies

Par action d'une radiation lumineuse, il est possible de faire passer une molécule d'un état d'énergie E_1 vers un état d'énergie supérieur E_2. Nous allons maintenant décrire les trois types de phénomènes radiatifs résultants de l'interaction entre rayonnement et matière sur lesquels sont basées les techniques de spectroscopie optique, à savoir l'absorption infrarouge, l'émission de fluorescence et la diffusion Raman. Le diagramme de Jablonski, présenté sur la figure 1.3, représente ces différents phénomènes en considérant les transitions électroniques induites par l'interaction de la lumière avec la matière.

1.2.2.1 Absorption et spectroscopie infrarouge

Il se produit un phénomène de résonance et d'absorption lorsque les mouvements des nuages électroniques ou les mouvements des liaisons entre atomes se font à la même fréquence[3] que l'onde électromagnétique incidente. Si ce mouvement modifie le moment dipolaire des liaisons de la molécule, une absorption infrarouge, caractérisant les états de vibration de la molécule, se produit.

Macroscopiquement, l'atténuation du rayonnement traversant un échantillon homogène d'épaisseur d s'exprime par :

$$\frac{I_t}{I_0} = exp(-Kd)$$

avec I_0 l'intensité du rayonnement incident, et I_t l'intensité du rayonnement transmis. Le coefficient K est le coefficient d'absorption de l'échantillon, il dépend de la longueur d'onde et il s'exprime en m^{-1}. La loi de Bouguer-Beer-Lambert définit l'absorbance $\mathcal{A}$ par :

$$\mathcal{A}(\lambda) = \log(\frac{I_0}{I_t}) = \varepsilon(\lambda)Cd$$

avec ε le coefficient d'extinction molaire s'exprimant en $mol^{-1}\ L\ cm^{-1}$ dont la valeur reflète la probabilité

[3] domaine s'étalant de l'ultraviolet au visible pour les mouvements des nuages électroniques, et domaine de l'infrarouge pour les mouvements des liaisons entre atomes

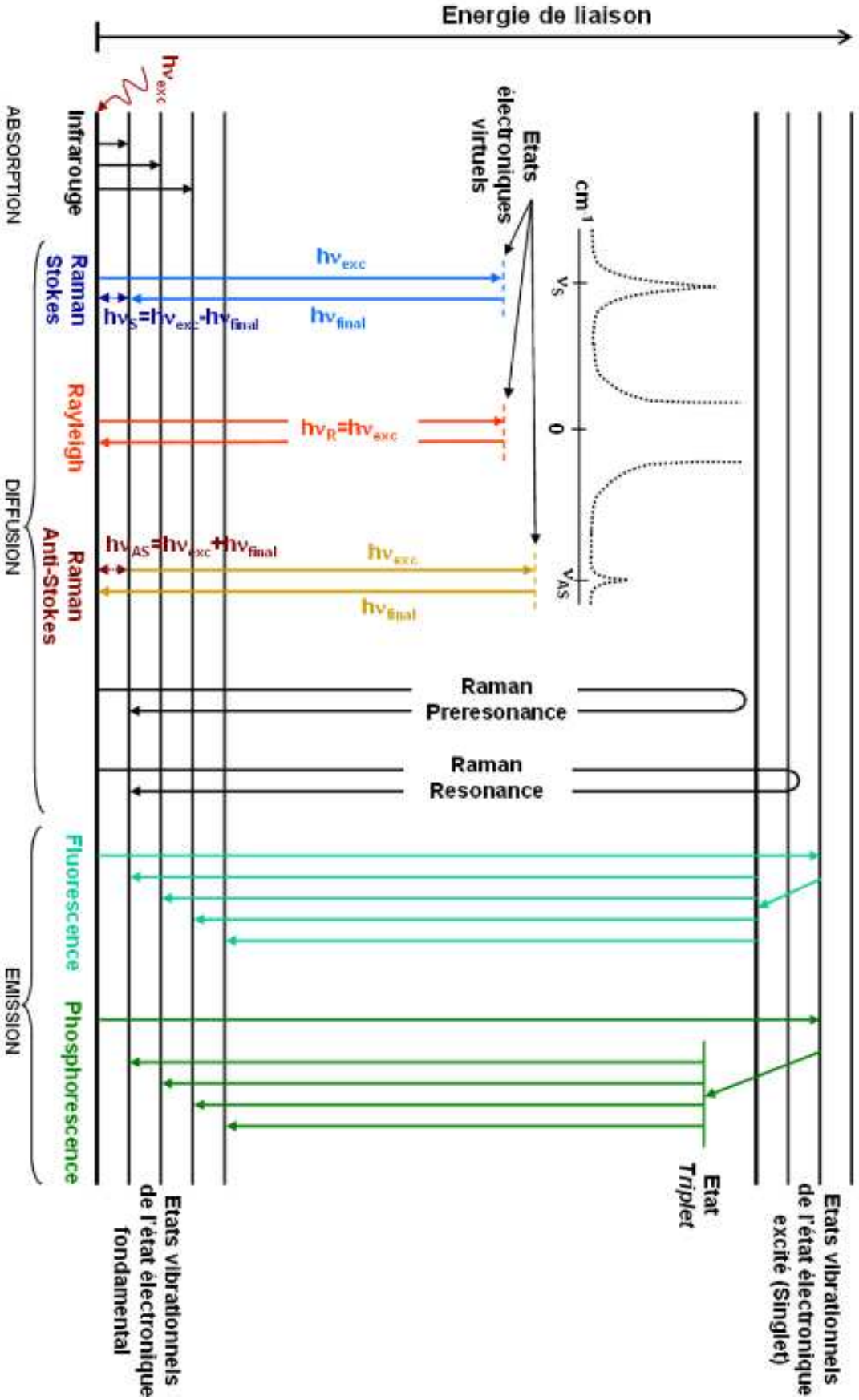


FIG. 1.3 – Diagramme de Jablonski

de la transition[4], et C la concentration en mol de la solution analysée. La proportionnalité apparente entre l'absorbance et la concentration permet d'utiliser les spectroscopies d'absorption comme méthode d'analyse quantitative, du moins dans la limite de linéarité de la loi de Bouguer-Beer-Lambert. Cette linéarité n'est vérifiée que pour des absorbances comprises entre 0 et 2.

Le spectre infrarouge présente l'absorbance $\mathcal{A}$ (ou la transmittance $\mathcal{T}$) en fonction du nombre d'onde $\bar{\nu}$ qui s'exprime en fonction de la longueur d'onde λ du rayonnement par l'équation :

$$\bar{\nu} = \frac{\nu}{c} = \frac{1}{\lambda} \tag{1.2}$$

avec ν la fréquence du rayonnement et c la vitesse de la lumière. Cette expression définie l'unité de $\bar{\nu}$ en m^{-1}. En spectroscopie, l'unité du nombre d'onde la plus fréquemment employée est le cm^{-1}. Le spectre infrarouge moyen s'étend de 400 à 4000 cm^{-1}.

1.2.2.2 Émission et spectroscopie de fluorescence

À l'inverse de l'absorption, un mouvement des nuages électroniques peut émettre une onde électromagnétique. Nous parlons alors d'émission de fluorescence qui correspond à des transitions entre les états électroniques de la molécule [128, chapitre 3] [1, chapitre 1]. La fluorescence prend naissance dans un phénomène radiatif en trois étapes.

La première est une phase d'absorption de l'onde lumineuse, cette fois avec transition électronique. Considérons une particule possédant deux niveaux d'énergie E_1 (configuration électronique la plus stable) et E_2 (structure électronique où un des électrons périphérique est porté dans un état excité). L'absorption est le phénomène qui porte la particule du niveau E_1 vers le niveau E_2 sous l'effet d'une onde lumineuse dont l'énergie $E = h\nu$ correspond exactement à la différence d'énergie entre les deux niveaux, c'est à dire que $E = h\nu = E_2 - E_1$. Le nombre de particules mises en jeu dans l'absorption (c'est-à-dire le nombre de particules qui vont être excitées et passer au niveau d'énergie E_2) dépend de la densité d'énergie du rayonnement incident et de la probabilité de la transition considérée.

La deuxième étape est une phase de relaxation partielle avec perte de chaleur par échange avec le milieu ambiant.

La troisième et dernière étape est une phase de désexcitation avec réémission d'une onde lumineuse d'énergie différente. Elle correspond au retour d'une particule se trouvant dans un état d'énergie élevée vers un niveau de plus basse énergie avec émission de photons par deux mécanismes distincts :

- *émission spontanée* : la désexcitation entre le niveau d'énergie E_2 et le niveau d'énergie E_1 s'accompagne de l'émission d'un photon d'énergie $E = h\nu = E_2 - E_1$ dont la polarisation et la direction de propagation sont aléatoires ;

[4] une forte valeur de ε caractérise une transition autorisée par les règles de sélection

- *émission induite* : la désexcitation entre le niveaux d'énergie E_2 et le niveau d'énergie E_1 a lieu sous l'effet d'un rayonnement incident d'énergie $E = h\nu = E_2 - E_1$. La polarisation et la direction de propagation du photon émis sont identiques à celle du photon incident.

La fluorescence prend donc naissance dans l'émission d'une onde radiative par un échantillon soumis à un rayonnement incident. Mais en fonction de la nature de l'échantillon, la fluorescence émise ne provient pas des mêmes phénomènes de désexcitation :

- **La fluorescence atomique** : pour un atome possédant n niveaux d'énergie et excité par un rayonnement monochromatique, la désexcitation des niveaux donne naissance à un rayonnement de fluorescence constitué de plusieurs raies ; trois types de rayonnement de fluorescence se distinguent :
  - ⋆ *La fluorescence résonnante* : elle résulte de la désexcitation d'un niveau atomique vers le niveau à partir duquel a eu lieu l'excitation radiative. Ce niveau de départ peut être soit le niveau fondamental, soit un niveau excité thermiquement. La longueur d'onde du rayonnement émis est identique à celle du rayonnement incident.
  - ⋆ *La fluorescence non résonnante de raie directe* : le niveau excité radiativement se désexcite directement vers un ou plusieurs niveaux d'énergie inférieure. Lorsque le niveau d'arrivée se situe à une énergie supérieure au niveau de départ, la longueur d'onde d'émission de fluorescence est supérieure à la longueur d'onde d'excitation. Nous parlons alors de fluorescence *Stokes*. Lorsque l'excitation a lieu à partir d'un niveau peuplé thermiquement et que le niveau vers lequel a lieu la désexcitation se situe à une énergie inférieure à celle du niveau de départ, la longueur d'onde d'émission de fluorescence est inférieure à la longueur d'onde d'excitation. La fluorescence est qualifiée d'*anti-Stokes*.
  - ⋆ *La fluorescence de cascade* : elle se produit lorsque l'émission a pour origine un niveau peuplé par collisions à partir du niveau excité radiativement. Le couplage collisionnel peut induire soit une augmentation d'énergie, soit une perte d'énergie.
- **La fluorescence moléculaire** : de la même façon qu'un atome absorbe ou émet de l'énergie par saut d'un de ses électrons d'un niveau discret à un autre, une molécule peut absorber ou émettre de l'énergie au cours des transitions d'un de ses électrons entre les divers niveaux d'énergie électronique E_e qu'il peut occuper dans la molécule. Contrairement à un atome, la présence de plusieurs noyaux atomiques dans une molécule amène à prendre en compte également une énergie vibrationnelle E_v et une énergie rotationnelle E_r. De plus amples détails sont fournis à ce sujet dans la section 1.2.1.2 ; l'ensemble de ces énergies incite donc à considérer les phénomènes d'absorption et d'émission comme résultants de transitions entre les différents niveaux d'énergie E_T définis par l'équation (1.1). Le nombre de niveaux d'énergie dans une molécule étant grand, un nombre important de transitions existe. À titre d'exemple, la figure 1.4 donne le diagramme des niveaux d'énergie pour une molécule diatomique. Les absorptions mettant en jeu des transitions entre niveaux électroniques se situent principalement dans l'ultraviolet et le visible. La structure la plus stable de l'état excité ne correspondant généralement pas à la structure la plus stable de

l'état fondamental, la transition d'absorption est la plus souvent suivie d'une conversion interne ramenant la molécule au plus bas niveau vibrationnel de l'état excité. De plus, dans la quasi-totalité des molécules, les états excités sont très rapprochés, conduisant presque toujours à un retour rapide au premier état excité (comme $S1$ sur la figure 1.4), même pour les molécules excitées à un niveau d'énergie élevé. L'analyse de la figure 1.4 nous montre que le spectre d'émission de fluorescence se situe toujours dans un domaine de longueurs d'onde supérieur à celui du spectre d'absorption en raison de la perte d'énergie due à la relaxation vibrationnelle des états excités. Outre l'émission radiative, la molécule se trouvant dans le plus bas niveau vibrationnel de l'état excité est susceptible de se désexciter par l'un des deux processus suivant :

- ⋆ *Conversion interne* : pour certaines molécules, la différence d'énergie entre le plus haut niveau vibrationnel de l'état fondamental et le plus bas niveau de l'état électronique excité est faible. Il peut alors se produire une désexcitation des espèces excitées par relaxation vibrationnelle ou rotationnelle.
- ⋆ *Passage à l'état triplet et phosphorescence* : un certain nombre de phénomènes peuvent provoquer le passage d'un état singulet[5] à un état triplet[6]. La désexcitation de l'état triplet ainsi peuplé peut avoir lieu par voie radiative ou non.

Il est important de noter que les molécules, à la différence des atomes ou des ions en phase vapeur, ne possèdent pas toutes des propriétés de fluorescence. L'émission de fluorescence est courante pour les espèces organiques riches en fluorophores[7] mais plus rare pour les espèces minérales.

1.2.2.3 Diffusion et spectroscopie Raman

Lorsque la fréquence de l'onde électromagnétique est très différente de toute fréquence de vibration de la molécule, il se produit un phénomène de diffusion Raman qui met en évidence les transitions électroniques entre les états vibrationnels (et rotationnels dans le cas des gaz) d'une molécule [7]. Ces énergies sont très faibles en comparaison de celles des photons incidents. Comme indiqué sur la figure 1.3, l'explication du phénomène Raman fait intervenir des niveaux énergétiques virtuels. Une molécule ne peut *a priori* absorber un photon pour passer à un état plus excité que si l'énergie de ce dernier correspond à l'écart entre le niveau énergétique actuel de la molécule et un niveau énergétique excité supérieur. Or en spectroscopie Raman, la molécule absorbe une partie du rayonnement incident pour passer à un autre niveau d'énergie possible. Du point de vue quantique, nous pouvons l'expliquer en utilisant la relation d'incertitude d'Heisenberg :

$$\Delta E \times \Delta t \geq \frac{h}{2\pi}$$

[5] un état singulet est un état énergétique d'une molécule dont les deux électrons formant une liaison ont leurs spins opposés

[6] un état triplet est un état énergétique d'une molécule dont les deux électrons formant une liaison ont leurs spins parallèles

[7] Un corp qui absorbe des photons s'appelle chromophore, et tout chromophore qui émet des photons s'appelle fluorophore

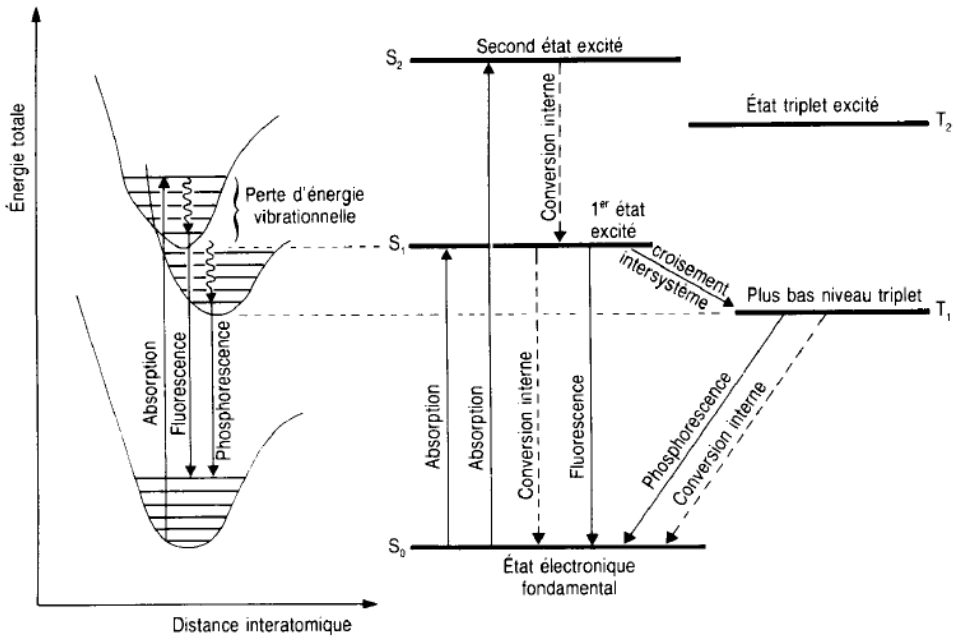


FIG. 1.4 – Diagramme des niveaux d'énergie pour une molécule diatomique

Comme le laps de temps Δt pendant lequel la molécule s'approprie l'énergie nécessaire est très court, l'incertitude sur l'énergie est très grande. La molécule peut donc utiliser une partie de l'énergie incidente pour effectuer une rotation ou vibrer. Le photon incident, qui a cédé une partie de son énergie, a maintenant une énergie $h\nu - \Delta E_{vib}$ où ΔE_{vib} représente un saut énergétique entre deux niveaux d'énergie permis (de rotation ou de vibration). Le photon émis est caractérisé par une fréquence plus faible : il s'agit de la diffusion Raman Stokes. À l'inverse, lorsqu'une molécule est dans un état excité, elle peut se désexciter en donnant de l'énergie à un photon incident. Le photon a alors une énergie égale à $h\nu + \Delta E_{vib}$. Le photon émis possède donc une fréquence plus élevée : il s'agit de la diffusion Raman anti-Stokes.

L'approche quantique précédente peut trouver une justification en utilisant un raisonnement classique.

La diffusion Raman est liée à la polarisabilité de la molécule, c'est à dire à la faculté du nuage électronique à acquérir un moment dipolaire induit sous l'effet du champ électrique de l'onde incidente. Supposons que l'onde électromagnétique incidente soit de fréquence ν_0. Son champ électrique peut alors s'écrire sous la forme :

$$E = E_0 \cos \nu_0 t.$$

Une molécule soumise à une telle onde possède un moment dipolaire induit (appelé aussi polarisabilité) qui traduit la faculté du nuage électronique à se déformer sous l'influence du champ électrique. Ce moment dipolaire induit est proportionnelle au vecteur champ électrique de l'onde incidente. Ce moment dipolaire induit se met sous la forme : $P = \alpha E_0 \cos \nu_0 t$ où le paramètre α est la polarisabilité de la molécule. Le moment dipolaire induit varie avec les mouvements de vibration de la molécule. Comme décrit dans le paragraphe 1.2.1.2 à la page 8, une molécule de n atomes possède $3n - 6$ degrés de liberté de vibrations.

La polarisabilité peut s'exprimer par un développement du premier degré autour de la position d'équilibre α_0 :

$$\alpha = \alpha_0 + \sum_{i=1}^{3n-6} \left(\frac{\partial \alpha}{\partial Q_i} \right) Q_i$$

où Q_i représente la coordonnée normale du mode de vibration i. Dans l'approximation harmonique, cette coordonnée normale s'écrit sous la forme :

$$Q_i = Q_0 \cos \nu_i t$$

pour une fréquence ν_i et une amplitude Q_0. En introduisant ces expressions dans l'équation du moment dipolaire induit, nous obtenons une équation qui englobe les différents termes de diffusion :

$$P = \alpha_0 (E_0 \cos \nu_0 t) + \frac{1}{2} E_0 \sum_{i=1}^{3n-6} \left(\frac{\partial \alpha}{\partial Q_i} \right) Q_0 [\cos(\nu_0 - \nu_i)t + \cos(\nu_0 + \nu_i)t] \tag{1.3}$$

Cette équation montre qu'un rayon monochromatique, d'énergie inférieure à l'énergie nécessaire à la transition entre deux niveaux électroniques d'une molécule, mais beaucoup plus intense que les énergies vibrationnelles et rotationnelles de cette molécule, peut interagir de deux façons différentes avec cette molécule polarisable. L'onde électromagnétique diffusée possède plusieurs composantes de différentes fréquences :

- Une onde diffusée à la même fréquence ν_0 que l'onde électromagnétique incidente est observée, c'est la diffusion Rayleigh ou diffusion élastique car il y a conservation de l'énergie. Cette onde correspond au premier terme de l'équation (1.3).
- Deux diffusions inélastiques aux fréquences $\nu_0 - \nu_i$ et $\nu_0 + \nu_i$ correspondantes au couplage des deux fréquences sont également observées. Ce sont ces deux diffusions inélastiques qui sont appelées diffusions Raman Stokes $(\nu_0 - \nu_i)$ et anti-Stokes $(\nu_0 + \nu_i)$. Elles sont prévues par le second terme de l'équation (1.3).

Une précision pratique est à donner sur l'intensité relative des raies Stokes et anti-Stokes. En pratique, seule la diffusion Raman Stokes est observée et enregistrée car elle est beaucoup plus intense que la diffusion Raman anti-Stokes. L'explication nécessite un rappel à une description quantique de la molécule oscillante. L'énergie de vibration est quantifiée en niveaux d'énergie discrets $E_i = h\nu_i$ avec $i = 0, 1, 2, \ldots$. À une température donnée, la répartition en niveaux d'énergie d'une molécule obéit à la loi de distribution de Maxwell-Boltzmann. Pour obtenir une diffusion anti-Stokes, il faut être en présence de molécules se trouvant dans un état d'énergie vibrationnel excité. Le rapport des intensités des raies Stokes et anti-Stokes s'écrit sous la forme [7] :

$$\frac{I_{anti-Stokes}}{I_{Stokes}} = \left(\frac{\nu_0 + \nu_i}{\nu_0 - \nu_i} \right)^4 \exp \left(- \frac{h\nu_i}{kT} \right)$$

où k est la constante de Boltzmann et T est la température. En conséquence, dans une expérience de spectroscopie Raman, seule la diffusion Stokes est détectée car elle est plus intense à température ambiante. En effet, l'état de départ de l'effet Stokes est plus peuplé en électrons car c'est un état de moindre énergie et, par conséquent, plus stable.

L'intensité de la lumière diffusée pour une transition Stokes entre les états moléculaires m et n est donnée par l'équation [105] :

$$I_{Stokes} = K(\nu_0 - \nu_i)^4 I_0 \sum_{jk} \left| \left(\frac{\partial \alpha_{jk}}{\partial Q} \right) mn \right|^2 \tag{1.4}$$

où K est une constante, Q correspond aux coordonnées normales de vibration, et $\frac{\partial \alpha_{jk}}{\partial Q}$ est la dérivée du tenseur de la polarisabilité α_{jk} selon les coordonnées normales de vibration Q. De cette expression de l'intensité, nous retiendrons que l'intensité de la diffusion Raman Stokes augmente avec l'intensité et la fréquence de l'onde incidente. Pour avoir un effet Raman Stokes plus intense, l'intensité de l'onde excitatrice peut être augmentée à condition de rester vigilant à ne pas dégrader l'échantillon. Par contre, nous sommes souvent limités en pratique à une fréquence d'excitation maximale car plus l'onde excitatrice est énergétique, plus le phénomène d'émission de fluorescence va masquer l'effet Raman.

1.2.3 Conditions d'interaction entre onde et matière

Les énergies des ondes et des liaisons sont quantifiées, c'est à dire qu'elles ne peuvent prendre que certaines valeurs discontinues ou discrètes. L'interaction d'une onde électromagnétique et d'une liaison atomique n'est possible que si plusieurs conditions se trouvent remplies :

- l'énergie E de l'onde électromagnétique correspond à la différence d'énergie ΔE entre deux niveaux d'énergie d'une transition possible de la liaison atomique ;
- la liaison est polaire (en spectroscopie infrarouge) ou polarisable (en spectroscopie Raman) ;
- la fréquence de la composante électrique de l'onde électromagnétique est égale à la fréquence propre de variation du moment dipolaire.

Enfin, il existe des règles de sélection quantiques qui déterminent les transitions possibles entre les différents niveaux vibrationnels excités de la molécule.

1.2.4 Règles de sélection

La spectroscopie de fluorescence repose sur les transitions des niveaux électroniques d'une molécule par excitation d'une onde électromagnétique. Pour qu'une molécule émette de la fluorescence, l'énergie du laser doit correspondre exactement à la différence d'énergie entre deux niveaux électroniques. Les règles de sélection de la fluorescence sont en fait les règles de sélection liées à la phase d'absorption d'un photon, nécessaire pour amener la molécule dans un état excité :

- des transitions entre états de multiplicité différentes sont interdites, c'est à dire que les transitions singulet-triplet et triplet-singulet ne sont pas autorisées. Seules sont permises les transitions singulet-singulet et triplet-triplet ;

- certaines conditions de symétrie ou d'antisymétrie de la molécule sont nécessaires, ou au contraire interdites. Ces considérations font appel à la théorie des groupes, théorie complexe qui ne sera pas présentée dans ce mémoire.

Examinons maintenant les cas plus complexes et complémentaires des spectroscopies Raman et infrarouge.

Supposons une molécule dont la structure reste invariante par des opérations de symétries (opérations de rotation ou de réflexion formant des groupes de symétrie). Dans la base des coordonnées normales, les énergies cinétique et potentielle exprimées dans le cas du traitement quantique sont des fonctions quadratiques. Ces énergies doivent rester invariantes pour toute opération de symétrie de la molécule. Il en résulte que chaque coordonnée normale doit être soit symétrique, soit anti-symétrique par rapport à ces opérations :

- en spectrométrie d'absorption infrarouge, un mode de vibration selon une coordonnée normale Q est actif si la dérivée du moment dipolaire $(\partial\mu/\partial Q)_0$ est non nulle ;
- en spectrométrie de diffusion Raman, un mode de vibration selon une coordonnée normale Q est actif si la dérivée de la polarisabilité $(\partial\alpha/\partial Q)_0$ est non nulle.

Le moment dipolaire étant un vecteur et la polarisabilité une matrice, ces conditions reviennent à ce que l'une au moins des composantes $(\partial\mu_i/\partial Q)_0$ ou $(\partial\alpha_{ij}/\partial Q)_0$ (avec i et j égaux aux coordonnées x, y et z) soit non nulle.

L'existence de symétries dans les molécules et la transformation possible des modes normaux comme une représentation irréductible du groupe de symétrie de la molécule concernée permet d'établir quelques règles importantes :

- Si la molécule possède un centre de symétrie, il n'existe aucune vibration commune aux spectres infrarouge et Raman :
  - ⋆ les vibrations symétriques par rapport à ce centre sont actives en Raman mais inactives en infrarouge ;
  - ⋆ les vibrations anti-symétriques par rapport à ce centre seront inactives en Raman mais actives en infrarouge.

  C'est la règle de l'exclusion mutuelle. Ainsi, la présence simultanée de modes de vibrations à la fois dans les spectres Raman et infrarouge indique de façon certaine l'absence de centre de symétrie.
- Par contre, certaines vibrations peuvent n'apparaître ni en infrarouge ni en Raman.
- Si la molécule possède au moins un axe de symétrie d'ordre supérieur à deux, des modes dégénérées apparaissent. Une dégénérescence double signifie que deux modes sont confondus en une seule raie. Une dégénérescence triple signifie que trois modes sont confondus en une seule raie. Ce phénomène réduit le nombre de raies ou de bandes apparentes, qui peut ainsi devenir inférieur au nombre $3n - 6$.

- Les vibrations totalement symétriques sont toujours actives en Raman, pour tous les groupes de symétrie. Les raies correspondantes sont polarisées et souvent intenses, ce qui permet de les repérer facilement dans le spectre Raman.
- Les autres modes de vibration (antisymétriques ou dégénérés) donnent, lorsqu'ils sont actifs en diffusion Raman, des raies dépolarisées.

Ces règles de sélection ont été établies dans l'approximation harmonique. Or, les spectres expérimentaux présentent parfois des bandes caractéristiques des harmoniques ou des combinaisons de modes. Ces bandes s'interprètent par des transitions entre niveaux vibrationnels non consécutifs.

De par ces règles de sélection, les spectres Raman et infrarouge fournissent des informations vibrationnelles différentes d'un même échantillon. Leur étude simultanée est clairement reconnue et la complémentarité des données Raman et infrarouge facilite l'interprétation des spectres expérimentaux [28].

Les informations fournies par la spectroscopie de fluorescence n'étant pas de l'ordre des énergies vibrationnelles ou rotationnelles mais étant de l'ordre des énergies électroniques, il n'existe pas de dualité entre Raman ou infrarouge et fluorescence. Au contraire, il est bien connu que la fluorescence est un effet parasite en spectroscopie Raman, et inversement. La spectroscopie de fluorescence est donc toujours utilisée indépendamment des spectroscopies Raman et infrarouge.

Cette section a présenté les aspects physiques théoriques sur lesquels reposent les spectroscopies infrarouge, de fluorescence et Raman. L'analyse d'échantillons par ces techniques nécessite tout d'abord l'élaboration d'une chaîne de mesure et d'acquisition qui va mettre en pratique les principes physiques introduits.

1.3 Chaîne d'acquisition et propriétés des spectroscopies optiques

Dans cette partie, nous allons nous intéresser particulièrement à l'instrumentation des microspectroscopies de fluorescence et Raman et à leurs propriétés puisque les applications décrites dans la suite de ce mémoire sont issues de ces deux techniques de spectrométrie.

1.3.1 Chaîne d'acquisition

L'architecture principale des microspectromètres est quasiment identique en microspectroscopie de fluorescence et Raman. Classiquement, toute installation de microspectrométrie comprend :

- *Une source* : un laser est utilisé comme source excitatrice. La longueur d'onde est choisie dans le visible ou le proche infrarouge pour la spectroscopie Raman, et dans l'ultraviolet ou le visible pour la spectroscopie de fluorescence. En spectroscopie Raman, le choix de la longueur d'onde d'exci-

tation résulte d'un compromis entre diffusion Raman et émission fluorescente. Comme présenté à la section 1.2.2.3, plus l'onde est énergétique (longueur d'onde faible ou fréquence élevée), plus la diffusion est intense, mais plus de transitions électroniques donnent naissance à un phénomène de fluorescence parasite.

- *Une platine porte échantillon* en spectroscopie Raman et *une cellule de mesure* en spectroscopie de fluorescence.
- *Un microscope optique* : il sert à focaliser le rayon laser sur l'échantillon à analyser. Le couplage entre le spectromètre et le microscope permet d'atteindre une résolution spatiale inférieure au micromètre.
- *Un jeu de filtres atténuateurs* permet de réduire, selon les besoins, l'intensité du rayonnement incident, ce qui évite dans certains cas une dégradation thermique de l'échantillon.
- *Une optique de collection* de la lumière diffusée (Raman) ou émise (fluorescence) et *une optique de couplage* : le signal émis est dirigé vers le spectromètre en passant au travers d'un trou confocal et d'un filtre notch. Le lecteur est renvoyé à la figure 1.5 où un exemple de microspectromètre Raman est présenté. Ce trou confocal permet de sélectionner le signal émis par un point particulier de l'échantillon et d'augmenter la résolution spatiale de l'analyse. Dans le cas d'un échantillon transparent multicouches, le trou confocal permet de sélectionner et d'analyser une couche de cet échantillon. Le filtre notch est utilisé pour sélectionner une bande spectrale étroite du signal polychromatique émis par l'échantillon.
- *Un système d'analyse spectrale (spectromètre)* : le système est équipé de réseaux dispersifs holographiques de 1800 $traits/mm$ ou de 600 $traits/mm$ selon la fenêtre spectrale désirée et selon la résolution spectrale souhaitée. Plus le réseau est fin (nombre de traits par mm plus élevé, donc réseau plus dispersif), meilleure est la résolution spectrale, mais plus étroite est la fenêtre spectrale d'analyse.
- *Un détecteur de rayonnement* : le spectre est analysé par un détecteur multicanal CCD (Coupled Charge Detector) de 1064×256 pixels. Ce système permet de mesurer simultanément l'intensité des différentes longueurs d'onde du spectre.
- *Une électronique d'acquisition.*

À tous ces équipements s'ajoute une platine motorisée pour déplacer l'échantillon. Cet équipement est indispensable pour l'enregistrement d'images spectrales à partir de spectres enregistrés en des points régulièrement espacés. Le pas de déplacement minimal est de 0,1 μm.

À titre d'exemple, la figure 1.5 décrit le principe du microspectromètre Raman.

1.3.2 Propriétés des spectroscopies optiques

Chaque technique spectroscopique décrite possède des propriétés qui lui sont singulières. Chaque application nécessite une étude préalable de l'échantillon à analyser et des conditions expérimentales dans

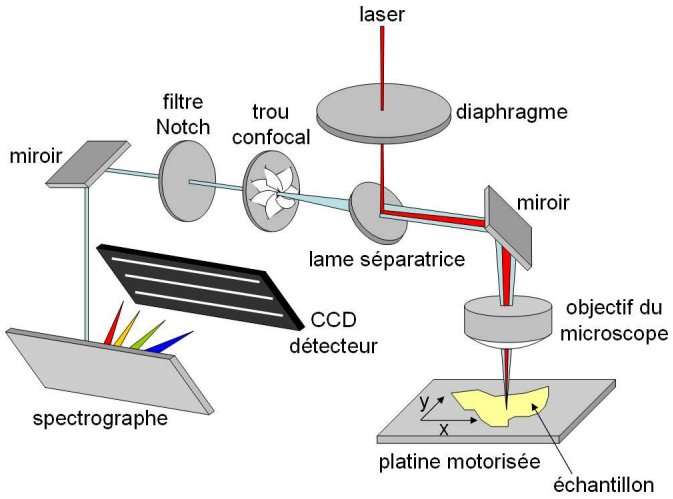


FIG. 1.5 – Description du microspectromètre Raman

lesquelles les spectres vont être acquis. De cette étude va dépendre le choix de la spectroscopie optique utilisée. Une liste des principaux avantages et propriétés des spectroscopies infrarouge, de fluorescence et Raman est établie ci-dessous.

1.3.2.1 Spectroscopie infrarouge

Les principaux avantages de la spectroscopie infrarouge sont :

- La rapidité d'acquisition : quelques secondes seulement sont nécessaires pour recueillir le spectre d'absorption d'un échantillon qui servira à l'étude de sa composition, en comparaison avec une analyse de composition chimique de très longue durée.
- Le faible coût : hormis l'investissement de départ dans le spectromètre et la constitution des calibrations pour chaque produit, le coût d'analyse des échantillons est très faible.
- La sensibilité et la résolution spectrale : contrairement à l'effet Raman, l'absorption infrarouge est de forte intensité, et les spectromètres sont très sensibles et possèdent une résolution spectrale de l'ordre du cm^{-1}. L'analyse quantitative et qualitative des spectres en est facilitée, et les composantes chimiques concentrées en faibles doses deviennent détectables.
- La résolution spatiale : l'utilisation d'un microscope couplé avec le spectromètre assure une résolution spatiale de l'ordre du micromètre.

Ses principaux inconvénients sont les suivants :

- La sensibilité à l'eau : l'analyse des solutions aqueuses est rendue quasiment impossible car l'eau est une espèce chimique très active en infrarouge.
- La largeur de ses bandes : de par l'origine des signaux infrarouge, les bandes spectrales sont larges et difficilement quantifiables sans des calibrations chimiométriques complexes. Un exemple

de spectre infrarouge enregistré à partir d'un échantillon de peau paraffinée est présenté sur la figure 1.6. Nous ne nous attarderons pas sur ce type de spectre puisque la spectroscopie infrarouge n'a pas été utilisée dans les applications mises en œuvre dans ce mémoire.

- La mise en forme des échantillons : les différentes techniques d'analyses en infrarouge requièrent des échantillons moulus et mélangés à des huiles minérales ou des poudres. Le mélange résultant est soit pulvérisé entre deux lamelles, soit compressé sous forme de pastille.

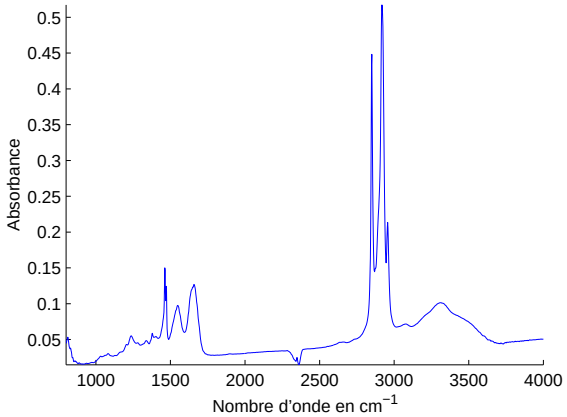


FIG. 1.6 – Exemple d'un spectre infrarouge acquis sur un échantillon de peau paraffinée

1.3.2.2 Spectroscopie de fluorescence

En spectroscopie de fluorescence, nous retiendrons principalement les propriétés suivantes :

- La très grande sensibilité : elle est de 100 à 1000 fois supérieure à celle de la spectroscopie infrarouge et donc beaucoup plus grande encore à celle de la spectroscopie Raman.
- Un large champ d'applications : les fluorophores sont présents dans de nombreuses espèces organiques.

Quelques problèmes limitent toutefois son utilisation :

- La sélectivité : le spectre d'émission dépend de la nature de la molécule, mais aussi des interactions mises en jeu entre cette molécule et son voisinage. L'étude de milieux complexes en est rendue difficile.
- La largeur des pics : les spectres de fluorescence sont constitués de bandes spectrales très larges, contrairement à la spectroscopie Raman. Chaque bande est difficilement attribuable à une seule espèce moléculaire. Les études quantitatives et qualitatives en deviennent approximatives. Un exemple de spectre de fluorescence est donné sur la figure 1.7. Ce spectre a été acquis sur l'extérieur de la couche à aleurone d'une coupe transversale de grain de blé. Il se compose clairement de deux bosses larges centrées aux longueurs d'onde 434 nm et 515 nm. Ces bosses sont respectivement attribuables à l'acide férulique lié et à l'acide férulique libre. Mais la présence connue d'acide para-coumarique n'est pas visible sur ce spectre. L'estimation des spectres de ces espèces et de

leurs profils de concentration au sein du grain de blé par des méthodes de séparation de sources fera l'objet du chapitre 3 de ce mémoire.

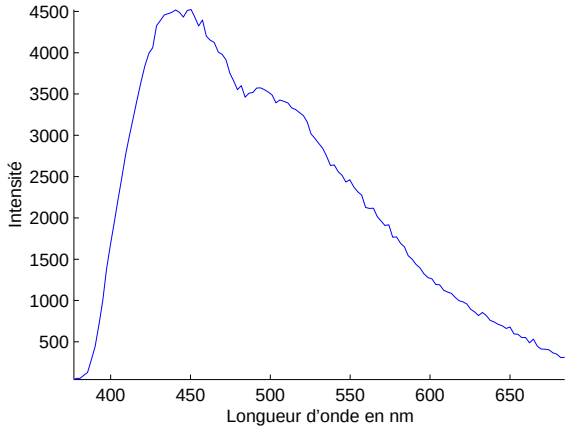


FIG. 1.7 – Exemple d'un spectre de fluorescence acquis sur l'extérieur de la couche à aleurone d'une coupe transversale de grain de blé

1.3.2.3 Spectroscopie Raman

La spectroscopie Raman présente les propriétés suivantes :

- La résolution spatiale : la taille des constituants de certains composés biologiques sont de l'ordre du micromètre. Il convient donc d'utiliser une technique dont la résolution spatiale est du même ordre de grandeur afin de pouvoir distinguer ces différents éléments. Le couplage du spectromètre et du microscope permet d'atteindre une résolution spatiale inférieure au micromètre.
- La spécificité moléculaire : la spectroscopie Raman fournit des informations sur les vibrations moléculaires de l'échantillon analysé. La structure chimique des composés de l'échantillon peut donc être déterminée. Le changement d'environnement des molécules liées ou interagissantes entre elles peut être mis en évidence.
- L'analyse non destructive : la faible puissance du laser excitateur utilisé en spectroscopie Raman permet de ne pas déstructurer les espèces ou les groupements moléculaires pendant l'analyse. L'étude de la relation entre structure et fonction des composants de l'échantillon est donc réalisable.
- L'analyse *in situ* : la spectroscopie Raman se prête facilement aux mesures en milieux hostiles et sous contraintes (température élevée, hautes pressions, atmosphère contrôlée, radioactivité).
- La facilité de mise en oeuvre : le laser est simplement dirigé vers l'échantillon qui n'a pas besoin d'une mise en forme particulière.
- La variété des matériaux analysables : la spectroscopie Raman est utilisable quel que soit l'état physique du matériau (solide amorphe ou cristallisé, liquide ou gazeux) et quelle que soit sa nature (organique ou non).

- L'utilisation de solutions aqueuses : l'eau est très peu active en Raman. Des échantillons humides ou en solutions aqueuses peuvent être analysés sans difficultés.
- L'étroitesse des pics Raman : les bandes Raman sont plus étroites que celles typiquement observées sur des spectres infrarouge. Ces bandes sont plus facilement exploitables pour une analyse quantitative. Un exemple de spectre Raman est proposé sur la figure 1.8. Ce spectre a été acquis sur un échantillon d'épiderme de mélanome paraffiné et fixé sur un support en fluorine. Chaque pic étroit est attribuable à une espèce constituante de l'échantillon, c'est-à-dire soit à la paraffine, soit à la fluorine, soit à la peau. Une méthode de déparaffinage numérique sera présentée au chapitre 4 de ce mémoire afin d'éliminer les fortes signatures Raman de la paraffine et de la fluorine des spectres enregistrés.

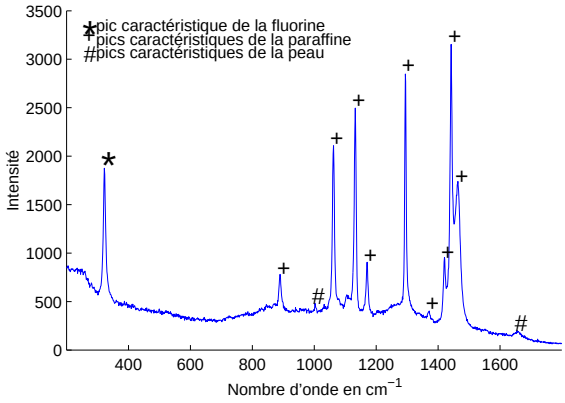


FIG. 1.8 – Exemple d'un spectre Raman acquis sur un échantillon d'épiderme de mélanome paraffiné et fixé sur un support en fluorine

Quelques limitations sont tout de même notables :

- La faiblesse du signal Raman : la spectroscopie Raman est basée sur la diffusion inélastique de la lumière interagissant avec un échantillon. Par définition, ce principe est de faible intensité. La spectroscopie Raman ne permet pas de mesurer des effets de faible ampleur. Elle n'est pas adaptée à la détection de composants chimiques se trouvant en faible concentration dans l'échantillon. Une augmentation du signal Raman enregistré est possible par une augmentation de la puissance du laser excitateur (une dégradation non souhaitable de l'échantillon est alors envisageable par échauffement ou photo décomposition, ainsi qu'un effet de fluorescence plus important) et/ou en utilisant des détecteurs très sensibles.
- La fluorescence : elle se manifeste lorsque l'échantillon à analyser est composé de certaines espèces chimiques particulières dont l'énergie du premier niveau électronique excité correspond exactement à l'énergie du photon incident. La longueur d'onde excitatrice est absorbée et la transition correspondante va jusqu'au premier niveau électronique excité au lieu de rester sur un niveau virtuel. La désexcitation de ce niveau entraîne une fluorescence dont l'intensité masque le faible phénomène de diffusion Raman. Cette fluorescence se manifeste fréquemment par l'ajout d'un

spectre de bandes larges sur le spectre Raman. Ce spectre superposé est gênant car il génère un bruit au niveau de la détection photoélectrique, ce qui rend difficile la détection des faibles raies Raman. Une augmentation de la longueur d'onde excitatrice vers le proche infrarouge est généralement suffisante pour limiter, voir éliminer, la fluorescence car les photons incidents, d'énergie plus faible, n'excitent plus les niveaux électroniques. En contrepartie, l'intensité des raies Raman étant inversement proportionnelle à la quatrième puissance de la longueur d'onde dans l'équation (1.4), le signal enregistré en proche infrarouge sera environ 20 fois plus faible que celui enregistré dans le visible.

- Le temps d'acquisition : la faible intensité de l'effet Raman impose une longue exposition de l'échantillon sous le faisceau incident. Plusieurs dizaines de secondes, voire quelques minutes, sont nécessaires pour l'acquisition d'un spectre Raman. L'enregistrement d'une image spectrale monopolise l'appareillage pendant plusieurs dizaines de minutes, voire plusieurs heures.

- La résolution spectrale qui est de l'ordre de 4 cm^{-1}.

Les principes physiques différents sur lesquels reposent les spectroscopies infrarouge, de fluorescence et Raman engendrent des avantages et des inconvénients spécifiques à chaque technique. La spectroscopie de fluorescence requière la présence de fluorophores dans l'échantillon à analyser. Elle sera donc une technique de choix pour l'étude d'échantillons biologiques, en particulier issus de l'agroalimentaire, tels que des grains de blé ou d'orge. Les spectroscopies Raman et infrarouge sont complémentaires et utilisées sur les mêmes types de tissus biologiques. Le développement de spectromètres Raman dédiés à l'analyse *in situ* et les recherches en miniaturisation des équipements pour proposer des spectromètres portatifs font de la spectroscopie Raman une technique privilégiée pour l'étude de tissus humains.

1.4 Spectres et données spectrales

Les spectroscopies de fluorescence et Raman informent sur la composition chimique d'un échantillon par la mesure des rayonnements émis et, respectivement, diffusés résultants de l'interaction entre un laser excitateur incident et le tissu à analyser. Le vecteur de cette information est un spectre que nous allons présenter dans cette section.

1.4.1 La notion de spectre

Dans le cadre des spectroscopies Raman et de fluorescence, la notion de spectre est différente. Ces différences trouvent leurs origines au cœur même des principes physiques régissant chacune de ces spectroscopies.

1.4.1.1 Spectre de fluorescence

Les spectres de fluorescence sont dépendants de la longueur d'onde λ des radiations émises. Contrairement à la spectroscopie Raman, la spectroscopie de fluorescence est fortement dépendante de la longueur d'onde λ^e du faisceau excitateur.

Ainsi, un spectre de fluorescence est dépendant de deux paramètres :

- la longueur d'onde excitatrice λ^e ;
- la longueur d'onde émise λ par l'échantillon après interaction avec le rayon incident.

Deux types différents de spectres sont ainsi exploitables en spectroscopie de fluorescence, en fonction du paramètre variable.

Définition :

- *Spectre de fluorescence d'excitation :* le spectre d'excitation traduit, pour une transition d'émission fixée, la variation de l'intensité du rayonnement de fluorescence en fonction de la longueur d'onde λ^e du rayonnement d'excitation.
- *Spectre de fluorescence d'émission :* le spectre d'émission est déterminé en enregistrant la variation de l'intensité du rayonnement de fluorescence en fonction de la longueur d'onde de fluorescence λ pour une longueur d'onde d'excitation λ^e fixée.

Un troisième type de spectre est également enregistrable en spectroscopie de fluorescence. Il s'agit d'une généralisation de la notion de spectres d'excitation et d'émission. Pour toutes les longueurs d'onde d'excitation, tous les spectres d'émission sont mesurés. Ces spectres sont appelés *spectres d'excitation-émission*. Mais l'acquisition de ces familles de spectres demande beaucoup de temps et de manipulations. De plus, en chaque point de mesure, une matrice de données est enregistrée. Si n longueurs d'onde d'excitation et m longueurs d'onde d'émission sont utilisées lors d'une expérimentation, n spectres d'émission ou m spectres d'excitation sont enregistrés en un point de mesure, c'est-à-dire $n \times m$ intensités de fluorescence. Par comparaison, un seul spectre Raman est enregistrable en un point de mesure. La visualisation des spectres est alors difficile lorsque les points de mesures quadrille une surface de l'échantillon. Deux dimensions supplémentaires sont à prendre en considération, à savoir l'axe de déplacement suivant les x et l'axe de déplacement suivant les y. Une pratique est alors couramment employée afin de limiter le volume des données. Pour une longueur d'onde d'excitation fixée, les intensités des longueurs d'onde d'émission comprises dans une bande spectrale sont intégrées. Les mesures sont répétées pour différentes longueurs d'onde d'excitation et différentes bandes d'intégration. Les bandes d'intégration sont techniquement réalisées par des filtres optiques transparent dans une bande spectrale bien définie. Chaque configuration expérimentale (longueur d'onde excitatrice et filtre) est numérotée et pour chacune l'intensité du rayonnement de fluorescence est mesurée. Un spectre dans ce cas est la variation de l'intensité du rayonnement de fluorescence en fonction de la configuration expérimentale. Une illustration est présentée

sur la figure 1.10. Afin de les différencier des spectres d'excitation-émission, ces spectres sont dénommés spectres hybrides.

Remarque : les spectres de fluorescence s'expriment en fonction des *longueurs d'onde* d'excitation λ^e et d'émission λ, contrairement aux spectres Raman qui utilisent les *nombres d'onde* $\bar{\nu}$.

Composition d'un spectre de fluorescence : À la fluorescence de l'échantillon s'ajoutent des phénomènes perturbateurs dus à la présence de nombreuses espèces dans le milieu de mesure. La composition classique d'un spectre de fluorescence est la suivante [91] :

- le spectre de fluorescence de l'espèce chimique analysée proprement dit ;
- le spectre de fluorescence parasite d'un constituant majoritaire du milieu de mesure dont les niveaux d'énergie sont très proches de ceux de l'élément à analyser ;
- les effets de filtres : certaines espèces chimiques présentes dans le milieu de mesure peuvent absorber le rayonnement d'excitation (on parle alors d'effet de préfiltre) ou le rayonnement de fluorescence (on parle alors d'effet de postfiltre) ;
- le spectre Raman : ce phénomène de faible intensité ne devrait pas gêner l'acquisition d'un spectre de fluorescence ; cependant, lorsque les constituants principaux du milieu de mesure sont très actifs en Raman, les signaux de fluorescence peuvent être perturbés ;
- des phénomènes instrumentaux :
  - ⋆ le courant noir[8] : ce signal correspond au bruit enregistré par les détecteurs en l'absence de signal ;
  - ⋆ la réponse spectrale de l'instrumentation : l'efficacité du détecteur et la transmission et/ou la réflexion des éléments optiques du spectromètre sont fonctions de la longueur d'onde. Le spectre enregistré est donc donné par la convolution entre le spectre Raman de l'échantillon et la réponse de l'appareil.

Remarque : les phénomènes parasites décrits n'affectent pas systématiquement tous les spectres enregistrés. Ils dépendent de l'instrumentation et de l'application considérée. Des prétraitements efficaces pour éliminer ces effets seront présentés à la section 2.3.4, page 48.

Exemples : La figure 1.9 présente le spectre d'émission à 365 nm de l'acide férulique libre. Cet acide est l'un des principaux constituants du cœur du grain de blé.

La figure 1.10 quant à elle illustre la notion de spectre hybride. L'intensité de fluorescence de la lignine 7 a été acquise pour différentes configurations expérimentales qui seront décrites dans le chapitre 3. La lignine 7 est l'une des principales espèces fluorescentes présente dans le grain d'orge.

Remarques : Peu de caractéristiques sont extractibles à la vue du spectre hybride de la lignine 7 car pour différentes configurations expérimentales le même laser et des filtres optiques à bande passante

[8] dark current en anglais

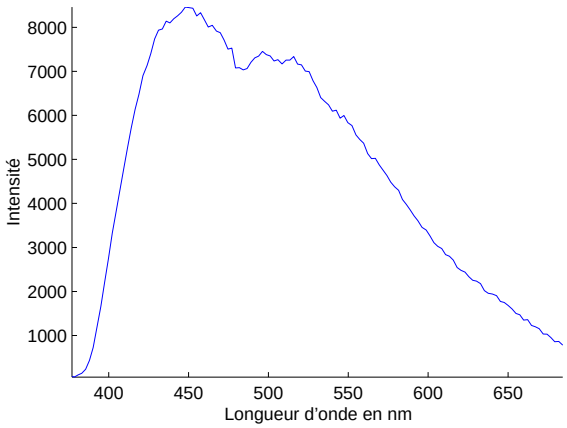


FIG. 1.9 – Spectre d'émission à 365 nm de l'acide férulique libre

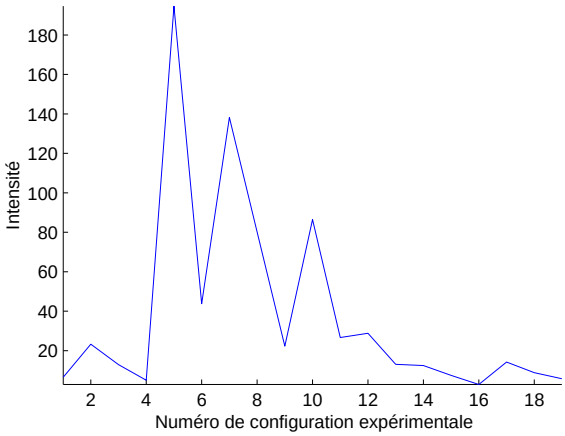


FIG. 1.10 – Spectre hybride de la lignine 7

se recouvrant ont pu être utilisé. L'information disponible est redondante et aucune logique n'est imposée à l'axe des abscisses de ces spectres. De plus, même si une logique se dégageait, le peu d'échantillon disponible par spectre limite les interprétations possibles.

Par contre, le spectre d'émission de l'acide férulique libre enregistré pour un laser excitateur de longueur d'onde de 365 nm autorise quelques remarques. Ce spectre est principalement caractérisé par une forme de type variations lentes que nous appellerons une "bosse". Contrairement à la spectroscopie Raman, des pics étroits ne sont pas les formes privilégiées de la spectroscopie de fluorescence. Par contre, la forme du spectre dans son ensemble amène l'identification du composé chimique analysé.

Une propriété commune évidente aux spectres hybrides et d'émission est la positivité de leurs intensités.

1.4.1.2 Spectre Raman

Un spectre Raman fournit des informations vibrationnelles et rotationnelles d'un échantillon. Mais ces informations ne sont pas mesurées directement. Lors de l'interaction entre l'échantillon et le faisceau incident, les électrons mis en jeu passent brièvement par un état virtuel d'énergie beaucoup plus forte que les transitions vibrationnelles et rotationnelles de l'échantillon. En quittant cet état, les électrons absorbent une partie de l'énergie du photon incident (effet Stokes) ou lui en restituent une partie (effet anti-Stokes). Les photons diffusés par le spécimen biologique n'ont plus la même énergie, donc la même fréquence, que les photons incidents. Mais ces photons diffusés ne caractérisent pas à eux seuls les états vibrationnels. L'information pertinente correspond à l'énergie absorbée ou restituée par rapport à celle du photon incident. Ainsi, la différence $\nu_{diff} = \nu_0 - \nu_i$ entre la fréquence des photons incidents ν_0 et la fréquence des photons diffusés ν_i est porteuse d'information sur les états vibrationnels. Et c'est parce que c'est une différence que les spectres Raman sont indépendants de la longueur d'onde des radiations

incidentes.

Remarque : les spectres Raman ne s'expriment pas en fonction de la fréquence, ou de la longueur d'onde comme les spectres de fluorescence, mais en fonction d'une grandeur qui est proportionnelle à la fréquence, à savoir le nombre d'onde $\bar{\nu}$. Cette grandeur est définie par l'équation (1.2), page 12, dans laquelle ν doit être remplacé par ν_{diff}, c'est-à-dire $\bar{\nu} = \frac{\nu_{diff}}{c}$.

Toutes ces informations sont synthétisées par un spectre Raman, dont une définition possible peut se formuler par :

Définition : Un spectre Raman représente l'évolution de l'intensité des ondes électromagnétiques diffusées par un échantillon biologique excité par une source laser en fonction de la différence entre les nombres d'ondes des photons incidents et les nombres d'ondes des photons diffusés.

Composition d'un spectre Raman : Un spectre Raman ne devrait être théoriquement constitué que des pics caractéristiques des énergies de vibrations et de rotations de la molécule ou du composé étudié. Or comme pour toute instrumentation et comme pour la spectroscopie de fluorescence, les spectres expérimentaux sont le résultat de plusieurs phénomènes [40] :

- le spectre Raman de l'espèce étudiée proprement dit ;
- le spectre du support sur lequel repose l'échantillon. Pour ne pas dénaturé le spectre Raman du spécimen, ce support sera choisi pour sa faible activité Raman, ou alors, le cas échéant, pour son activité Raman localisée dans des bandes spectrales différentes de celles du spectre d'intérêt ;
- un fond du fluorescence : il correspond à la fluorescence parasite de l'échantillon lui-même ;
- des phénomènes instrumentaux :
    - ⋆ le courant noir ;
    - ⋆ la réponse spectrale de l'instrumentation ;
    - ⋆ les rayons cosmiques : lors de l'enregistrement d'un spectre Raman, des raies très intenses à une seule longueur d'onde apparaissent de façon aléatoire.

Remarque : les phénomènes qui parasitent un spectre Raman sont éliminés par des prétraitements ou par l'adaptation des conditions d'enregistrement. Les méthodes usuellement employées pour mettre en forme les spectres Raman seront développées à la section 2.4.4, page 60.

Exemples : Afin d'imager la définition ci-dessus, quelques exemples de spectres Raman sont donnés ci-dessous. Ces spectres correspondent aux spectres de référence des principaux produits utilisés dans les applications qui seront décrites dans le chapitre 4.

Les supports de mesure : Dans toute expérimentation de spectroscopie Raman, l'échantillon à analyser est placé sur un support dont la nature diffère d'une application à l'autre. Les enregistrements

de spectres sont réalisés sur des supports transparents (pour laisser passer la lumière diffusée) et possédant une faible activité Raman (pour ne pas cacher l'information vibrationnelle de l'échantillon étudié). Pour les spécimens organiques, l'utilisation des supports de fluorine (CaF_2) ou de fluorure de baryum (BaF_2) est privilégiée. Comme visible sur la figure 1.11, la fluorine possède un pic Raman étroit et unique centré au nombre d'onde 325 cm^{-1}.

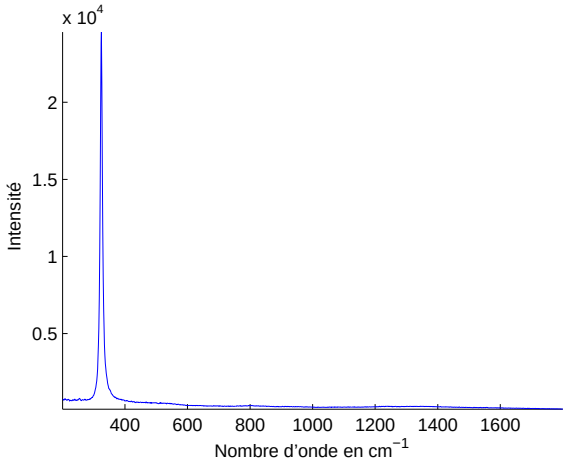


FIG. 1.11 – Spectre Raman de référence de la fluorine (CaF_2)

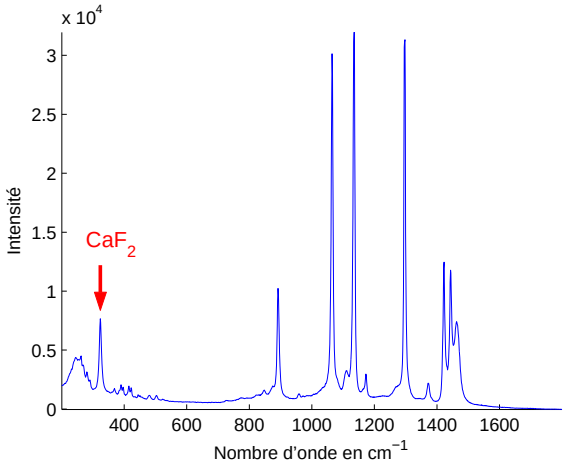


FIG. 1.12 – Spectre de paraffine sur support de fluorine

La paraffine : En médecine, la paraffine est couramment employée pour enrober les échantillons biologiques. La paraffine est un excellent conservateur et permet ainsi de répertorier les échantillons dans des bibliothèques, sans crainte de leur destruction par le temps. Une acquisition sur support en fluorine (CaF_2) du spectre caractéristique de la paraffine est montré sur la figure 1.12. L'influence du support est visible au nombre d'onde 325 cm^{-1} et repéré par une flèche rouge sur la figure. La paraffine est constituée quant à elle de pics étroits. Les principales vibrations de ce composé sont traduites par les pics localisés aux nombres d'onde 890 cm^{-1}, 1063 cm^{-1}, 1133 cm^{-1}, 1172 cm^{-1}, 1296 cm^{-1}, 1418 cm^{-1}, 1441 cm^{-1} et 1463 cm^{-1}.

Remarques : Les exemples présentés ci-dessus illustrent les caractéristiques communes à la plupart des spectres Raman. Ils sont composés de pics étroits localisés aux nombres d'onde auxquels la molécule considérée vibre. Pour certains composés, la fluorescence est beaucoup plus forte que le signal Raman sous-jacent. Les informations vibrationnelles sont obsolètes dans ce cas sans traitements supplémentaires.

Une autre caractéristique des spectres est la *positivité* de leurs intensités. En effet, l'intensité réceptionnée à une longueur d'onde donnée est proportionnelle au nombre de photons diffusés par la matière analysée.

Les propriétés intrinsèques à la spectroscopie Raman seront décrites et analysées à la section 2.4.3, page 59.

1.4.2 Les différents types de mesures spectrales

En fonction de l'application considérée, le type de mesures effectuées diffère. Trois principaux groupes de mesure coexistent :

- *L'enregistrement d'un spectre unique* : la signature spectrale d'une espèce chimique pure est enregistrée afin de répertorier cette espèce. Ce spectre sert généralement de référence pour des expériences sur des composés complexes où cette espèce est connue comme présente. Des comparaisons entre spectres ou des opérations de régression linéaire sur ces spectres serviront à éliminer les informations liées à l'espèce chimique de référence et ainsi à isoler l'information complémentaire sur le composé complexe.

- *La spectroscopie résolue dans le temps* : des spectres d'un système moléculaire en évolution sont enregistrés à plusieurs instants de temps successifs afin de suivre la dynamique de ce système.

- *L'imagerie spectrale* : l'analyse d'un échantillon global est réalisée en mesurant des spectres en plusieurs points de l'échantillon. La localisation de défauts ou l'étude chimique d'un matériau sont rendues possibles par l'ajout des dimensions de surface. L'enregistrement de telles images se fait selon trois approches standards :

  ⋆ Le balayage point par point pour lequel un spectre est enregistré successivement en différents points de l'échantillon. Ce mode d'imagerie est assez lent. Une répartition uniforme des points d'enregistrement suivant l'axe horizontal et l'axe vertical est généralement utilisée pour collecter l'information disponible sur l'ensemble de l'échantillon.

  ⋆ Le balayage ligne par ligne qui demande moins de temps d'acquisition que le balayage point par point car plusieurs spectres sont enregistrés en même temps.

  ⋆ L'imagerie directe pour laquelle une image complète de l'échantillon est collectée en une unique longueur d'onde. Le processus est répété à des intervalles spectraux réguliers. Le temps d'acquisition est donc optimisé.

  Les données issues de ces techniques peuvent être considérées comme des cubes d'intensité spectrale en fonction de la longueur d'onde des rayons diffusés ou émis et des axes spatiaux. L'imagerie spectrale connaît un essor constant avec des applications nombreuses en environnement, industrie, agroalimentaire, semiconducteurs et pharmacie.

Les applications présentées dans les chapitres 3 et 4 s'appuient sur la manipulation d'images spectrales de fluorescence et Raman enregistrées en mode point par point sur des échantillons biologiques de grains de céréales et de tissus de peau humaine. Ces applications nouvelles s'ajoutent aux nombreuses autres qui vont être brièvement décrites dans la section suivante, afin de donner un aperçu de la puissance des spectroscopies optiques et de l'intérêt de l'exploitation de leurs propriétés.

1.5 Applications

Il y a une vingtaine d'années, les champs d'applications de la spectrométrie optique n'étaient pas nombreux du fait de l'encombrement des équipements, de la complexité de leur utilisation, de la faible sensibilité des spectromètres et de la lenteur d'acquisition des spectres. La spectrométrie optique était restreinte à la recherche fondamentale dans les laboratoires universitaires. L'information vibrationnelle enregistrée a permis d'expliquer la structure et la conformation moléculaires de nombreux composés chimiques. Depuis une dizaine d'années, les innovations technologiques telles que les détecteurs multicanaux ou les filtres holographiques, l'amélioration des sources laser, le développement et les avancées de l'informatique, la miniaturisation des équipements, l'augmentation de la sensibilité et la simplification des protocoles expérimentaux n'ont cessé de démocratiser la spectrométrie optique. Il en a résulté une multiplication des domaines d'applications. Un tel succès repose sur la généralité de ces approches, sur leur capacité à analyser n'importe quel type de matériau, sur la nature vibrationnelle et rotationnelle ou électronique des informations qu'elles mesurent, et sur leur simplicité.

1.5.1 Le contrôle en industrie

Avec l'amélioration des instruments, leur miniaturisation et l'apparition des fibres optiques, la spectrométrie optique s'est bien implantée dans l'industrie, aussi bien pour le contrôle-qualité que pour le contrôle en ligne dans les unités de fabrication. Les principaux domaines industriels où la spectrométrie optique est devenue un outil quotidien et indispensable sont donnés à titre d'exemple [7] :

- l'industrie pétrolière : analyse quantitative en ligne et à distance par fibres optiques, de mélanges d'hydrocarbures pour la conduite d'unité de séparation ;
- l'industrie agroalimentaire : suivi à distance de réactions de fermentation ;
- l'industrie pharmaceutique : analyse quantitative en ligne de principes actifs dans des pastilles et comprimés ;
- l'industrie électronique : contrôle en ligne de la qualité de dépôt de carbone sur les têtes de lecture et surfaces de disques durs d'ordinateur.

1.5.2 Étude et contrôle des matériaux

L'étude de l'information vibrationnelle d'un matériau révèle sa structure moléculaire. Le contrôle de la qualité des matériaux peut donc être réalisé par microspectroscopie optique pour [7] :

- les céramiques : analyse des grains et des inclusions dans les matériaux céramiques ;
- les polymères : étude des modifications de conformation, de cristallinité et de stéréorégularité des chaînes polymériques à l'intérieur du polymère ;
- la microélectronique : contrôle des matériaux semi-conducteurs et des microcircuits électroniques.

1.5.3 Environnement

La microspectroscopie optique est utilisée depuis une quinzaine d'année et de plus en plus souvent pour les contrôles de pollution des usines thermiques, chimiques et pétrolières. Les sols ou les eaux avoisinant ces usines sont analysés par spectroscopie optique afin de caractériser leur composition chimique, déterminer les agents polluants ainsi que leur dosage dans les sols ou les eaux [16]. Les particules les plus facilement identifiables sont les cendres volantes produites par les centrales thermiques, les silicates, les oxydes et sulfures métalliques, les carbonates, les phosphates, les sulfates, les insecticides, les pesticides et certains métaux non ferreux.

Quelques applications sont également attribuables à l'étude des grains de sédiments naturels. Les constituants du sédiments (quartz, calcite, oxydes et sulfures de fer) ont pu être identifiés ainsi que leurs répartitions à la surface des grains [17].

1.5.4 Médecine

Depuis quelques années seulement, la microspectroscopie optique connaît une large application en biologie et en médecine. La principale raison de cette utilisation massive est l'opportunité offerte par la spectroscopie Raman en particulier à pouvoir être utilisé *in vivo* ou *in situ* au niveau cellulaire, mais également le développement de supports permettant l'utilisation de l'effet Raman SERS (Surface Enhanced Raman Spectroscopy) et le développement des spectromètres Raman à transformée de Fourier.

L'effet Raman SERS permet d'obtenir des informations sur la structure de molécules adsorbées par une surface rugueuse. Il est basé sur l'exaltation des champs électromagnétiques liés à la présence des rugosités sur une surface métallique et sur le transfert de charge entre la molécule adsorbée et la surface. Ces effets contribuent à intensifier l'effet Raman d'un facteur 10^6.

Les spectromètres Raman à transformée de Fourier sont une adaptation des spectromètres infrarouge à transformée de Fourier. Ils sont basés sur l'utilisation d'un interféromètre de Michelson pour moduler le faisceau laser incident. Le signal enregistré à la sortie du détecteur est donc un interférogramme qui est la somme de toutes les fréquences du faisceau collecté par le détecteur. Une transformée de Fourier de cet interférogramme donne accès au spectre Raman de l'échantillon analysé.

Ces deux méthodes permettent de s'affranchir à la fois de la puissance du laser qui risque de détruire les échantillons biologiques qui sont extrêmement fragiles, et du phénomène de fluorescence qui parasite complètement les spectres enregistrés.

Les applications biomédicales les plus significatives de la spectroscopie Raman sont l'identification des structures moléculaires au niveau cellulaire (aussi bien en biologie qu'en pathologie), l'étude quantitative de la répartition de drogues antitumorales sur des cellules cancéreuses afin d'étudier les problèmes de résistance aux médicaments antitumoraux [9], l'établissement de diagnostics par analyse et comparaison

des spectres de tissus sains et pathologiques [44].

La spectroscopie de fluorescence est une méthode de choix pour l'analyse des tissus d'origine biomédicale puisque de nombreux fluorophores les composent naturellement. Par exemple, l'identification de bactéries est facilitée grâce à la signature spectrale unique de chaque bactérie [75]. La spectroscopie de fluorescence a permis de prouver l'influence du cycle menstruel sur le diagnosis d'un cancer du col de l'utérus [23].

1.5.5 Autres applications

Les applications de la spectroscopie optiques ne se limitent pas aux domaines fondamentaux décrits juste avant. De nombreuses applications existent et donnent une preuve irréfutable de l'importance et de la généralité de l'application de la spectroscopie optique.

Parmi ces exemples, nous pouvons citer l'examen et le contrôle des objets et des œuvres d'arts, l'étude des catalyseurs à base d'oxydes métalliques supportés, l'étude de la composition et de la densité des fluides profonds pour comprendre les mécanismes de formation des roches en géologie, l'analyse des inclusions dans les gemmes en gemmologie, etc.

1.6 Conclusion

Les spectroscopies optiques sont des techniques d'analyse de la structure moléculaire d'échantillons de toutes natures. Les trois principales interactions possibles entre une onde électromagnétique et la matière sont à l'origine des trois grandes techniques de spectroscopie optique. La spectroscopie infrarouge repose sur l'absorption de photons dans l'infrarouge par la matière. La spectroscopie Raman exploite la diffusion inélastique de la lumière. Ces deux spectroscopies sondent les transitions vibrationnelles et rotationnelles d'une molécule. Des règles de sélection affirment qu'une molécule est active en infrarouge si elle est polarisable et active en Raman si sa polarisabilité est variable sous l'action d'un champ électrique. Des règles sur les symétries ou antisymétries régissent l'activité ou l'inactivité Raman et infrarouge. Ces règles sont antagonistes entre la diffusion Raman et l'absorption infrarouge. Des modes de vibrations actifs en Raman sont généralement inactifs en infrarouge et inversement. Spectroscopies Raman et infrarouge sont alors unanimement reconnues comme des techniques d'analyse complémentaires. Le troisième type de spectroscopie est la fluorescence. L'absorption d'une onde électromagnétique très énergétique amène la molécule dans un état électronique excité. Des conversions internes amènent la molécule au plus bas niveau vibrationnel de l'état excité et un photon est émis pour ramener la molécule dans son état le plus stable. Des conditions sur le spin et les symétries de la molécule régissent l'émission de fluorescence.

La chaîne d'acquisition, constituée principalement d'un laser, de jeux de filtres, d'un spectromètre et d'un détecteur, livre à l'utilisateur un ou plusieurs spectres de la matière à analyser en fonction de

l'application. Ces spectres représentent l'intensité des ondes diffusées par l'échantillon en fonction du nombre d'onde des vibrations excitées en Raman, et l'intensité des ondes émises en fonction du nombre d'onde des états électroniques excités et du nombre d'onde du laser excitateur en fluorescence. L'imagerie spectrale exploite les deux dimensions spatiales disponibles sur un échantillon afin d'en étudier la structure et la répartition des constituants principaux.

La puissance d'analyse des techniques spectroscopiques en fait un outil aux multiples applications en industrie pour le contrôle qualité, en environnement et en médecine. Chaque champ d'application possède ses propres caractéristiques qui guident l'utilisateur vers l'emploi d'une technique d'analyse en particulier et vers un ensemble de méthodes de traitements numériques spécifiques. Dans le deuxième chapitre, nous allons nous concentrer sur l'étude des propriétés des spectres de fluorescence et Raman afin de proposer une description des principales méthodes utilisées pour analyser ces spectres dans le cadre de la biologie.

Chapitre 2

Caractéristiques et traitements spectroscopiques de données biologiques

Sommaire

2.1 Introduction

Parmis les nombreuses applications possibles des spectroscopies optiques, une se détache par ses enjeux. La biologie s'affaire à percer les secrets des mécanismes de la vie. Pour l'aider dans sa tâche, les spectroscopies optiques lui fournissent des informations sur la structure moléculaire d'échantillons biologiques, sous la forme de spectres plus ou moins complexes qu'il est nécessaire d'interpréter pour accéder aux informations individuelles de chaque espèce chimique formant l'échantillon à analyser. Ce chapitre concerne l'étude de méthodes classiques utilisées pour effectuer cette interprétation.

Dans une première partie, nous allons introduire plus explicitement le domaine de la biologie, et plus particulièrement les problèmes qu'elle rencontre et les solutions apportées par les spectroscopies de fluorescence et Raman. La complexité des spectres enregistrés, corrélée à la complexité des échantillons biologiques, nous met face à un problème de séparation de sources, à savoir estimer les spectres des espèces chimiques pures afin de les identifier et en déduire leurs concentrations au sein de l'échantillon. Les caractéristiques communes et les dissimilitudes des spectres de fluorescence et Raman vont respectivement conduire à la définition d'un modèle de mesure commun et montrer la nécessité d'utiliser des techniques d'analyses multivariées différentes pour chaque type de spectroscopie. La section 2.3 présentera les propriétés structurelles et d'acquisition propres à la spectroscopie de fluorescence. Les prétraitements usuellement appliqués aux spectres de fluorescence seront décrits et les techniques classiques d'analyse numérique étudiées. Dans la section 2.4, le même schéma d'étude sera repris pour le traitement des spectres Raman.

2.2 Biosignaux

Cette section a pour but de présenter l'application des spectroscopies optiques au domaine de la biologie qui nous intéressera tout au long de ce mémoire. La quête d'informations structurelles sur les échantillons biologiques amène à exploiter les propriétés des spectroscopies de fluorescence et Raman pour sonder les structures biologiques au niveau moléculaire. La complexité structurelle des échantillons biologiques étudiés conduit à la définition d'un problème général d'analyse de leurs composantes chimiques. Les biosignaux acquis par spectroscopies optiques possèdent des propriétés dérivées de l'instrumentation et des lois physiques régissant les interactions entre onde et matière exposées à la section 1.2, page 6. Une modélisation des signaux acquis en est déduite. Cependant les différents types d'interactions onde-matière mesurés par les spectroscopies de fluorescence et Raman se traduisent par des formes différentes, présentées à la section 1.4.1, page 25, entre spectres de fluorescence et spectres Raman. Ceci induit l'élaboration de prétraitements et de traitements des données multidimensionnelles spécifiques à chaque type de spectre.

2.2.1 Biologie et spectroscopies optiques

La biologie cherche principalement à caractériser, quantifier et analyser des matières organiques afin d'en comprendre les structures, les dynamiques, les réactions, les propriétés structurelles, etc. La chimie partage certains de ces objectifs et est de ce fait largement employée en biologie. Cependant, la lenteur des procédés chimiques et la complexité de ses protocoles opératoires restent un obstacle. Une dégradation des échantillons à analyser est souvent une condition nécessaire en chimie pour réaliser une analyse et il est difficile d'établir la structure d'éléments fugaces et de quantifier des espèces en très faible concentration. En particulier, dans certaines réactions chimiques (impliquant des organométalliques par exemple), les espèces chimiques intermédiaires ont une durée de vie tellement courte qu'il est difficile de les identifier par des techniques chimiques telles que des séparations et des purifications [24]. Dans d'autres exemples, les produits de la réaction sont instables et ont des structures difficilement observables [137]. L'utilisation de suppositions et d'interpolations donnent naissance à des incertitudes qui amoindrissent toute tentative d'interprétation des résultats.

L'utilisation de techniques palliant ces difficultés est nécessaire. Il est important de détecter les espèces chimiques présentes en faible concentration et de quantifier l'ensemble des espèces, majoritaires ou minoritaires. Les échantillons doivent rester intacts après l'analyse et la méthode de caractérisation doit être suffisament rapide afin de capturer les informations liées à la présence furtive d'espèces intermédiaires ou instables. Comme décrit dans le chapitre précédent, les techniques de spectroscopie optique possèdent tous ses avantages et donnent accès à toutes ces informations moléculaires. De plus, couplées avec un microscope, les échantillons sont observables au niveau cellulaire. Les méthodes d'imagerie spectrale associent également une dimension spatiale qui assure la simplification de l'analyse structurelle de l'échantillon. Les protocoles expérimentaux étant simplifiés, l'analyse en bandes spectrales aboutit à l'identification rapide des composés. Tous ces avantages sont à l'origine de l'utilisation des méthodes de spectroscopies optiques pour l'analyse biologique d'échantillons.

La spectroscopie de fluorescence est souvent utilisée en biologie. Elle montre des propriétés de discrimination entre un tissu sain et un tissu cancéreux [102]. Elle s'avère être un outil précieux de diagnostic précoce du cancer du col de l'utérus [23]. L'identification et la discrimination de bactéries au niveau du genre, de l'espèce et de la souche est possible grâce aux spectres de fluorescence qui représentent leur empreinte digitale [75].

La spectroscopie Raman assure l'étude d'échantillons biologiques *in vitro* et *in vivo*. La composition et la structure moléculaire des différentes couches de la peau humaine ont ainsi été étudiées *in vitro* [21]. Les spectres Raman *in vivo* de tissus normaux et dysplasiques de palais de rats ont été enregistrés et analysés [5]. La dysplasie désigne une lésion résultant d'une anomalie de développement d'un tissu. Elle est souvent considérée comme une lésion précancéreuse. Les différences spectrales entre ces enregistrements rendent compte des différences structurelles entre tissu sain et tissu malsain. Cette étude a prouvé la faisabilité de la détection précoce d'un cancer à partir de l'analyse de spectres Raman *in vivo*.

La complexité des spectres de fluorescence et Raman est fonction de la composition de l'échantillon à analyser. L'augmentation du nombre d'espèces chimiques pures présentes dans le spécimen rend difficile l'étude des spectres. Ces applications partagent un point commun. Elles tendent toutes vers un même objectif de séparation et d'identification que nous allons détailler dans la section suivante.

2.2.2 Formulation du problème

Les informations fournies par les spectres de fluorescence et Raman enregistrés à partir de mélanges de plusieurs molécules ou espèces chimiques différentes sont complexes. Le principal problème auquel a du faire face la chimiométrie en spectroscopie optique était l'analyse quantitative des spectres enregistrés. Le but de cet analyse est souvent double :

- D'une part, les spectres des différentes espèces chimiques présentes dans le mélange doivent être isolés. Chaque espèce étant identifiée de façon unique par sa signature spectrale, chaque spectre isolé permet alors de leur mettre une étiquette. Le mélange ou la réaction chimique peuvent donc être caractérisés par leur composition chimique ou par l'évolution des espèces présentes.
- D'autre part, les concentrations de chaque espèce chimique doivent être quantifiées, soit pour caractériser leur répartition spatiale, soit pour connaître leur évolution temporelle. L'étude des intensités des spectres isolés des espèces chimiques pures permet d'acquérir ces connaissances puisque les intensités des bandes spectrales d'un spectre sont directement reliées à la concentration des espèces dans le mélange [16, 88].

Mais ce problème ne peut pas être résolu sans l'utilisation de connaissances *a priori* sur les spectres enregistrés et sur la nature du tissu à analyser. Les connaissances *a priori* sur les spectres regroupent les propriétés physiques liées au type de spectroscopie optique employée, et les propriétés statistiques des spectres. Les propriétés physiques permettent de modéliser les données enregistrées en fonction des spectres des différentes espèces chimiques pures, et des concentrations de ses espèces. Les propriétés statistiques mènent à l'utilisation d'une technique de chimiométrie ou de traitement du signal spécifique à l'application considérée.

Il est possible de décrire un modèle général des données enregistrées par spectroscopie optique sur la base de leurs propriétés physiques. Les propriétés statistiques des spectres Raman et de fluorescence étant différentes, elles feront l'objet de deux sections séparées, 2.3 et 2.4. Par contre, les lois physiques régissant leur principe de fonctionnement sont très similaires et vont être étudiées dans la suite de cette section.

2.2.3 Propriétés physiques

Dans la plupart des applications spectroscopiques en biologie, le but est l'étude des compositions chimiques et des structures biologiques des échantillons analysés. La diversité des spectres enregistrés en

plusieurs points de l'échantillon facilite la détermination des constituants chimiques en recherchant par exemple des points où seulement une espèce est présente. Les structures biologiques sont physiquement délimitées et l'enregistrement de spectres en différents points de l'échantillon peut faire ressortir cette structure. C'est pourquoi la notion de surface d'acquisition est toujours implicitement liée à ces applications. Le recours à l'imagerie spectrale semble donc naturel. Mais l'acquisition des spectres par imagerie se matérialise sous la forme d'un cube de données $\mathcal{X}$ dont chaque dimension correspond au déplacement x selon la longueur de l'échantillon, au déplacement y selon la largeur de l'échantillon, et au nombre d'onde $\overline{\nu}$ de la vibration mise en jeu en spectroscopie Raman (voir section 1.4.1.2, page 29) ou à la longueur d'onde λ de la transition électronique considérée en spectroscopie de fluorescence (voir section 1.4.1.1, page 27). Ce cube de données $\mathcal{X}$ se définit par :

$$\mathcal{X} = \{\mathcal{X}_{xy\Lambda} \mid 1 \leq x \leq N_x \ , \ 1 \leq y \leq N_y \ , \ 1 \leq \Lambda \leq N_\Lambda\}$$

où N_x, N_y et N_Λ sont respectivement les nombres de spectres acquis selon les dimensions x, y et le nombre de longueurs d'onde λ, pour la spectroscopie de fluorescence, ou le nombre de nombres d'onde $\bar{\lambda}$, pour la spectroscopie Raman, enregistré pour chaque spectre. La grandeur Λ représente le nombre d'onde $\bar{\lambda}$ ou la longueur d'onde λ selon que la spectroscopie considérée est Raman ou de fluorescence. Comme en chaque point de mesure nous voulons identifier uniquement la présence et la concentration d'un composant, nous n'avons pas besoin de relation spatiale entre les différents points de mesure. Le cube de données $\mathcal{X}$ peut être déplié. Il peut s'écrire sous une forme plus attractive en concaténant toutes les lignes du cube les unes à la suite des autres.

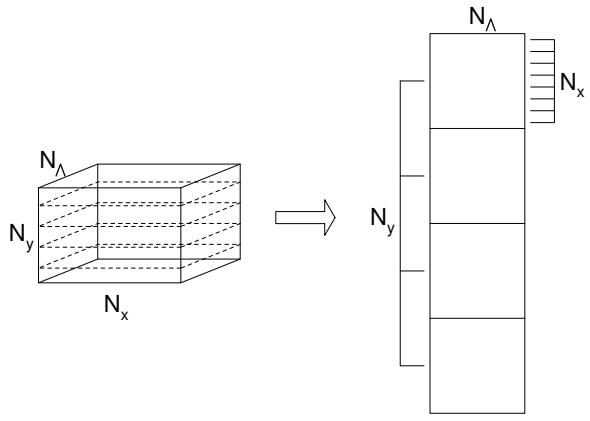


FIG. 2.1 – Représentation schématique du déploiement d'un cube de données sous forme d'une matrice

La figure 2.1 présente cette opération de manière schématique. Il en découle une représentation des données sous forme d'une matrice $\mathbf{X}$. La dimension i de cette matrice représente une combinaison linéaire des dimensions x et y, et l'autre dimension reste le Λ. Cette matrice $\mathbf{X}$ s'exprime sous forme mathématique :

$$\mathbf{X} = \{x_{i\Lambda} \mid 1 \leq i \leq N_{xy} \ , \ 1 \leq \Lambda \leq N_\Lambda \ \} \tag{2.1}$$

où $x_{i\Lambda} = \mathcal{X}_{xy\Lambda}$, $i = x + N_x(y-1)$ et $N_{xy} = N_x N_y$.

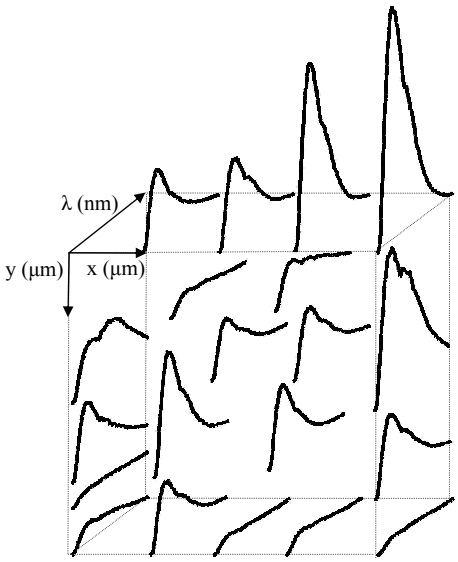


FIG. 2.2 – Exemple d'un cube de données $\mathcal{X}$ en imagerie spectrale de fluorescence : x représente la position en μm selon la longueur de l'échantillon, y la position en μm suivant la largeur et λ la longueur d'onde en nm

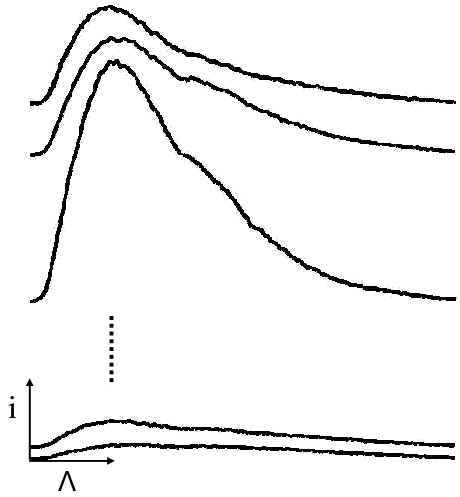


FIG. 2.3 – Le déploiement du cube $\mathcal{X}$ en une matrice de données en imagerie spectrale de fluorescence : i est l'indice de concaténation et Λ la longueur d'onde en nm de l'équation (2.1)

Un exemple de cube de données $\mathcal{X}$ est fournit sur la figure 2.2 lors de l'enregistrement de spectres de fluorescence à la surface d'un grain de blé pour des points de mesure respectant un quadrillage suivant la longueur et la largeur de l'échantillon. Le remaniement du cube $\mathcal{X}$ sous la forme d'une matrice de données $\mathbf{X}$ est représenté sur la figure 2.3. C'est sous cette forme que tous les cubes de données utilisés dans ce mémoire seront agencés.

Spectres Raman : Les lois physiques régissant la spectroscopie Raman sont connues comme étant *linéaires*. Il a été observé dans de nombreux travaux [88, 129] qu'en un point i de l'échantillon à analyser le spectre mesuré $\mathbf{x}_i$ résulte de la superposition pondérée par les coefficients a_{ij} des spectres Raman des espèces chimiques pures $\mathbf{s}_j$ constituant cet échantillon, ce qui est défini mathématiquement par : $\mathbf{x}_i = \sum_{j=1}^{p} a_{ij}\mathbf{s}_j$, où p est le nombre d'espèces chimiques pures constituant l'échantillon. La figure 2.4 illustre cette stratification théorique des spectres enregistrés pour un échantillon virtuel de peau paraffinée sur support de fluorine. Par virtuel, nous entendons qu'à des fins illustratives, le spectre d'un échantillon de peau proposé sur la figure 2.4 n'est rien d'autre que la somme pondérée des trois spectres de référence de la peau, de la paraffine et du support en fluorine. Pour une mesure effectuée en un autre point de l'échantillon, le spectre résultant sera vraisemblablement différent car les espèces chimiques constitutives de l'échantillon sont présentes en des proportions différentes d'un point de mesure à un autre. Pour une longueur d'onde donnée et pour une espèce chimique donnée, l'intensité du rayonnement diffusé sera directement proportionnelle à la concentration de cette espèce [88], et ceci reste valable pour toutes les

longueurs d'onde et pour toutes les espèces présentes. Ainsi les concentrations des espèces chimiques, qui sont indépendantes du nombre d'onde, sont traduites par les intensités des spectres enregistrés.

Une autre propriété physique des spectres Raman découle de la nature des signaux manipulés. La spectroscopie optique est basée sur des phénomènes radiatifs mettant en jeu les transitions vibrationnelles d'une molécule (voir section 1.2.2.3, page 14). Ces transitions n'étant pas continues, les photons diffusés par l'échantillon possèdent des énergies discrètes caractérisées par leur nombre d'onde. L'appareillage de réception et d'enregistrement est calibré pour éviter tout décalage en fréquence et toute distorsion des signaux réceptionnés. Les spectres acquis sont donc *instantanés* en nombre d'onde.

Spectres de fluorescence : En spectroscopie de fluorescence, une relation de *linéarité* existe entre l'intensité du spectre d'une molécule et sa concentration [16]. Les spectres d'excitation et d'émission mesurés sur un échantillon sont supposés être le résultat de la superposition pondérée des spectres des espèces chimiques constitutives de l'échantillon [95]. Les coefficients de pondération restent les concentrations relatives de ces espèces. De plus, pour un raisonnement semblable à celui de la spectroscopie Raman mais reposant cette fois sur le phénomène d'émission de lumière par un corps, les spectres enregistrés sont également supposés *instantanés* mais en longueur d'onde cette fois ci.

Modélisation commune : En conclusion, les spectroscopies Raman et de fluorescence fournissent des spectres qui suivent un modèle linéaire et instantané, à savoir que les spectres des espèces en présence dans le mélange s'additionnent, le spectre de chaque espèce étant pondéré par un coefficient proportionnel à la concentration de cette espèce. Par abus de langage, dans toute la suite de ce mémoire, les coefficients de pondération seront dénommés comme les concentrations des espèces. Ce modèle linéaire et instantané s'écrit sous la forme d'un produit matriciel :

$$\mathbf{X} = \mathbf{AS}. \tag{2.2}$$

où $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_{N_{xy}}]^T \in \mathbb{R}^{N_{xy} \times N_\Lambda}$ est la matrice des données spectrales acquises sur l'échantillon, avec $\mathbf{x}_i = [x_{i1}, \ldots, x_{i\Lambda}, \ldots, x_{iN_\Lambda}]^T \in \mathbb{R}^{N_\Lambda}$ le spectre enregistré au point de mesure i, et $x_{i\Lambda}$ l'acquisition en Λ (la longueur d'onde pour la spectroscopie de fluorescence ou au nombre d'onde pour la spectroscopie Raman) au point de mesure i. $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_j, \ldots, \mathbf{s}_p]^T \in \mathbb{R}^{p \times N_\Lambda}$ est la matrice des spectres des p espèces chimiques de base (ou espèces pures), avec $\mathbf{s}_j = [s_{j1}, \ldots, s_{j\Lambda}, \ldots, s_{jN_\Lambda}]^T \in \mathbb{R}^{N_\Lambda}$ le spectre pure de l'espèce chimique j, et $s_{j\Lambda}$ l'intensité du spectre pure j en Λ. $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_j, \ldots, \mathbf{a}_p] \in \mathbb{R}^{N_{xy} \times p}$ est la matrice des concentrations des espèces pures, avec $\mathbf{a}_j = [a_{1j}, \ldots, a_{ij}, \ldots, a_{N_{xy}j}]^T \in \mathbb{R}^{N_{xy}}$ le profil de concentration dans l'échantillon de l'espèce chimique pure j, et a_{ij} la concentration de l'espèce chimique j au point de mesure i. L'opérateur ^{T} définit l'opération de transposition d'un vecteur ou d'une matrice.

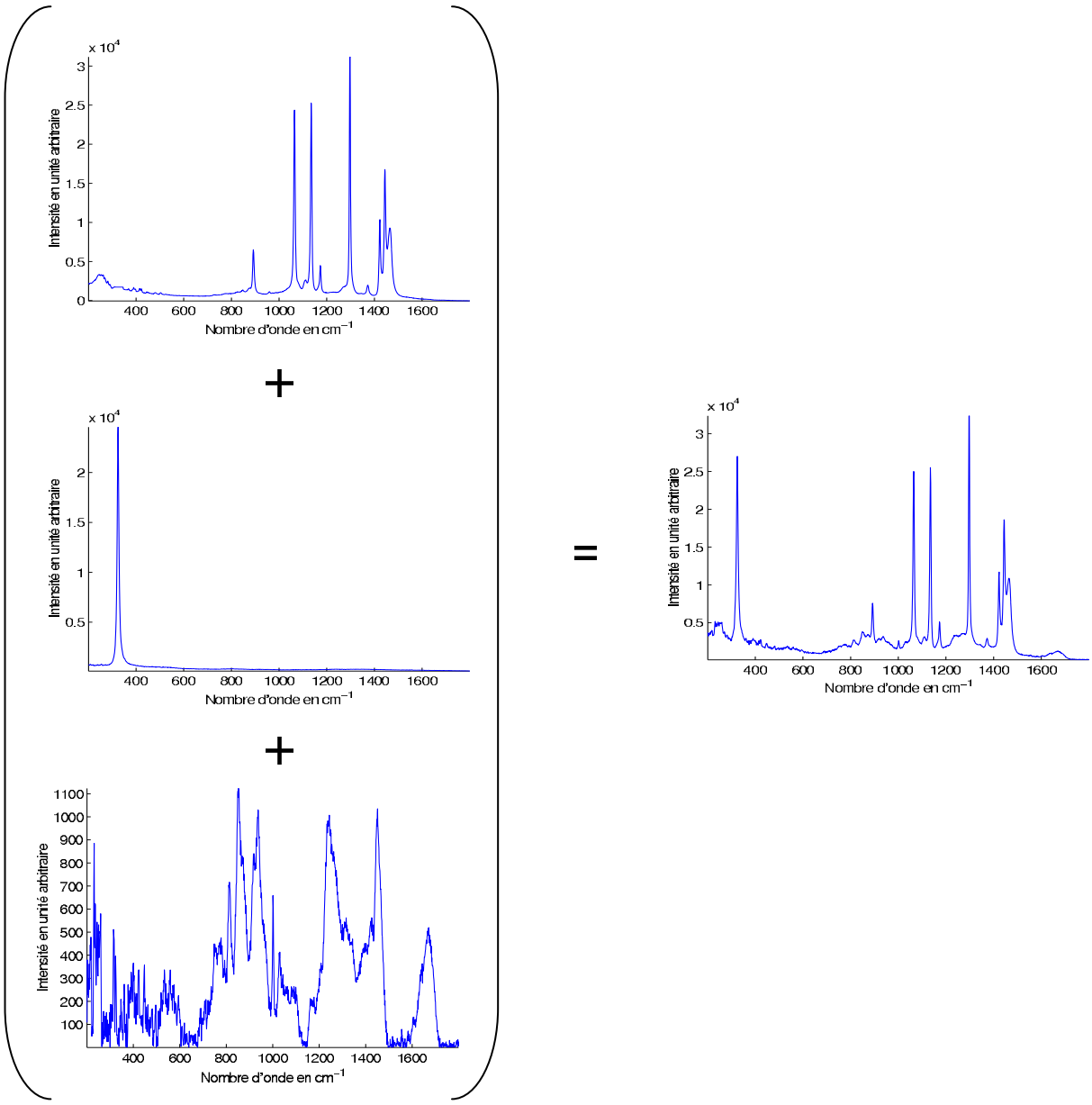


FIG. 2.4 – Décomposition d'un spectre virtuel de peau en fonction de ses constituants. À gauche et de haut en bas : spectre de référence de la paraffine, spectre de référence du support de fluorine, spectre de référence du collagène. À droite : spectre résultant d'une somme pondérée des trois spectres précédents

2.2.4 Dissimilitudes

Les spectroscopies de fluoresence et Raman se différencient cependant sur certains points. Les paramètres d'acquisition sont très différents d'une technique à l'autre. En spectroscopie de fluorescence le laser a une fréquence beaucoup plus grande qu'en Raman pour que les photons excitateurs aient l'énergie suffisante pour exciter les transitions électroniques des molécules de l'échantillon (voir section 1.2.2.2, page 12). Les plages spectrales en spectroscopie de fluorescence sont beaucoup plus étendues que celles en spectroscopie Raman car les énergies mises en jeu sont beaucoup plus fortes en spectroscopie de fluorescence.

L'effet Raman agit localement, c'est à dire qu'il agit au niveau de l'énergie vibrationnelle d'une molécule (voir section 1.2.2.3, page 14). La fluorescence sonde la molécule excitée ainsi que son environnement puisque des processus d'interactions entre la molécule excitée et une autre molécule de son voisinage sont en concurrence avec la désexcitation intrinsèque de la molécule cible (voir la section 1.4.1.1, page 27, et les références [56, 106]).

Les formes caractéristiques des spectres Raman et de fluorescence sont différentes. Les spectres Raman sont caractérisés par des bandes spectrales étroites très énergétiques par rapport au reste du spectre, et les spectres de fluorescence sont d'allure plus régulière. Cette dernière constatation a une conséquence directe sur le nombre de longueurs d'onde d'acquisition des spectres. La spectroscopie Raman nécessite un grand nombre de longueurs d'onde d'enregistrement car l'information liée à une molécule se situe souvent dans une bande spectrale très étroite. En revanche, le profil des spectres de fluorescence étant à variations lentes, un faible nombre de longueurs d'onde d'acquisition suffit. De plus, ceci permet de limiter le temps d'acquisition en ne conservant par exemple qu'un seul pixel en longueur d'onde sur 8. Ce pixel n'est autre que la moyenne de l'intensité de ses 8 voisins. Ce regroupement de pixel augmente la vitesse de lecture du détecteur CCD et divise le temps d'acquisition.

Les différences entre ces deux types de spectroscopies, allant du principe physique, déjà exposé à la section 1.2.2 à la page 10, en passant par les paramètres d'acquisition jusqu'aux allures des spectres enregistrés, nous incitent à séparer dans la suite de ce mémoire les traitements numériques qui leur sont dédiés afin de clarifier leur présentation. Par la suite, certains prétraitements ou certaines techniques d'analyse multidimensionnelle restent communs aux spectroscopies Raman et de fluorescence. Dans ce cas, ils seront décrits pour la spectroscopie de fluorescence et simplement rappelés pour la spectroscopie Raman.

2.3 Traitements des spectres de fluorescence

Les traitements des spectres de fluorescence dépendent de leurs propriétés. Afin d'illustrer ce fait, une étude des spécificités des spectres de fluorescence va être menée dans cette section. De cette étude

seront déduites des techniques de prétraitements ayant pour but d'éliminer tous les signaux parasites des spectres et également de mettre en forme les spectres débruités. Les méthodes d'analyse usuellement employées en spectroscopie de fluorescence seront ensuite décrites.

2.3.1 Paramètres d'acquisition

Bien qu'au sein même de la spectroscopie de fluorescence les paramètres d'acquisition diffèrent d'une application à l'autre, certains sont presque toujours constants :

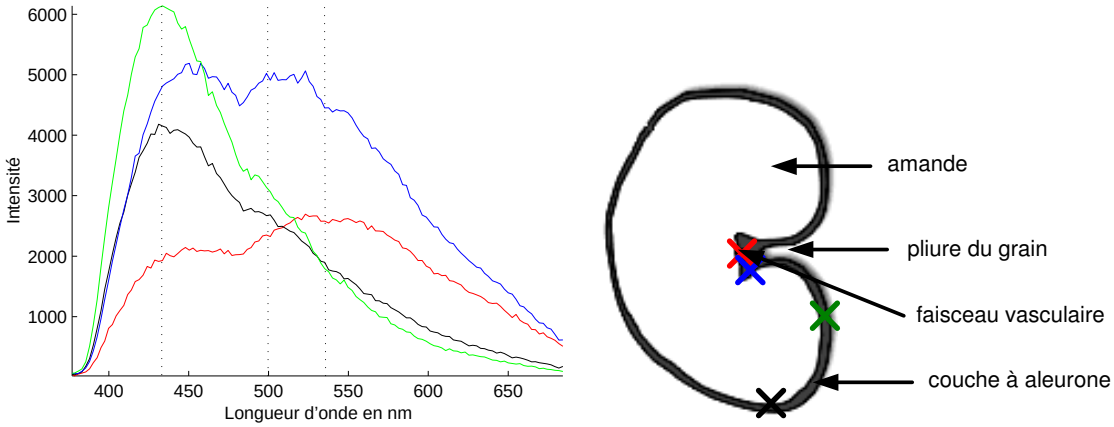
- Le laser excitateur possède généralement une longueur d'onde comprise entre 300 nm et 450 nm, c'est à dire de l'ultraviolet au visible (jusqu'au bleu voire vert). Les photons excitateurs possèdent alors une forte énergie qui va faire transiter une molécule de son état électronique stable à un état électronique excité.
- Les spectres d'émission sont enregistrés suivant une centaine de longueurs d'onde différentes généralement réparties dans un intervalle allant de 250 nm à 700 nm.

2.3.2 Exemples

Des spectres de fluorescence enregistrés en différents points d'une coupe transversale d'un grain de blé sont présentés sur la figure 2.5. Ces spectres ont été acquis pour une longueur d'onde excitatrice de 365 nm et pour des longueurs d'onde d'émission comprises dans l'intervalle spectral de 350 nm à 670 nm. Le spectre rouge a été acquis au niveau du faisceau vasculaire du grain, le spectre bleu au niveau de l'enveloppe entourant le faisceau, le spectre vert dans la couche à aleurone et le spectre noir à l'interface entre la couche à aleurone et l'amande. Ces spectres possèdent une grande diversité et sont composés de bosses, dont les trois principales sont centrées en 436 nm, 500 nm et 530 nm. Les variations décorrélées de ces trois bosses suggèrent l'existence d'au moins trois espèces chimiques différentes. Ceci est dû au fait que les spectres ont été enregistrés sur des parties différentes du grain de blé. Chaque partie est donc composée d'une espèce chimique majoritaire mélangée à d'autres espèces présentes en plus faibles concentrations. Afin de caractériser la composition de la structure du grain de blé, il convient d'en identifier les espèces chimiques principales par leurs spectres et d'en déduire leurs concentrations. À cette fin, diverses méthodes d'analyse multidimensionnelle ont été développées [56, 74]. Toutes reposent sur les propriétés fondamentales des spectres de fluorescence que nous allons décrire dans la partie suivante.

2.3.3 Propriétés et caractéristiques

Il apparaît évident à la vue des spectres de la figure 2.5 et des principes physiques sur lesquels repose la spectroscopie de fluorescence que les spectres enregistrés ont des intensités positives, et ceci quelle que soit la substance analysée. Ainsi, quels que soient la longueur d'onde Λ, le point d'acquisition i et



(a) (b)

FIG. 2.5 – (a)Spectres de fluorescence enregistrés en différents points d'un grain de blé, (b) illustration sur une coupe transversale de grain de blé des points d'acquisition des spectres

l'espèce chimique j considérés, nous pouvons assurer que les éléments $x_{i\Lambda}$ de la matrice de données $\mathbf{X}$ et les éléments $s_{j\Lambda}$ de la matrice des spectres sources $\mathbf{S}$ de l'équation (2.2) sont positifs. Cette première propriété s'écrit mathématiquement sous la forme :

$$x_{i\Lambda} \geq 0 \quad \text{et} \quad s_{j\Lambda} \geq 0, \quad \forall \Lambda, \ \forall i, \ \forall j. \tag{2.3}$$

Les concentrations des espèces chimiques sont des grandeurs positives. Ainsi, une deuxième propriété peut être énoncée :

$$a_{ij} \geq 0, \quad \forall i, \ \forall j. \tag{2.4}$$

Une troisième propriété des spectres de fluorescence est liée à leur forme. Les bosses étant extrêmement larges [16], elles induisent une forte corrélation entre les différents spectres de fluorescence. Les méthodes d'analyse multidimensionnelle basées sur des hypothèses de décorrélation des sources recherchées, telles que l'Analyse en Composantes Principales (ACP), sont alors inefficaces pour estimer les spectres des espèces chimiques pures constituant l'échantillon analysé et leurs profils de concentrations. Par contre, dans un but unique d'analyse, ces techniques peuvent fournir des résultats interprétables car elles permettent d'isoler certaines singularités liées à une seule espèce [56]. Les variations lentes dictant la forme des spectres de fluorescence inhibent quasi-totalement les propriétés statistiques utilisables pour la séparation des spectres des espèces chimiques pures.

Des propriétés liées à la structure de l'échantillon analysé sont alors indispensables pour réussir à estimer correctement le modèle de l'équation (2.2). Également, des propriétés dépendantes de l'acquisition des données sont utiles pour faciliter la résolution du problème de séparation. Cependant, ces types de propriétés sont liées à l'application considérée et ne sont donc pas générales. Nous en reparlerons plus en détail dans le chapitre 3 dédié à l'application de techniques de séparation de sources en spectroscopie de

fluorescence.

Pour le moment, les propriétés générales des spectres de fluorescence sont posées. Intéressons-nous maintenant à la mise en forme de ces spectres par des prétraitements adaptés et indispensables avant toute application d'une technique d'analyse multivariée.

2.3.4 Prétraitements

Les spectres acquis par spectroscopie de fluorescence résultent d'une acquisition complexe. Certains phénomènes, tels que des fluctuations ou des dérives de la chaîne d'acquisition, peuvent masquer l'information liée à la fluorescence de l'échantillon et apparaissent comme gênants dans diverses applications. Des prétraitements sont appliqués aux jeux de données pour améliorer la forme des spectres et faciliter l'utilisation ultérieure d'analyses multidimensionnelles. Ces prétraitements, devenus classiques en spectroscopie de fluorescence et exposés dans la suite, sont l'élimination du courant noir, la correction de la réponse spectrale du système, l'élimination de la diffusion Raman et la normalisation des spectres.

2.3.4.1 Élimination du courant noir

Comme nous l'avons décrit dans le chapitre précédent, le courant noir[9] est le courant mesuré sur les détecteurs CCD dans une obscurité totale, en l'abscence de laser excitateur et d'échantillon. L'intensité de ce bruit est dépendante de la température du CCD mais également du temps d'exposition des pixels. Pour s'affranchir de ce courant, un spectre $\mathbf{s^{cf}} \in \mathbb{R}^{N_\Lambda}$ est enregistré à vide grâce au CCD, c'est à dire qu'aucune lumière n'est dispersée sur le détecteur. Ce spectre du courant noir est simplement soustrait au spectre $\mathbf{x}_i$ enregistré en chaque point de mesure i de l'échantillon à analyser [23] :

$$\mathbf{x}_i = \mathbf{x}_i - \mathbf{s^{cf}}, \quad i = 1, \ldots, N_{xy}.$$

2.3.4.2 Correction de la réponse spectrale non uniforme du spectromètre

L'efficacité du détecteur CCD ainsi que la transmission et/ou la réflexion de la lumière par les éléments optiques du monochromateur d'émission dépendent de la longueur d'onde du rayonnement émis. Afin de rendre les mesures indépendantes de l'appareillage, et ainsi uniquement dépendantes de la structure moléculaire de l'échantillon analysé, cette réponse de l'instrumentation doit être corrigée. Les fabriquants de spectromètres de fluorescence fournissent dans la documentation de l'appareil des coefficients correcteurs qui doivent être appliqués sur les spectres de fluorescence mesurés afin de corriger la réponse spectrale non uniforme du spectromètre [38]. Une autre méthode consiste à enregistrer les spectres de fluorescences de sources calibrées. Ces spectres étant connus, des facteurs de correction de la réponse spectrale du

[9] dark current en anglais

spectromètre, dépendants de la longueur d'onde d'émission, peuvent en être déduits. Ces facteurs sont, comme dans la première méthode, appliqués à chaque spectre mesuré afin d'en éliminer la contribution non uniforme du système de détection [82, 23].

2.3.4.3 Élimination de la diffusion Raman

Dans certaines applications, si une substance majoritaire est très active en Raman, le rayonnement diffusé peut interférer avec le signal de fluorescence, voire le masquer complètement. Par exemple, Wentzell et al [136] ont montré que les molécules du solvent (du méthanol dans cette expérience) diffusent de manières élastique (diffusion Rayleigh) et inélastique (diffusion Raman) dans le voisinage de la longueur d'onde excitatrice. Ce phénomène, concurrent de l'émission de fluorescence, s'observe dans la matrice d'excitation-émission par une bande diagonale. Un exemple, tiré des travaux de Wentzell [136], est montré sur la figure 2.6. Sur les spectres d'excitation-émission représentés sur la figure 2.6 sous forme de courbes d'équi-intensité, une bande diagonale apparaît à partir de la longueur d'onde d'excitation de 360 *nm* et pour des longueurs d'onde d'émission qui lui sont voisines, comme prédit par la théorie des phénomènes radiatifs. Cette zone est entourée par une ellipse rouge sur la figure 2.6.

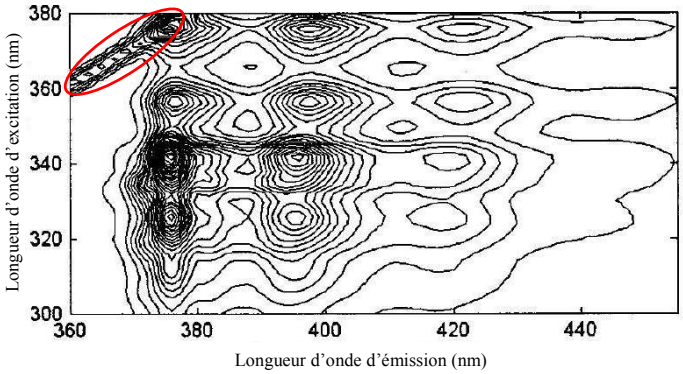


FIG. 2.6 – Pollution d'une matrice d'excitation-émission par la diffusion Raman d'un solvant (figures tirées de [136])

Plusieurs statégies existent pour éliminer cette diffusion. Premièrement, une solution est offerte en soustrayant de la matrice d'excitation-émission le fond de diffusion Raman enregistré à partir d'un échantillon de la matière diffusante seule [136]. Cependant, les coefficients de soustractions doivent être minutieusement estimés afin de ne laisser aucun résidu, et le spectre enregistré de la matière diffusante doit être exactement identique à la partie polluante de la matrice d'excitation-émission. Les conditions d'acquisition de ce spectre et de la matrice doivent être rigoureusement les mêmes.

La deuxième méthode consiste simplement à écarter, de l'analyse multidimensionnelle qui va suivre, les portions de la matrice d'excitation-émission affectées par le spectre Raman de la molécule diffusante. Nous éliminons donc les zones de la matrice où longueur d'onde d'excitation et longueur d'onde d'émission

sont approximativement identiques. Mais, dans certaines applications, cette méthode a l'inconvénient de réduire fortement l'information disponible pour l'analyse multidimensionnelle .

Une troisième méthode non-supervisée a été développée dans [136]. Les matrices d'excitation-émission enregistrées sont concaténées selon l'une des deux dimensions afin de former une matrice unique $\mathbf{X}$. Une Analyse en Composantes Principales (ACP) est utilisée pour décomposer cette matrice. Les composantes relatives à la diffusion Raman sont écartées, et seules les composantes significatives sont conservées. Sous forme mathématique, cette décomposition s'écrit :

$$\mathbf{X} = \mathbf{X}^{\mathbf{signal}} + \mathbf{X}^{\mathbf{bruit}}$$

où $\mathbf{X}^{\mathbf{signal}}$ est le sous-espace signal obtenu par ACP et $\mathbf{X}^{\mathbf{bruit}}$ le sous-espace bruit composé de l'influence Raman. La matrice est reconstruite à partir des composantes principales retenues et définissant le sous-espace $\mathbf{X}^{\mathbf{signal}}$. L'influence du spectre Raman de l'espèce diffusante a maintenant disparue. Finalement, cette matrice est redéployée pour donner les matrices d'excitation-émission originales mais nettoyées de l'effet Raman.

2.3.4.4 Normalisation des spectres

Le prétraitement le plus communément appliqué aux spectres de fluorescence est la normalisation des spectres. Dans la plupart des applications de spectroscopie de fluorescence nécessitant un traitement numérique des signaux enregistrés, cette opération est réalisée. Cette normalisation a pour effet d'accorder à tous les spectres enregistrés la même importance informative [14], c'est-à-dire de réduire les effets liés à l'intensité lors de l'application de techniques d'analyse multivariée. En effet, lors de l'enregistrement de matrices d'excitation-émission, l'intensité du laser excitateur, et donc l'intensité des rayonnements émis, et l'efficacité de l'optique d'excitation sont dépendantes de la longueur d'onde excitatrice. Une normalisation de la matrice d'excitation-émission corrige cette dépendance à la longueur d'onde excitatrice et annule les disparités d'intensité. La priorité est ainsi donnée à la forme des spectres et non à leurs intensités relatives. Plusieurs types de normalisations existent.

Le premier style de normalisation est la réduction de l'aire des spectres de fluorescence à l'unité [75], c'est à dire que la transformation suivante est réalisée sur chaque vecteur $\mathbf{x}_i$ de la matrice des données $\mathbf{X}$:

$$\mathbf{x}_i = \frac{\mathbf{x}_i}{\sum_{\Lambda=1}^{N_\Lambda} x_{i\Lambda}} \quad \text{pour} \quad i = 1, \ldots, N_{xy} \tag{2.5}$$

avec N_Λ le nombre total de longueurs d'onde d'émission dont l'intensité est enregistrée.

La deuxième forme de normalisation est réalisée par rapport à l'intensité maximale de chaque spectre enregistré [101]. Cette transformation se traduit donc sous la forme :

$$\mathbf{x}_i = \frac{\mathbf{x}_i}{max(\mathbf{x}_i)} \quad \text{pour} \quad i = 1, \ldots, N_{xy}. \tag{2.6}$$

Une troisième possibilité consiste à forcer les spectres à un écart type unitaire, c'est à dire à leur appliquer la transformation suivante :

$$\mathbf{x}_i = \frac{\mathbf{x}_i}{(\mathbf{x}_i^T \mathbf{x}_i)^{\frac{1}{2}}} \quad \text{pour} \quad i = 1, \ldots, N_{xy}. \tag{2.7}$$

À la suite de ces transformations, les spectres de fluorescence sont prêts à être traités par des techniques d'analyses multidimensionnelles qui estimeront les spectres des espèces pures ainsi que leurs profils de concentrations.

Dans la suite de ce mémoire, ces trois types de normalisation seront successivement utilisés en fonction de la méthode de séparation numérique des spectres employée.

2.3.5 Méthodes classiques d'analyse et de traitement

Les méthodes classiquement appliquées en spectroscopie de fluorescence sont basées sur l'inspection visuelle et l'Analyse en Composantes Principales (ACP).

2.3.5.1 Inspection visuelle directe

Une méthode simple d'analyse des spectres de fluorescence est leur inspection visuelle directe. Cette méthode repose sur l'observation et la comparaison des formes des spectres pour différentes configurations de l'échantillon, par exemple pour une analyse spectrale résolue dans le temps, ou pour différents échantillons afin d'en étudier les différences moléculaires. Cette technique s'applique principalement lorsqu'un seul fluorophore émet majoritairement dans l'échantillon, ou alors lorsque tous les fluorophores émetteurs sont liés à la même structure ou à la même fonction d'une structure de l'échantillon. Il est rare d'analyser visuellement des spectres de fluorescence résultant de l'activité simultanée de plusieurs fluorophores ayant des fonctions complètement différentes au sein de l'échantillon puisque la spectroscopie de fluorescence est réputée pour les bosses larges associées à une fonction moléculaire [16]. Ainsi, les spectres de différentes fonctions moléculaires se superposent, et un mélange de ces spectres risque de rendre impossible une analyse visuelle. Lors d'une analyse visuelle, les fonctions des molécules fluorescentes sont déduites des formes des spectres, ou des variations entre spectres, puisque chaque molécule possède une signature spectrale qui lui est propre et qui est caractérisée principalement par l'intensité et la position de son ou de ses maximums. Par conséquence, les paramètres importants à observer sont les intensités des bandes et leur longueur d'onde.

Dans [106] par exemple, les auteurs étudient les propriétés des spectres de fluorescence de cellules normales, métastatiques et non-métastatiques obtenues sur différentes tumeurs malignes du rat et de l'homme. Des spectres tirés de ces travaux sont fournis sur la figure 2.7. Une étude visuelle permet de différencier, sur des spectres acquis pour une longueur d'onde d'excitation de 310 nm, deux bandes

spectrales différentes centrées à 340 nm et à 450 nm. Chaque bande a été reconnue comme associée à une molécule unique. La bande à 340 nm traduit la présence de tryptophan, un acteur majeur de la photolyse des protéines et des enzymes dans la région des UV. La bande à 450 nm est liée à la présence de nicotinamide-adénine dinucléotide sous forme réduite (NADH), une molécule importante en recherche biochimique et biomédicale qui est impliquée dans les réactions d'oxydoréduction des cellules de l'organisme. L'étude de ces spectres a révélé que l'intensité de ces deux bandes est beaucoup plus forte pour des cellules non-métastatiques que pour des cellules métastatiques. Ces deux espèces moléculaires sont donc des témoins de la santé d'une cellule et peuvent être utilisées pour prévenir de formation de tumeurs malignes dans différents types de cellules.

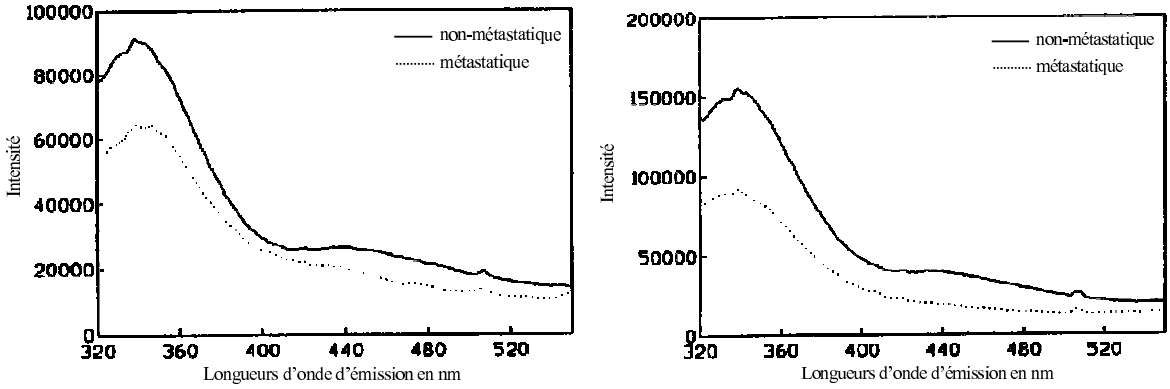


FIG. 2.7 – Spectres d'émission excité à 310 nm enregistrés sur (a) des cellules de carcinome d'un poumon humain (b) des cellules de rhabdomyosarcome de rat (figures tirées de [106])

Cette méthode est utilisée lors de l'analyse de matériaux simples. Dès que le nombre de fluorophores présents dans l'échantillon augmente, des méthodes d'analyse numériques lui sont préférées. L'Analyse en Composantes Principales (ACP) et les méthodes par enveloppes en sont des exemples et sont présentées ci-après.

2.3.5.2 Analyse en Composantes Principales

La technique d'analyse multidimensionnelle la plus populaire pour traiter et classifier les données spectrales de fluorescence est sans contestation l'Analyse en Composantes Principales (ACP). Cette méthode requiert la connaissance ou l'estimation possible des statistiques d'ordres 2 d'un vecteur aléatoire $\mathbf{x} = [x_1, \ldots, x_i, \ldots, x_{N_{xy}}]^T \in \mathbb{R}^{N_{xy}}$.

Aucune hypothèse supplémentaire sur la fonction de densité de probabilité de $\mathbf{x}$ n'est nécessaire. Les éléments du vecteur $\mathbf{x}$ doivent être mutuellement corrélés, et donc présenter une redondance mutuelle des informations [66]. L'ACP transforme linéairement le vecteur $\mathbf{x}$, à composantes mutuellement corrélées, en un vecteur $\mathbf{v} = [v_1, \ldots, v_k, \ldots, v_{N_v}]^T$ de dimensions $N_v \leq N_{xy}$, à composantes mutuellement décorrélées.

Ce but est achevé en appliquant une projection $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k, \ldots, \mathbf{w}_{N_v}]^T \in \mathbb{R}^{N_v \times N_{xy}}$ au système de coordonnées de $\mathbf{x}$ pour trouver un système de coordonnées orthogonales dans lequel les éléments de $\mathbf{v}$ sont décorrélés :

$$E\{v_k v_l\} = 0 \text{ si } k \neq l. \tag{2.8}$$

Les éléments v_k , $k = 1, \ldots, N_v$ s'expriment mathématiquement par :

$$v_k = \sum_{i=1}^{N_{xy}} w_{ki} x_i = \mathbf{w}_k^T \mathbf{x}$$

où les vecteurs $\mathbf{w}_k = [w_{k1}, \ldots, w_{ki}, \ldots, w_{kN_{xy}}]^T$ pour $k = 1, \ldots, N_v$ représentent les axes du système de coordonnées de $\mathbf{v}$. De plus, les variances des projections de $\mathbf{x}$ sur les axes de coordonnées de $\mathbf{v}$ sont maximisées.

Nous cherchons donc à estimer dans un premier temps le vecteur $\mathbf{w}_1$ qui maximise le critère :

$$Q^{ACP}(\mathbf{w}_1) = E\{(v_1 - E\{v_1\})^2\} = \mathbf{w}_1^T \mathbf{R}_\mathbf{x} \mathbf{w}_1$$

sous la contrainte $\|\mathbf{w}_1\| = (\mathbf{w}_1^T \mathbf{w}_1)^{\frac{1}{2}} = 1$ pour limiter sa norme à l'unité. La matrice $\mathbf{R}_\mathbf{x} = E\{(x - E\{x\})(x - E\{x\})^T\} \in \mathbb{R}^{N_{xy} \times N_{xy}}$ est la matrice de covariance de $\mathbf{x}$.

Les autres vecteurs $\mathbf{w}_l$, $l = 2, \ldots, N_v$ sont recherchés pour maximiser le critère $Q^{ACP}(\mathbf{w}_l)$ sous la contrainte $\|\mathbf{w}_l\| = (\mathbf{w}_l^T \mathbf{w}_l)^{\frac{1}{2}} = 1$ et sous la contrainte de décorrélation (2.8) entre v_l et les composantes principales estimées précédemment :

$$E\{v_l v_k\} = \mathbf{w}_l^T \mathbf{R}_\mathbf{x} \mathbf{w}_k = 0.$$

La considération simultanée du critère Q^{ACP} pour tout vecteur $\mathbf{w}_k$, $k = 1, \ldots, N_v$, des contraintes d'unicité de la norme et de décorrélation entre les composantes principales permet facilement de montrer [36] que :

$$\mathbf{W} = [\mathbf{e}_1, \ldots, \mathbf{e}_k, \ldots, \mathbf{e}_{N_v}]^T \tag{2.9}$$

où les vecteurs $\mathbf{e}_k = [e_{k1}, \ldots, e_{ki}, \ldots, e_{kN_{xy}}]^T \in \mathbb{R}^{N_{xy}}$, $k = 1, \ldots, N_v$ sont les N_v premiers vecteurs propres associés aux N_v plus grandes valeurs propres de la décomposition de la matrice $\mathbf{R}_\mathbf{x}$ en valeurs propres :

$$\mathbf{E}\mathbf{R}_\mathbf{x}\mathbf{E}^T = \mathbf{D}$$

où $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_i, \ldots, \mathbf{e}_{N_{xy}}]^T \in \mathbb{R}^{N_{xy} \times N_{xy}}$ est la matrice des vecteurs propres $\mathbf{e}_i$, $i = 1, \ldots, N_{xy}$, et où $\mathbf{D} \in \mathbb{R}^{N_{xy} \times N_{xy}}$ est une matrice diagonale dont les éléments diagonaux d_{ii} sont égaux aux variances des projections de $\mathbf{x}$ sur les vecteurs propres $\mathbf{e}_i$.

Le vecteur des composantes principales $\mathbf{v}$ est estimé par :

$$\mathbf{v} = \mathbf{W}\mathbf{x}. \tag{2.10}$$

Lorsqu'une matrice de données $\mathbf{X}$, définie suivant l'équation (2.1), est traitée à la place d'un vecteur $\mathbf{x}$, ses colonnes représentent différentes réalisations du vecteur aléatoire $\mathbf{x}$ et l'estimation de la matrice des composantes principales $\mathbf{V}$ est la généralisation de l'estimation de l'équation (2.10) à toutes les colonnes de $\mathbf{X}$:

$$\mathbf{V} = \mathbf{W}\mathbf{X}. \tag{2.11}$$

De nombreux travaux relatent l'utilisation intensive de l'ACP en biologie afin de réduire les dimensions des observations à quelques composantes principales explicatives de la structure interne des données ou pour faire de la classification [51, 18, 101, 75].

2.3.5.3 Méthodes par enveloppes

L'une des premières méthodes développée en 1971 pour la séparation de spectres entrelacés est la méthode de Lawton et Sylvestre [74]. Bien que simpliste dans son approche, elle n'en reste pas moins assez puissante dans l'extraction de formes. Elle repose sur des considérations physiques générales de la spectroscopie et sur l'Analyse en Composantes Principales (ACP), mais elle est limitée à la séparation de deux sources.

Considérons des mélanges $\mathbf{x}_i$ de deux sources $\mathbf{s}_1$ et $\mathbf{s}_2$:

$$\mathbf{x}_i = a_{i1}\mathbf{s}_1 + a_{i2}\mathbf{s}_2 \tag{2.12}$$

où les a_{ij} sont les coefficients de concentration de la source j dans le mélange i.

Les hypothèses de travail sont les suivantes :

- les sources $\mathbf{s}_j$ sont supposées non-négatives ;
- les sources $\mathbf{s}_j$ sont supposées linéairement indépendantes ;
- les sources $\mathbf{s}_j$ sont supposées normalisées à l'aire unité par la transformation décrite par l'équation (2.5) ;
- les sources $\mathbf{s}_j$ ne doivent pas s'annuler en même temps ;
- les coefficients a_{ij} sont supposés non-négatifs.

Si les signaux observés proviennent bien de mélanges de deux sources, alors ces signaux peuvent être exprimés en fonction des deux premières composantes principales $\mathbf{v}_1$ et $\mathbf{v}_2$ de la matrice des données et estimées par l'ACP selon l'équation (2.11) pour une décomposition en deux composantes principales.

Remarque : contrairement à l'utilisation classique de l'ACP qui requiert des signaux à moyenne nulle, elle est appliquée dans cette méthode sur les données non-centrées. Dans ce cas, la première composante principale est composée de coefficients positifs. Les autres composantes principales ne sont pas exclusivement constituées de coefficients positifs afin d'assurer la décorrélation entre les différentes composantes principales.

Cependant, les sources recherchées $\mathbf{s}_j$ peuvent s'exprimer chacune en fonction des signaux observés $\mathbf{x}_i$, donc en fonction des vecteurs $\mathbf{v}_1$ et $\mathbf{v}_2$ sous la forme :

$$\mathbf{s}_j = n_{j1}\mathbf{v}_1 + n_{j2}\mathbf{v}_2. \tag{2.13}$$

Trouver les sources $\mathbf{s}_j$ devient alors équivalent à retrouver les coefficients n_{jk} pour tous les indices j et tous les indices $k = \{1, 2\}$.

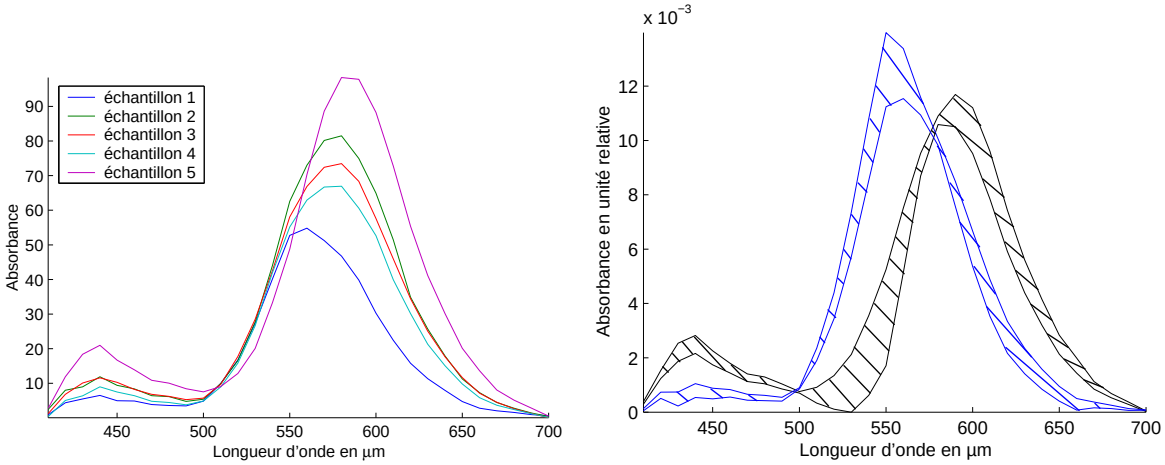
A partir de la reformulation des contraintes appliquées au modèle en fonction des coefficients n_{jk} à trouver, Lawton et Sylvestre déduisent des conditions géométriques sur ces coefficients dans le plan des coefficients $\mathbf{w}_1$ et $\mathbf{w}_2$ qui sont les vecteurs de la matrice $\mathbf{W}$ estimée par l'équation (2.11). La combinaison de toutes ces conditions géométriques conduit à l'isolement de deux régions (morceaux de droites) dans lesquelles les coefficients de l'équation (2.13) remplissent les hypothèses de départ. Chaque zone définit une bande de solutions possibles pour une des sources recherchées. Ainsi, la source réelle recherchée se situe dans la bande de solution. Le problème est alors résolu sans hypothèses fortes sur les densités de probabilité des sources.

Pour illustrer l'utilisation de cette méthode, nous reprendrons l'exemple utilisé dans l'article original [74]. Un chimiste enregistre les courbes spectrophotométriques de cinq échantillons de matière provenant d'un même processus de production industrielle. Ces spectres sont représentés sur la figure 2.8(a). Ces échantillons devraient normalement être identiques puisque obtenus par le même processus. Or de fortes variations sont visibles d'un échantillon à l'autre. Le chimiste aimerait connaître l'origine de ces variations pour pouvoir les éliminer. Pour cela, il désire connaître les formes des sources sous-jacentes à ces variations. L'application de la méthode décrite ci-dessus fournit les bandes de solutions pour les deux sources présentes. Elles sont visibles sur la figure 2.8(b). Grâce à ces résultats, le chimiste en conclut qu'une des sources est bien l'un des composés utilisés pour la fabrication des matériaux, mais que la seconde source ne correspond pas au deuxième composé utilisé. Des impuretés se sont introduites dans le second composé. Des vérifications des conditions de stockage ou d'acheminement du second composé sont donc nécessaires.

Bien que s'étant révélée d'une aide précieuse dans l'exemple cité juste avant, cette méthode n'en reste pas moins limitée. Tout d'abord, seules deux sources peuvent être supposées. De plus, elle fournit des bandes de solutions qui sont très sensibles à des mesures éventuellement erronées.

Des extensions ont été développées, tout d'abord pour des mélanges à trois sources par Ohta [97], puis pour le cas multidimensionnel par Sasaki et ses collaborateurs [117]. Mais ces techniques ne sont réellement efficaces que pour des mélanges de deux ou trois sources de par la forte augmentation de la dimension des variables de décision lorsque le nombre de sources recherchées s'accroît. La recherche du minimum de la fonction objectif est alors laborieuse.

La méthode originale de Lawton et Sylvestre a succité de nombreuses recherches et a donné naissance à un ensemble plus large de méthodes d'analyse multidimensionnelle regroupées sous l'appellation des



(a) (b)

FIG. 2.8 – Applications de l'algorithme de Lawton et Sylvestre : (a) Courbes spectrophotométriques enregistrées sur cinq échantillons de matière (b) Bandes spectrales sources estimées : les zones hachurées en bleu et en noir correspondent aux bandes spectrales autorisées pour la première et la deuxième source spectrale respectivement (exemple tiré de [74])

méthodes par automodélisation (Self-Modeling Curve Resolution ou SMCR en anglais) [74]. De nombreuses extensions, dont nous allons décrire très brièvement le fonctionnement des principales, ont été appliquées à la spectroscopie.

Afin de réduire l'intervalle possible de solutions, des contraintes ont été ajoutées au modèle de Lawton et Sylvestre, comme par exemple des contraintes physiques ou chimiques, ou sur la forme particulière des spectres recherchés, ou sur des dissimilitudes maximales entre spectres.

Dans [115], les auteurs utilisent une contrainte physique liée à leur application de la spectroscopie de fluorescence, à savoir la contrainte de la constante de Stern-Volmer. Cette contrainte stipule que, dans un système chimique pour lequel les spectres de la matrice de données $\mathbf{X}$ sont acquis en faisant varier la concentration d'un quencheur[10] (aussi appelé suppresseur en français), la constante de Stern-Volmer de chaque conformère est indépendante de la longueur d'onde d'excitation pour un mélange équilibré des conformères. La fusion de la SMCR et de ces contraintes a conduit à l'estimation unique des spectres des deux espèces chimiques pures présentes en solution.

Dans [123], l'un des spectres sources est contraint à une forme gaussienne. Pour une paire de coefficients de mélanges, ce spectre source est généré. Il est ensuite approximé par une gaussienne. Le minimum de l'erreur de reconstruction correspond exactement à la paire des coefficients de concentration qui génère la source de forme gaussienne recherchée.

[10] chromophore non fluorescent qui capte la fluorescence émise par le reste de la solution et qui la dissipe sous forme de chaleur

Les techniques d'analyse multivariée s'appuient essentiellement sur la positivité des spectres de fluorescence et des concentrations des espèces pures pour séparer les données originales préalablement mises en forme par des prétraitements. Les propriétés statistiques des spectres de fluorescence ne sont pas exploitables, contrairement à celles de la spectroscopie raman, comme nous allons le voir dans la section suivante.

2.4 Traitements des spectres Raman

Cette section s'appuie sur la même présentation que pour la section précédente sur les traitements classiques des spectres de fluorescence.

2.4.1 Paramètres d'acquisition

En spectroscopie Raman, contrairement à la spectroscopie de fluorescence, une seule longueur d'onde excitatrice est utilisée pour une expérimentation. Cette longueur d'onde dépend de l'application et doit proposer un compromis afin d'avoir un signal Raman assez intense tout en limitant l'émission parasite de fluorescence par l'échantillon analysé. Les longueurs d'onde courtes favorisent l'intensité de la diffusion Raman, mais privilégient en même temps l'émission de fluorescence par les impuretés fluorescentes de l'échantillon. Inversement, les longueurs d'onde élevées atténuent la diffusion Raman, mais limitent également l'émission de fluoresence jusqu'à son élimination totale pour des longueurs d'onde très grandes.

La puissance du laser est fonction de l'échantillon analysé. Pour des échantillons biologiques fragiles, elle ne devra pas dépassé une dizaine de mW pour conserver l'intégrité de l'échantillon qui risquerait d'être abimé par destruction thermique ou photochimique. Cette puissance pourra descendre à quelques mW pour des instrumentations constituées de CCD extrêmement sensibles. En pratique, les puissances utilisées varient de 5 mW à 300 mW en fonction de l'application et surtout de l'échantillon étudié.

La gamme des lasers excitateurs s'étend du visible au proche infrarouge. Les plus utilisés sont :

- Les lasers He-Ne (hélium-néon) qui sont accordables sur une large gamme spectrale s'étalant de 540 nm à 640 nm. Pour éviter une émission de fluorescence trop forte, la longueur d'onde de 632.8 nm est préférée.
- Les lasers Nd-YAG (Neodymium-doped Yttrium Aluminum Garnet) à 1064 nm.
- Les laser à diode GaAlAs (arséniure de gallium et d'aluminium) opérant dans le proche infrarouge (750 nm à 850 nm).

En général, les spectres Raman sont enregistrés sur de larges gammes spectrales. La localisation et la largeur de cette bande sont tributaires du type d'échantillon analysé. Toute molécule possède des fonctions principales qui se traduisent en spectroscopie Raman par des bandes spectrales caractéristiques. La gamme

spectrale enregistrée lors d'une expérimentation est variable à la fois en position mais également en largeur, et dépend fortement des fonctions que l'expérimentateur cherche à mettre en évidence. Par exemple, la caractérisation des tissus cancéreux de la peau par spectroscopie Raman nécessite l'enregistrement de spectres allant de 600 cm^{-1} à 1800 cm^{-1} puisque les principales fonctions discriminantes entre tissu sain et tissu cancéreux sont la bande amide I en 1650 cm^{-1} et la bande amide III en 1260 cm^{-1} [31]. La bande en 850 cm^{-1}, due à des changements de la structure des polysaccharides, montre des différences d'intensité en fonction de la pathologie du tissu, et peut être également considérée comme discriminante [45].

2.4.2 Exemples

Pour des systèmes biologiques complexes, les spectres Raman sont très riches en information. Cette richesse est indispensable à l'étude des structures et fonctions biologiques des échantillons, mais peut vite se révéler désordonnée et inexploitable. Les principales raisons en sont :

- Soit des espèces chimiques différentes qui possèdent des pics Raman localisés aux même nombres d'onde, c'est-à-dire des spectres Raman qui se chevauchent.
- Soit des espèces chimiques caractérisées par des spectres Raman composés de nombreux pics Raman ; il devient difficile d'attribuer chaque pic à une espèce chimique particulière.
- Soit des espèces chimiques très actives en Raman mais de faible intérêt d'étude et qui masquent les informations d'autres espèces sous-jacentes d'intérêt primordial à l'étude d'une structure biologique.

Des phénomènes parasites, dus à la fluorescence et étudiés à la partie 2.4.4.2, s'ajoutent à ces problèmes. La lisibilité des spectres et leur comparaison entre différents points d'acquisition d'un même échantillon en sont rendues plus difficiles.

L'instrumentation d'acquisition influence également la forme des spectres par l'ajout de la réponse spectrale du système ou des décalages fréquentiels.

Un spectre Raman résulte donc d'une somme de phénomènes biologiques et instrumentaux qui lui confèrent une forme spécifique, comme nous pouvons l'observer sur la figure 2.9. Ces spectres ont été enregistrés en différents points d'un échantillon de peau paraffinée fixé sur un support de fluorine (CaF_2). Des pics Raman intenses sont localisés en 325 cm^{-1}, 890 cm^{-1}, 1063 cm^{-1}, 1133 cm^{-1}, 1172 cm^{-1}, 1296 cm^{-1}, 1418 cm^{-1}, 1441 cm^{-1} et 1463 cm^{-1}. Comme décrit dans la partie 1.4.1.2, ces pics étroits sont caractéristiques du support de fluorine et de la paraffine. Le spectre Raman théorique est donc d'intensité nulle hormis quelques pics étroits représentatifs des éléments chimiques présents dans l'échantillon. Or sur l'ensemble des spectres de la figure 2.9, mais plus particulièrement sur les faibles nombres d'onde, un signal à variation lente est superposé. Ce signal est caractéristique d'une émission de fluorescence et sera commenté dans la suite de ce chapitre et dans le chapitre 4. Les autres bosses et pics de très faibles intensité sur ces spectres traduisent la contribution de la peau dans l'échantillon analysé.

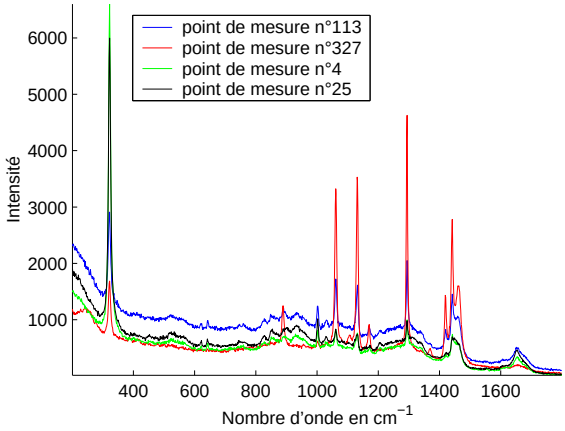


FIG. 2.9 – Exemples de 4 spectres Raman sur un ensemble de 425 spectres acquis sur un échantillon de peau paraffinée fixée sur un support en fluorine

Accéder exclusivement à l'information associée à la peau afin d'en déduire le degré de normalité nécessite l'élimination de tous ses signaux perturbateurs. Les propriétés intrinsèques des spectres Raman doivent être déterminées afin de les exploiter dans des méthodes d'analyse multidimensionnelle qui estimeront le spectre sous-jacent de la peau. Ces propriétés font l'objet du paragraphe suivant.

2.4.3 Propriétés et caractéristiques

La première propriété, définie sur des bases théoriques et expérimentales tout comme pour la spectroscopie de fluorescence, est la positivité des spectres Raman. Ainsi l'équation (2.3) est vérifiée en spectroscopie Raman également.

La deuxième propriété concerne les concentrations relatives des espèces chimiques pures. Elles sont regroupées dans la matrice des concentrations $\mathbf{A}$ du modèle linéaire instantané de l'équation (2.2) et sont des grandeurs également positives respectant l'équation (2.4).

La troisième propriété est la parcimonie des spectres Raman. La spectroscopie Raman se caractérise des autres spectroscopies optiques par la forme des spectres qu'elle enregistre. Comme décrit dans la partie 1.4.1.2, un spectre Raman peut être approximativement décomposé comme une succession de pics intenses et très étroits. Pour beaucoup de molécules, seuls quelques pics sont présents et le reste du spectre est d'intensité négligeable. Les spectres de la paraffine et de la fluorine, présentés sur les figures 1.11 et 1.12, sont deux exemples parmis tant d'autres de cette struture parcimonieuse des spectres. Cette propriété de parcimonie des spectres sources pourra servir soit comme contrainte dans certaines méthodes d'analyse multidimensionnelle, soit comme critère visuel de qualité de séparation des spectres.

2.4.4 Prétraitements

Comme dans toute expérimentation, des paramètres impondérables surviennent lors des enregistrements et polluent les spectres Raman de façon plus ou moins forte. La théorie prédit un ensemble de phénomènes parasites, facilitant ainsi leur élimination. Dans cette partie, ces divers phénomènes vont être étudiés en même temps que les méthodes pour les soustraire.

Certains prétraitements sont communs à la spectroscopie de fluorescence et ne seront donc pas réétudiés dans ce paragraphe. Il s'agit de l'élimination du courant noir étudié à la section 2.3.4.1, de la calibration de la réponse spectrale du système à la section 2.3.4.2 et de la normalisation des spectres donnée à la section 2.3.4.4.

2.4.4.1 Calibration en nombre d'onde

Le but de la calibration des nombres d'onde sur des systèmes dispersifs est d'attribuer le nombre d'onde à chaque canal individuel du détecteur CCD. Des spectres de produits tels que le benzène, le cyclohexane et le naphthalène ont été enregistrés dans 6 laboratoires internationaux différents. Un certain nombre de bandes fines ont été localisées avec précision. Des spectres d'une ou de plusieurs de ces espèces sont enregistrés sur l'appareil à calibrer et la position de leurs bandes sont comparées à celles des références. L'alignement des raies sur celles de référence permet de calibrer l'appareil [40].

2.4.4.2 Élimination du fond de fluorescence

Le fond de fluorescence se traduit par l'ajout au spectre $\mathbf{x}_i$, acquis au point de mesure i du spécimen à analyser, d'un spectre $\mathbf{s}_i^{\mathbf{ff}}$ dénommé *ligne de base* qui est formé de bandes larges (de type variations lentes). Plusieurs méthodes d'estimation ou d'élimination de la ligne de base de fluorescence existent et sont introduites dans [92]. Nous allons en faire une rapide description dans ce qui suit. Pour toutes ces méthodes, le spectre corrigé $\mathbf{x}_i^{\mathbf{c}}$ est obtenu en soustrayant le fond de fluorescence estimé $\mathbf{s}_i^{\mathbf{ff}}$ du spectre original $\mathbf{x}_i$:

$$\mathbf{x}_i^{\mathbf{c}} = \mathbf{x}_i - \mathbf{s}_i^{\mathbf{ff}}, \quad i = 1, \ldots, N_{xy}.$$

La transformée en ondelette discrète a été appliquée à la spectroscopie dans le but de séparer le fond de fluorescence $\mathbf{s}_i^{\mathbf{ff}}$ du spectre Raman proprement dit de l'échantillon analysé [34]. La décomposition d'un spectre $\mathbf{x}_i$ enregistré au point d'acquisition i de l'échantillon par cette méthode est illustrée sur la figure 2.10. La ligne de base de fluorescence $\mathbf{s}_i^{\mathbf{ff}} \in \mathbb{R}^{N_\Lambda}$ est compressée dans le vecteur $\mathbf{c}_1 = [c_{11}, \ldots, c_{1k}, \ldots, c_{1n_i}]^T \in \mathbf{R}^{n_i}$ des coefficients d'approximation de la transformée, où n_i est le nombre d'éléments de $\mathbf{c}_1$, puisqu'il est usuellement établi que ce fond de fluorescence est défini par des variations lentes :

$$\mathbf{s}_i^{\mathbf{ff}} = \phi^{1T}\mathbf{c}_1$$

où $\phi^1 = [\phi^1_1, \ldots, \phi^1_\Lambda, \ldots, \phi^1_{N_\Lambda}] \in \mathbb{R}^{n_i \times N_\Lambda}$ est composée par les vecteurs $\phi^1_\Lambda = [\phi^1_{1\Lambda}, \ldots, \phi^1_{k\Lambda}, \ldots, \phi^1_{n_i\Lambda}]^T \in \mathbb{R}^{n_i}$ et $\phi^1_{k\Lambda} = 2^{\frac{1}{2}}\phi(2\Lambda - k)$ où ϕ est une fonction d'échelle. Cette association du fond de fluorescence à cette transformation est repérée sur la figure 2.10 par un cercle rouge. Le reste du spectre, c'est-à-dire les informations utiles et indispensables à l'analyse spectrale, est quant à lui transformé dans les vecteurs $\mathbf{d}_l = [d_{l1}, \ldots, d_{lk}, \ldots, d_{ln_i}]^T \in \mathbb{R}^{n_i}, l = 1, \ldots, N_{max}$ de coefficients de détails, où N_{max} représente le nombre total de décompositions successives appliquées sur $\mathbf{x}_i$, puisque les pics Raman sont des signaux composés de pics étroits. Le signal $\mathbf{x}^\mathbf{c}_i$ corrigé de la ligne de base est reconstruit par :

$$\mathbf{x}^\mathbf{c}_i = \sum_{l=1}^{N_{max}} \psi^{lT}\mathbf{d}_l$$

où $\psi^l = [\psi^l_1, \ldots, \psi^l_\Lambda, \ldots, \psi^l_{N_\Lambda}] \in \mathbb{R}^{n_i \times N_\Lambda}$ est composée par les vecteurs $\psi^l_\Lambda = [\psi^l_{1\Lambda}, \ldots, \psi^l_{k\Lambda}, \ldots, \psi^l_{n_i\Lambda}]^T \in \mathbb{R}^{n_i}$ et $\psi^l_{k\Lambda} = 2^{\frac{l}{2}}\psi(2^l\Lambda - k)$ où ψ est l'ondelette mère. Cette reconstruction est schématisée sur la figure 2.10 par les coefficients contenus dans l'ellipse bleue. Les transformées sont calculées par la transformée en ondelette rapide et les ondelettes de Sturm-Liouville de type II ont été utilisées dans [34] pour réaliser 8 décompositions successives du signal. Pour que cette méthode fonctionne, les transformées du fond de fluorescence et du spectre utile doivent être nettement séparées [34]. L'efficacité de ces méthodes dépend donc de la méthode de décomposition retenue et de la forme du fond de fluorescence. Cette méthode de correction de la ligne de base reste marginale puisqu'à notre connaissance seul l'article [34] l'exploite sur un jeu de données simulées. Les résultats qui y sont présentés ne sont d'ailleurs pas très convaincants de son efficacité.

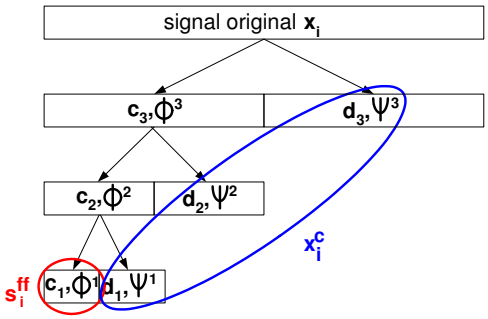


FIG. 2.10 – Décomposition d'un spectre Raman $\mathbf{x}_i$ en un spectre de fond de fluorescence $\mathbf{s}^\mathbf{ff}_i$ et en un spectre corrigé $\mathbf{x}^\mathbf{c}_i$ par la transformée en ondelettes

Lorsque les contenus fréquentiels de la ligne de base et du signal original sont distinctement séparés, un filtrage passe-haut permet d'éliminer l'information fréquentielle de la ligne de base. Or en pratique, les spectres Raman activent l'ensemble du domaine de Fourier, mélangeant des informations à basses fréquences et des informations à hautes fréquences. De même la ligne de base possède des informations fréquentielles sur l'ensemble du domaine de Fourier, bien que les composantes à hautes fréquences soient de faibles intensités. Le retrait de la ligne de base par des méthodes fréquentielles est incorrect puisque de l'information du spectre original est éliminée, et de l'information de la ligne de base est laissée.

Un autre type de méthode décompose les spectres en deux parties orthogonales par projection des spectres sur la matrice des concentrations. La partie résidante dans l'espace de la matrice des concentrations correspond au spectre corrigé, et la partie située dans l'espace orthogonal représente le fond de fluorescence. Comme indiqué dans [84], cette méthode peut être vue comme un outil d'élimination de la ligne de base mais dépend de la composition de l'échantillon au point de mesure. Dans [84], il est clairement observable que l'élimination du fond de fluorescence dépend fortement de la concentration du principe actif. Cette méthode recquiert la réalisation de mesures de spectres sur des échantillons dont la matrice de concentrations est parfaitement connue.

D'autres méthodes se basent sur l'allure de courbes à variations lentes, caractéristique du fond de fluorescence. Cette forme typique est modélisable par un polynôme :

$$\mathbf{s}_i^{\mathbf{ff}} = \mathbf{M}^T \mathbf{a}^i \tag{2.14}$$

où le vecteur $\mathbf{a}^i = [a_1^i, \ldots, a_l^i, \ldots, a_{L+1}^i]^T \in \mathbb{R}^{L+1}$ définit le vecteur des coefficients du polynôme, et L est l'ordre du polynôme. La matrice $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_\Lambda, \ldots, \mathbf{m}_{N_\Lambda}] \in \mathbb{R}^{(L+1)\times N_\Lambda}$ est la matrice des puissances de nombres d'onde, avec $\mathbf{m}_\Lambda = [\bar{\nu}_\Lambda^0, \ldots, \bar{\nu}_\Lambda^l, \ldots, \bar{\nu}_\Lambda^L]^T \in \mathbb{R}^{L+1}$. L'ordre de ce polynôme est fonction de l'application et de l'intensité du fond de fluorescence. Les plus couramment rencontrés sont les ordres $L = 4$ ou $L = 5$ [118, 54]. L'estimation des coefficients de ce polynôme doit s'appuyer sur les ensembles $\mathbf{I}^i = \{I_1^i, \ldots, I_n^i, \ldots, I_N^i\} = \{\Lambda \mid x_{i\Lambda} = s_{i\Lambda}^{ff}\ ,\ 1 \leq \Lambda \leq N_\Lambda\} \in \mathbb{R}^N$ de points appartenant effectivement au fond de fluorescence réel $\mathbf{s}_i^{\mathbf{ff}} = [s_{i1}^{ff}, \ldots, s_{i\Lambda}^{ff}, \ldots, s_{iN_\Lambda}^{ff}]^T$ et $N < N_\Lambda$. La restriction $\mathbf{x}_i^r = [x_{iI_1^i}, \ldots, x_{iI_n^i}, \ldots, x_{iI_N^i}]^T \in \mathbb{R}^{N^i}$ du vecteur $\mathbf{x}_i$, et la restriction $\mathbf{M}^r = [\mathbf{m}_{I_1^i}, \ldots, \mathbf{m}_{I_n^i}, \ldots, \mathbf{m}_{I_N^i}] \in \mathbb{R}^{L\times N^i}$ de la matrice $\mathbf{M}$ à l'ensemble $\mathbf{I}$ sont utilisées pour estimer le vecteur des coefficients $\mathbf{a}^i$ par des moindres carrés [50] :

$$\mathbf{a}^i = \left(\mathbf{M}^r \mathbf{M}^{rT}\right)^{-1} \mathbf{M}^r \mathbf{x}_i^r \tag{2.15}$$

Le fond de fluorescence $\mathbf{s}_i^{\mathbf{ff}}$ est ensuite estimé par l'équation (2.14). En d'autres termes, les pics Raman caractéristiques de l'échantillon analysé ne doivent pas influencer cette estimation. La définition de l'ensemble $\mathbf{I}^i$ est réalisée par l'utilisateur. Les caractères ingrat et long de cette tâche, lorsqu'elle doit être réitérée pour de nombreux spectres, ont crédité les recherches vers l'automatisation de la sélection des points ou vers le développement de fonctions coût insensibles à la présence des pics Raman.

Dans les travaux de Lieber [81], la procédure classique d'estimation polynômiale par moindres carrés de l'équation (2.15) est complétée par une étape de réassignation qui rend inutile toute intervention de l'utilisateur. Lors d'une première itération $t = 1$ de l'algorithme par les équations (2.15) et (2.14) calculées pour la matrice $\mathbf{M}$ et le vecteur $\mathbf{x}_i$, un vecteur ${}^t\mathbf{s}_i^{\mathbf{ff}}$ est estimé. Puisque l'estimation est réalisée par des moindres carrés, les pics Raman influencent fortement cette estimation et le polynôme correspond à une estimation grossière du fond de fluorescence. Ce constat est représenté sur la figure 2.11. C'est pourquoi les points du spectre à corriger de plus forte intensité que celle du fond de fluorescence vont être réassignés

aux valeurs du fond de fluorescence :

$$ {}^{t}x_{i\Lambda} = \begin{cases} {}^{t}s_{i\Lambda}^{ff} & \text{si } {}^{t}x_{i\Lambda} > {}^{t}s_{i\Lambda}^{ff} \\ {}^{t}x_{i\Lambda} & \text{sinon} \end{cases} \quad \text{pour } i = 1, \ldots, N_{xy} \text{ et } \Lambda = 1, \ldots, N_{\Lambda}. \tag{2.16} $$

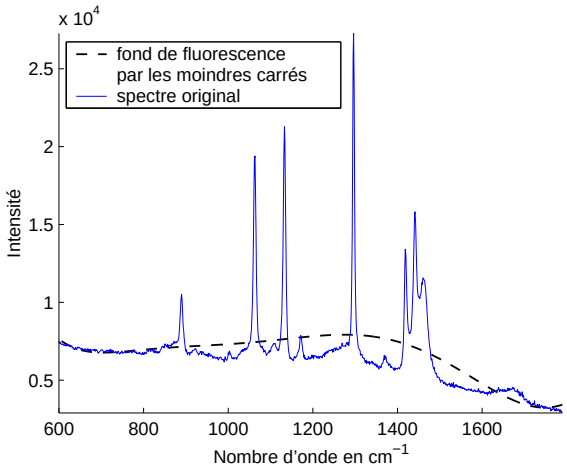


FIG. 2.11 – Estimation du fond de fluorescence par les moindres carrés sur le spectre original à corriger

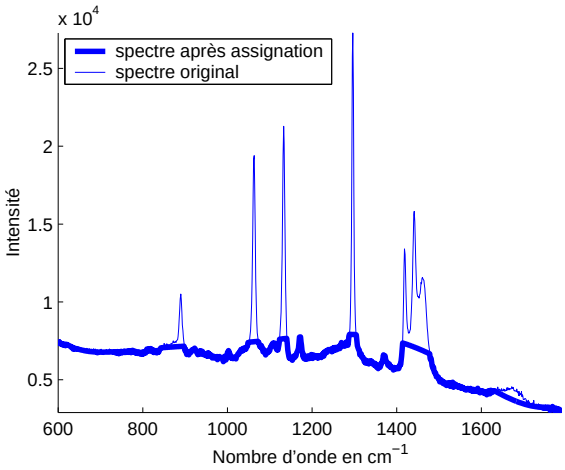


FIG. 2.12 – Nouvelle assignation du spectre à corriger en fonction de l'estimation des moindres carrés du fond de fluorescence

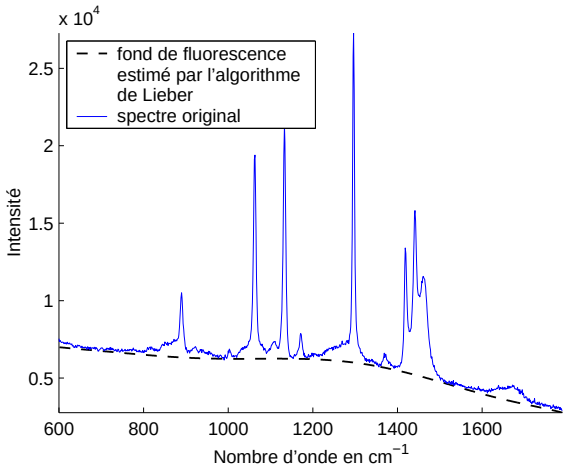


FIG. 2.13 – Estimation finale du fond de fluorescence par l'algorithme de Lieber

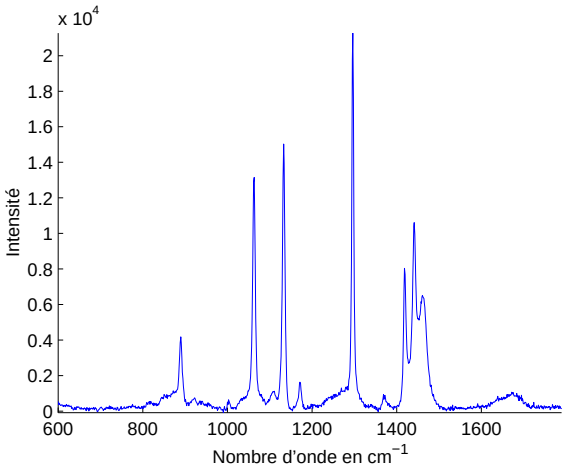


FIG. 2.14 – Spectre Raman corrigé de la ligne de base estimée par l'algorithme de Lieber

La figure 2.12 rend compte de cette transformation. Une deuxième itération $t = 2$ est alors appliquée sur ce nouveau spectre ${}^{t}\mathbf{x}_i$ et un deuxième fond de fluorescence ${}^{t}\mathbf{s}_i^{\mathbf{ff}}$ est estimé par les équations (2.15) et (2.14). Une nouvelle correction est appliquée sur ${}^{t}\mathbf{x}_i$ par l'équation (2.16). Ainsi, les pics Raman intenses sont écartés petit à petit de la procédure d'estimation et le polynôme tend vers une estimation de plus en plus réelle du fond de fluorescence. Une centaine d'iterations suffit pour assurer la convergence de

l'algorithme. Le polynôme final estimé par cette procédure est présenté sur la figure 2.13 et prouve que cet algorithme basé sur des principes simples estime le fond de fluorescence avec une grande précision. Le spectre Raman corrigé par soustraction de la ligne de base estimée est présenté sur la figure 2.14. Les fluctuations du spectre ont été annulées et la base du spectre est confondue avec la droite d'intensité nulle.

2.4.4.3 Élimination des rayons cosmiques

Les détecteurs CCD enregistrent parfois des raies d'une extrême intensité mais localisées en une seule longueur d'onde. Cette longueur d'onde varie d'un spectre à un autre, et tous les spectres ne présentent pas ces raies. Ce phénomène aléatoire est connu sous le nom de rayons cosmiques. Ces artéfacts doivent être éliminés des enregistrements car ils gênent l'interprétation et le traitement des spectres enregistrés. La figure 2.15 présente quelques exemples de raies dus à l'enregistement de rayons cosmiques par les CCD. Ces spectres ont été acquis sur un échantillon de peau paraffinée sur support de fluorine. Ces raies sont identifiés par des flèches rouges situées au dessus d'eux.

Une méthode courante consiste à éliminer de l'analyse tous les spectres traduisant la présence de rayons cosmiques [29]. Cette méthode radicale s'avère dangereuse puisque les rayons cosmiques apparaissent fréquemment dans les enregistrements. La matrice des données à disposition peut donc être considérablement emputée et une grande quantité d'informations utiles sera négligée.

Une méthode moins drastique consiste à enregistrer N spectres $\mathbf{x}_i^n = [x_{i1}^n, \ldots, x_{i\Lambda}^n, \ldots, x_{iN_\Lambda}^n]^T \in \mathbb{R}^{N_\Lambda}$, $n = 1, \ldots, N$ en un même point de mesure i. Les rayons cosmiques apparaissant de façon aléatoire en nombre d'onde et en intensité, la probabilité qu'un rayon cosmique localisé en un même nombre d'onde soit enregistré en plusieurs spectres est quasiment nulle. Le spectre médian $\mathbf{x}_i = [x_{i1}, \ldots, x_{i\Lambda}, \ldots, x_{iN_\Lambda}]^T \in \mathbb{R}^{N_\Lambda}$ est calculé pour chaque nombre d'onde $\Lambda = 1, \ldots, N_\Lambda$ sur l'ensemble des spectres issus du même point de mesure i :

$$x_{i\Lambda} = \begin{cases} x_{i\Lambda}^{P\left(\frac{N+1}{2}\right)} & \text{si } N \text{ est impair} \\ \frac{x_{i\Lambda}^{P\left(\frac{N}{2}\right)} + x_{i\Lambda}^{P\left(\frac{N+1}{2}\right)}}{2} & \text{si } N \text{ est pair} \end{cases}$$

où $P(x)$ est une fonction qui réarrange les éléments $x_{i\Lambda}^n$ par ordre croissant pour les différents n. Lorsqu'un seul enregistrement par point est disponible, cette dernière technique n'est pas applicable. De même lorsque seuls deux spectres ont été acquis en un même point de mesure, seule l'intensité la plus faible est conservée pour chaque nombre d'onde entre les deux spectres. Un filtrage médian par fenêtrage peut s'avérer efficace. Une fenêtre de largeur m échantillons balaie le spectre au voisinage des pics cosmiques. En chaque position de la fenêtre, le médian de la fenêtre est calculé et affecté en son point milieu. Les pics cosmiques disparaissent grâce à leur très petite largeur spectrale. Par contre, cette méthode ne doit surtout pas être appliquée en des bandes spectrales caractérisées par des pics Raman très étroits car l'effet passe-bas du filtre médian atténue fortement ces pics [122].

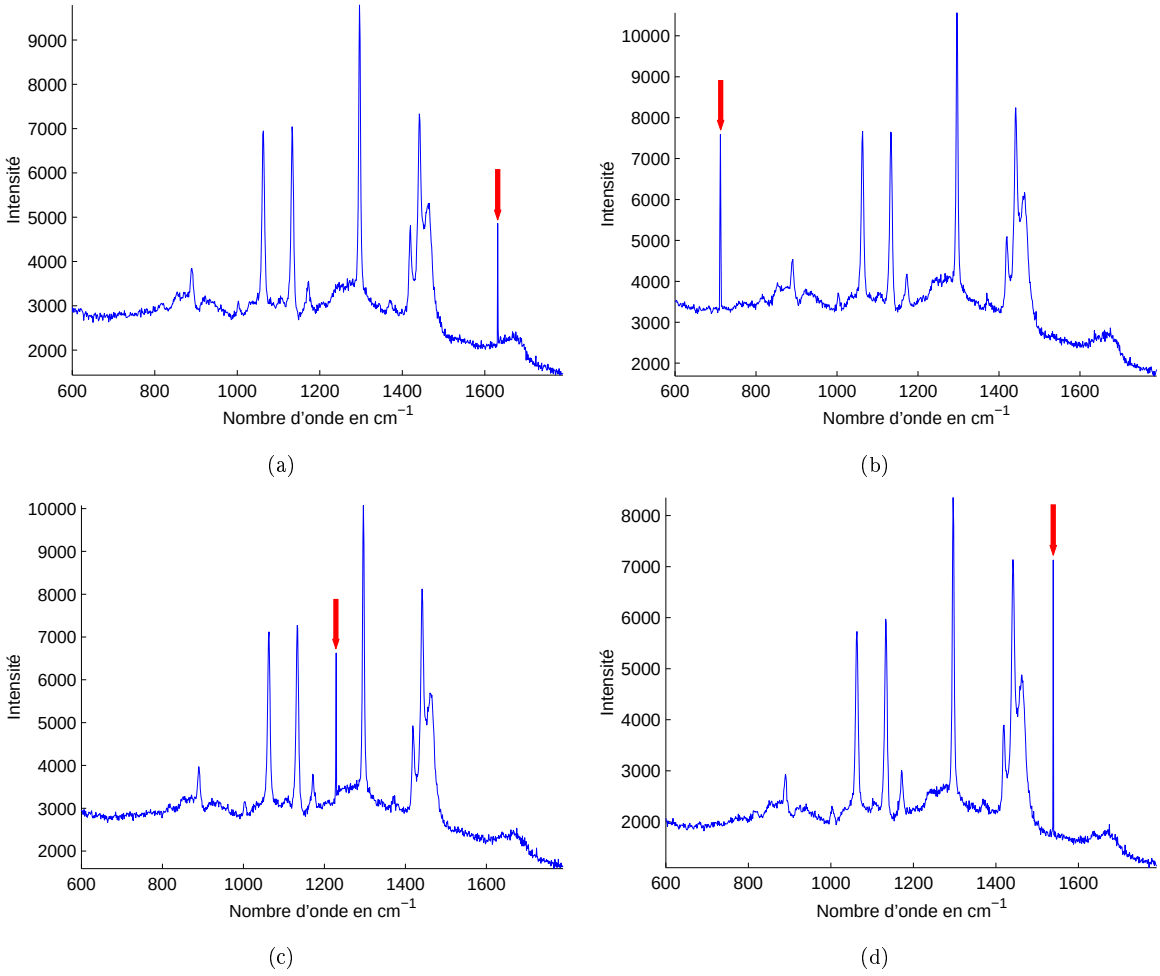


FIG. 2.15 – Spectres Raman d'un échantillon de peau paraffiné sur support de fluorine bruités par des rayons cosmiques signalés par des flèches rouges

2.4.4.4 Élimination du spectre du support de fixation de l'échantillon

La correction du fond spectral du support de fixation de l'échantillon est couramment réalisée par les spectroscopistes en effectuant une simple soustraction. Le spectre du support, mesuré seul lors d'une manipulation indépendante, est tout simplement soustrait au spectre enregistré de l'échantillon à analyser [118].

En microbiologie, les micro-organismes sont élevés sur des milieux de culture afin d'être nourris. Ces supports sont actifs en Raman et gênent le clinicien dans son diagnostic. L'élimination du spectre $\mathbf{s^s}$ des supports est nécessaire afin d'isoler le spectre $\mathbf{x_i^o}$ de la bactérie et faciliter son identification. Dans [90], les auteurs proposent une méthode d'élimination du spectre $\mathbf{s^s}$ du milieu de culture basée sur une représentation vectorielle des spectres. Le vecteur du spectre enregistré $\mathbf{x}_i$ possède une composante $\mathbf{x_i^p}$ parallèle au vecteur du milieu de culture $\mathbf{s^s}$, et une composante $\mathbf{x_i^o}$ qui lui est orthogonale. Cette

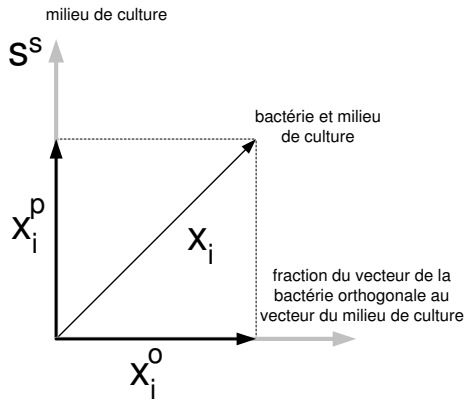


FIG. 2.16 – Représentation vectorielle du spectre du milieu de culture et du spectre enregistré

représentation est schématisée sur la figure 2.16. En projetant le spectre original $\mathbf{x}_i$ sur le spectre du milieu de culture $\mathbf{s^s}$, la composante parallèle $\mathbf{x}_i^\mathbf{p}$ du spectre original $\mathbf{x}_i$ est calculée :

$$\mathbf{x}_i^\mathbf{p} = \frac{\mathbf{x}_i^T \mathbf{s^s}}{\mathbf{s^s}^T \mathbf{s^s}} \mathbf{s^s}.$$

En soustrayant cette composante $\mathbf{x}_i^\mathbf{p}$ au spectre original $\mathbf{x}_i$, la composante orthogonale $\mathbf{x}_i^\mathbf{o}$ est cette fois-ci déterminée :

$$\mathbf{x}_i^\mathbf{o} = \mathbf{x}_i - \mathbf{x}_i^\mathbf{p}.$$

Mais cette méthode a plusieurs inconvénients. Tout d'abord, le spectre exacte des bactéries n'est pas estimé puisque rien ne laisse prédire que le spectre pur de la bactérie et le spectre du milieu de culture sont orthogonaux. En pratique, ils ne le sont jamais. Et enfin, cette méthode recquiert l'enregistrement préalable du spectre du milieu de culture lors d'une expérience indépendante.

2.4.5 Méthodes classiques d'analyse et de traitement

Une fois les spectres corrigés des effets parasites et indésirables, il ne reste que les informations vibrationnelles des espèces chimiques présentes dans l'échantillon. La séparation des spectres sources des espèces pures doit être achevée afin d'identifier ces espèces, et leur concentrations respectives doivent en être déduites pour analyser la structure de l'échantillon. Parmi les techniques classiques de séparation des spectres Raman, certaines sont communes à la séparation des spectres de fluorescence. L'inspection visuelle, développée à la section 2.3.5.1, est la première méthode utilisée lors de l'analyse de spectres Raman puisque la présence de certains pics traduit la présence de certaines espèces dans le mélange [41]. L'ACP de la section 2.3.5.2, de par la généralité et la souplesse de ses hypothèses de travail, sert très souvent à séparer les informations des espèces chimiques pures [31]. Les méthodes par enveloppes, étudiées à la section 2.3.5.3, sont également utilisées pour estimer les spectres Raman des espèces pures puisque les contraintes de positivités sur lesquelles elles reposent sont naturelles en spectroscopie Raman

[6]. D'autres méthodes spécifiques au traitement des spectres Raman existent. Nous présentons dans la suite de cette section quelques unes de ces techniques.

2.4.5.1 Calcul de la dérivée des spectres

Un traitement classique en spectroscopie Raman est le calcul des dérivées premières ou secondes des spectres à analyser. L'intérêt d'une telle transformation est double.

Comme décrit dans la section 2.4.4.2, le fond de fluorescence est limité aux variations lentes des signaux Raman. Il est universellement connu que la dérivation est équivalente à un filtrage passe-haut. De ce fait, les variations lentes sont atténuées par ce filtrage [90, 84].

La seconde utilité de la dérivation réside dans son pouvoir séparateur. Les petites différences entre deux spectres très corrélés ne sont pas visible par une analyse visuelle. La dérivation les accentue et les révèle à l'analyste. En particulier, la dérivation sert à mettre en évidence les légères différences de positions de pics Raman sur des spectres différents, c'est à dire de séparer des pics qui se recouvrent [84].

Le désavantage de cette méthode est qu'elle est sensible au bruit qui est caractérisé par des variations rapides.

2.4.5.2 SIMPLISMA

Une méthode générale a été développée dans les années 90 par Windig [138] : SIMPLe to use Interactive Self-modeling Mixture Analysis (SIMPLISMA). Contrairement aux méthodes par enveloppes, cette technique commence par estimer la matrice de mélange, puis en déduit les spectres des sources originales. Nous pouvons remarquer que si pour un nombre d'onde donnée Λ_1 les spectres de toutes les sources, exceptée la source $\mathbf{s}_l$, sont égales à zéro, alors nous obtenons une estimation des concentrations de cette source dans chaque mélange. En effet, les mélanges s'expriment sous la forme de l'équation 2.2 :

$$x_{i\Lambda} = \sum_{j=1}^{p} a_{ij} s_{j\Lambda}.$$

Pour le nombre d'onde Λ_1, ces équations se simplifient alors sous la forme :

$$x_{i\Lambda_1} = a_{il} s_{l\Lambda_1}.$$

À $s_{l\Lambda_1}$ près, la l-ème colonne $\mathbf{a}_l$ de la matrice de mélange $\mathbf{A}$ est estimée. En déterminant $p-1$ autres nombres d'onde pour lesquels une seule source (différente des sources dont la colonne correspondante de la matrice de mélange a déjà été déterminée) à la fois est active, la matrice de mélange entière peut être estimée. Mais le problème est de développer une technique capable d'estimer ces nombres d'onde si particuliers.

Détermination du premier nombre d'onde caractéristique : Remarquons que, pour un nombre d'onde donné, si toutes les sources sont actives en même temps, alors la moyenne des données capteurs calculée pour cette longueur d'onde est généralement importante, et l'écart-type peut être quelconque. Par contre, lorsqu'une seule source est active pour un nombre d'onde donné, alors la moyenne est plus faible, et l'écart-type est relativement important si une grande diversité de mélanges est utilisée. Ainsi, en calculant l'écart-type relatif (c'est-à-dire le rapport entre l'écart-type et la moyenne) pour chaque nombre d'onde, les nombres d'onde pour lesquels une seule source est active présenteront une valeur d'écart-type relatif importante. L'estimation du premier nombre d'onde pur va donc se faire en déterminant le nombre d'onde pour lequel l'écart-type relatif est maximum. Le calcul s'effectue de la manière suivante :

$$r_1(\Lambda) = \frac{\sigma(\Lambda)}{\mu(\Lambda)} \text{ pour } 1 \leq \Lambda \leq N_\Lambda.$$

La moyenne $\mu(\Lambda)$ est calculée sur chaque nombre d'onde par :

$$\mu(\Lambda) = \frac{\sum_{i=1}^{N_{xy}} x_{i\Lambda}}{N_{xy}}$$

et l'écart-type $\sigma(\Lambda)$ par :

$$\sigma(\Lambda) = \left(\frac{\sum_{i=1}^{N_{xy}} (x_{i\Lambda} - \mu(\Lambda))^2}{N_{xy} - 1} \right)^{\frac{1}{2}}.$$

Détermination du deuxième nombre d'onde caractéristique : Une fois le premier nombre d'onde pur déterminé, il faut être capable de déterminer un deuxième nombre d'onde mais qui soit le plus différent possible du premier. Par différent, il est sous-entendu que l'écart-type relatif pour le deuxième nombre d'onde pur soit le plus décorrélé possible du premier. En fait, les écarts-type relatifs calculés précédemment vont être modulés par un coefficient traduisant la corrélation entre le premier nombre d'onde pur et le nombre d'onde considéré. Ainsi, si les deux nombres d'onde sont fortement corrélés, il est désiré que ce facteur modulant soit égal à 0, et par contre, si les nombres d'onde sont fortement décorrélés, alors il doit être égal à 1 pour conserver l'écart-type relatif intact à ce nombre d'onde. D'où l'utilisation de la matrice de corrélation autour de l'origine des données. Mais avant de calculer cette matrice, un changement d'échelle des données est nécessaire. En effet, afin de donner autant de poids à tous les nombres d'onde, les données vont être normalisées par rapport à la racine carré de la moyenne de l'énergie contenue au nombre d'onde considéré. Les données s'expriment alors sous la forme :

$$x_{i\Lambda} = \frac{x_{i\Lambda}}{(\frac{1}{N_{xy}} \sum_{i=1}^{N_{xy}} x_{i\Lambda}^2)^{\frac{1}{2}}}.$$

La matrice de corrélation entre nombres d'onde peut alors être calculée par :

$$\mathbf{C} = \frac{\mathbf{X}^T \mathbf{X}}{N_{xy}} \in \mathbb{R}^{N_\Lambda \times N_\Lambda}.$$

Et le coefficient de modulation est défini par :

$$w_2(\Lambda) = \begin{vmatrix} c_{\Lambda\Lambda} & c_{\Lambda p_1} \\ c_{p_1\Lambda} & c_{p_1 p_1} \end{vmatrix} \text{ pour } \Lambda = 1, \ldots, N_\Lambda \tag{2.17}$$

où les coefficients $c_{kl},\ k = 1, \ldots, N_\Lambda$ et $l = 1, \ldots, N_\Lambda$ sont les éléments de la matrice $\mathbf{C}$. L'indice p_1 représente l'indice du premier nombre d'onde pur défini par :

$$p_1 = \arg\max_{\Lambda} \big(r_1(\Lambda)\big).$$

Les mesures de l'écart-type relatif pour déterminer le deuxième nombre d'onde pur doivent maintenant tenir compte du facteur modulant $w_2(\Lambda)$:

$$r_2(\Lambda) = r_1(\Lambda) w_2(\Lambda).$$

L'indice p_2 pour lequel cette expression est maximale est alors recherché.

Une généralisation de l'expression (2.17) est nécessaire pour déterminer les facteurs de modulation quel que soit le nombre de nombres d'onde purs déjà estimés :

$$w_k(\Lambda) = \begin{vmatrix} c_{\Lambda\Lambda} & c_{\Lambda p_1} & \cdots & c_{\Lambda p_{k-1}} \\ c_{p_1\Lambda} & c_{p_1 p_1} & \cdots & c_{p_1 p_{k-1}} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p_{k-1}\Lambda} & c_{p_{k-1} p_1} & \cdots & c_{p_{k-1} p_{k-1}} \end{vmatrix}.$$

Estimation des sources : Une fois les indices des nombres d'onde purs estimés, la matrice de mélange $\mathbf{A}$ peut être formée en y reportant les colonnes de la matrice $\mathbf{X}$ des données correspondantes aux nombres d'onde purs. Connaissant la matrice des données $\mathbf{X}$ et la matrice des mélanges $\mathbf{A}$, la matrice des spectres purs peut être formée par :

$$\mathbf{S} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{X}.$$

Cette méthode est assez simple à programmer. Cependant son utilisation reste parfois hasardeuse puisqu'elle demande l'intervention de l'utilisateur pour spécifier quel nombre d'onde doit être choisi dans certains cas litigieux. Une bonne connaissance du phénomène à étudier est nécessaire. Les résultats varient énormément pour des choix parfois peu différents de nombre d'onde. Cette méthode exige l'existence de plages de nombres d'onde pour lesquels une seule source est active à la fois, ce qui n'est pas toujours physiquement exact.

2.4.5.3 BTEM

Une deuxième technique issue de la chimiométrie est basée sur la minimisation de l'entropie dans une bande fréquentielle cible (en anglais Band-Target Entropy Minimization ou BTEM) [24, 137, 109]. L'entropie est par définition une mesure de dispersion d'une densité de probabilité. Ainsi, la minimisation de l'entropie d'une bande spectrale va permettre d'estimer un spectre pur ayant la caractéristique spectrale la plus simple (la plus étroite) dans la bande spectrale spécifiée.

Considérons les spectres enregistrés des mélanges. Ces derniers sont regroupés dans la matrice $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_{N_{xy}}]^T \in \mathbb{R}^{N_{xy} \times N_\Lambda}$, avec N_{xy} le nombre de mélanges, N_Λ le nombre de nombres d'onde considérés, et $\mathbf{x}_i = [x_{i1}, \ldots, x_{i\Lambda}, \ldots, x_{iN_\Lambda}]^T \in \mathbb{R}^{N_\Lambda}$ le spectre enregistré au point de mesure i. Ces mélanges sont modélisés par l'équation (2.2) en respectant les mêmes notations.

Décomposition en valeurs singulières : Une décomposition en valeurs singulières (ou DVS) est réalisée sur la matrice des spectres mesurés $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

où $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_i, \ldots, \mathbf{u}_{N_{xy}}] \in \mathbb{R}^{N_{xy} \times N_{xy}}$ avec $\mathbf{u}_i = [u_{1i}, \ldots, u_{ji}, \ldots, u_{N_{xy}i}]^T \in \mathbb{R}^{N_{xy}}$, où $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_\Lambda, \ldots, \mathbf{d}_{N_\Lambda}] \in \mathbb{R}^{N_{xy} \times N_\Lambda}$ avec $d_\Lambda = [0, \ldots, 0, d_{\Lambda\Lambda}, 0, \ldots, 0]^T \in \mathbb{R}^{N_{xy}}$, et enfin où $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_\Lambda, \ldots, \mathbf{v}_{N_\Lambda}] \in \mathbb{R}^{N_\Lambda \times N_\Lambda}$ avec $\mathbf{v}_\Lambda = [v_{1\Lambda}, \ldots, v_{j\Lambda}, \ldots, v_{N_\Lambda\Lambda}]^T \in \mathbb{R}^{N_\Lambda}$. La matrice $\mathbf{D}$ est une matrice diagonale dont les éléments diagonaux $d_{\Lambda\Lambda}$ sont les valeurs singulières de $\mathbf{X}$ et sont rangés par ordre décroissant, c'est-à-dire que $d_{11} \geq d_{22} \geq \ldots \geq d_{\Lambda\Lambda} \geq \ldots$. Mais pour s'affranchir des non-linéarités introduites dans les mesures (translation de la position des bandes caractéristiques, modifications de la forme de certaines bandes), au lieu de ne conserver que les p premiers vecteurs singuliers (méthodologie identique à l'étape de blanchiment de certains algorithmes d'ACI afin d'éliminer les composantes du bruit et améliorer la convergence), les auteurs conservent un plus grand nombre de vecteurs singuliers. Plus d'information est gardée donc plus de caractéristiques spectrales sont conservées. En effet, une bande spectrale translatée à différents nombres d'onde pour différents points de mesure introduit de nouvelles composantes singulières. Si exactement p vecteurs singuliers sont conservés, alors l'information correspondante à la bande non translatée sera perdue. Une fois les vecteurs singuliers les plus significatifs choisis, ils sont projetés dans un espace de dimension plus petite pour reconstruire les spectres purs un par un.

Minimisation d'entropie : Si tous les spectres purs et leur concentrations respectives sont indépendants, alors le rang de la matrice $\mathbf{X}$ est exactement égal au nombre p d'espèces présentes dans les mélanges. Ainsi, seulement les p premiers vecteurs singuliers sont conservés. Ils recouvrent l'espace vectoriel entier pour les spectres purs $\mathbf{s}_j$, $j = 1, \ldots, p$. La décomposition DVS s'écrit alors sous la forme :

$$\mathbf{X} = \tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}^T$$

où $\tilde{\mathbf{U}} = [\mathbf{u}_1, \ldots, \mathbf{u}_p] \in \mathbb{R}^{N_{xy} \times p}$, $\tilde{\mathbf{D}} = [\mathbf{d}_1, \ldots, \mathbf{d}_p] \in \mathbb{R}^{p \times p}$, et $\tilde{\mathbf{V}} = [\mathbf{v}_1, \ldots, \mathbf{v}_p] \in \mathbb{R}^{N_\Lambda \times p}$. Une rotation $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_j, \ldots, \mathbf{t}_p]^T \in \mathbb{R}^{p \times p}$ avec $\mathbf{t}_j = [t_{1j}, \ldots, t_{lj}, \ldots, t_{pj}]^T \in \mathbb{R}^p$ permet de transformer les vecteurs $\mathbf{v}_j$ en estimation des spectres sources $\mathbf{s}_j$, $j = 1, \ldots, p$. L'équation précédente peut alors se réécrire sous la forme :

$$\mathbf{X} = \tilde{\mathbf{U}}\tilde{\mathbf{D}}\mathbf{T}^{-1}\mathbf{T}\tilde{\mathbf{V}}^T.$$

Les expressions de la matrice des concentrations et de la matrice des spectres purs en sont déduites :

$$\mathbf{A} = \tilde{\mathbf{U}}\tilde{\mathbf{D}}\mathbf{T}^{-1} \tag{2.18}$$

$$\mathbf{S} = \mathbf{T}\tilde{\mathbf{V}}^{T}. \tag{2.19}$$

Cette matrice $\mathbf{T}$ est obtenue par un algorithme de minimisation d'entropie basé sur une procédure d'optimisation globale non-linéaire de recuit simulé [27]. Afin de s'affranchir des problèmes des non-linéarités spectrales, un nombre $q > p$ de vecteurs singuliers est peut être utilisé dans les équations (2.18) et (2.19).

L'approche par minimisation de l'entropie dans une bande spectrale cible effectue l'estimation des spectres purs un par un, de la même manière que celle décrite par l'équation (2.19) :

$$\mathbf{s}_j = \tilde{\mathbf{V}}\mathbf{t}_j.$$

Cependant, elle offre plus de liberté à l'utilisateur. En effet, l'inspection visuelle des vecteurs singuliers permet à l'utilisateur de choisir une bande spectrale caractéristique d'une source (forte intensité du vecteur singulier dans cette bande). Cette bande va ensuite être conservée par l'algorithme qui va chercher à reconstruire la source la moins dispersée autour de cette bande spectrale.

Fonction objectif : La procédure de reconstruction étant décrite, nous allons maintenant introduire la fonction objectif utilisée. Comme spécifié précédemment, cette fonction doit minimiser l'entropie des sources estimées. De plus, deux contraintes fondamentales au bon fonctionnement de l'algorithme sont nécessaires : la positivité des sources et des concentrations. La fonction objectif de BTEM s'écrit donc :

$$Q^{BTEM} = -\sum_{\Lambda=1}^{N_\Lambda} h_\Lambda ln(h_\Lambda) + P(\mathbf{s}_j, \mathbf{a}_j)$$

où $-\sum_{\Lambda=1}^{N_\Lambda} h_\Lambda ln(h_\Lambda)$ représente le terme de minimisation de l'entropie, et la fonction P est la fonction de pénalité assurant la positivité des concentrations et des spectres. Nous ne développerons pas ici les expressions de h_Λ et de P qui sont complexes et qui peuvent varier en fonction de l'application considérée [24, 137].

Bien que cette technique permette dans certains cas d'estimer des sources assez fortement recouvertes spectralement, elle possède de gros inconvénients. L'utilisateur doit intervenir dans la procédure d'estimation en indiquant à l'algorithme quelle bande spectrale utilisée pour l'estimation d'un spectre pur et le nombre de sources à estimer. De plus, la fonction de pénalisation s'écrit à l'aide de six paramètres. Le réglage de ceux-ci pour une bonne estimation peut être particulièrement rébarbatif et difficile.

2.5 Conclusion

Les méthodes chimiques classiques d'analyse d'échantillons biologiques présentent des limitations dues à leur lenteur et à la complexité de leurs protocoles. Les spectroscopies de fluorescence et Raman per-

mettent de pallier ces difficultés et fournissent des informations vibrationnelles fondamentales sur la nature des molécules composant les échantillons. Cependant, les spectres sont enregistrés à partir de mélanges de plusieurs espèces chimiques différentes. L'analyse quantitative de ces spectres est donc complexe. Les spectres des espèces chimiques pures constituant l'échantillon biologique doivent néanmoins être extraits de la masse considérable d'information enregistrée par les spectroscopies optiques afin de pouvoir en identifier les composantes chimiques. De ces spectres purs doivent être estimés les profils de concentration de ces espèces en chaque point de l'échantillon analysé.

Les propriétés physiques communes des spectroscopies de fluorescence et Raman ont conduit à l'élaboration d'une modèle linéaire et instantané des spectres enregistrés. Ces spectres possèdent des propriétés structurelles et d'acquisition différentes. Certains prétraitements sont donc exclusifs à une spectroscopie, comme par exemple l'élimination du fond de fluorescence sur les spectres Raman. Mais certains prétraitements de mise en forme des spectres sont communs comme la normalisation des spectres.

La séparation des spectres originaux est réalisée par des méthodes d'analyse multivariées qui s'appuient sur les propriétés propres à chaque spectroscopie. En fluorescence, l'ACP et les méthodes par enveloppes s'adaptent particulièrement bien aux formes évasives des spectres. Par contre, la structure de pics étroits et intense entrecoupés de zones sans signal en spectroscopie Raman privilègie des techniques basées sur la parcimonie et sur la détection de pics particuliers, comme réalisées par SIMPLISMA et BTEM. Mais le manque de précision de certaines méthodes (en particulier les méthodes par enveloppes), les informations *a priori* très restrictives, ou le manque d'autonomie des méthodes proposées (intervention nécessaire d'un expert au cours du processus d'estimation pour BTEM) rendent ces méthodes peu attractives. L'exploitation de propriétés plus générales des signaux peut mener vers des algorithmes plus généraux et autonomes. Par exemple, l'exploitation de la positivité des spectres purs et de leur profil de concentration conjointement à la minimisation de l'erreur de reconstruction des spectres mènent vers l'élaboration de techniques générales pour l'estimation des sources et des concentrations des espèces chimiques pures en spectroscopie de fluorescence dans le chapitre suivant. De même, en spectroscopie Raman, l'exploitation de l'indépendance statistique mutuelle des sources recherchées va nous mener à nous concentrer sur les techniques d'Analyse en Composantes Indépendantes (ACI) dans le dernier chapitre de ce mémoire.

Chapitre 3

Application de la Factorisation en Matrices Non-négatives (FMN) à la spectroscopie de fluorescence

Sommaire

3.1 Introduction

Les spectres Raman et les spectres d'émission de fluorescence possèdent certaines propriétés différentes. Le traitement et l'analyse de ces spectres doit faire appel à des techniques de traitement numérique du signal basées sur des hypothèses différentes. Dans ce chapitre, des méthodes spécifiques de traitement des spectres de fluorescence vont être proposées.

Dans la partie 3.2, les propriétés fondamentales des spectres de fluorescence vont justifier l'étude d'un ensemble de méthodes particulières basées sur des contraintes de positivité des sources et des mélanges recherchés. La section 3.3 se concentrera sur l'utilisation de ces méthodes depuis une vingtaine d'années. La Factorisation en Matrices Non-négatives (FMN) sera l'objet la partie 3.4. Ses hypothèses, ses principaux algorithmes, et ses nombreuses applications y seront présentés. La partie 3.5 traitera de l'application de ces méthodes à la spectroscopie de fluorescence sur deux exemples issus de l'agronomie : l'étude de la composition chimique d'un grain de blé et l'analyse de la structure d'un grain d'orge. Une précision doit être faite sur le caractère novateur de cette application. Elle n'est pas novatrice en spectroscopie puisque des résultats ont déjà été publiés sur l'application des méthodes de FMN en spectroscopie de résonance magnétique nucléaire [114]. Notre travail s'insère dans un autre champ de spectroscopie, à savoir la spectroscopie de fluorescence où l'application des méthodes de FMN est nouvelle à notre connaissance.

3.2 FMN et spectroscopie de fluorescence

Le chapitre précédent nous a révélé l'existence en nombre limité de techniques de traitements spécifiques des signaux enregistrés par spectroscopie de fluorescence. L'analyse visuelle montre rapidement ses faiblesses d'analyse liées aux limites physiques de l'homme à extraire des informations mélangées. L'Analyse en Composantes Principales (ACP), qui est la méthode la plus exploitée, n'utilise pas les propriétés physiques et statistiques des spectres de fluorescence. Elle est dédiée à l'extraction de composantes orthogonales qui expliquent le plus largement possible la variance des données originales. Or, toutes les propriétés des spectres de fluorescence suggèrent l'inverse, à savoir que les composantes cachées dans les mélanges sont corrélées. Le dernier groupe de méthodes dédiées au traitement des spectres de fluorescence, les méthodes par enveloppes, exploitent plus profondément les propriétés des spectres, mais deviennent rapidement lourdes à implémenter et trop restrictives dans la mesure où les contraintes appliquées au modèle sont très fortes et limitées à l'application à un cas pratique unique. L'efficacité de ses méthodes est fortement liée à l'application considérée qui dicte les contraintes à exploiter.

Aucune de ces méthodes utilise conjointement les propriétés fortes caractérisant les spectres de fluorescence, à savoir leur forme caractéristique à variations lentes ainsi que leur positivité. L'association de ces deux contraintes est nécessaire pour mieux résoudre le modèle proposé au chapitre 2 par l'équation (2.2), page 43. C'est ce que nous nous proposons de faire dans les sections suivantes en introduisant les

travaux de Shen [119] et de Paatero [98, 99, 100] qui ont fortement influencés la Factorisation en Matrices Non-négatives ou FMN[11]. Cette technique s'accorde parfaitement avec les données issues de spectroscopie de fluorescence.

3.3 Historique de la FMN

La Factorisation en Matrices Non-négatives (FMN) résulte de recherches en sciences de l'environnement [55, 112, 57, 119, 100, 99, 98].

3.3.1 Importance de la positivité

En 1984, un sévère mais réaliste constat de Ronald Henry [55] montre les limites des algorithmes d'analyse factorielle pour l'étude des données atmosphériques. Il pointe du doigt le manque d'interprétabilité physique des solutions estimées par ces méthodes.

En effet, en sciences de l'environnement, les experts cherchent à représenter la matrice des données $\mathbf{X} \in \mathbb{R}^{N_{xy} \times N_\Lambda}$ par un produit matriciel de deux matrices $\mathbf{A}$ et $\mathbf{S}$:

$$\mathbf{X} = \mathbf{AS}. \tag{3.1}$$

Dans ce modèle, les matrices $\mathbf{S} \in \mathbb{R}^{p \times N_\Lambda}$ et $\mathbf{A} \in \mathbb{R}^{N_{xy} \times p}$ sont respectivement la matrice des sources et la matrice des intensités (ou concentrations relatives) des sources. Ce modèle est exactement identique à celui présenté dans l'équation (2.2), page 43 dans le cadre des spectroscopies Raman et de fluorescence. Les mêmes notations et définitions que pour cette équation sont conservées pour les matrices $\mathbf{X}$, $\mathbf{S}$ et $\mathbf{A}$.

Mais les méthodes d'analyse multivariée, usuellement employées ou spécialement développées pour les sciences de l'environnement, s'appuient sur l'ACP[12] qui décompose la matrice des données $\mathbf{X}$ en un produit de deux matrices $\mathbf{W} \in \mathbb{R}^{N_v \times N_{xy}}$ et $\mathbf{V} \in \mathbb{R}^{N_v \times N_\Lambda}$:

$$\mathbf{X} = \mathbf{W}^T\mathbf{V}. \tag{3.2}$$

En sciences de l'environnement, la matrice des données $\mathbf{X}$ est utilisée par l'ACP dans sa forme brute, c'est-à-dire sans centrage des données. La matrice $\mathbf{W}$ est alors estimée comme étant la matrice des vecteurs propres de la matrice de corrélation $\mathbf{R_x} = E\{XX^T\}$ des données $\mathbf{X}$. La matrice $\mathbf{V}$ est quant à elle calculée par inversion matricielle de l'équation (3.2) $\mathbf{V} = \mathbf{WX}$. Lorsque les mélanges, représentés par les lignes de la matrice $\mathbf{X}$, sont connus ou estimés comme étant formés de p sources expliquant la majorité de la variance des données originales $\mathbf{X}$ et de $N_{xy} - p$ sources liées au bruit dans le système, supposé décorrélé et blanc, seuls les p vecteurs propres associés aux p plus grandes valeurs propres de $\mathbf{R_x}$

[11] Non-negative Matrix Factorization ou NMF ou NNMF en anglais

[12] Analyse en Composantes Principales, voir le paragraphe 2.3.5.2, page 52

sont conservés pour décrire le jeu de données $\mathbf{X}$:

$$\mathbf{X} \approx \tilde{\mathbf{X}} = \tilde{\mathbf{W}}^T \tilde{\mathbf{V}} \tag{3.3}$$

avec $\tilde{\mathbf{W}} \in \mathbb{R}^{p \times N_{xy}}$ et $\tilde{\mathbf{V}} \in \mathbb{R}^{p \times N_\Lambda}$. Les seules hypothèses exploitées par l'ACP étant l'orthogonalité des composantes principales, la décorrélation et le caractère blanc du bruit, les sources estimées par cette méthode ne reflètent pas forcément la réalité physique du phénomène étudié. Elles manquent d'interprétabilité physique ou chimique.

La comparaison entre les équations (3.1) et (3.3) prouve que les matrices $\mathbf{S}$ et $\tilde{\mathbf{V}}$ représentent des bases différentes d'un même espace des sources. En réalité, la matrice $\tilde{\mathbf{X}}$ étant définie à partir de versions tronquées des matrices $\mathbf{W}$ et $\mathbf{V}$, les matrices $\mathbf{S}$ et $\tilde{\mathbf{V}}$ représentent des bases différentes d'espaces voisins. Une matrice de rotation $\mathbf{T}$ de dimensions $p \times p$ assure donc le passage d'une base à l'autre par les transformations :

$$\begin{aligned} \mathbf{S} &\approx \hat{\mathbf{S}} = \mathbf{T}\tilde{\mathbf{V}} \\ \mathbf{A} &\approx \hat{\mathbf{A}} = \tilde{\mathbf{W}}^T \mathbf{T}^{-1}. \end{aligned}$$

L'estimation de la matrice de passage $\mathbf{T}$ est suffisante pour estimer le modèle original (3.1).

Cependant, toute matrice de dimensions $p \times p$ et non singulière est une estimation possible de la matrice $\mathbf{T}$. Cette infinité de solutions possibles peut être partiellement levée en contraignant les propriétés de la matrice $\mathbf{S}$. La méthode Target Transformation Factor Analysis (TTFA) [112, 57] est une solution possible pour l'estimation de $\mathbf{T}$. Elle teste la proximité de sources tests $\mathbf{s_j^t} = [s_{j1}^t, \ldots, s_{j\Lambda}^t, \ldots, s_{jN_\Lambda}^t]^T \in \mathbb{R}^{N_\Lambda}$, $j = 1, \ldots, p$ à l'espace des sources réelles $\mathbf{S}$. Dans ce but, la matrice $\mathbf{T}$ est estimée par les moindres carrés comme la matrice de rotation entre la matrice des sources tests $\mathbf{S^t} = [\mathbf{s_1^t}, \ldots, \mathbf{s_j^t}, \ldots, \mathbf{s_p^t}]^T \in \mathbb{R}^{p \times N_\Lambda}$ et la matrice $\mathbf{V}$. La matrice des concentrations $\mathbf{A^t}$ en est déduite elle aussi par les moindres carrés calculés à partir de la matrice des données originales $\mathbf{X}$ et la matrice des sources tests $\mathbf{S^t}$. La distance euclidienne entre la matrice des données originales $\mathbf{X}$ et la matrice des données reconstruites $\mathbf{X^r} = \mathbf{A^t S^t}$ est utilisée comme critère de proximité entre les matrices $\mathbf{S}$ et $\mathbf{S^t}$. Si la proximité entre $\mathbf{X}$ et $\mathbf{X^r}$ est confirmée, les matrices $\mathbf{S^t}$ et $\mathbf{A^t}$ sont considérées comme de bonnes estimations des matrices $\mathbf{S}$ et $\mathbf{A}$. Dans le cas contraire, des affinements doivent être faits sur la matrices des sources tests $\mathbf{S^t}$. De plus, les signes des coefficients de la matrice $\mathbf{A^t}$ donnent une première indication sur la qualité des sources tests de la matrice $\mathbf{S^t}$. La connaissance approximative des profils des sources est nécessaire pour assurer le succès de la méthode. Cette contrainte est trop forte pour permettre la généralisation de son application.

Des contraintes naturelles sur les données de sciences de l'environnement ont été énoncées par Henry dans [55] afin que les algorithmes de traitement estiment des transformations physiquement et chimiquement réalistes. Parmis ces contraintes, les plus importantes sont :

(a) Le modèle proposé des données originales doit représenter ces données avec le maximum de précision possible.

(b) La matrice estimée $\mathbf{S}$ des sources doit présenter des éléments non-négatifs : $s_{j\Lambda} \geq 0$, $\forall j \in \{1, \ldots, p\}$, $\forall \Lambda \in \{1, \ldots, N_\Lambda\}$.

(c) La matrice estimée $\mathbf{A}$ des intensités doit exhiber des éléments non-négatifs : $a_{ij} \geq 0$, $\forall i \in \{1, \ldots, N_{xy}\}$, $\forall j \in \{1, \ldots, p\}$.

La prise de conscience de la nécessité de ses contraintes pour l'estimation d'un modèle cohérent des données issues des sciences de l'environnement s'est suivie du développement de plusieurs algorithmes basés sur ces contraintes par plusieurs équipes de recherche dans les années 80 et 90. Nous n'allons que succinctement décrire ceux qui ont marqué leur temps et influencé l'algorithme de FMN.

3.3.2 Analyse Factorielle avec Transformation Non-négative

En 1989, Shen et Israël décrivent dans [119] une méthode générale d'estimation des profils des sources et des intensités des sources, appelée Analyse Factorielle avec Transformation Non-négative (AFTN)[13], et basée sur des contraintes de non-négativité. Cette méthode se décompose en 5 étapes :

1. normalisation des moyennes des colonnes de la matrice des données originales $\mathbf{X}$ à l'unité afin d'accorder à chaque échantillon la même concentration relative ;
2. décomposition de la matrice $\mathbf{X}$ par ACP en deux matrices $\mathbf{W}$ et $\mathbf{V}$;
3. initialisation de l'algorithme par application d'une transformation $\mathbf{T}$ sur la matrice des sources estimées $\tilde{\mathbf{V}}$ afin d'obtenir une matrice $\tilde{\mathbf{V}}^1 = \mathbf{T}\tilde{\mathbf{V}}$ composée d'éléments non-négatifs ; la matrice des concentrations résultante $\tilde{\mathbf{W}}^1 = (\mathbf{T}^{-1})^T\tilde{\mathbf{W}}$ n'est quant à elle pour le moment pas composée d'éléments non-négatifs ;
4. procédure itérative[14] de transformation des matrices $\tilde{\mathbf{V}}^i = \mathbf{T}\tilde{\mathbf{V}}^{i-1}$ et $\tilde{\mathbf{W}}^i = (\mathbf{T}^{-1})^T\tilde{\mathbf{W}}^{i-1}$ par rotation des composantes de $\tilde{\mathbf{V}}^{i-1}$ conduisant à la plus forte réduction de négativité dans la matrice $\tilde{\mathbf{W}}^i$; cette étape conduit à une extension de l'espace des sources estimées ; cette extension doit être régulièrement compensée pour éviter une divergence continue entre l'espace des sources réelles et l'espace des sources estimées ;
5. contraction de l'espace représenté par les nouvelles composantes $\tilde{\mathbf{V}}^i$ estimées à l'étape précédente ;
6. réitération des étapes 4 et 5 jusqu'à convergence de l'espace des sources estimées vers l'espace des sources réelles.

Cet algorithme s'appuie essentiellement sur la couverture par les sources estimées de la matrice $\tilde{\mathbf{V}}^i$ de l'espace définit par les données originales de la matrice $\mathbf{X}$. L'estimation du modèle est d'autant meilleure que les données originales présentent une forte variabilité, c'est-à-dire qu'elles couvrent le maximum de l'espace défini par les sources réelles. De plus, la procédure itérative de cet algorithme dépend de deux paramètres α et β définissant les taux respectifs d'extension et de contraction des espaces des sources estimées $\tilde{\mathbf{V}}^i$. Ces paramètres sont à fixer par l'utilisateur en fonction de l'application considérée afin

[13] Factor Analysis with Non-negative Transformation ou FANT en anglais
[14] i pour l'index de l'itération

d'assurer un compromis entre la rapidité d'exécution de l'algorithme et la précision de l'estimation. Il incombe à l'utilisateur d'assurer la décroissance du paramètre β d'une itération à l'autre afin de limiter le temps de calcul de l'algorithme et d'assurer la convergence de l'algorithme.

3.3.3 Factorisation en Matrices Positives

Au début des années 90, Paatero et Tapper mènent une recherche active pour l'élaboration d'un algorithme général d'analyse de données positives par leur factorisation en matrices elles-mêmes positives [100].

Modèle : Les données sont modélisées suivant l'équation :

$$\mathbf{X} = \mathbf{AS} + \mathbf{E}. \tag{3.4}$$

Les matrices $\mathbf{A}$ et $\mathbf{S}$ ont la même signification que dans les sections antérieures. La matrice $\mathbf{E} = \{e_{ij}\ ,\ i = 1, \ldots, N_{xy}\ ,\ j = 1, \ldots, N_\Lambda\} \in \mathbb{R}^{N_{xy} \times N_\Lambda}$ est la matrice d'erreur de reconstruction ou erreur de modélisation. Elle se déduit simplement de l'équation (3.4) par :

$$\mathbf{E} = \mathbf{X} - \mathbf{AS}.$$

Un objectif possible serait de minimiser l'erreur quadratique de reconstruction des données à partir du modèle estimé, c'est-à-dire de satisfaire la condition (a) de Henry (voir page 76).

Cependant, en sciences de l'environnement, les enregistrements sont effectués plusieurs fois. En effet, de fortes variations sont observables dans certaines applications à cause de l'apparition d'un élément perturbateur des conditions d'enregistrement. Des changements radicaux des conditions atmosphériques peuvent par exemple invalider certaines mesures [78]. Plusieurs enregistrements permettent de mesurer la variabilité des différents échantillons d'un jeu de données. L'originalité des travaux de Paatero et al a été d'étoffer la première condition de Henry pour y inclure cette variabilité des mesures. Les écarts-type de chaque mesure, collectés dans la matrice $\mathbf{\Sigma} = \{\sigma_{ij}\ ,\ i = 1, \ldots, N_{xy}\ ,\ j = 1, \ldots, N_\Lambda\} \in \mathbb{R}^{N_{xy} \times N_\Lambda}$, instruit l'algorithme sur la confiance à accorder à ces mesures.

La méthode de Paatero respecte évidemment les conditions (b) et (c) de Henry (voir page 76), à savoir que les éléments de la matrice des sources $\mathbf{S}$ et de la matrice des concentrations ou intensités $\mathbf{A}$ sont positifs.

Résolution : La combinaison de la minimisation de l'erreur quadratique de reconstruction des données, conditionnée par la précision des données enregistrées, et de la positivité des éléments s_{kj} et a_{ik} des matrices $\mathbf{S}$ et $\mathbf{A}$ conduit à la résolution du problème suivant :

$$\{\mathbf{A}, \mathbf{S}\} = \arg\min_{\mathbf{A},\mathbf{S}} Q(\mathbf{E}) = \arg\min_{\mathbf{A},\mathbf{S}} \sum_{i=1}^{N_{xy}} \sum_{j=1}^{N_\Lambda} \left(\frac{e_{ij}}{\sigma_{ij}} \right)^2$$

sous les contraintes :

$$a_{ik} \geq 0 \quad \forall i \in \{1,\ldots,N_{xy}\}\ ,\ \forall k \in \{1,\ldots,p\} \tag{3.5}$$

$$s_{kj} \geq 0 \quad \forall k \in \{1,\ldots,p\}\ ,\ \forall j \in \{1,\ldots,N_{\Lambda}\}. \tag{3.6}$$

La minimisation du système ci-dessus est un problème des moindres carrés pondérés non-linéaire de par les contraintes de non-négativité (3.5) et (3.6), et par le produit des deux matrices à estimer $\mathbf{S}$ et $\mathbf{A}$. Paatero propose un algorithme qui assure la minimisation de ce problème dans [98] et [99] , et qui est appelé Factorisation en Matrices Positives ou FMP[15].

Leur algorithme va chercher à minimiser la fonction objectif suivante :

$$\begin{aligned} Q^n(\mathbf{E},\mathbf{A},\mathbf{S}) &= Q(\mathbf{E}) + P(\mathbf{A}) + P(\mathbf{S}) + R(\mathbf{A}) + R(\mathbf{S}) \\ &= \sum_{i=1}^{N_{xy}}\sum_{j=1}^{N_{\Lambda}}\left(\frac{e_{ij}}{\sigma_{ij}}\right)^2 - \alpha\sum_{i=1}^{N_{xy}}\sum_{k=1}^{p}\log a_{ik} - \beta\sum_{k=1}^{p}\sum_{j=1}^{N_{\Lambda}}\log s_{kj} \\ &\quad +\gamma\sum_{i=1}^{N_{xy}}\sum_{k=1}^{p}a_{ik}^2 + \delta\sum_{k=1}^{p}\sum_{j=1}^{N_{\Lambda}}s_{kj}^2. \end{aligned} \tag{3.7}$$

L'optimisation de cette fonction se réalise en deux étapes.

La partie principale de Q^n, à savoir $Q(\mathbf{E})$, est optimisée seule dans une première étape qui est basée sur une décomposition de Cholesky.

Une deuxième étape va cette fois-ci chercher à minimiser la fonction objectif Q^n sur la base des matrices $\mathbf{A}$ et $\mathbf{S}$ estimées à l'étape précédente. Une matrice de rotation $\mathbf{T}$ va transformer $\mathbf{A}$ et $\mathbf{S}$ en $\mathbf{AT}^{-1}$ et $\mathbf{TS}$ de manière à ce que ces nouvelles matrices minimisent la fonction objectif Q^n. Cette étape est approximativement équivalente à chercher la matrice de rotation $\mathbf{T}$ qui va minimiser la quantité $P(\mathbf{A}) + P(\mathbf{S}) + R(\mathbf{A}) + R(\mathbf{S})$, avec $\mathbf{A}$ et $\mathbf{S}$ estimées dans la première étape. Les termes $P(\mathbf{A})$ et $P(\mathbf{S})$ sont des termes qui pénalisent des éléments négatifs des matrices $\mathbf{A}$ et $\mathbf{S}$. Les termes $R(\mathbf{A})$ et $R(\mathbf{S})$ sont des termes qui régularisent l'intensité des éléments des matrices $\mathbf{A}$ et $\mathbf{S}$.

Ces deux étapes sont réitérées jusqu'à convergence de l'algorithme vers un minimum de la fonction objectif Q^n.

Limitations : Un inconvénient majeur de la FMP est la lourdeur de sa programmation qui limite les dimensions des matrices à manipuler. Dans [99], ce point est clairement dévoilé par les auteurs qui stipulent que l'algorithme est efficace jusqu'à des dimensions maximales approximativement égales à 30 observations, 300 échantillons et 10 sources sous-jacentes. La diminution de l'une de ces dimensions permet l'augmentation d'une autre. Une autre limitation est l'existence de plusieurs minima locaux pour

[15] Positive Matrix Factorization ou PMF en anglais

la fonction objectif Q^n. La convergence de l'algorithme est réputée pour être lente pour des problèmes de larges dimensions.

Malgré ces quelques inconvénients, la FMP s'est révélée efficace lors de son utilisation dans de nombreuses applications différentes [78, 108, 103]. La véracité de ses estimées et la lenteur de sa convergence ont incité la recherche de méthodes basées sur les mêmes objectifs mais exploitant des outils d'optimisation plus performants. C'est ce que nous allons présenter dans la partie suivante qui est consacrée à la Factorisation en Matrices Non-négatives (FMN).

3.4 La Factorisation en Matrices Non-négatives

Deux chercheurs de chez Bell Laboratories, Daniel Lee et Sebastian Seung, se sont inspirés des travaux sur les contraintes de positivité appliquées à une décomposition bilinéaire d'une matrice de données, mais aussi des travaux sur la perception visuelle et l'encodage des données visuelles par le cerveau. Leur constat de départ est que des études psychologiques et physiologiques ont prouvé que le cerveau décompose un objet en ses différentes parties pour faciliter sa représentation mentale. La question se pose alors de savoir comment cette représentation d'un tout en parties est réalisée [76]. La représentation d'un objet en parties signifie que l'objet se sépare en plusieurs parties. Un objet d'intensités strictement positives sera donc décomposé en parties d'intensités strictement positives.

Pour reconstruire cet objet, les parties sont recollées ensemble. Le recollage des morceaux signifie qu'il faut réunir toutes les parties pour retrouver le tout. Ainsi, les parties sont sommées pour retrouver l'objet initial. De plus, lorsqu'un même objet est observé sous diverses conditions lumineuses, ses parties sont toujours les mêmes, mais avec des coefficients de pondération différents qui modélisent la luminosité ambiante. Ces coefficients sont forcément positifs.

3.4.1 Modélisation du problème

La représentation d'un objet par ses parties induit que seules des sommes, pondérées par des coefficients positifs, entre parties sont autorisées pour reconstruire l'objet initial, et que si l'objet est positif, alors ses parties le sont également. Sous forme mathématique, ce problème, nommé Factorisation en Matrices Non-négatives ou FMN par Lee et Seung, est formulé [77] :
étant donné une matrice non-négative $\mathbf{X} \in \mathbb{R}^{N_{xy} \times N_\Lambda}$, *trouver les matrices non-négatives* $\mathbf{A} \in \mathbb{R}^{N_{xy} \times p}$ *et* $\mathbf{S} \in \mathbb{R}^{p \times N_\Lambda}$ *telles que :*

$$\mathbf{X} \approx \mathbf{AS}. \tag{3.8}$$

Par *matrice non-négative* nous entendons une matrice dont tous les éléments sont non-négatifs, comme spécifié par les équations (3.5) et (3.6). Ce modèle est évidemment identique à celui de la FMP. Dans ce mémoire, nous distinguons FMP de FMN par les méthodes d'optimisation utilisées. La FMP s'appuie

sur une optimisation par moindres carrés alternés qui se révèle lourde à mettre en oeuvre, tandis que la FMN s'est inspirée des algorithmes EM (Expectation Maximisation) [33] pour déduire des lois de mise à jour simples et rapides à calculer.

3.4.2 Algorithmes

Les fonctions objectifs initialement optimisées par Lee et Seung dans [77] sont les premières de la famille de la FMN. Le schéma d'optimisation simple et efficace qu'ils ont utilisé a été exploité par d'autres pour proposer des extensions de la FMN. Les hypothèses sur les matrices $\mathbf{A}$ et $\mathbf{S}$ diffèrent d'une méthode à une autre, conduisant nécessairement à des lois de mise à jour différentes qui conservent cependant les avantages algorithmiques instaurés par Lee et Seung. Quatre méthodes sont développées dans la suite de ce chapitre : la minimisation de la distance euclidienne, la minimisation de la divergence, la minimisation de la distance euclidienne sous contraintes de parcimonie et la minimisation de la divergence sous contraintes de localisation spatiale des parties.

3.4.2.1 Distance euclidienne

La première fonction objectif utilisée pour estimer le modèle de l'équation (3.8) est le carré de la distance euclidienne entre la matrice à factoriser $\mathbf{X}$ et le produit des matrices $\mathbf{A}$ et $\mathbf{S}$ [77]. Cette fonction, qui peut être vue comme l'erreur quadratique de reconstruction des données, s'écrit :

$$Q_1^{FMN} = \sum_{i=1}^{N_{xy}} \sum_{j=1}^{N_\Lambda} \left(x_{ij} - \sum_{k=1}^{p} (a_{ik} s_{kj}) \right)^2 . \tag{3.9}$$

Il apparaît évident que la fonction objectif Q_1^{FMN} a 0 pour borne inférieure, et qu'elle s'annule si et seulement si $\mathbf{X} = \mathbf{AS}$. Elle doit être optimisée sous les contraintes de positivité des éléments des matrices $\mathbf{A}$ et $\mathbf{S}$. Ces contraintes sont définies par les équations (3.5) et (3.6). L'optimisation de ce problème peut se faire par diverses techniques. Cependant, l'utilisation des méthodes par descente de gradient est à éviter car elles requièrent la gestion d'un pas d'adaptation dont le choix conditionne la convergence de l'algorithme. Lee et Seung ont alors cherché une méthode d'optimisation qui serait libre de tout paramètre. C'est ce qu'ils ont réussi à faire en proposant des lois de mise à jour *multiplicatives* des matrices à estimer $\mathbf{A}$ et $\mathbf{S}$, et non additives comme c'est le cas pour les algorithmes par descente de gradient. Cette méthode a été inspirée par les approches des algorithmes d'EM (Expectation Maximisation) [33] et est basée sur la manipulation de fonctions auxiliaires. Les démonstrations de la convergence vers un minimum local de la fonction Q_1^{FMN} est fournie dans [77].

Nous allons maintenant présenter l'architecture de l'algorithme. Les matrices $\mathbf{A}$ et $\mathbf{S}$ sont tout d'abord initialisées aléatoirement par des matrices qui respectent les contraintes de positivité établies par les équations (3.5) et (3.6). Cette étape reste valable pour tous les algorithmes de FMN que nous allons

développer dans la suite.

Puis la minimisation de la fonction Q_1^{FMN} se fait en deux étapes itératives. Tout d'abord, pour $\mathbf{S}$ fixée, la matrice $\mathbf{A}$ est recherchée. Ensuite, pour $\mathbf{A}$ fixée, la matrice $\mathbf{S}$ est calculée. Les lois de mises à jour des matrices $\mathbf{A}$ et $\mathbf{S}$ sont :

$$\begin{aligned} a_{ik} &= a_{ik} \frac{\sum_{j=1}^{N_\Lambda}(x_{ij}s_{kj})}{\sum_{l=1}^{p}\left(a_{il}\sum_{j=1}^{N_\Lambda}(s_{lj}s_{kj})\right)} \\ s_{kj} &= s_{kj} \frac{\sum_{i=1}^{N_{xy}}(a_{ik}x_{ij})}{\sum_{l=1}^{p}\left(s_{lj}\sum_{i=1}^{N_{xy}}(a_{ik}a_{il})\right)}. \end{aligned}$$

Les expressions précédentes sont simplifiables en considérant que les sommes des équations ci-dessus représentent les éléments de produits matriciels. Les éléments d'un tel produit entre deux matrices $\mathbf{U}$ et $\mathbf{V}$ seront notés par $(\mathbf{UV})_{gh}$ où les indices g et h représenteront respectivement l'indice de ligne et l'indice de colonne de la matrice résultante du produit des deux matrices $\mathbf{U}$ et $\mathbf{V}$. La même représentation sera utilisée pour un produit de trois matrices.

Les lois de mises à jour pour assurer la minimisation de la fonction Q_1^{FMN} se simplifient donc en [77] :

$$a_{ik} = a_{ik} \frac{(\mathbf{XS}^T)_{ik}}{(\mathbf{ASS}^T)_{ik}} \tag{3.10}$$

$$s_{kj} = s_{kj} \frac{(\mathbf{A}^T\mathbf{X})_{kj}}{(\mathbf{A}^T\mathbf{AS})_{kj}} \tag{3.11}$$

Les lois de mises à jour des équations (3.10) et (3.11) ont une forme simple et attractive qui assure la facilité de leur programmation.

3.4.2.2 Divergence

Un autre algorithme développé par Lee et Seung cherche également à minimiser une mesure de distance entre les données et le modèle, mais cette fois-ci en considérant comme fonction objectif la divergence entre les matrices $\mathbf{X}$ et $\mathbf{AS}$:

$$Q_2^{FMN} = \sum_{i=1}^{N_{xy}}\sum_{j=1}^{N_\Lambda}\left(x_{ij}\log\frac{x_{ij}}{\sum_{k=1}^{p}(a_{ik}s_{kj})} - x_{ij} + \sum_{k=1}^{p}(a_{ik}s_{kj})\right). \tag{3.12}$$

Elle se réduit simplement à la divergence de Kullback-Leibler entre $\mathbf{X}$ et $\mathbf{AS}$ si la somme des éléments de $\mathbf{X}$ est normalisée à 1, et de même pour la matrice $\mathbf{AS}$, c'est-à-dire si $\sum_{i=1}^{N_{xy}}\sum_{j=1}^{N_\Lambda} x_{ij} = 1$ et $\sum_{i=1}^{N_{xy}}\sum_{j=1}^{N_\Lambda}\sum_{k=1}^{p}(a_{ik}s_{kj}) = 1$.

De même que dans le paragraphe précédent, cette fonction objectif a 0 pour borne inférieure, s'annule si et seulement si $\mathbf{X} = \mathbf{AS}$, et doit être contrainte à la positivité des éléments des matrices $\mathbf{A}$ et $\mathbf{S}$. L'optimisation est à nouveau réalisée par des lois multiplicatives dont la convergence vers un minimum local de Q_2^{FMN} est prouvée dans [77].

Ces lois de mises à jour des matrices $\mathbf{A}$ et $\mathbf{S}$ s'expriment par :

$$\begin{aligned} a_{ik} &= a_{ik} \frac{\sum_{j=1}^{N_\Lambda} \left(\frac{x_{ij} s_{kj}}{\sum_{l=1}^{p} (a_{il} s_{lj})} \right)}{\sum_{j=1}^{N_\Lambda} s_{kj}} \\ s_{kj} &= s_{kj} \frac{\sum_{i=1}^{N_{xy}} \left(\frac{x_{ij} a_{ik}}{\sum_{l=1}^{p} (a_{il} s_{lj})} \right)}{\sum_{i=1}^{N_{xy}} a_{ik}}. \end{aligned}$$

La simplification d'écriture par des produits matriciels conduit alors aux lois de mise à jour suivantes [77] :

$$a_{ik} = a_{ik} \frac{\sum_{j=1}^{N_\Lambda} \left(\frac{x_{ij} s_{kj}}{(\mathbf{AS})_{ij}} \right)}{\sum_{j=1}^{N_\Lambda} s_{kj}} \tag{3.13}$$

$$s_{kj} = s_{kj} \frac{\sum_{i=1}^{N_{xy}} \left(\frac{x_{ij} a_{ik}}{(\mathbf{AS})_{ij}} \right)}{\sum_{i=1}^{N_{xy}} a_{ik}}. \tag{3.14}$$

Tout comme dans le paragraphe précédent, les lois des équations (3.13) et (3.14) sont simples à mettre en oeuvre.

3.4.2.3 Parcimonie et non-négativité

Patrik Hoyer s'est inspiré des travaux sur la FMN et sur le codage parcimonieux de données pour proposer un algorithme qui recherche un codage parcimonieux non-négatif (CPN) de données [58]. Il ajoute à la fonction objectif de l'équation (3.9) un terme de pénalité sur la matrice $\mathbf{S}$ afin d'assurer un compromis entre la précision de la reconstruction des données et la parcimonie des lignes de $\mathbf{S}$:

$$Q^{CPN} = \sum_{i=1}^{N_{xy}} \sum_{j=1}^{N_\Lambda} \left(x_{ij} - \sum_{k=1}^{p} (a_{ik} s_{kj}) \right)^2 + \mu \sum_{i=1}^{N_{xy}} \sum_{j=1}^{N_\Lambda} s_{ij}.$$

La constante de pénalisation μ est contrainte à la positivité. Elle contrôle le poids de la contrainte de parcimonie $\sum_{i=1}^{N_{xy}} \sum_{j=1}^{N_\Lambda} s_{ij}$. Si $\mu = 0$, la parcimonie des lignes de $\mathbf{S}$ n'est pas requise. Si μ est grande, alors les lignes de $\mathbf{S}$ doivent être fortement parcimonieuses. Les éléments des matrices $\mathbf{A}$ et $\mathbf{S}$ sont évidemment contraints à la non-négativité. Afin d'éviter la convergence de l'algorithme vers la solution évidente où $\mathbf{A}$ croît sans borne et $\mathbf{S}$ tend vers 0, une contrainte supplémentaire est que les colonnes de $\mathbf{A}$ sont de norme unité, c'est à dire que $\sum_{i=1}^{N_{xy}} a_{ik} = 1$, $\forall k \in \{1, \ldots, p\}$.

La minimisation de cette fonction se fait en deux étapes, tout comme pour les fonctions objectif précédentes. Tout d'abord pour $\mathbf{A}$ fixée, la matrice $\mathbf{S}$ qui minimise la fonction Q^{CPN} est recherchée. La règle de mise à jour qui en résulte est [58] :

$$s_{kj} = s_{kj} \frac{(\mathbf{A}^T \mathbf{X})_{kj}}{(\mathbf{A}^T \mathbf{AS})_{kj} + \mu} \tag{3.15}$$

L'équation (3.15) diffère simplement de l'équation (3.11) par l'ajout du terme de pénalisation μ au dénominateur.

Considérons maintenant que la matrice **S** est fixée. La minimisation de la fonction Q^{CPN} suivant **A** ne permet pas de dériver une loi de mise à jour de forme multiplicative à cause de la contrainte de normalisation à 1 des colonnes de **A**. Hoyer propose donc une descente de gradient classique, suivi d'une projection de la mise à jour de **A** sur l'espace des contraintes [58]. Cette procédure étant classique, nous ne nous attarderons pas dessus.

La répétition de ses deux étapes forme le cœur de l'algorithme qui converge vers un minimum local de la fonction Q^{CPN}.

Une généralisation de cette méthode est proposée dans [59]. L'utilisateur peut maintenant contraindre les lignes de **S** et/ou les colonnes de **A** à des valeurs de parcimonie choisies. La parcimonie des lignes de **S** est calculée par [59] :

$$P(\mathbf{s}_j) = \frac{\sqrt{N_\Lambda} - \frac{\sum_{\Lambda=1}^{N_\Lambda} |s_{j\Lambda}|}{\sqrt{\sum_{\Lambda=1}^{N_\Lambda} s_{j\Lambda}^2}}}{\sqrt{N_\Lambda} - 1} \tag{3.16}$$

où $P(\mathbf{s}_j)$ est la parcimonie de la j-ème ligne $\mathbf{s}_j$ de la matrice **S**. La parcimonie des colonnes de **A** est calculée par [59] :

$$P(\mathbf{a}_j) = \frac{\sqrt{N_{xy}} - \frac{\sum_{i=1}^{N_{xy}} |a_{ij}|}{\sqrt{\sum_{i=1}^{N_{xy}} a_{ij}^2}}}{\sqrt{N_{xy}} - 1} \tag{3.17}$$

où $P(\mathbf{a}_j)$ est la parcimonie de la j-ème colonne $\mathbf{a}_j$ de la matrice **A**.

Cette méthode est implémentée par un algorithme de gradient projeté, avec des contraintes de parcimonie, proche de celui présenté dans [58] si des contraintes de parcimonie sont imposées. Sinon, les règles de mise à jour sont identiques aux équations (3.10) et (3.11). C'est ce second algorithme que nous utiliserons dans la suite de ce chapitre.

3.4.2.4 Localisation spatiale

Dans [76], l'algorithme de FMN découlant de la minimisation de la divergence de l'équation (3.12) a été appliqué à une base de données d'images de visages humains. Les lignes de la matrice **S** représentaient donc les images de la base estimée à partir des images originales stockées suivant les lignes de **X**. Chaque ligne de **S** modélisait une partie bien définie du visage, comme par exemple une bouche, un œil, un nez. Cet exemple prouvait donc l'efficacité de cet algorithme de FMN à représenter un tout (un visage) en parties (la bouche, l'œil ou le nez).

Dans [80], Li et al appliquent ce même algorithme à une autre base de données de visages. La décomposition en parties distinctes des visages n'est pas observable sur cet exemple. Les résultats en sont même très différents puisque les images de la base estimée s'avoisinent à celles estimées par ACP. Cette différence d'estimation entre les deux bases de données s'explique par le fait que les images de visages utilisées dans [76] ont été correctement alignées les unes par rapport aux autres, c'est-à dire que par

exemple les yeux sont situés approximativement sur les mêmes pixels pour toutes les images de la base de données.

Face à ce constat, de nouvelles contraintes sont imposées au modèle de la FMN afin d'assurer la décomposition d'un objet en parties spatiallement localisées quel que soit l'alignement des données [80] :

- Normalisation des lignes de $\mathbf{S}$ à l'unité. Des contraintes étant ajoutées au modèle, cette dernière évite l'estimation de solutions basées sur l'indétermination d'échelle des matrices à estimer. Mathématiquement elle s'exprime par : $\sum_{j=1}^{N_\Lambda} s_{kj} = 1\ ,\ \forall k \in \{1,\ldots,p\}$.
- Maximisation de la parcimonie de la matrice $\mathbf{A}$. Cette contrainte est utile pour que seules quelques composantes soient actives pour générer l'objet entier. Ceci suppose que le maximum d'information soit disponible dans les différentes composantes de la matrice $\mathbf{S}$. Ainsi, chaque partie doit être entièrement représentée par une seule composante. Cette contrainte peut être imposée en cherchant à maximiser le nombre d'éléments non nuls de chaque ligne de la matrice $\mathbf{S}$, c'est-à-dire en minimisant l'énergie associée à chaque ligne de $\mathbf{S}$. Sous forme mathématique, ceci équivaut à minimiser la quantité $\sum_{l=k=1}^{p}(\mathbf{S}\mathbf{S}^T)_{kl}$.
- Maximisation de l'orthogonalité entre les lignes de $\mathbf{S}$. Cette contrainte vise à éviter la redondance des informations dans plusieurs composantes de la base estimée. Elle va favoriser l'attribution d'un élément de l'objet à une seule composante. Cette contrainte est achevée en minimisant la quantité $\sum_{k=1}^{p}\sum_{l=1;l\neq k}^{p}(\mathbf{S}\mathbf{S}^T)_{kl}$.
- Maximisation de l'activité des composantes. Cette contrainte est en étroite relation avec la deuxième. Maximiser l'activité des composantes signifie réduire le nombre de composantes nécessaires pour décrire les données. Cette contrainte va associer à chaque composante le maximum d'information. Cette contrainte va être réalisée en maximisant l'énergie d'activation liée à chaque composante, c'est-à-dire en maximisant la somme des carrés des éléments de chaque colonne de la matrice $\mathbf{A}$. Mathématiquement, nous allons maximiser la quantité $\sum_{l=k=1}^{p}(\mathbf{A}^T\mathbf{A})_{kl}$.

Pour assurer une décomposition d'un objet en parties localisées, la fonction objectif suivante doit être minimisée [80] :

$$Q^{FMNL} = \sum_{i=1}^{N_{xy}}\sum_{j=1}^{N_\Lambda}\left(x_{ij}\log\frac{x_{ij}}{\sum_{k=1}^{p}(a_{ik}s_{kj})} - x_{ij} + \sum_{k=1}^{p}(a_{ik}s_{kj})\right) + \alpha\sum_{k=1}^{p}\sum_{l=1}^{p}(\mathbf{S}\mathbf{S}^T)_{kl} - \beta\sum_{l=k=1}^{p}(\mathbf{A}^T\mathbf{A})_{kl} \tag{3.18}$$

où α et β sont des constantes strictement positives. La fonction précédente est soumise aux contraintes de positivité et de normalisation suivantes :

$$a_{ik} \geq 0\ ,\ s_{kj} \geq 0\ ,\ \forall i \in \{1,\ldots,N_{xy}\}\ ,\ \forall j \in \{1,\ldots,N_\Lambda\}\ ,\ \forall k \in \{1,\ldots,p\}$$

$$\sum_{j=1}^{N_\Lambda} s_{kj} = 1\ ,\ \forall k \in \{1,\ldots,p\}.$$

Des règles de mise à jour ont été dérivées des équations (3.13) et (3.14) afin de minimiser localement

la fonction objectif Q^{FMNL} :

$$a_{ik} = \sqrt{a_{ik} \sum_{j=1}^{N_\Lambda} \left(\frac{x_{ij} s_{kj}}{(\mathbf{AS})_{ij}} \right)} \tag{3.19}$$

$$s_{kj} = s_{kj} \frac{\sum_{i=1}^{N_{xy}} \left(\frac{x_{ij} a_{ik}}{(\mathbf{AS})_{ij}} \right)}{\sum_{i=1}^{N_{xy}} a_{ik}} \tag{3.20}$$

$$s_{kj} = \frac{s_{kj}}{\sum_{j=1}^{N_\Lambda} s_{kj}}. \tag{3.21}$$

Nous pouvons observer que les principales différences entre les règles d'apprentissage (3.13) et (3.19) de $\mathbf{A}$ sont l'absence de normalisation de la mise à jour par la norme de la ligne de $\mathbf{S}$ associée à la colonne de $\mathbf{A}$ qui est en train d'être actualisée, et la considération non pas de la mise à jour ainsi obtenue, mais de sa racine carrée. De même, nous constatons que les mises à jour (3.14) et (3.20) de $\mathbf{S}$ sont identiques. Les lignes de $\mathbf{S}$ sont cependant forcées à une norme unité par l'équation (3.21). Les similitudes entre ces règles d'apprentissage et celles trouvées par Lee et Seung s'expliquent par l'utilisation de simplifications pour s'affranchir des paramètres α et β dans l'expression des lois de mise à jour.

3.4.2.5 Positivité des grandeurs estimées

Pour toutes les fonctions objectifs à minimiser qui ont été présentées dans cette section, les contraintes de positivité des éléments des matrices $\mathbf{A}$ et $\mathbf{S}$ ne sont pas explicitement incluses. Mais la mise à jour de chaque élément de ces matrices est obtenue par multiplication de cet élément par un facteur qui est positif si les matrices $\mathbf{A}$ et $\mathbf{S}$ sont initialement définies par des éléments strictement positifs. Ces matrices sont donc assurées d'être positives à chaque itération de l'algorithme.

3.4.2.6 Convergence des algorithmes

Chaque méthode a été prouvée comme convergente vers un minimum local des fonctions objectif définies. Cette limitation s'explique en partie à cause de la non convexité des contraintes appliquées au modèle. Un même algorithme peut donc converger vers des solutions différentes pour des conditions initiales différentes. Il est recommandé de lancer les algorithmes plusieurs fois en différents points d'initialisation et de conserver parmi toutes les solutions, la solution qui donne la fonction objectif minimale.

Cependant, certains travaux commencent à voir le jour sur les conditions nécessaires et suffisantes à l'unicité de la solution estimée par la FMN. Parmi eux, Donoho et Stodden proposent dans [37] trois règles que doivent remplir la base de données à analyser pour assurer l'unicité de la solution estimée par l'algorithme de Lee et Seung proposé dans [76] et minimisant la fonction objectif Q_2^{FMN} de l'équation (3.12). Ces trois règles sont :

- la matrice de données $\mathbf{X}$ est générée par le modèle $\mathbf{X} = \mathbf{AS}$ pour des matrices $\mathbf{A}$ et $\mathbf{S}$ non-négatives ;
- les composantes recherchées, c'est-à-dire les lignes de la matrice $\mathbf{S}$, sont linéairement indépendantes ;
- la matrice de données est complète, c'est-à-dire que toutes les combinaisons linéaires autorisées des composantes recherchées forment la matrice des données $\mathbf{X}$.

Si la matrice de données $\mathbf{X}$ ne respectent pas toutes ses conditions, l'unicité des solutions n'est plus vraie.

L'algorithme de FMN sous contraintes de parcimonie de Hoyer présenté dans [59] possède lui aussi une première preuve théorique d'unicité de sa solution [127] à condition que les sources de la matrice $\mathbf{S}$ ne soient pas dégénérées, c'est-à-dire que les sources ne soient pas identiques, à un facteur d'échelle près.

3.4.2.7 Applications courantes

La FMN repose sur un modèle génératif général, sur des contraintes réalistes pour de nombreux domaines scientifiques, et sur une implémentation simple de ses lois d'estimation. Le seul paramètre indispensable à gérer et qui conditionne fortement la qualité des résultats est le nombre p de composantes génératrices des données. De nombreux problèmes dans diverses branches de recherche ont trouvé une solution grâce à l'application de la FMN au jeu de données à analyser.

La première application répertoriée de la FMN s'est faite sur des images de visages humains afin de les décomposer en parties [76]. Cette application est devenue classique dans le sens où elle est toujours utilisée pour comparer les différents algorithmes de FMN qui existent [59, 80].

Cette méthode s'est vue appliquée avec succès à l'analyse sémantique de textes. Dans [76], chaque élément x_{ij} de la matrice de données représente le nombre d'occurrence du i-ème mot dans le j-ème texte. Les résultats prouvent que non seulement la FMN est capable de retrouver les mots appartenant au champ sémantique du texte, mais qu'elle réussi à différencier plusieurs significations d'un même mot en les associant à différents mots forts du contexte du texte analysé.

L'analyse de fichiers sonores a aussi été prouvée comme une application possible de la FMN. Dans [121], la FMN est appliquée sur les spectres de puissance calculés par transformée de Fourier fenêtrée sur de petits intervals temporels d'un enregistrement audio pour construire un système automatique de transcription de musiques polyphoniques.

Mais la principale utilisation des algorithmes de FMN se fait dans le biomédical où de nombreux systèmes d'imagerie sont exploités pour étudier le corps humain. Des données acquises par imagerie spectroscopique du cerveau ont été traitées par la FMN dans [114]. Les spectres sources estimés sont associés au cerveau et aux muscles avoisinant. Leurs cartes de répartition, représentées par les colonnes de la matrice $\mathbf{A}$, valident ces résultats par leur véracité physique. Dans [79], des images de tomographie à

émission de positron de cœur de chien sont segmentées par la FMN en estimant les différents constituants anatomiques du cœur que sont le ventricule droit, le ventricule gauche et le myocarde. Leur activité temporelle a aussi été estimée efficacement.

Dans la suite de ce mémoire, nous nous proposons de présenter une nouvelle application de la FMN en imagerie de fluorescence afin de caractériser la structure de grains de blé et d'orge.

3.5 Application de la FMN à l'imagerie de fluorescence

3.5.1 Étude sur un grain de blé

Le grain de blé est un fruit sec qui ne s'ouvre pas spontanément. Ses parois sont soudés à la graine. Un schéma des différentes structures du grain de blé est fournie sur la figure 3.1. Il est nécessaire de moudre le grain pour détacher les enveloppes, appelées sons, et récupérer l'amande farineuse.

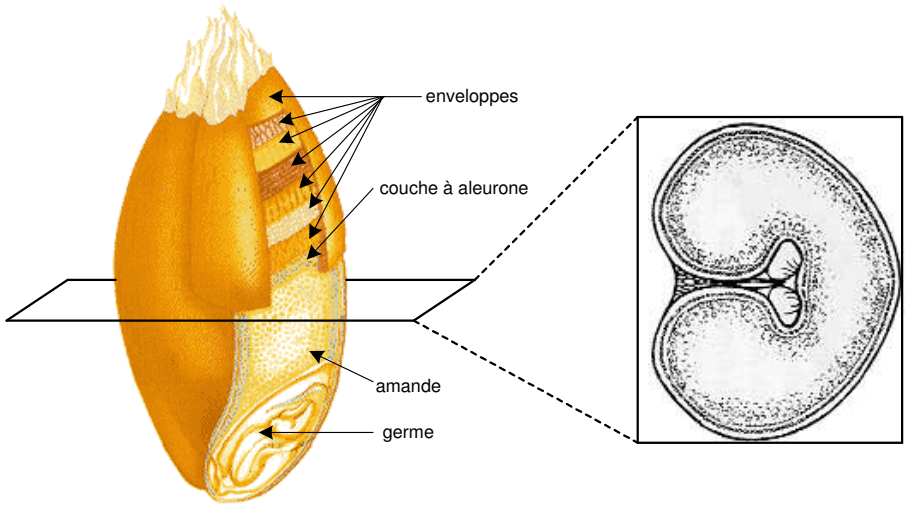


FIG. 3.1 – Représentation schématique des différentes structures d'un grain de blé

Le processus de transformation du blé en farine, correspondant à la mouture ou "cracking" des grains, n'a fait l'objet que de très peu d'études fondamentales. Il se décompose en deux processus clés :

- Le fractionnement qui correspond à la séparation entre l'amande du grain et ses enveloppes. Un fractionnement idéal doit fournir le maximum de farine sans contamination des tissus périphériques formant les sons. Le siège du fractionnement entre l'amande et les enveloppes se situe au niveau de la couche à aleurone qui est le tissu formant l'interface entre l'amande farineuse est les enveloppes.
- La fragmentation qui correspond au broyage de l'amande formée de granules d'amidons jusqu'à l'obtention de particules de farine de granulométrie suffisamment fine (inférieure à 150 μm).

L'aptitude à séparer l'amande de ses enveloppes, d'une part, et la facilité à broyer le grain, d'autre part, sont les deux critères qui déterminent la valeur meunière du blé.

L'importance qu'occupe le blé au niveau de l'industrie agroalimentaire mondiale et l'évolution des modes alimentaires poussent à optimiser la valeur meunière du blé. Cette valeur étant dépendante du fractionnement et de la fragmentation des grains, des méthodes pour caractériser ces processus ont du être développées.

Une étude sur ce sujet a consisté à caractériser les constituants de la couche de cellules à aleurone formant l'interface entre amande et enveloppes, et les espèces moléculaires (protéines et lipides) responsables de la cohésion de l'amande du grain par détection des espèces phénoliques, telles que les acides férulique et para-coumarique, par microspectroscopie d'émission de fluorescence [113].

Une fois ces bases moléculaires identifiées, la valeur meunière d'un grain de blé peut être déterminée à partir de critères spectroscopiques. Une étude similaire a été menée à partir de la spectroscopie Raman [105] mais sur la caractérisation, non pas des enveloppes, mais du contenu protéique du blé.

Un autre challenge est de mesurer la pureté des farines. Cette pureté dépend du processus de fractionnement du grain et se situe au niveau de la couche à aleurone. Un fractionnement idéal doit aboutir à un rendement maximum en farine, sans aucune contamination par les tissus de l'enveloppe. La qualité de la farine est corrélée à la teneur en sons de cette farine. La mesure de la contamination des farines par les enveloppes peut être déterminée par deux méthodes :

- L'analyse de marqueurs non spécifiques des tissus de l'enveloppe telle que le taux de cendres représentant la quantité de matières minérales, principalement localisées dans l'enveloppe du grain. Mais cette mesure ne distingue pas les matières minérales de l'albumen et celles de l'enveloppe. Le taux de cendres n'évalue donc pas réellement la pureté des fractions de mouture [35] ;
- L'analyse de marqueurs spécifiques des tissus de l'enveloppe telle que le dosage des constituants pariétaux de la couche à aleurone. Des travaux de quantification des acides phénoliques par spectroscopie de fluorescence ont montré que ces espèces moléculaires de la couche à aleurone était un meilleur marqueur de la contamination des farines par les enveloppes [107, 113].

3.5.1.1 But de l'étude

Les différentes structures biologiques du grain de blé sont caractérisables par les acides phénoliques présents dans le grain, à l'exception de l'amidon qui ne les contient qu'en très faibles concentrations. De plus, les acides phénoliques caractéristiques de la couche à aleurone sont des marqueurs efficaces de la contamination des farines par les enveloppes et ces acides sont les éléments les plus auto-fluorescents du grain de blé. Une analyse par spectroscopie de fluorescence se révèle un moyen efficace de ne sélectionner que l'émission d'information liée à ces acides phénoliques. Les spectres enregistrés sont des mélanges des spectres d'émission de fluorescence liés à chaque acide phénolique individuel.

Dans la suite de ce chapitre, nous nous proposons de présenter une méthode automatique d'extraction des spectres de fluorescence des acides phénoliques dans le grain de blé, ainsi que la cartographie des concentrations de ces espèces dans le grain de blé. Cette étude vise à aider les biologistes à analyser la composition chimique d'un grain de blé, à étudier la provenance biologique des divers constituants, et à faciliter la mesure de qualité des farines.

Notre étude diffère de celle proposée par Piot [105] puisque nous allons chercher à caractériser les enveloppes du grain de blé. Dans ses travaux, Piot a cherché à étudier le contenu protéique du blé, c'est-à-dire l'amidon du grain. Il a mis en évidence l'existence d'informations moléculaires de l'amidon à l'échelle microscopique par une analyse non destructive. Il analyse la qualité des farines par leur contenu protéique, alors que nous cherchons à mesurer le taux de contamination des farines par les sons.

3.5.1.2 Considérations expérimentales

Des échantillons de grains de blé ont été choisis parmi une série de *Triticum durum*[16] utilisée pour évaluer l'efficacité des moutures à l'INRA de Montpellier. Des sections transverses, présentées schématiquement sur la figure 3.1, cryogénisées, de $60\mu m$ d'épaisseur ont été obtenues par le procédé suivant. Les grains de blé ont été trempés pendant approximativement 4 heures dans de l'eau distillée. Ils ont ensuite été congelés à une température de -20°C. Des tranches de $60\mu m$ d'épaisseur ont enfin été coupées à -20°C en utilisant un cryostat (modèle 2800 de chez Jung). Les sections ainsi obtenues ont été placées sur des supports en quartz pour sécher à température ambiante.

Les spectres d'émission de fluorescence sont enregistrés à partir d'une section transversale d'un grain de blé par microspectrofluorimétrie confocale. Le microspectrofluorimètre est couplé à un microscope optique équipé d'un objectif $4\times$. Un laser argon ionique émet dans l'ultraviolet (UV) à 365 nm. L'échantillon est placé sous un objectif adapté pour une transmission totale des radiations UV et la position de l'échantillon est assurée par une platine motorisée. La fluorescence émise par l'échantillon au point de mesure est collectée et focalisée sur un trou confocal. Le signal d'émission est ensuite dévié vers la fente du spectrographe où le signal d'émission est projeté sur un détecteur CCD (voir le schéma de la figure 1.5 à la page 21).

Les spectres sont enregistrés dans l'intervalle spectral allant de 377 nm à 684 nm. Un total de 128 longueurs d'onde différentes est enregistré dans cet intervalle. Une image formée de 20×20 points a été acquise. La concaténation des lignes de cette image suivant la méthode proposée au paragraphe 2.2.3, à la page 40, amène finalement à considérer une matrice de données de dimensions 400×128.

La figure 3.2(a) présente quelques exemples de ces spectres. Afin de faire ressortir les diverses formes des spectres acquis, ces spectres ont été normalisés à une intensité maximale de 1, en utilisant l'équation

[16] nom scientifique donné au blé dur, qui se différencie du blé tendre, *Triticum aestivum*, par son grain à albumen vitreux, la dureté de son amande et sa plus haute teneur en protéines

(2.6) de la page 50, et sont visibles sur la figure 3.2(b). Sur la figure 3.2(c), les points d'acquisition de ces 4 spectres sont précisés par des croix de même couleur que le spectre correspondant. Le spectre rouge a été acquis au niveau du faisceau vasculaire du grain, le spectre bleu au niveau de l'enveloppe entourant le faisceau, le spectre vert dans la couche à aleurone et le spectre noir à l'interface entre la couche à aleurone et l'amande.

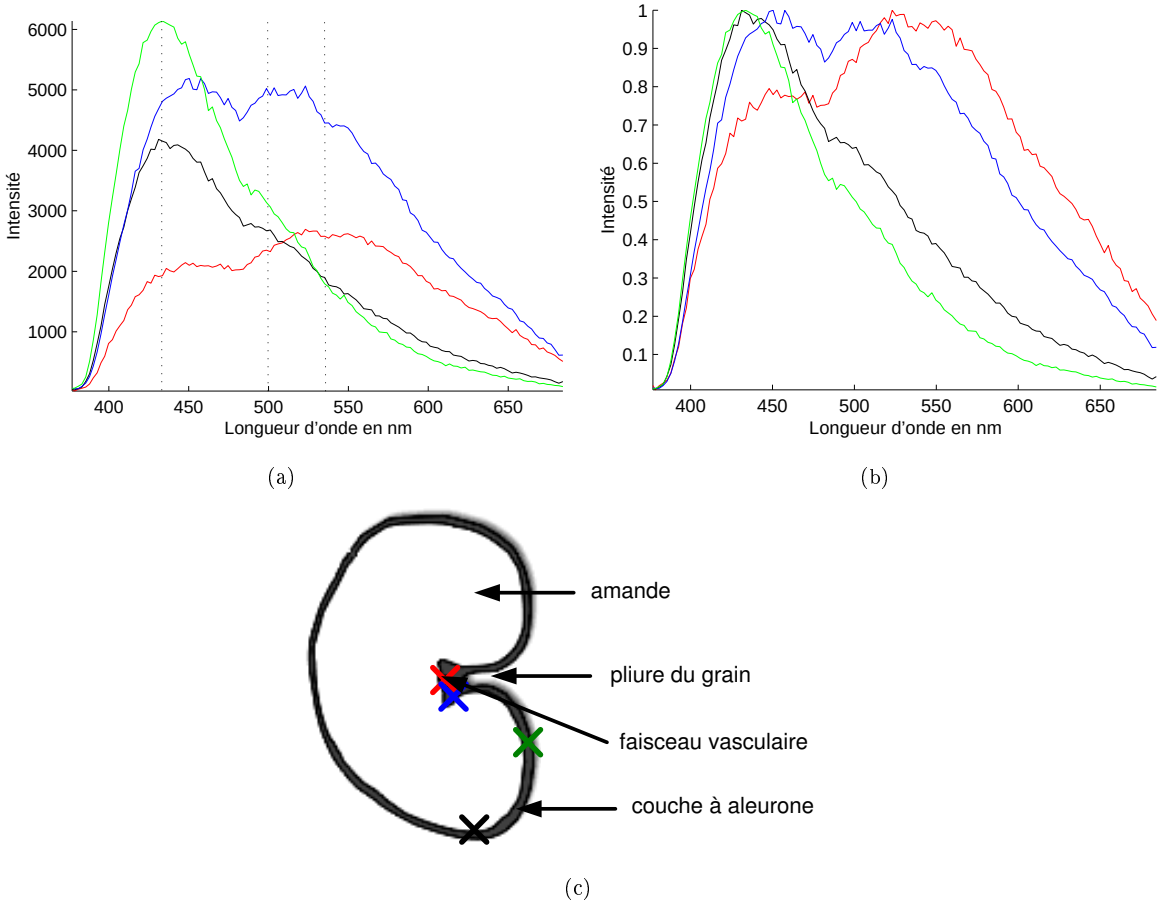


FIG. 3.2 – Exemple de 4 spectres d'émission de fluorescence représentatifs des 400 enregistrés sur une coupe transversale de grains de blé (a) dans leur version brute, (b) dans leur version normalisée à une intensité maximale égale à 1 ; (c) localisation des points d'acquisition de ces spectres sur la coupe transversale

3.5.1.3 Modélisation et propriétés

Dans le paragraphe 2.2.3, page 40, la linéarité et l'instantanéité des spectres de fluorescence ont conduit à l'élaboration du modèle défini par l'équation (2.2). Par commodité de lecture, ce modèle est rappelé ici. La matrice $\mathbf{X}$ des spectres d'émission de fluorescence, de dimensions 400×128, est décomposable en

deux matrices **A** et **S** suivant la factorisation matricielle suivante :

$$\mathbf{X} = \mathbf{AS}$$

où **A** est la matrice des profils de concentration des acides phénoliques et **S** la matrice des spectres d'émission de fluorescence des acides phénoliques. Ces deux matrices sont respectivement de dimensions $400 \times p$ et $p \times 128$, où p représente le nombre d'acides phénoliques différents présents dans le grain de blé. Cette dimension reste à fixer. Comme indiqué au paragraphe 2.3.3, page 46, les matrices **X**, **A** et **S** sont composées exclusivement d'éléments positifs.

La modélisation de notre application est identique à celle du problème de FMN présentée au paragraphe 3.4.1, page 80. La FMN semble donc être une méthode de traitement de données efficace pour estimer les matrices **A** et **S** sur les seules connaissances de la matrice d'enregistrement **X**, de la positivité des matrices **X**, **A** et **S**, et du nombre p de composantes sous-jacentes au modèle. Il est maintenant nécessaire de fixer ce nombre p et de choisir la méthode de FMN la plus efficace sur ce jeu de données.

3.5.1.4 Spectres de référence

Les travaux de Saadi [113] ont prouvé l'existence majoritaire de deux acides phénoliques dans le grain de blé : l'acide férulique et l'acide para-coumarique. Les spectres de référence d'émission de fluorescence de ces espèces ont été mesurés sur des cristaux purs d'acides férulique et para-coumarique. Ces spectres sont fournis respectivement sur les figures 3.3(c) et 3.3(b). Cependant, l'étude des spectres expérimentaux laissait suggérer l'existence d'une troisième espèce majoritaire avec un maximum à 434 nm qui ne coïncidait pas avec ceux des acides férulique à 450 nm et 515 nm, et para-coumarique à 455 nm et 544 nm. Des études complémentaires ont permis de prouver que cette espèce correspondait à une forme particulière d'acide férulique lié à des hydrates de carbone. Dans la suite, l'acide férulique libre se reportera à la forme de l'acide férulique présentée sur la figure 3.3(c), et l'acide férulique lié se réfèrera à sa forme liée à des hydrates de carbone et dont un spectre d'émission de fluorescence est montré à la figure 3.3(a).

Le spectre de l'acide férulique lié de la figure 3.3(a) se caractérise par une large bosse centrée à la longueur d'onde 434 nm. Son intensité augmente rapidement pour les faibles longueurs d'onde jusqu'à atteindre son maximum à 434 nm, puis décroît plus lentement pour les longueurs d'onde plus élevées. Le spectre de l'acide para-coumarique de la figure 3.3(b) a la forme d'une bosse à large base centrée à 544 nm. S'y ajoute une épaule de plus faible intensité centrée à 455 nm. Le spectre de l'acide férulique libre de la figure 3.3(c) se décrit quant à lui comme une juxtaposition de deux larges bosses centrées en 450 nm et 515 nm avec une croissance rapide de l'intensité pour les faibles longueurs d'onde et une décroissance rapide pour les fortes longueurs d'onde.

Ces spectres vont être utilisés pour apprécier la qualité des résultats estimés par la FMN. Une remarque doit être faite à ce sujet. Il est à noter que la qualité de l'estimation ne sera pas d'autant meilleure que les spectre estimés des acide seront proches de leurs spectres de référence. Dans le grain de blé, ces acides

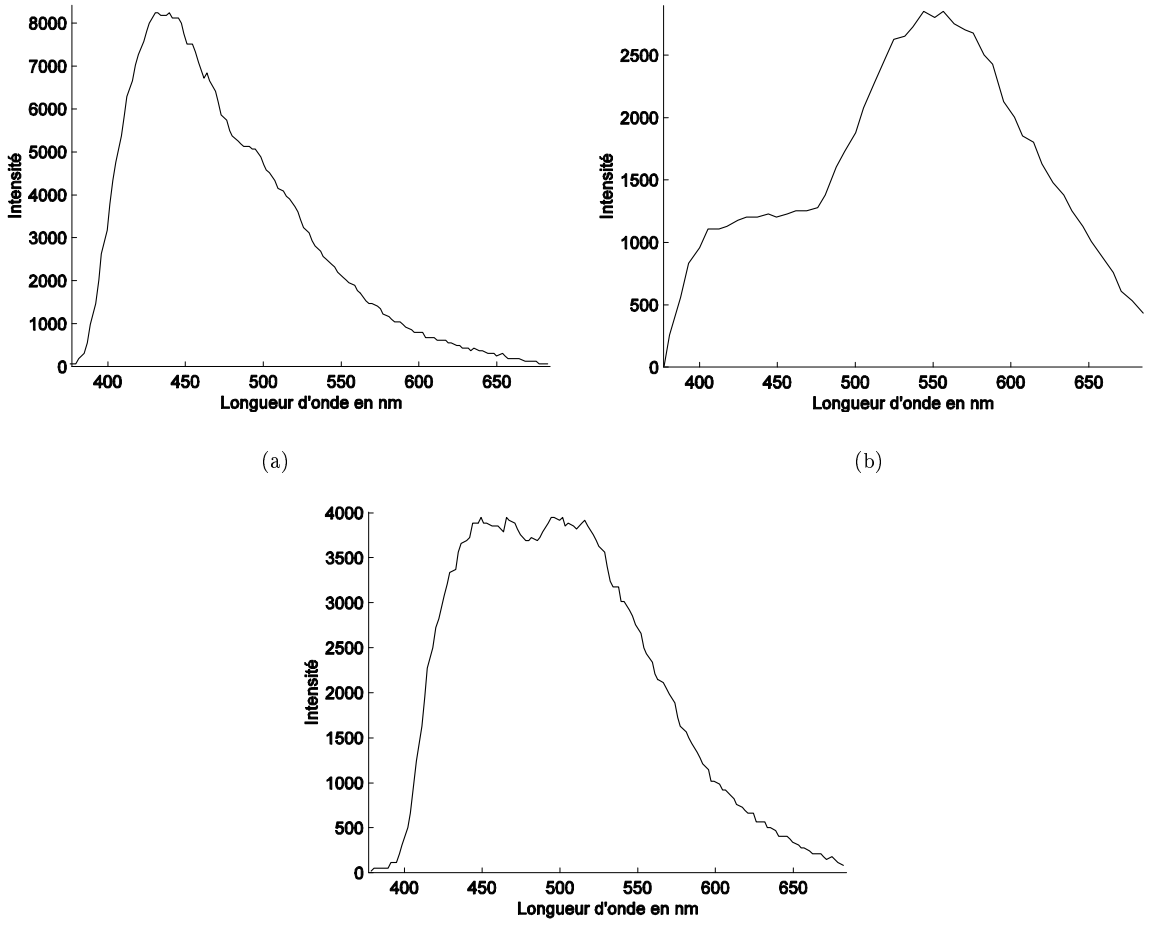


FIG. 3.3 – Spectres de référence (a) de l'acide férulique lié, (b) de l'acide para-coumarique, (c) de l'acide férulique libre

sont au contact d'autres espèces chimiques et structures biologiques. La spectroscopie de fluorescence enregistre l'information non seulement des espèces fluorescentes mais également de leur voisinage [56, 106]. Cependant, l'estimation de spectres approximativement ressemblants à ceux des acides férulique et para-coumarique de la figure 3.3 sera pour nous gage de qualité des résultats.

3.5.1.5 Choix des algorithmes de FMN

Discussion sur l'algorithme à contraintes de parcimonie :

Les spectres de fluorescence sont connus pour leur forme diffuse sur une large bande spectrale [128, page 34] . Cette remarque est particulièrement valable pour les spectres des acides phénoliques de la figure 3.3. La parcimonie n'est donc pas une hypothèse réaliste attribuable aux spectres d'émission de fluorescence.

Dans une étude précédente [113], il a été montré que le grain de blé peut être décomposé en régions où un seul acide phénolique est présent majoritairement, ce qui exprime une certaine parcimonie dans la structure du grain de blé donc dans les profils de concentration des acides phénoliques. L'algorithme de Hoyer présenté dans [59] et à la page 83 peut être appliqué à notre jeu de données, sans parcimonie pour les spectres sources à estimer, et pour une parcimonie non nulle, mais dont la valeur reste à déterminer, pour les profils de concentrations.

Discussion sur l'algorithme à contraintes de localisation spatiale :

Problème direct : L'algorithme de FMN basé sur la recherche de sources spatiallement localisées et présenté au paragraphe 3.4.2.4, page 84, semble ne pas être applicable à notre jeu de données. La principale raison est la recherche par cette méthode de sources maximalement décorrélées entre elles. Or, la forme diffuse des spectres d'émission de fluorescence induit une corrélation évidente entre les spectres des acides phénoliques.

Problème transposé : Un moyen évident pour contourner ce problème est de considérer le problème transposé et dans ce cas, la factorisation de la matrice $\mathbf{X}$ s'écrit :

$$\mathbf{X}^T = \mathbf{S}^T \mathbf{A}^T. \tag{3.22}$$

Les rôles initialement joués par les matrices $\mathbf{A}$ et $\mathbf{S}$ sont inversés. Les contraintes imposées par la méthode de FMN basée sur la recherche de sources spatiallement localisées impliquent que les profils de concentration associés à chaque acide phénolique, et représentés par les lignes de $\mathbf{A}^T$, soient décorrélés. Cette hypothèse est réaliste à la vue de la remarque en début de paragraphe sur le partitionnement des acides dans le grain de blé. Un problème provient de l'hypothèse de parcimonie de la matrice $\mathbf{S}^T$ qui n'est pas respectée à cause de la forme spécifique des spectres d'émission de fluorescence.

Tests sur le jeu de données : Des tests ont été réalisés sur le jeu de données afin de confirmer ces hypothèses. Il est à noter que dans l'équation (3.18), deux paramètres α et β pondèrent respectivement la contrainte de décorrélation des sources $\mathbf{S}$ et de parcimonie des colonnes de $\mathbf{A}$, et la contrainte de forte énergie des colonnes de $\mathbf{A}$. Des simplifications apportées dans [80] dans l'expression des règles de mise à jour initiales ont mené aux règles des équations (3.19),(3.20) et (3.21) qui sont devenues indépendantes des paramètres α et β. Dans l'algorithme de Li [80], ces paramètres ne servent qu'au calcul de la fonction objectif à chaque itération, et influencent faiblement le calcul du critère d'arrêt de l'algorithme. Par exemple, pour des paramètres égaux à 0, l'algorithme devrait devenir équivalent à celui de Lee et Seung, ce qui n'est pas le cas. Nous avons tester l'algorithme pour des valeurs de $\alpha \in \{0, \ldots, 10000\}$ et de $\beta \in \{0, \ldots, 10000\}$. Le critère d'arrêt de l'algorithme a été calculé comme le taux de variation relative de la fonction objectif. Sous forme mathématique, ce critère d'arrêt est calculé par l'équation suivante :

$$\mathcal{C}_t = \frac{|Q_t^{FMNL} - Q_{t-1}^{FMNL}|}{|Q_{t-1}^{FMNL}|} \tag{3.23}$$

où t représente l'itération courante, et $t-1$ l'itération précédente. Les essais ont été réalisés pour un critère d'arrêt de 10^{-5}. Tous nos essais se sont conclus par l'estimation de sources fortement décorrélées

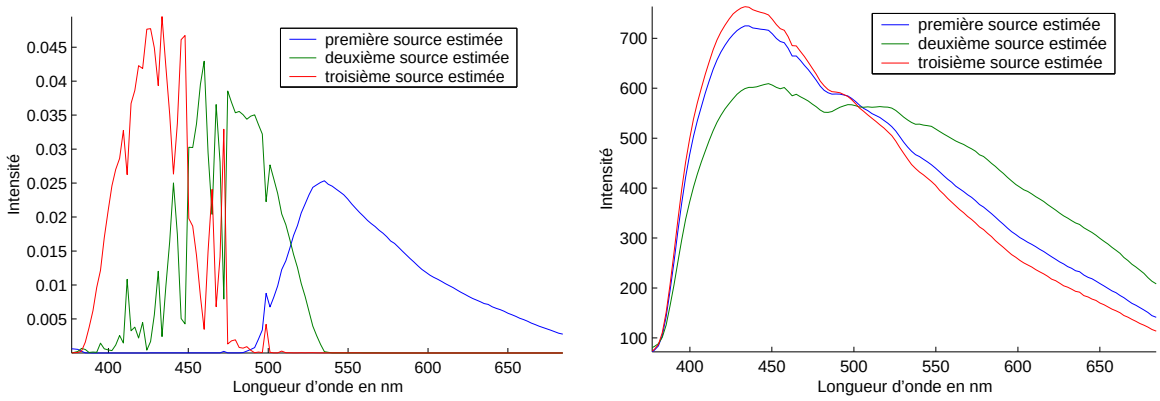


FIG. 3.4 – Spectres sources estimés par la FMN avec contrainte de localisation spatiale sur un modèle à 3 sources avec $\alpha = 1000$ et $\beta = 100$ pour (a) le problème non transposé et (b) le problème transposé

et de profils de concentrations très corrélés. Cette corrélation des colonnes de $\mathbf{A}$ est une conséquence de la décorrélation des sources. Les sources n'étant actives que pour certaines longueurs d'onde, la génération d'un spectre nécessite l'activation de toutes les sources. Un spectre original de faible intensité implique des coefficients de concentration tous faibles. Tandis qu'un spectre de forte intensité réclame des coefficients de concentration tous intenses. Pour tous les essais, des sources très similaires ont été estimées quels que soient les coefficients de pondération α et β, ce qui signifie que le choix de ces coefficients n'est que de faible importance sur l'estimation de la solution. Un exemple d'estimation est fourni sur la figure 3.4(a) pour les paramètres $\alpha = 1000$ et $\beta = 100$. La décorrélation des sources est clairement visibles par l'introduction d'intensités nulles dans ces sources. Cette décorrélation forcée des sources se fait au détriment de leur interprétabilité physique.

La transposition du problème, représentée par l'équation (3.22), conduit à l'estimation cette fois-ci de profils de concentration décorrélés, et de sources fortement corrélées. Ceci s'explique par le fait qu'une seule source est autorisée à être active pour reconstruire un spectre original, et que les spectres originaux sont de formes similaires. Les spectres sources estimés représentent donc des sortes de spectres moyens des spectres originaux. La figure 3.4(b) illustre ce propos en présentant les spectres sources estimés par transposition du problème, pour $\alpha = 1000$ et $\beta = 100$ et pour un critère d'arrêt de 10^{-5}.

L'algorithme de FMN contraint à une localisation spatiale des sources estimées et décrit à la section 3.4.2.4 ne sera pas utilisé dans notre étude du fait des problèmes décrits ci-dessus lors de son application sur nos jeux de données.

Algorithmes choisis :

Les algorithmes basés sur la distance euclidienne à la section 3.4.2.1, sur la divergence à la section 3.4.2.2, et sur les contraintes de parcimonie à la section 3.4.2.3 ne posent pas ces problèmes puisqu'ils sont

moins resctrictifs que l'algorithme de la section 3.4.2.4. La transposition du problème selon l'équation (3.22) n'affecte pas les estimations des matrices $\mathbf{A}$ et $\mathbf{S}$ puisque seule la positivité des matrices $\mathbf{A}$ et $\mathbf{S}$ est requise par les algorithmes basés sur la distance euclidienne et sur la divergence. L'algorithme de Hoyer diffère des deux précédents par l'ajout de contraintes de parcimonie sur les matrices $\mathbf{A}$ et $\mathbf{S}$. La transposition du problème n'affecte en rien l'algorithme. Il suffit simplement d'affecter à $\mathbf{A}^T$ la parcimonie choisie pour $\mathbf{A}$ et à $\mathbf{S}^T$ la parcimonie imposée à $\mathbf{S}$. Notre étude va donc se limiter à l'utilisation sur le problème direct[17] des algorithmes initialement proposés par Lee et Seung dans [77] et proposés aux sections 3.4.2.1 et 3.4.2.2, et de la méthode d'estimation parcimonieuse développée par Hoyer dans [59] et expliquée dans la section 3.4.2.3.

3.5.1.6 Résultats

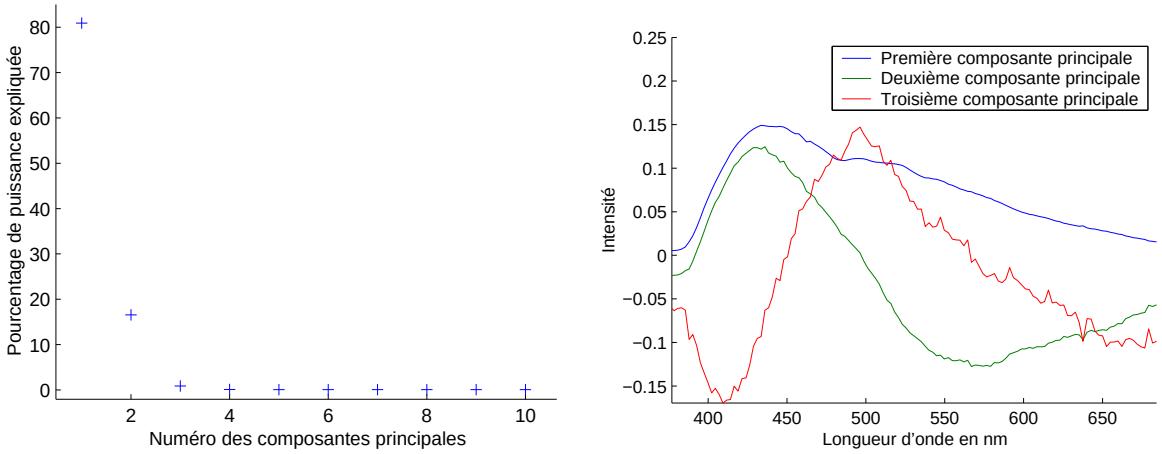
Estimation du nombre d'espèces chimiques : Les études menées par Saadi dans [113] ont été basées sur des connaissances biologiques, et les espèces *majoritairement* fluorescentes ont été recensées.

Nous avons estimé le nombre d'espèces chimiques par ACP. Les pourcentages de puissance associés aux 10 premières composantes principales sont fournis sur la figure 3.5(a). Les deux premières composantes expliquent à elles-seules environ 97.5% de la puissance totale des signaux. En prenant en compte la troisième composante, 98.3% de la puissance est représentée. La conservation de composantes supplémentaires n'apporte pas beaucoup plus d'information. La considération des 10 premières composantes explique 98.8% de la puissance, c'est-à-dire qu'une faible amélioration de 0.5% de puissance est obtenue en utilisant 7 composantes supplémentaires. Ainsi, 3 composantes principales participent majoritairement à la construction des données. Ces trois composantes sont dessinées sur la figure 3.5(b). Dans la suite, nous estimerons donc des modèles de 3 sources. Cependant, comme remarqué précédemment, cette valeur doit être vérifiée par application des algorithmes de FMN pour différents nombres de sources puisqu'il n'est pas impossible qu'une quatrième composante, même si elle est de faible énergie, soit physiquement viable.

Nous pouvons noter que l'ACP n'est pas une méthode viable pour estimer le modèle sous-jacent à nos données puisque seule la première composante présente une ressemblance avec les spectres de référence de la figure 3.3. De plus, les deux autres composantes ne sont pas physiquement interprétables puisqu'elles possèdent des intensités à la fois positives et négatives. Elles ne modélisent donc pas les spectres d'émission de fluorescence des espèces chimiques présentes dans le grain de blé. Des techniques d'Analyse en Composantes Indépendantes (ACI) ont été appliquées sur ce jeu de données. Les résultats ne sont pas satisfaisants et sont présentés dans l'annexe F.

L'ACP suggère donc l'existence de trois espèces chimiques majoritairement présentes dans le grain de blé. Cependant, différentes applications des algorithmes de FMN pour différents nombres de sources sous-jacentes ont été réalisées sur les spectres de fluorescence acquis sur le grain de blé. Les spectres

[17] par direct nous entendons le problème initial, c'est-à-dire non transposé



(a) (b)

FIG. 3.5 – ACP : (a) Pourcentages de puissance associés aux 10 premières composantes principales, (b) Les 3 premières composantes principales

estimés pour un modèle à deux sources par minimisation de l'erreur quadratique (voir la section 3.4.2.1) sont présentés sur la figure 3.6(a). Ce modèle n'est pas viable puisqu'aucun de ces deux spectres estimés ne révèle la présence d'acide férulique libre qui est l'un des constituants majeurs du grain de blé. Un modèle basé sur plus de spectres sources doit être estimé.

Les estimations par minimisation de l'erreur quadratique d'un modèle à 4 sources sont proposées sur la figure 3.6(b). Deux problèmes relèvent de ces estimations. Le premier est la décomposition du spectre de l'acide férulique en deux spectres distincts (les spectres rouge et vert). Ce problème se révèle mineur puisque cette décomposition du spectre de l'acide férulique en deux spectres peut s'expliquer par le fait que le spectre de fluorescence de l'acide férulique diffère en fonction de son voisinage. Le deuxième problème est la variabilité importante introduite dans les spectres estimés par l'application des algorithmes de FMN pour différentes conditions initiales des matrices $\mathbf{A}$ et $\mathbf{S}$. Les algorithmes de FMN ne sont pas du tout reproductibles pour un modèle à 4 sources. De plus, les deux sources attribuables à l'acide férulique et estimées par minimisation de la divergence présentent des formes différentes de celles des sources estimées par minimisation de la distance euclidienne. Le nombre d'espèces chimiques constituant le grain de blé est donc naturellement choisi égal à 3 aux vues des prédictions de l'ACP, des analyses biologiques, et des expérimentations sur des modèles basées sur 2 ou 4 sources.

Applications des algorithmes de Lee et Seung : L'application de l'algorithme de Hoyer basé sur la minimisation de la distance euclidienne avec des contraintes de parcimonie (voir la section 3.4.2.3) doit être appliqué aux données sans contrainte de parcimonie sur les sources à estimer, mais avec une contrainte de parcimonie sur les colonnes de la matrice $\mathbf{A}$ comme nous l'avons expliqué dans le paragraphe 3.5.1.5 par l'existence d'une structure biologique ordonnée et parcimonieuse des grains de céréales. La valeur de la parcimonie à imposer aux colonnes de $\mathbf{A}$ doit être estimée. Cette valeur sera déterminée dans

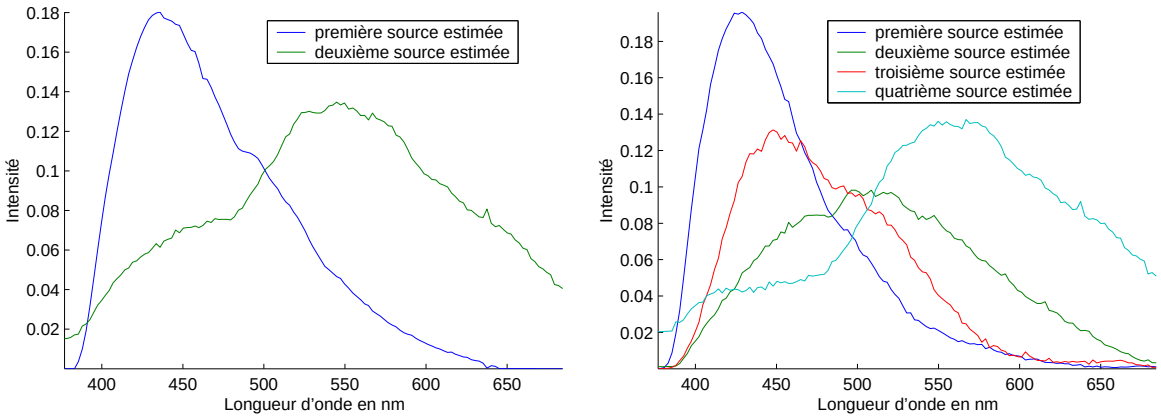


FIG. 3.6 – Sources estimées par minimisation de l'erreur quadratique pour un modèle (a) à 2 sources, (b) à 4 sources

la suite de ce chapitre à partir des résultats de séparation, que nous allons présenter maintenant, par les algorithmes de FMN basés sur la mesure de distance euclidienne et sur la mesure d'une divergence.

Chacun des algorithmes a été lancé pour 200 points aléatoires d'initialisation des matrices $\mathbf{A}$ et $\mathbf{S}$ pour un modèle à 3 sources. Les 200 spectres estimés de chaque source obtenus par l'algorithme de minimisation de la distance euclidienne et par l'algorithme de minimisation de la divergence sont présentés respectivement sur les figures 3.7 et 3.8.

Comme prévu par la théorie de la FMN, les algorithmes convergent vers des minima locaux des fonctions objectif à minimiser. Les solutions estimées par ces méthodes sont dépendantes des conditions initiales. Une certaine variance est observable au sein des sources estimées et il est nécessaire de choisir la solution la plus pertinente. Nous avons considéré deux méthodes classiques.

La première méthode consiste à conserver la solution menant à la fonction objectif de plus faible intensité. La valeur finale de la fonction objectif après convergence des deux algorithmes a été sauvegardée pour les 200 essais. Les spectres estimés impliquant la plus faible valeur de la fonction objectif sont affichés sur la figure 3.9(a) pour la méthode basée sur la distance euclidienne, et sur la figure 3.9(b) pour la méthode par calcul de divergence. Les minima atteints se situent dans un voisinage de rayon petit. Mais la présence de données bruitées dans la matrice de mesure n'a pas été prise en compte dans les différents algorithmes présentés à la section 3.4.2. À cause de ce bruit, des spectres estimés pour une plus forte valeur finale de la fonction objectif sont tout aussi pertinents pour décrire le modèle. La méthode de sélection de la solution par recherche de la valeur minimale de la fonction objectif sur l'ensemble des essais n'est donc pas juste.

Une deuxième méthode est la considération de toutes les solutions estimées par un moyennage de l'ensemble de ces solutions. Les spectres résultants de cette méthode sont visibles sur la figure 3.10. Les

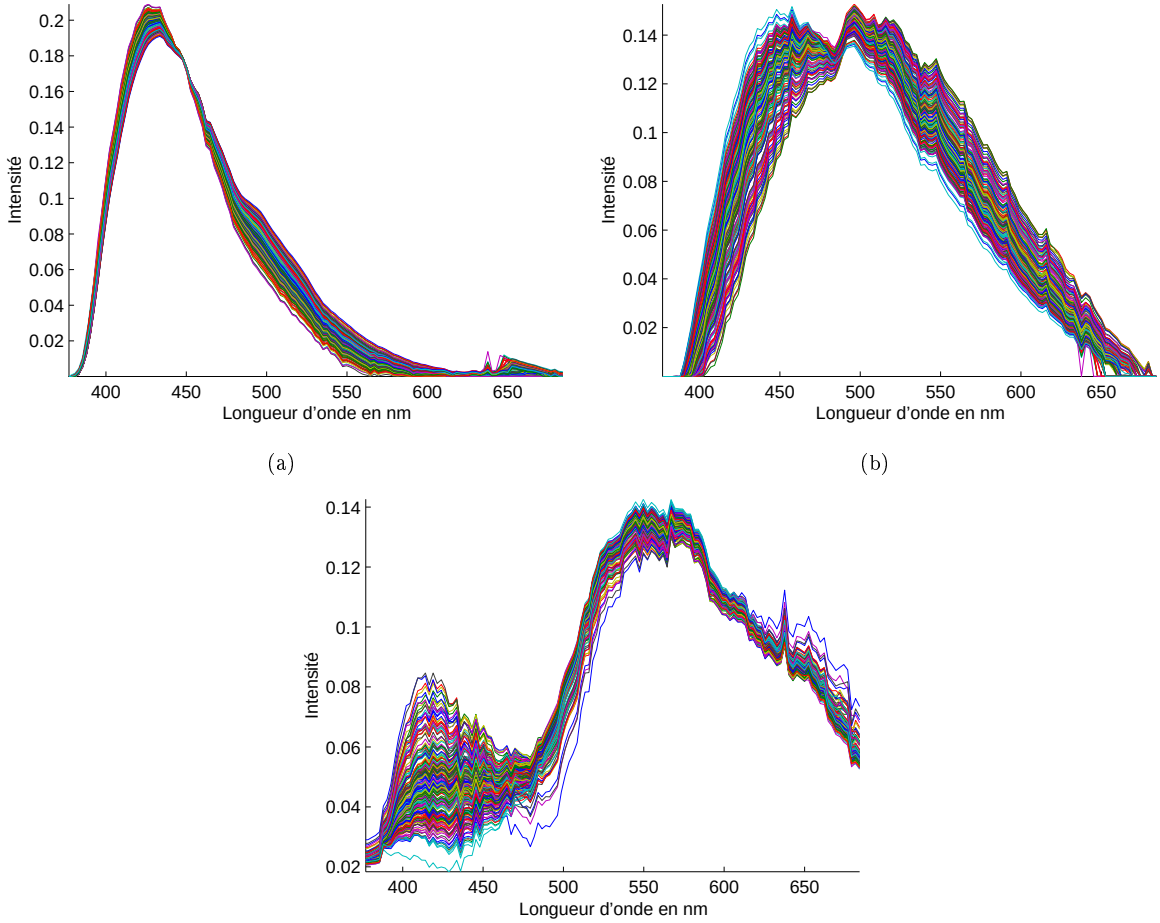


FIG. 3.7 – Estimations par minimisation de distance euclidienne des spectres sources pour 200 essais : (a) première source estimée, (b) deuxième source estimée, (c) troisième source estimée

spectres ainsi estimés des espèces chimiques pures possèdent une forme similaire à celle des spectres de références des acides phénoliques présents dans le grain de blé de la figure 3.3. La comparaison entre les spectres estimés et les spectres de référence sera faite dans le paragraphe **Discussion**, page 105, dans la suite de ce chapitre.

Chaque spectre source possède un profil de concentration localisé dans une région précise du grain de blé d'après [113], et différente d'un spectre à l'autre, et les profils de concentration sont parcimonieux. L'application de l'algorithme de Hoyer, basé sur des contraintes de parcimonie, est donc possible sur le jeu de données.

Application de l'algorithme de Hoyer : L'application de l'algorithme de Hoyer, décrit dans la section 3.4.2.3, nécessite la connaissance *a priori* de la valeur de la contrainte de parcimonie à imposer sur les colonnes de $\mathbf{A}$. Un choix inopportun de cette valeur conduit assurément vers l'estimation d'un mauvais modèle. La figure 3.11 illustre ce propos pour des contraintes de parcimonie égales à 0.1 et à 0.5.

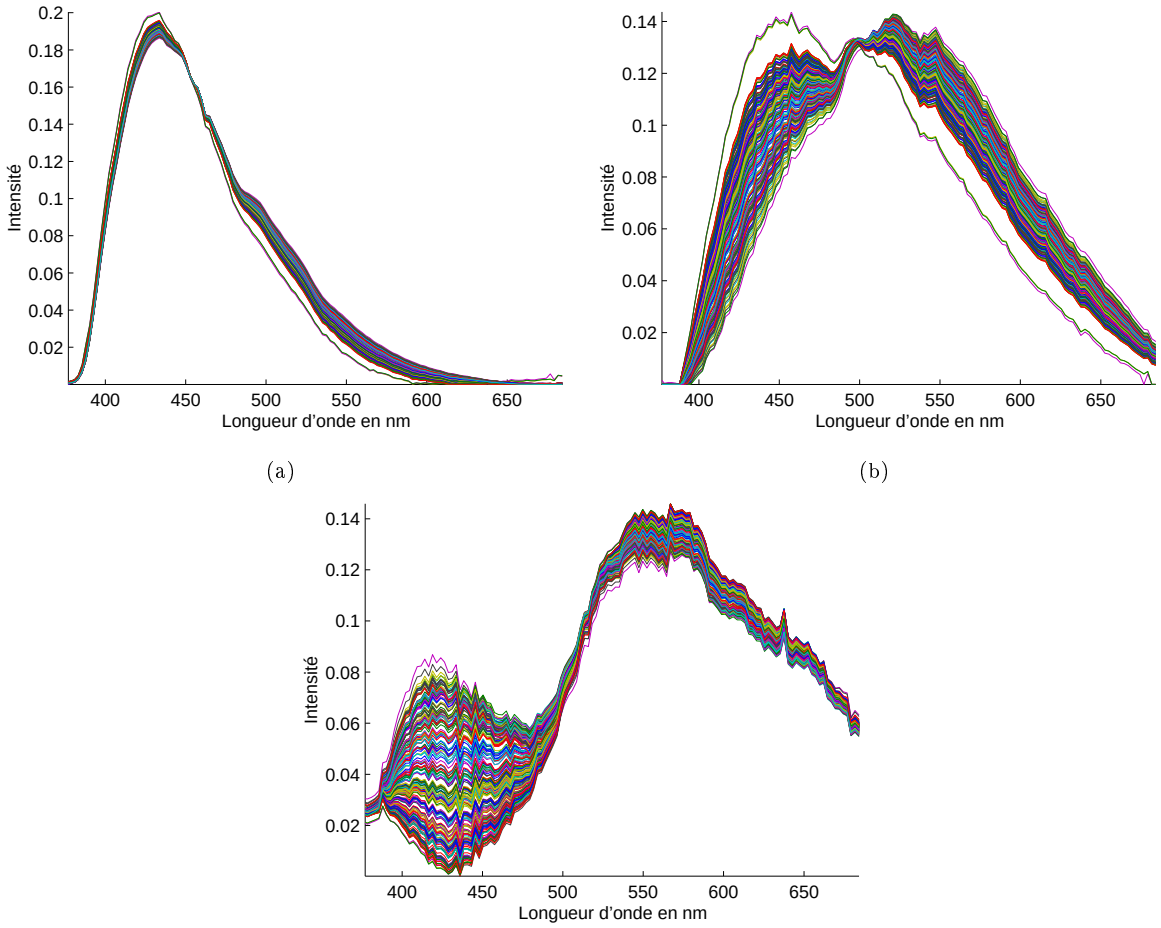


FIG. 3.8 – Estimations par minimisation de la divergence des spectres sources pour 200 essais : (a) première source estimée, (b) deuxième source estimée, (c) troisième source estimée

Une parcimonie trop faible des colonnes de la matrice $\mathbf{A}$ conduit à l'estimation de spectres parcimonieux. Une parcimonie trop forte des colonnes de la matrice $\mathbf{A}$ implique l'estimation de spectres traduisant le spectre moyen des spectres originaux.

La parcimonie à imposer aux colonnes de $\mathbf{A}$ a été calculé grâce aux spectres estimés dans le paragraphe précédent par les méthodes de Lee et Seung par la formule tirée de [59] :

$$P(\mathbf{a}_j) = \frac{\sqrt{N_{xy}} - \frac{\sum_{i=1}^{N_{xy}} |a_{ij}|}{\sqrt{\sum_{i=1}^{N_{xy}} a_{ij}^2}}}{\sqrt{N_{xy}} - 1} \tag{3.24}$$

où $P(\mathbf{a}_j)$ est la parcimonie de la j-ème colonne $\mathbf{a}_j$ de la matrice $\mathbf{A}$.

La parcimonie moyenne sur les colonnes de $\mathbf{A}$ pour 200 essais est répertoriée dans le tableau pour chacune des deux méthodes de Lee et Seung employées précédemment.

L'algorithme de Hoyer est paramétré par la parcimonie P de l'équation (3.24) à imposer sur toutes les

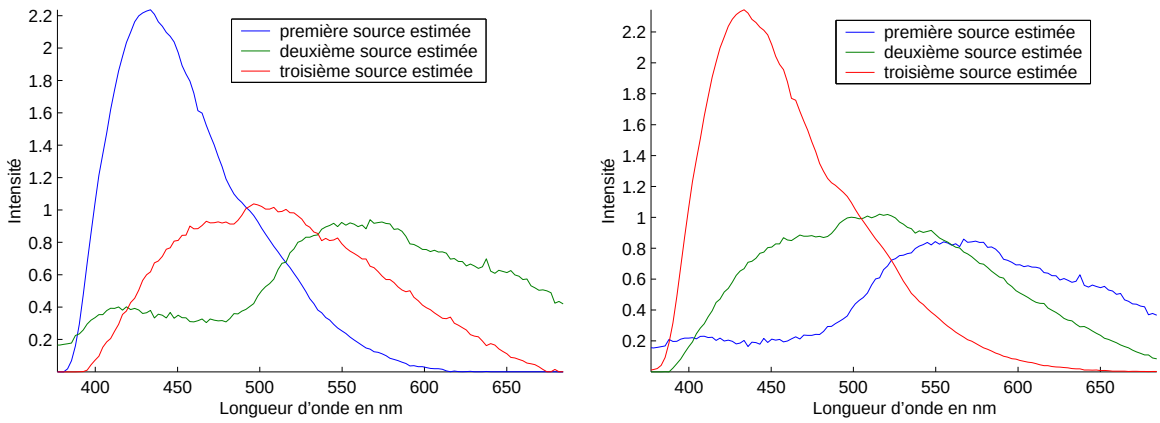


FIG. 3.9 – Spectres estimés donnant les minima des fonctions objectifs : (a) de l'erreur quadratique de reconstruction des données et (b) de la divergence entre les données et leurs reconstructions

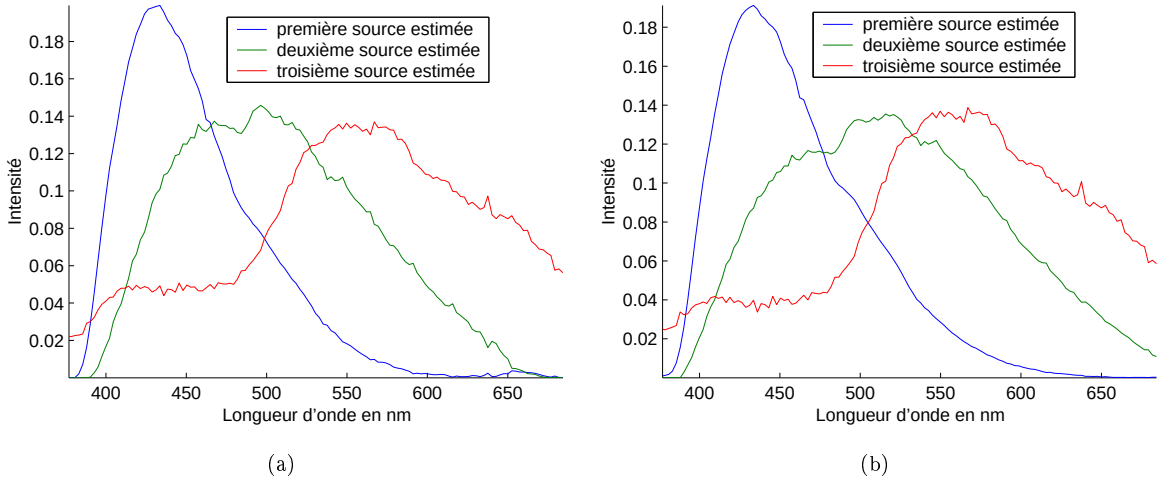
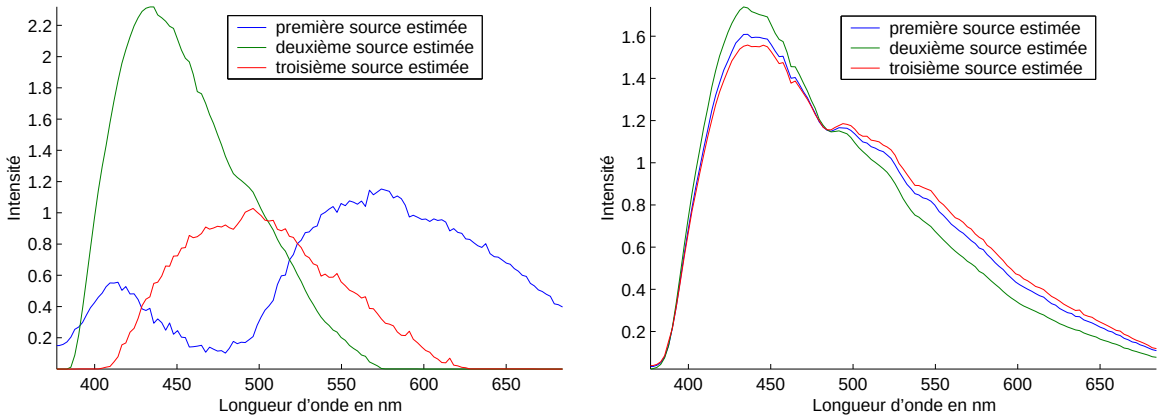


FIG. 3.10 – Spectres moyens de l'ensemble des solutions pour (a) l'erreur quadratique de reconstruction des données et (b) la divergence entre les données et leurs reconstructions



(a) (b)

FIG. 3.11 – Spectres estimés par la méthode de Hoyer pour une parcimonie *imposée* aux colonnes de $\mathbf{A}$ égale à (a) 0.1 (b) 0.5

colonnes de la matrice de $\mathbf{A}$. Pour nos expérimentations, ce paramètre a été choisi égal à la moyenne des parcimonies calculées sur chaque colonne de la matrice A estimée au cours des 200 essais. Cette valeur est classée dans le tableau 3.1 et vaut 0.1763. Les sources estimées pour 200 essais qui diffèrent par les conditions initiales imposées à l'algorithme sont représentées sur la figure 3.12

Une comparaison de cette figure avec les figures 3.7 et 3.8, obtenues par les algorithmes de Lee et Seung, montre que les spectres sources estimés par la méthode de Hoyer présentent pour chaque longueur d'onde une variance d'estimation beaucoup plus faible. La contrainte de parcimonie imposée aux colonnes de $\mathbf{A}$ réduit l'espace des solutions possibles. La fonction coût Q^{CPN} exhibe moins de minima locaux. Il est cependant à remarquer qu'une condition nécessaire et suffisante de convergence de cet algorithme est la non-dégénérescence des vrais spectres sources qui doivent être estimés [127]. Les spectres des acides phénoliques à estimer sont évidemment non-dégénérés. Or l'algorithme ne converge pas vers une solution unique. En fait, dans [127], la démonstration de l'unicité de la solution estimée par l'algorithme de Hoyer est réalisée dans le cas d'une reconstruction parfaite de la matrice de donnée par les matrices $\mathbf{A}$ et $\mathbf{S}$ estimées. Or dans notre étude, l'erreur de reconstruction n'est pas nulle puisque le critère d'arrêt de l'équation (3.23) est fixé à $C_t = 10^{-5}$. L'espace des solutions n'est donc plus un seul point, mais un espace de petit volume.

	$A_{.1}$	$A_{.2}$	$A_{.3}$	Moyenne totale
distance euclidienne	0.10187	0.08264	0.26883	0.1511
divergence	0.08916	0.13438	0.35810	0.2015
moyenne	0.09551	0.10851	0.31347	0.1763

TAB. 3.1 – Estimation de la parcimonie, moyennée sur les 200 essais, des colonnes de la matrice A estimée par les algorithmes de Lee et Seung

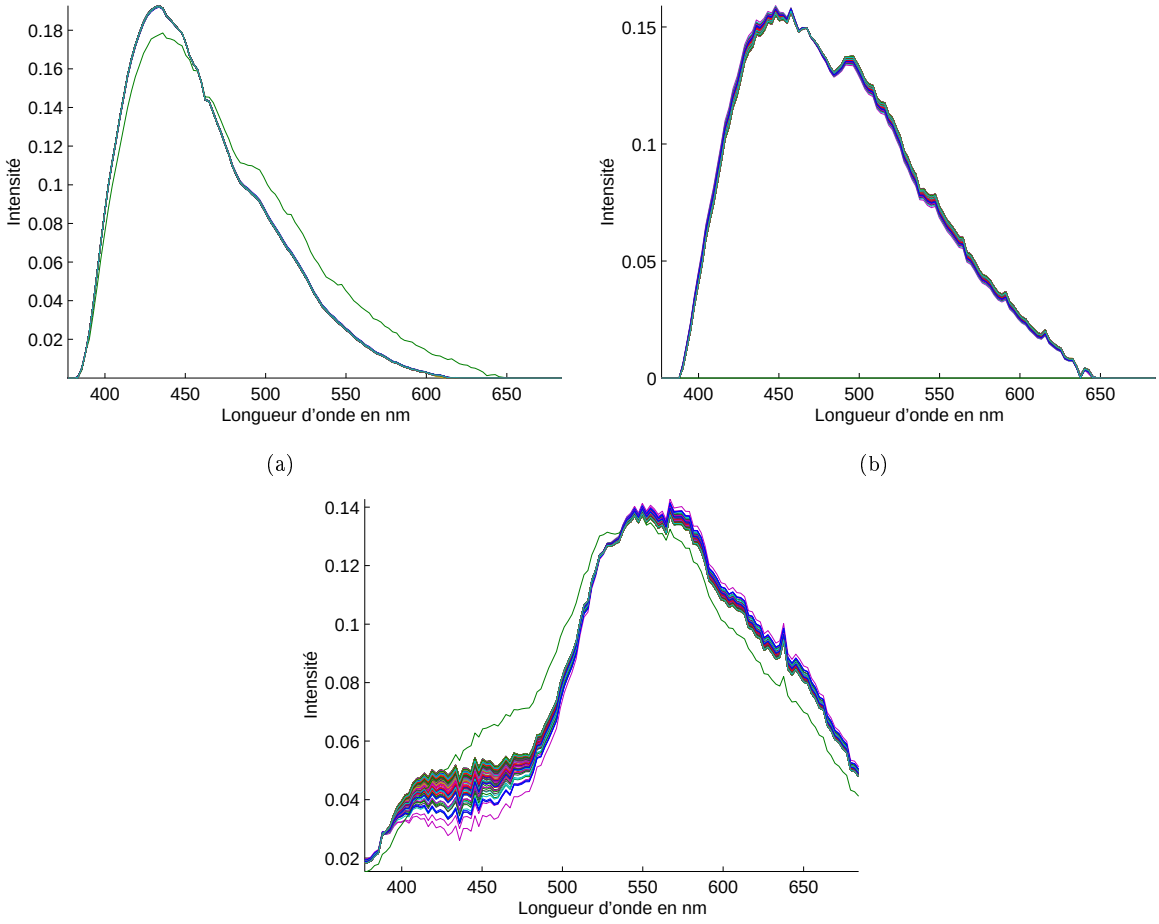


FIG. 3.12 – Estimations par minimisation de la distance euclidienne sous contraintes de parcimonie des colonnes de **A** pour les mêmes 200 essais : (a) première source, (b) deuxième source, (c) troisième source

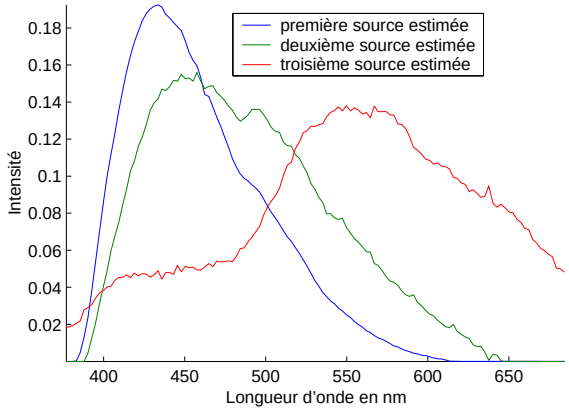


FIG. 3.13 – Spectres moyens de l'ensemble des solutions estimées sur les 200 essais par minimisation de la distance euclidienne sous contrainte de parcimonie des colonnes de **A**

Les spectres sources moyens sont représentés sur la figure 3.13. Une discussion sera menée dans le paragraphe **Discussion** qui va suivre afin de confronter les spectres et leurs profils de concentrations estimés aux spectres et profils de concentrations de référence.

La connaissance de la valeur de la parcimonie des colonnes de la matrice **A** mène à l'estimation de spectres sources quasiment uniques pour des essais lancés de conditions initiales différentes. Cependant, cette valeur est imposée comme identique pour toutes les colonnes de **A**. Or la distribution de chaque espèce dans le grain de blé n'est pas identiquement parcimonieuse. Nous avons donc modifié l'algorithme de Hoyer pour qu'il accepte des parcimonies différentes des colonnes de **A**. Des tests ont été refaits en contraignant les colonnes de **A** à des parcimonies ayant les valeurs de la dernière ligne du tableau 3.1 et les résultats sont similaires à ceux précédemment exposés.

Remarques : Trois caractéristiques importantes sont à prendre en considération lorsque des algorithmes de FMN sont appliqués à des spectres d'émission de fluorescence. La première est le temps de calcul des divers algorithmes. Les nombres d'itérations nécessaires pour une convergence des algorithmes à un critère d'arrêt de 10^{-5} dans le cadre de 200 essais sont présentés sur la figure 3.14. Les nombres moyens d'itérations pour les méthodes par minimisation de la divergence, de la distance euclidienne et de la distance euclidienne sous contraintes de parcimonie sont respectivement égaux à 1048, 881 et 630. La méthode par contraintes de parcimonie est en moyenne celle qui converge le plus rapidemment lorsque les valeurs de la parcimonie sont connues, sinon elle devient la plus longue lorsqu'il faut estimer ces valeurs. La deuxième caractéristique importante est le nombre de minima locaux de la fonction objectif à minimiser. Comme expliqué précédemment, seul l'algorithme par contraintes de parcimonie possède une preuve de convergence vers le minimum global. Mais l'estimation d'une solution unique ne se fait que sous une connaissance *a priori* qui est la parcimonie des profils de concentration des espèces chimiques à estimer et qui représente la troisième caractéristique. Un compromis entre ces trois exigences doit être fait pour choisir l'algorithme le mieux adapté à l'application considérée. De toute évidence, l'algorithme

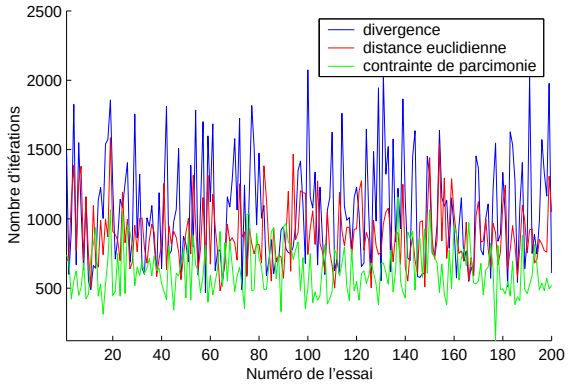
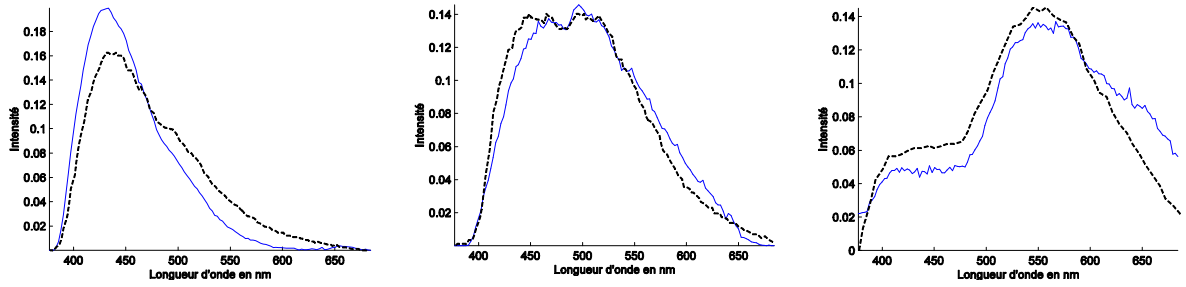


FIG. 3.14 – Évolution du nombre d'itérations nécessaires à la convergence pour 200 essais par minimisation de la divergence (courbe bleue), la distance euclidienne (courbe rouge) et la distance euclidienne sous contraintes de parcimonie (courbe verte)

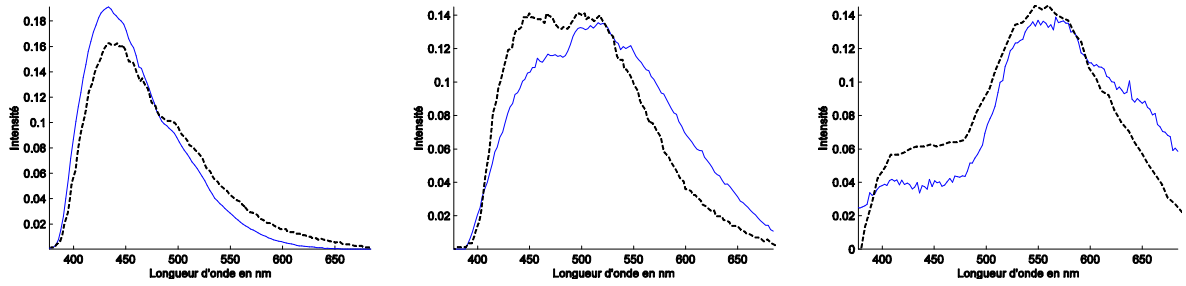
de Hoyer est celui qui offre le plus de garanties dès que la valeur de la parcimonie des colonnes de la matrice $\mathbf{A}$ est connue. Les algorithmes de Lee et Seung proposent quant à eux une grande simplicité de programmation.

Discussion : Les spectres étant estimés, il est indispensable de les comparer aux spectres de référence des acides phénoliques du grain de blé pour apprécier la qualité des estimations. Sur la figure 3.15 sont représentées les comparaisons entre les spectres de référence et les spectres estimés par distance euclidienne. La figure 3.16 propose les comparaisons entre les spectres de référence et les spectres estimés par divergence. La figure 3.17 compare quant à elle les spectres de référence et les spectres estimés par contraintes de parcimonie. Sur toutes ces figures, les spectres estimés des espèces pures sont normalisés à une aire unité. Bien que l'estimation ne soit pas parfaite, la forme de chaque spectre estimé suit les principales variations des spectres de références. Les dissimilitudes des formes proviennent évidemment des convergences vers des minima locaux des fonctions objectif à minimiser, mais également des origines différentes des spectres de référence et des spectres de la matrice de données. Dans le premier cas, les spectres ont été acquis sur des cristaux d'acides purs. Dans le second cas, les spectres ont été enregistrés sur des échantillons biologiques de blé. Les différents acides phénoliques sont insérés dans un environnement biologique complexe. Chaque acide se trouve en contact avec une multitude d'espèces chimiques différentes. Les spectres de fluorescence ne reflètent pas la fluorescence émise par les acides phénoliques, mais plutôt la fluorescence émise par les acides et leurs environnements. Les environnements très différents entre les deux cas peuvent expliquer les différences observées entre les spectres de référence et les spectres estimés. Malgré ces différences, chaque spectre estimé est facilement attribuable aux différents acides phénoliques du grain de blé et ont été reconnus par des biophysiciens : les spectres des figures 3.15(a), 3.16(a) et 3.17(a) sont associés à l'acide férulique lié, ceux des figures 3.15(b), 3.16(b) et 3.17(b) à l'acide férulique libre, et enfin ceux des figures 3.15(c), 3.16(c) et 3.17(c) à l'acide para-coumarique.



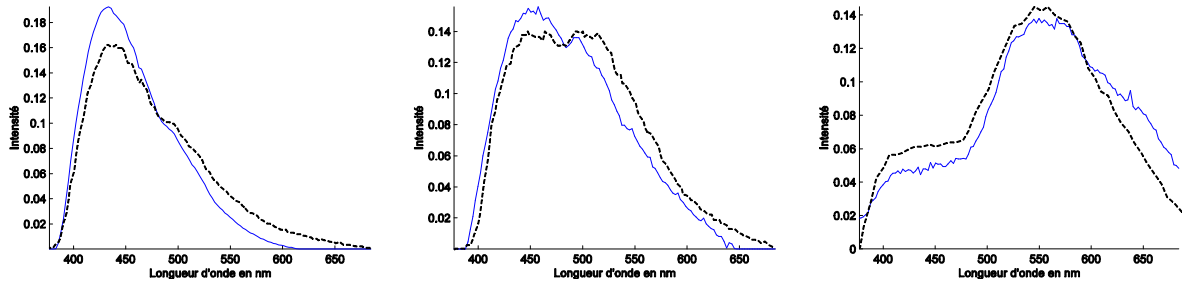
(a) (b) (c)

FIG. 3.15 – Comparaison des spectres de référence (en tirets noirs) des acides phénoliques purs et des spectres estimés (en continu bleu) sur le grain de blé par distance euclidienne : (a) acide férulique lié, (b) acide férulique libre, (c) acide para-coumarique



(a) (b) (c)

FIG. 3.16 – Comparaison des spectres de référence (en tirets noirs) des acides phénoliques purs et des spectres estimés (en continu bleu) sur le grain de blé par divergence : (a) acide férulique lié, (b) acide férulique libre, (c) acide para-coumarique



(a) (b) (c)

FIG. 3.17 – Comparaison des spectres de référence (en tirets noirs) des acides phénoliques purs et des spectres estimés (en continu bleu) sur le grain de blé par contraintes de parcimonie : (a) acide férulique lié, (b) acide férulique libre, (c) acide para-coumarique

Les spectres estimés par les algorithmes de FMN ayant été validés par des biophysiciens, et chaque spectre ayant été attribué à un acide phénolique du grain de blé, il reste à étudier leurs profils de concentration au sein du grain. Ces profils sont accessibles par les colonnes de la matrice de mélange A estimée par les méthodes de FMN. Chaque colonne de cette matrice est associée à une ligne de la matrice S, donc à un acide phénolique. La première colonne de $\mathbf{A}$ est attachée à l'acide férulique lié, la deuxième à l'acide férulique libre, et la troisième à l'acide para-coumarique.

Les profils de concentration moyens ont été calculés et mis sous forme d'images pour restituer une information d'espace qui avait été perdue par la concaténation des lignes du cube spectral original. Ces images sont visibles sur les figures 3.18 pour les profils moyens estimés par distance euclidienne, 3.19 pour ceux estimés par divergence, et sur 3.20 pour ceux estimés par contraintes de parcimonie.

Remarque : Comme expliqué au paragraphe 3.5.1.2, page 90, les spectres expérimentaux $\mathbf{x}_i,\ i = 1, \ldots, N_{xy}$ de la matrice $\mathbf{X}$ ont été normalisés par rapport à leur maximum. Chaque mélange normalisé s'écrit :

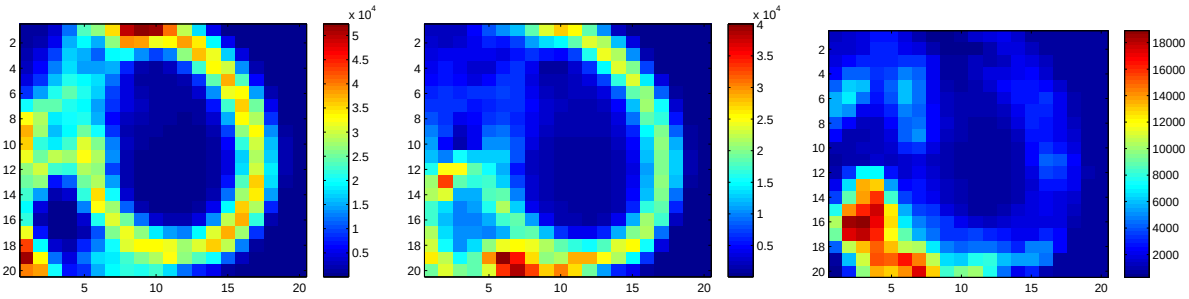
$$\frac{\mathbf{x}_i}{max(\mathbf{x}_i)} = \sum_{j=1}^{p} \frac{a_{ij}}{max(\mathbf{x}_i)} \mathbf{s}_j.$$

Les méthodes de FMN ont permis d'estimer les coefficients $\tilde{a}_{ij} = \frac{a_{ij}}{max(\mathbf{x}_i)}$. Or ces concentrations doivent être exprimées relativement à l'intensité des spectres expérimentaux originaux pour modéliser les concentrations relatives de chaque espèce en chaque point de mesure. Il convient donc de multiplier chaque concentration estimée $\tilde{a}_{ij}$ par $max(\mathbf{x}_i)$. Ce sont ces coefficients $a_{ij} = \tilde{a}_{ij} max(\mathbf{x}_i)$ qui sont représentés sur les figures 3.18, 3.19 et 3.20.

Quelle que soit la méthode de factorisation privilégiée, les profils de concentration restituent des informations similaires sur la répartition des espèces phénoliques au sein du grain de blé. Chaque espèce se trouve confinée dans des régions spécifiques du grain de blé et traduisent la structure interne du grain. Il est à noter que contrairement à l'algorithme de Hoyer, les algorithmes de Lee et Seung n'étaient contraint qu'à la positivité des matrices $\mathbf{S}$ et $\mathbf{A}$ à estimer. Pourtant, les profils de concentration des acides restitués par ces deux algorithmes présentent une certaine parcimonie qui a été répertoriée dans le tableau 3.1. Ces deux algorithmes sont ainsi capables d'estimer les parties d'un tout sans contraintes supplémentaires sur les données.

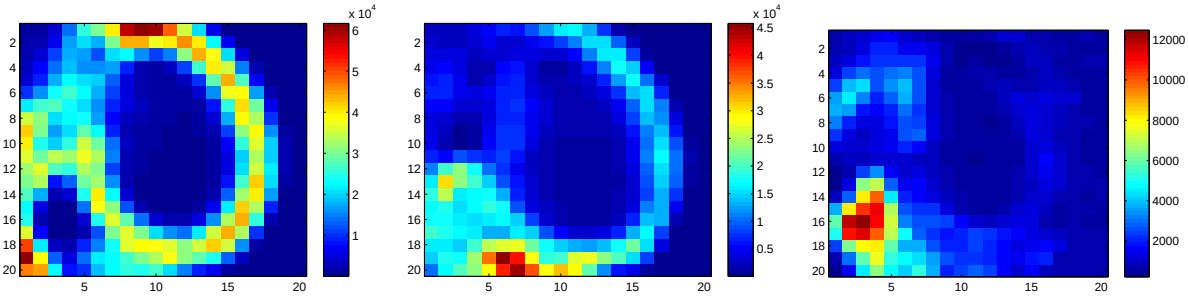
Sur les figures 3.18, 3.19 et 3.20, l'image (a) représente la carte de la répartition chimique de l'acide férulique dans la coupe transversale du grain de blé. L'image (b) est associée à la répartition chimique de l'acide férulique libre et l'image (c) à celle de l'acide para-coumarique. L'étude de ces images montre que les acides phénoliques sont distribués de manière organisée dans le grain de blé. L'analyse de ces images par des experts a conduis aux conclusions suivantes :

- L'acide férulique libre est largement réparti sur l'ensemble du grain de blé à la vue des images (b). Cependant, sa concentration est faible dans l'amande et forte dans la couche à aleurone et dans les couches les plus externes qui entourent la pliure du grain.
- L'acide para-coumarique possède une répartition biologique beaucoup plus binaire. Sa présence



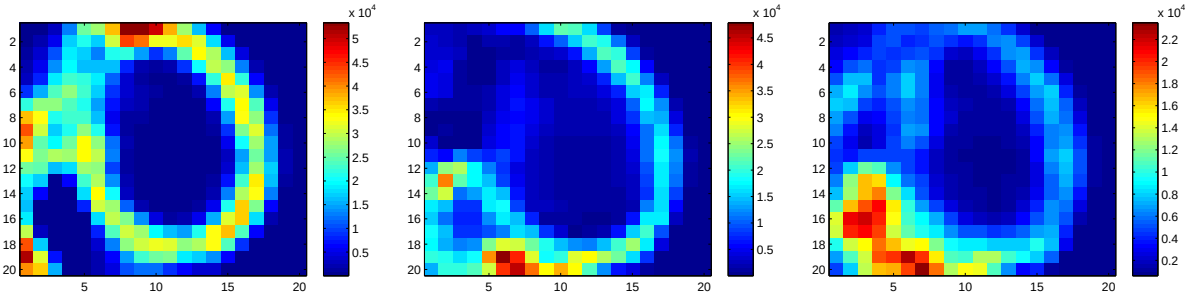
(a) (b) (c)

FIG. 3.18 – Profils de concentration moyens estimés par minimisation de la distance euclidienne : (a) acide férulique lié, (b) acide férulique libre, (c) acide para-coumarique



(a) (b) (c)

FIG. 3.19 – Profils de concentration moyens estimés par minimisation de la divergence : (a) acide férulique lié, (b) acide férulique libre, (c) acide para-coumarique



(a) (b) (c)

FIG. 3.20 – Profils de concentration moyens estimés par minimisation de la distance euclidienne sous contraintes de parcimonie des colonnes de A : (a) acide férulique lié, (b) acide férulique libre, (c) acide para-coumarique

est forte au centre de la pliure du grain et faible dans le reste du grain comme indiqué sur les images (c).

- L'acide férulique lié se condense exclusivement dans la couche à aleurone et dans la zone entourant la pliure, comme visible sur les images (a). Cette localisation de l'acide férulique lié suggère l'utilisation de cet acide comme un indicateur des tissus non-endospermiques du grain. Cet indicateur peut être utilisé par exemple pour estimer la contamination de farines par les couches externes du grain de blé. La qualité des farines peut être mesuré en fonction de la concentration de cet indicateur dans les farines.

Ces travaux ont conduit à la rédaction de deux communications [47, 46] sur l'apport de la FMN et des contraintes de positivité à la séparation de spectres de fluorescence acquis sur un grain de blé.

3.5.2 Étude sur un grain d'orge

3.5.2.1 Objectif

Une étude similaire à la précédente a été menée sur un grain d'orge. Des images multispectrales ont été enregistrées à partir d'une coupe transversale de grain d'orge. Le procédé d'estimation des sources et des profils de concentration est basé sur le même principe que pour l'étude du grain de blé. La présentation de cette application va être rapide, afin de ne pas répéter les mêmes explications que dans la section précédente. Les grains d'orge sont constitués de plusieurs tissus qui se superposent. La partie centrale contient de l'amidon et les couches externes servent de protection au grain. Les tissus externes ont la propriété d'être naturellement autofluorescents. L'objectif de l'étude est d'identifier les différents tissus *in situ* grâce aux constituants fluorescents présents dans le grain d'orge. La structure d'un grain d'orge est présentée sur la figure 3.21. L'analyse des grains est réalisée par microscopie multispectrale en fluorescence. Le but final est de pouvoir utiliser les propriétés de fluorescence mises en évidence en milieu structuré pour pouvoir suivre le devenir des tissus lors de procédés de transformation tels que le broyage.

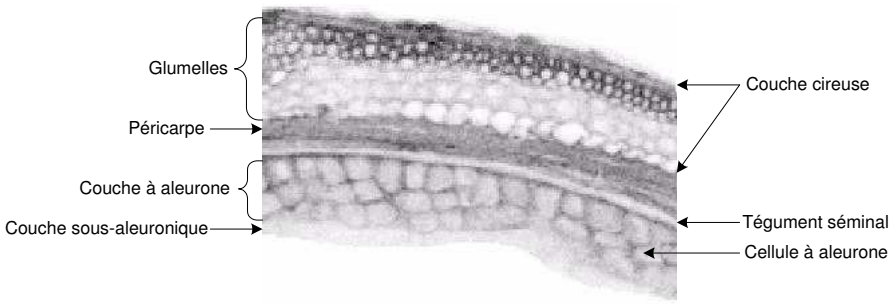


FIG. 3.21 – Identification des structures principales d'un grain d'orge

Identifiant de l'acquisition	Longueur d'onde d'excitation en nm	Longueur d'onde d'émission en nm
(a)	633	$[665, \infty]$
(b)	543	$[570, \infty]$
(c)	543	$[575, 640]$
(d)	543	$[665, \infty]$
(e)	488	$[515, \infty]$
(f)	488	$[515, 565]$
(g)	488	$[570, \infty]$
(h)	488	$[575, 640]$
(i)	488	$[665, \infty]$
(j)	364	$[397, \infty]$
(k)	364	$[450, 490]$
(l)	364	$[515, \infty]$
(m)	364	$[515, 565]$
(n)	364	$[570, \infty]$
(o)	364	$[575, 640]$
(p)	364	$[665, \infty]$
(q)	488	$[510, 525]$
(r)	364	$[400, 435]$
(s)	364	$[510, 525]$

TAB. 3.2 – Conditions d'acquisition des images spectrales du grain d'orge

3.5.2.2 Conditions d'acquisition

La microscopie de fluorescence confocale à balayage laser permet de visualiser l'autofluorescence des parois végétales. Quatre longueurs d'onde d'excitation à 364 nm (UV), à 488 nm (bleu), à 543 nm (vert) et à 633 nm (rouge), et 9 filtres d'émission permettent de définir 19 conditions d'acquisitions, donc 19 images, dans ce cas. Ces conditions sont listées dans le tableau 3.2. Chaque image est enregistrée pour une longueur d'onde d'excitation donnée et un filtre passe-haut ou passe-bande qui définit les plages de longueurs d'onde d'émission pour lesquelles la fluorescence émise est captée. Les essais ont porté sur un grain d'orge de la variété Clarine. Une coupe transversale au milieu du grain d'orge a été observée avec un objectif 10×. Les images mesurées sont de taille 512×512 pixels. La surface observée est de 1270 μm sur 1270 μm. Les 19 images acquises sont représentées sur la figure 3.22.

Dans les conditions d'acquisitions décrites, seules les couches externes du grain d'orge sont visibles en fluorescence, puisque l'albumen amylacé, ou l'amande, n'est pas fluorescent. Les images de la figure 3.22 montrent que pour une condition d'acquisition donnée, plusieurs tissus fluorescent simultanément et que la fluorescence d'un tissu peut être due à la présence de plusieurs composés biochimiques. Le but de

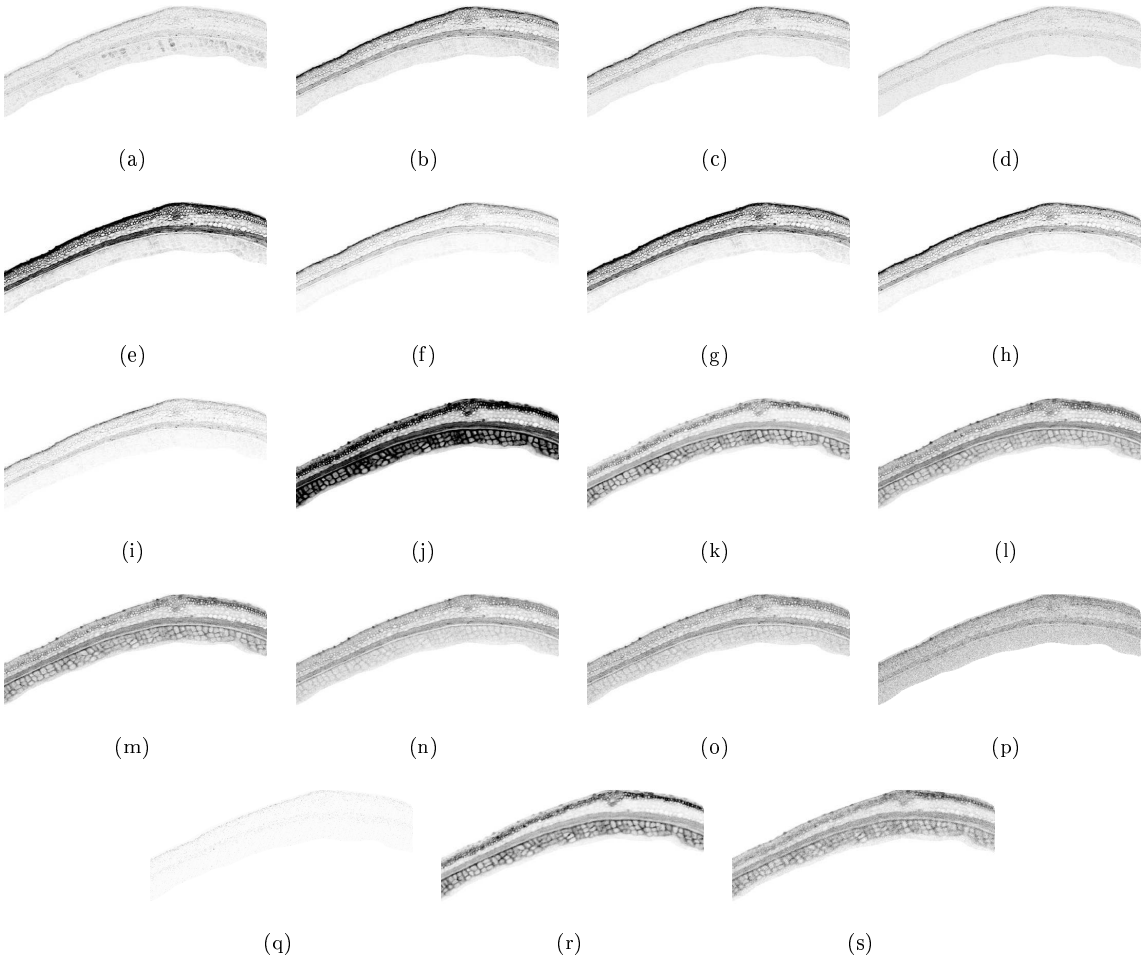


FIG. 3.22 – Séquence des 19 images spectrales acquises sur un grain d'orge à partir des 19 conditions d'acquisition répertoriées dans le tableau 3.2

cette étude est d'isoler sur des images différentes la fluorescence émise par les divers composés chimiques présents dans le grain d'orge.

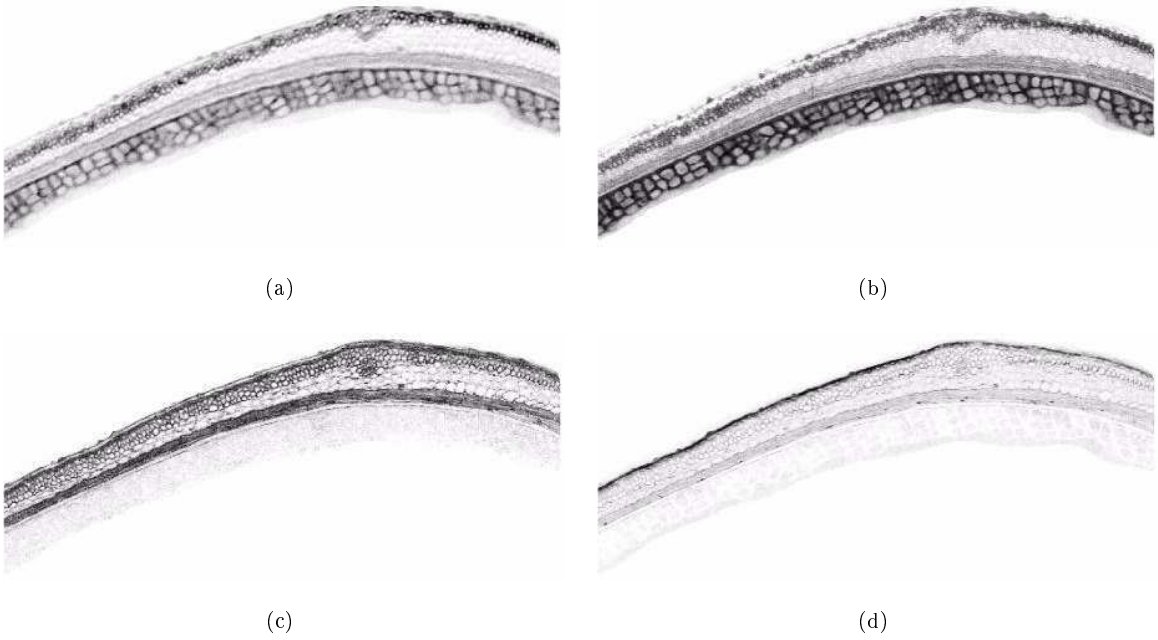
3.5.2.3 Traitement numérique

Le cube de données est de dimensions $512 \times 512 \times 19$. Après concaténation de ce cube suivant ses lignes, une matrice de données de dimensions 262144×19 est à disposition. Or, les images de la figure 3.22 sont composées essentiellement de pixels de faibles intensités localisés dans la partie représentant l'amidon sur les images. L'étude cherchant à mettre en évidence les structures des enveloppes du grain d'orge et non l'amidon, et afin de limiter au maximum les temps de calcul lors de la factorisation matricielle, ces pixels ont été écartés de l'analyse. Au final, une matrice de taille 47021×19 est traitée numériquement. Cette matrice $\mathbf{X}$ est décomposable selon le modèle de l'équation (3.8). De plus, les objets manipulés étant des images, et les parties sous-jacentes à en extraire étant également des images, les matrices de facteurs $\mathbf{A}$ et

S doivent être composées exclusivement d'éléments non-négatifs. Nous sommes donc à nouveau confronté au problème de la FMN à savoir estimer des matrices positives qui factorisent au mieux la matrice de données **X**. La même démarche que dans la section 3.5.1 a été menée sur ce jeu de données [46]. Le nombre d'images sources sous-jacentes au modèle est pris égal à 4 comme suggéré par une ACP préliminaire.

3.5.2.4 Résultats

Les répartitions spatiales des 4 espèces chimiques fluorescentes et présentes dans le grain d'orge sont présentées sur la figure 3.23. Seules, ces distributions spatiales ne représentent rien. Il est indispensable de pouvoir attacher à chaque distribution une espèce chimique. Ceci est rendu possible par l'étude des signatures spectrales (ou spectres hybrides selon la dénomination choisie à la section 1.4.1.1, page 26) estimées par FMN et visibles sur la figure 3.24. Les figures 3.23(a), 3.23(b), 3.23(c) et 3.23(d) sont les distributions spatiales des espèces chimiques dont les signatures spectrales sont estimées respectivement par les figures 3.24(a), 3.24(b), 3.24(c) et 3.24(d). La comparaison de ces signatures et des signatures de référence des éléments connus comme autofluorescents dans le grain d'orge va nous permettre d'associer des noms à chaque signature estimée et donc à chaque distribution spatiale.



(a) (b)

(c) (d)

FIG. 3.23 – Répartitions spatiales des espèces chimiques pures estimées par minimisation de la divergence

Les spectres hybrides de référence de la lignine et de l'acide férulique sont disponibles sur les figures 3.24(a), 3.24(b) et 3.24(b), et sont représentés respectivement par des tirets et des pointillés. Ces deux espèces sont connues comme constituants des couches externes du grain d'orge. Les deux spectres hybrides estimés des figures 3.24(a) et 3.24(b) reconstituent à eux deux le spectre hybride de l'acide férulique quasiment parfaitement. Le spectre hybride de la lignine est lui aussi très bien estimé par le spectre hybride de la figure 3.24(c).

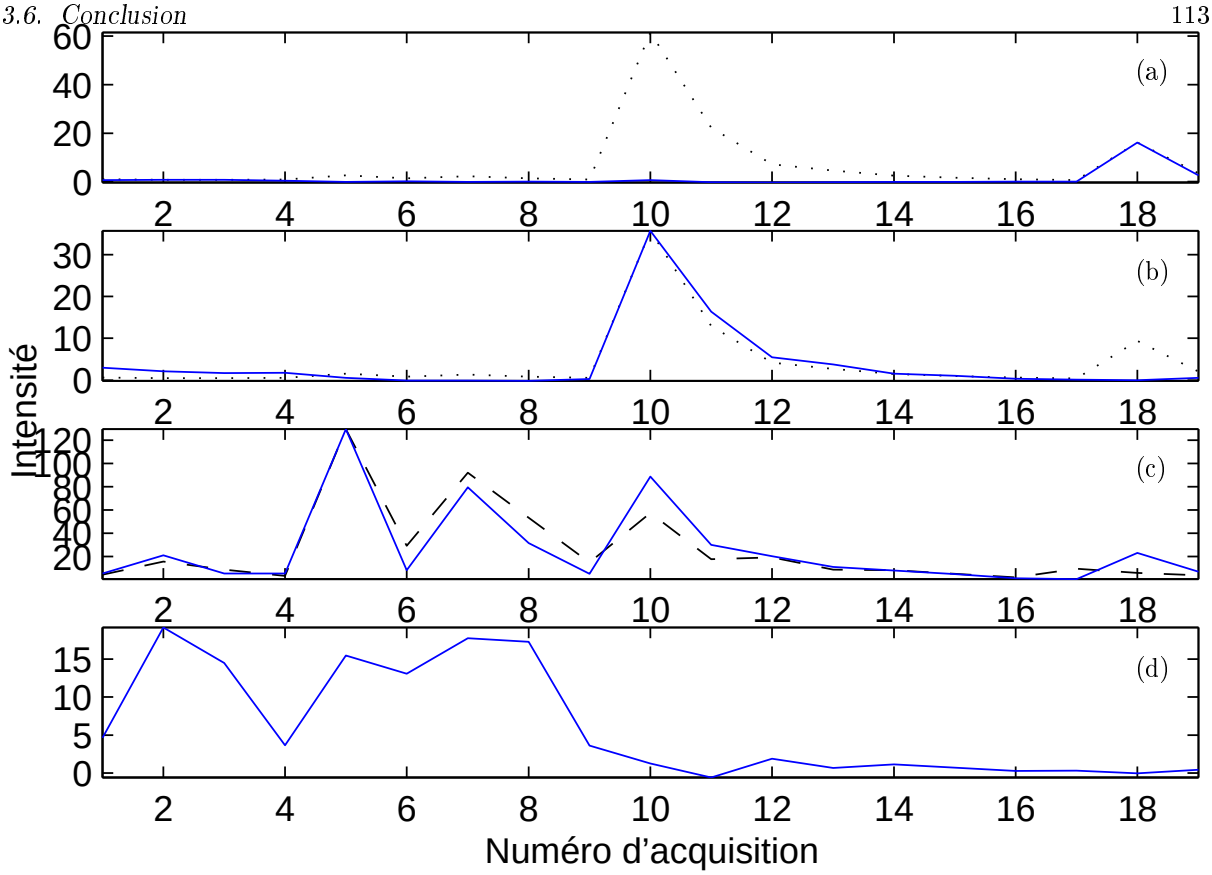


FIG. 3.24 – Spectres hybrides estimés des espèces chimiques présentes dans le grain d'orge (en bleu) et spectres hybrides de référence de l'acide férulique (en pointillés) et de la lignine (en tirets)

La cutine est la troisième espèce fluorescente composant le grain d'orge. Nous ne disposons pas de spectre hybride de référence pour la cutine. Mais la répartition spatiale de cette espèce, disponible à la figure 3.23(d), montre qu'elle est concentrée sur la partie supérieure de la couche cireuse, tout comme la cutine. Le spectre estimé à la figure 3.24(d) est donc associé à la cutine. La lignine, quant à elle, est un constituant majeur du péricarpe et de la partie inférieure de la couche cireuse. L'acide férulique est présent en majorité dans la couche à aleurone et les glumelles.

Cette analyse numérique par FMN des spectres de fluorescence enregistrés sur un grain d'orge et l'interprétation de ces résultats a donné lieu à un article de conférence [46].

3.6 Conclusion

Face à l'inexistence de méthodes adaptées à l'analyse réaliste de données positives factorisables en matrices elles-mêmes positives, un nouveau champ de recherche en séparation de sources positives a été développé. Depuis 20 ans, le potentiel de contraintes de positivité appliquées à un modèle de factorisation de données est étudié et a mené au développement d'algorithmes basés sur des règles de mise à

jour multiplicatives et appelés algorithmes de Factorisation en Matrices Non-négatives. Quatre approches principales ont été développées. Les deux premières cherchent à minimiser une mesure de distance entre les données et leur approximation par le modèle : une méthode exploite la distance euclidienne et l'autre utilise la divergence. La troisième approche cherche à minimiser l'erreur de reconstruction mais en imposant des contraintes de parcimonie sur les matrices de la factorisation à estimer. Une quatrième cherche à représenter des composantes positives spatiallement localisées pour réaliser une décomposition d'un objet en parties. Seule la méthode par contraintes de parcimonie possède une preuve théorique de convergence vers un minimum global de la fonction objectif. Une précaution lors de l'utilisation de ces algorithmes est de les lancer plusieurs fois en différents points d'initialisation.

Cependant, la représentation évidente des spectres de fluorescence par le modèle de la FMN et l'interprétation physique aisée des estimations de ces méthodes nous ont conduit à les appliquer sur des spectres de fluorescence enregistrés sur des grains de céréales pour étudier la composition chimique de leurs structures biologiques. Pour un grain de blé, les spectres estimés ont été associés aux acides férulique lié, para-coumarique et férulique libre. Les profils de concentration estimés de chacune de ces espèces ont permis d'identifier la localisation de chacune de ces espèces dans le grain de blé : l'acide férulique libre est fortement concentré dans la couche à aleurone, l'acide para-coumarique se concentre principalement dans la pliure du grain, et l'acide férulique lié se condense dans la couche à aleurone et dans la zone entourant la pliure du grain. L'application de la FMN sur des spectres acquis sur un grain d'orge conduit à l'estimation des spectres hybrides de la lignine, de l'acide férulique et de la cutine. Chacune de ces espèces a été respectivement localisée dans le péricarpe et la couche cireuse supérieure, dans la couche à aleurone et les glumelles, et dans la partie supérieure de la couche cireuse. La FMN s'est ainsi révélée un outil efficace de traitement des spectres de fluorescence et a permis de prouver que la couche à aleurone, indicatrice de la qualité d'une farine, d'un grain de blé et d'un grain d'orge est composée d'acide férulique qui est exclusivement concentré dans cette couche. Cet acide pourra donc être utilisé comme indicateur de la contamination d'une farine par les sons.

Chapitre 4

Application de l'Analyse en Composantes Indépendantes à la spectroscopie Raman

Sommaire

4.1 Introduction

Les méthodes classiques de traitement des spectres Raman sont supervisées et n'exploitent pas les propriétés statistiques de ces signaux (voir section 2.4.5, page 66), or celles-ci sont exploitables par l'Ana-

lyse en Composantes Indépendantes ou ACI. Cette technique de Séparation Aveugle de Sources utilise l'hypothèse d'indépendance statistique mutuelle des sources recherchées afin de proposer une estimation de ces sources. Il s'avère que les spectres Raman des espèces chimiques pures composant un échantillon biologique respectent cette hypothèse d'indépendance de par leur forme parcimonieuse composée de pics étroits. Dans une première partie, l'ACI est proposée comme un outil de résolution de la Séparation Aveugle de Sources (SAS). Ses principes, ses hypothèses, ses prétraitements, ses fonctions objectifs, ses algorithmes classiques d'optimisation et ses nombreuses applications sont ensuite proposés.

Une nouvelle application de l'ACI à la spectroscopie Raman est étudiée dans une deuxième partie. Il s'agit du déparaffinage numérique d'échantillons de tissus enrobés d'une couche de paraffine à des fins de stockage. Cette méthode s'appuie sur la combinaison de la spectroscopie Raman pour acquérir les informations vibrationnelles, véritables empreintes digitales de l'échantillon paraffiné à étudier, et de l'ACI pour numériquement nettoyer les spectres acquis de l'influence de la paraffine qui est gênante en spectroscopie Raman. Cette méthode est appliquée à des échantillons de peau. Le spectre de la peau est parfaitement restitué malgré la présence de paraffine qui se décompose quant à elle en trois sources distinctes, contrairement aux méthodes classiques qui la modélise par un spectre Raman unique. Des études sur des blocs de paraffine seule ont confirmé ce modèle à trois sources. La méthode de déparaffinage numérique a été ensuite testée sur des échantillons de mélanomes et de nævi. Ces applications démontrent sa robustesse à restituer les informations spectrales, même les plus délicates, liées à la peau. Ceci conduit à une discrimination entre mélanome et nævus après l'étape de déparaffinage numérique par l'ACI.

4.2 ACI et spectroscopie Raman

Dans la partie 2.4.5, page 66, des méthodes de traitement numérique des spectres Raman ont été étudiées. Elles cherchent à estimer les spectres sources des espèces chimiques de la solution ou de l'échantillon biologique analysé. Bien que potentiellement efficaces pour certaines applications, ces méthodes restent supervisées. Un expert informe l'algorithme sur les bandes spectrales qui respectent la modélisation des données imposée par la méthode. La gestion minutieuse de paramètres d'optimisation par l'utilisateur rend ces techniques tributaires d'un analyste qui maîtrise parfaitement le comportement de l'algorithme et la physique du phénomène étudié.

Le développement d'une méthode d'estimation des spectres sources reposant sur une propriété globale du spectre Raman, plutôt que sur une propriété locale, permet d'automatiser la procédure d'estimation. Les spectres Raman des espèces chimiques pures sont souvent composés de quelques pics étroits et caractéristiques de l'espèce considérée, le reste des spectres avoisine l'intensité nulle. D'une espèce à une autre, les spectres ne se recouvrent que très peu spectralement. L'association de ces caractéristiques traduit implicitement l'indépendance mutuelle des spectres sources. Cette propriété des spectres Raman sera étudiée de manière plus approfondie à la section 4.4.4. Cette propriété d'indépendance est suffisante

pour estimer les spectres sources à partir des spectres originaux en utilisant des techniques de Séparation Aveugle de Sources (SAS). Les méthodes d'Analyse en Composantes Indépendantes ou ACI[18], qui sont des techniques de SAS, sont spécifiquement désignées pour cette tâche, comme nous le verrons à la section 4.4. Une étude comparative avec l'application de l'ACP, technique couramment employée en analyse de spectres Raman et présentée à la section 2.3.5.2, page 52, est faite à l'annexe A. Cette étude montre les limites de l'ACP appliquée à des spectres Raman. Les résultats estimés par cette méthode ne sont pas physiquement interprétables. De même, les techniques de FMN étudiées dans le chapitre 3 ont été appliquées mais les hypothèses de positivité des sources et des mélanges ne sont pas suffisantes pour estimer des résultats cohérents. Cette étude fait l'objet de l'annexe E.

4.3 L'Analyse en Composantes Indépendantes

4.3.1 Le modèle des mélanges

En spectroscopie Raman, les signaux enregistrés suivent un modèle linéaire et instantané comme expliqué au paragraphe 2.2.3, page 40, et les sources sont considérées comme stationnaires. Le modèle de séparation à étudier dans ce chapitre doit donc lui aussi être instantané.

La théorie de l'Analyse en Composantes Indépendantes a été développée pour des variables aléatoires. Afin de faciliter les explications de cette théorie, nous nous placerons dans ce cas de figure.

Soit un ensemble de variables aléatoires observées *connus* x_i avec $i \in \{1, \ldots, N\}$ où $N \in \mathbb{N}$ est le nombre de variables observées. Le terme *connus* traduit la connaissance d'une réalisation des processus x_i. Leurs densités de probabilité ne sont pas disponibles et seules leurs statistiques sont estimables. Les variables observées s'écrivent sous forme vectorielle $\mathbf{x}(t) = [x_1(t), \ldots, x_N(t)]^T$. Ces variables x_i sont supposées générées par des combinaisons linéaires de variables aléatoires *inconnues* s_j avec $j \in \{1, \ldots, p\}$ où $p \in \mathbb{N}$ est le nombre de variables cachées, *inconnues* signifiant qu'aucune information n'est disponible sur ces variables. Par la suite, ce terme perdra son sens original pour signifier plus exactement que très peu d'informations sont supposées connues sur ces variables. Ces variables sont appelées les sources puisqu'elles sont à l'origine des variables aléatoires observées x_i. Les sources peuvent toutes être regroupées sous forme de vecteur $\mathbf{s} = [s_1, \ldots, s_p(t)]^T$.

Ce modèle génératif s'écrit de la manière suivante :

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{4.1}$$

où $\mathbf{A} = \{a_{ij} \mid i \in \{1, \ldots, N\}\ ,\ j \in \{1, \ldots, p\}\}$ est une matrice de mélange. Ce modèle est identique au modèle génératif des spectres Raman de l'équation (2.2) à la page 43 où une longueur d'onde représente une réalisation des variables aléatoires.

[18] Independent Component Analysis ou ICA en anglais

La définition ci-dessus du modèle de mélange ne suffit évidemment pas à estimer de façon unique les sources sous-jacentes de ce modèle. L'équation 4.1 est vérifiée pour tout couple de matrices $\{\mathbf{A}, \mathbf{A}^{-1}\mathbf{x}\}$, $\mathbf{A}$ étant une matrice non-singulière. De nouvelles hypothèses doivent compléter la modélisation des données pour proposer une solution unique au problème de séparation de sources.

L'utilisation d'hypothèses supplémentaires sur la structure des mélanges est possible dans certaines applications. Mais à moins de connaître parfaitement la matrice $\mathbf{A}$, ces hypothèses ne feront que réduire l'espace des solutions du problème représenté par l'équation (4.1). Pour conserver une forme générale au problème et lui désigner une méthode de résolution générale, des hypothèses cette fois-ci sur les sources $\mathbf{s}(t)$ doivent être posées.

4.3.2 L'indépendance statistique

En 1985, Hérault et Jutten [60] se sont retrouvés face à ce problème dans le cadre de la séparation de l'information sur l'étirement d'une articulation et de l'information de la vitesse de cet étirement à partir d'enregistrements de l'activité de fibres nerveuses. Dans [70], ils proposent un algorithme efficace de séparation basé sur une architecture neuromimétique. L'indépendance statistique mutuelle des sources recherchées s'avère être l'hypothèse fondamentale responsable du succès de cet algorithme. Bien que peu contraignante, elle se révèle suffisamment puissante pour proposer une solution unique au modèle décrit précédemment.

L'indépendance statistique mutuelle entre p variables aléatoires s_i pour $i \in \{1, \ldots, p\}$ signifie que la connaissance des valeurs prises par certaines variables n'informe en aucune manière sur les valeurs prises par les autres variables. Considérons $\mathbf{s} = [s_1, \ldots, s_p]^T$ comme étant un vecteur aléatoire prenant ses valeurs dans $\mathbb{R}^p$, et supposons que sa densité de probabilité $f(\mathbf{s})$ existe. Le vecteur $\mathbf{s}$ a des composantes mutuellement indépendantes si et seulement si :

$$f(\mathbf{s}) = \prod_{i=1}^{p} f_i(s_i). \tag{4.2}$$

Cette equation traduit le fait que des composantes sont indépendantes si la densité de probabilité conjointe $f(\mathbf{s})$ est factorisable par les densités de probabilité marginales $f_i(s_i)$.

L'indépendance doit être différenciée de la décorrélation qui est une indépendance à l'ordre 2 uniquement. En effet, la décorrélation entre deux variables aléatoires s_i et s_j se traduit par :

$$E\{s_i s_j\} - E\{s_i\}E\{s_j\} = 0, \text{ pour } i \neq j.$$

L'indépendance est une hypothèse beaucoup plus exigeante que la décorrélation puisqu'elle suppose l'annulation de tous les cumulants croisés d'ordre supérieur à 2. Elle peut également s'exprimer sous la forme [66, chapitre 2] :

$$E\{g_1(s_i)g_2(s_j)\} - E\{g_1(s_i)\}E\{g_2(s_j)\} = 0, \text{ pour } i \neq j$$

et pour toutes fonctions g_1 et g_2. L'indépendance apparaît comme une condition beaucoup plus stricte que la décorrélation.

L'ACI regroupe un ensemble de méthodes à philosophie commune basée sur l'estimation de sources indépendantes.

4.3.3 Définition de l'ACI

Définition : L'ACI propose un ensemble de méthodes basées sur les mêmes principes. Une définition générale de l'ACI peut être la suivante [64] :
l'ACI d'un vecteur aléatoire $\mathbf{x} \in \mathbb{R}^N$ *consiste à estimer le modèle génératif des données* $\mathbf{x} = \mathbf{As}$*, avec* $\mathbf{A} \in \mathbb{R}^{N \times p}$*, en déterminant une transformation linéaire* $\mathbf{s} = \mathbf{Wx}$*, avec* $\mathbf{W} \in \mathbb{R}^{p \times N}$*, de telle manière que les composantes de* $\mathbf{s} \in \mathbb{R}^p$ *soient aussi indépendantes que possible par maximisation d'une fonction* $F(\mathbf{s})$ *qui est une mesure de l'indépendance statistique.*

Un modèle plus exacte de l'ACI introduit du bruit dans les mesures. Afin de simplifier le problème, ce bruit est omis dans la suite de ce mémoire puisque le problème d'ACI sans bruit de mesure est assez complexe en lui-même. De plus, ce modèle sans bruit suffit dans de nombreuses applications [66].

Les sources s_i ne sont pas observées directement. Leurs densités de probabilité $f_i(s_i)$ ne sont évidemment pas connues car il est rare en pratique d'avoir une connaissance étendue de ces fonctions. Il est donc impossible en général d'utiliser la définition de l'indépendance de l'équation (4.2) pour estimer des sources indépendantes.

Les méthodes d'ACI reposent donc sur une mesure d'indépendance qui reste à définir. Les différentes approches se différencient par la mesure d'indépendance retenue et par la méthode d'optimisation choisie. Mais toute méthode d'ACI repose sur les mêmes hypothèses et restrictions que nous allons répertorier dans le paragraphe suivant.

Hypothèses : Les composantes s_i, $i \in \{1, \ldots, p\}$, sont supposées statistiquement indépendantes. Cette hypothèse est fondamentale pour garantir l'estimation du modèle direct $\mathbf{x} = \mathbf{As}$ de l'ACI.

Les composantes indépendantes doivent avoir des distributions non-gaussiennes. Toutes les informations des variables aléatoires gaussiennes sont contenues dans la matrice de covariance dont l'exploitation peut conduire au mieux à la décorrélation. Or l'ACI exploite les informations contenues ailleurs que dans la matrice de covariance, notamment celles contenues dans les tenseurs définis par les cumulants croisés. Il s'avère qu'au plus une composante indépendante peut avoir une distribution gaussienne.

Le nombre N de composantes du vecteur $\mathbf{x}$ doit être supérieur ou égal au nombre p de composantes indépendantes du vecteur $\mathbf{s}$. Dans le cas contraire, le mélange est dit sous-déterminé et le problème n'est pas soluble sans connaissances *a priori* supplémentaires sur les sources puisque même si la matrice de

mélange $\mathbf{A}$ est connue, elle n'est pas inversible.

Les méthodes d'ACI exploitées dans la suite de ce mémoire divisent le problème de séparation en deux, à savoir une étape d'identification de la matrice de mélange $\mathbf{A}$ et une étape d'extraction des sources [13]. Mais d'autres méthodes telles que l'algorithme de Hérault et Jutten [70] mettent à jour les sorties en premier plutôt que les mélanges, mais elles ne seront pas utilisées ici. Dans la suite de ce mémoire, nous nous placerons dans le cas plus simple où le nombre de sources est inférieur ou égal au nombre de signaux mélanges observés.

La matrice de mélange $\mathbf{A}$ doit être inversible c'est-à-dire que ses colonnes doivent être linéairement indépendantes. Les mélanges redondants peuvent être éliminés de l'analyse tout en s'assurant que le nombre de lignes de $\mathbf{A}$ reste supérieur ou égal à son nombre de colonnes pour éviter le cas sous-déterminé.

Sous ces hypothèses, le modèle de l'ACI présenté dans le paragraphe précédent est identifiable. La matrice de mélange $\mathbf{A}$ et le vecteur des sources $\mathbf{s}$ sont ainsi estimables sur la seule connaissance du vecteur des observations $\mathbf{x}$. Mais quelques indéterminations inhérentes à la modélisation de l'ACI subsistent.

Indéterminations : Le modèle linéaire et instantané de l'ACI se réécrit facilement sous la forme :

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_{j=1}^{p} \mathbf{a}_j s_j$$

où le vecteur $\mathbf{a}_j$ représente la j-ème colonne de la matrice de mélange $\mathbf{A}$. Cette équation est strictement équivalente à :

$$\mathbf{x} = \sum_{j=1}^{p} \left(\frac{1}{\alpha_j}\mathbf{a}_j\right)(\alpha_j s_j)$$

où les α_j sont des constantes non nulles. Ainsi la multiplication d'une source s_j par toute constante α_j peut être annulée par la division de la colonne correspondante $\mathbf{a}_j$ de $\mathbf{A}$ par la même constante α_j sans influer sur les hypothèses posées sur le modèle de l'ACI. Cette indétermination traduit l'impossibilité d'estimer les énergies des composantes indépendantes. C'est pourquoi les sources s_j seront supposées être de variance unité, sans perte de généralité, c'est-à-dire que $\mathbf{R_s} = E\{(\mathbf{s} - E\{\mathbf{s}\})(\mathbf{s} - E\{\mathbf{s}\})^T\} = \mathbf{I_p}$ où $\mathbf{I_p}$ est la matrice identité de dimensions $p \times p$. Le signe de chaque composante est également impossible à déterminer puisque le choix $\alpha_j = -1$ est possible.

Le modèle de l'ACI reste valide sous la transformation linéaire suivante :

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \mathbf{A}\mathbf{P}^{-1}\mathbf{P}\mathbf{s}$$

où la matrice $\mathbf{P}$ est une matrice de permutation de dimensions $p \times p$. Dans la somme $\sum_{j=1}^{p} \mathbf{a}_j s_j$, les termes $\mathbf{a}_j s_j$ peuvent être permutés librement. L'ordre des composantes indépendantes n'est donc pas déterminable.

La prise en compte de ces deux indéterminations signifie que les sources indépendantes ne sont extractibles qu'à une matrice $\mathbf{\Delta} = \mathbf{D}\mathbf{P}$ près, où la matrice $\mathbf{D}$ est diagonale, d'éléments α_j pour $j \in \{1, \ldots, p\}$

et de rang plein, et la matrice $\mathbf{P}$ est une matrice de permutation. Des prétraitements sur le vecteur des données, tels que le centrage et le blanchiment, permettent de s'affranchir de certaines indéterminations et de simplifier le problème de séparation.

4.3.4 Prétraitements

4.3.4.1 Centrage

Afin de simplifier la théorie et les algorithmes d'ACI, les composantes du vecteur des observations $\mathbf{x}$ sont forcées à une moyenne nulle sans perte de généralité. Ceci est lié aux cumulants croisés et à leurs estimateurs qui sont d'expression beaucoup plus simple dans le cas de variables aléatoires centrées [72]. Cette étape est réalisée par une simple soustraction de la moyenne des différentes composantes estimée à partir de leurs réalisations. Le vecteur $\mathbf{x}$ est ainsi transformé en :

$$\mathbf{x^c} = \mathbf{x} - E\{\mathbf{x}\}.$$

Par cette opération, les composantes indépendantes sont elles-aussi centrées puisque :

$$E\{\mathbf{s^c}\} = \mathbf{A}^{\#} E\{\mathbf{x^c}\} = 0$$

où le sigle $^{\#}$ définit l'inversion matricielle de Moore-Penrose, encore appelée pseudo inverse, qui est utilisée lorsque la matrice $\mathbf{A}$ n'est pas carrée. Lorsqu'elle est carrée, l'inversion matricielle classique est utilisée et l'équation s'écrit : $E\{\mathbf{s^c}\} = \mathbf{A}^{-1} E\{\mathbf{x^c}\} = 0$. La matrice de mélange $\mathbf{A}$ n'est pas modifiée par cette opération qui peut donc toujours être appliquée aux données sans affecter l'estimation de $\mathbf{A}$.

4.3.4.2 Blanchiment

Cette étape consiste à décorréler et à imposer une variance unité aux variables du vecteur $\mathbf{x^c}$. Cette transformation linéaire est réalisée en multipliant $\mathbf{x^c}$ par une matrice $\mathbf{V}$ de dimensions $p \times N$ afin d'obtenir le vecteur

$$\mathbf{z} = \mathbf{V}\mathbf{x^c}$$

de dimensions $p \times 1$ où les composantes de $\mathbf{z}$ sont décorrélées et de variance unité, ce qui signifie que la matrice de covariance de $\mathbf{z}$ vaut

$$\mathbf{R_z} = E\{\mathbf{z}\mathbf{z}^T\} = \mathbf{I_p} \tag{4.3}$$

où $\mathbf{I_p}$ est la matrice identité de dimensions $p \times p$.

Diverses méthodes ont été développées pour réaliser cette transformation et l'une des plus populaires est l'utilisation de l'ACP. La décomposition en valeurs propres de la matrice de covariance de $\mathbf{x^c}$ est définie par :

$$\mathbf{R_x} = E\{\mathbf{x^c}\mathbf{x^c}^T\} = \mathbf{E}\mathbf{D}\mathbf{E}^T$$

où la matrice $\mathbf{D}$ est une matrice diagonale de dimensions $N \times N$ dont les éléments diagonaux d_i avec $i \in \{1, \ldots, N]$ sont les valeurs propres de la matrice $\mathbf{R_x}$ et sont rangés par ordre décroissant, c'est-à-dire que $d_1 \geq d_2 \geq \ldots \geq d_\Lambda \geq \ldots$. La matrice $\mathbf{E}$ de dimensions $N \times N$ est la matrice orthogonale des vecteurs propres de la matrice $\mathbf{R_x}$. Dans le cas où le nombre de sources à estimer est plus petit que le nombre de signaux observés, c'est-à-dire dans le cas où $p < N$, et où le modèle est supposé sans bruit, seules p valeurs propres sont non nulles. En présence de bruit, l'ACP permet de projeter les données dans un espace de dimension p engendré par les sources. Seules les p premières valeurs propres et les vecteurs propres associés sont conservés. Notons $\tilde{\mathbf{D}}$ et $\tilde{\mathbf{E}}$ les versions tronquées respectives des matrices $\mathbf{D}$ et $\mathbf{E}$. Les matrices $\tilde{\mathbf{D}}$ et $\tilde{\mathbf{E}}$ sont de dimensions respectives $p \times p$ et $N \times p$. En définissant la matrice diagonale $\tilde{\mathbf{D}}^{-\frac{1}{2}}$ de dimensions $p \times p$ par :

$$\tilde{\mathbf{D}}^{-\frac{1}{2}} = \begin{pmatrix} \tilde{d_1}^{-\frac{1}{2}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \tilde{d_p}^{-\frac{1}{2}} \end{pmatrix}$$

la matrice de blanchiment $\mathbf{V}$ est calculée par :

$$\mathbf{V} = \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{E}}^T \tag{4.4}$$

et est effectivement de dimensions $p \times N$.

Observons les conséquences de ce blanchiment sur le modèle génératif de l'équation (4.1) :

$$\mathbf{z} = \mathbf{V}\mathbf{x^c} = \mathbf{V}\mathbf{A}\mathbf{s^c}.$$

En définissant la matrice $\mathbf{U} = \mathbf{VA}$ de dimensions $p \times p$ le modèle s'écrit sous la forme :

$$\mathbf{z} = \mathbf{U}\mathbf{s^c}. \tag{4.5}$$

La matrice de covariance $\mathbf{R_z}$ de $\mathbf{z}$ peut se mettre sous la forme :

$$\mathbf{R_z} = \mathbf{U}E\{\mathbf{s^c}\mathbf{s^{c}}^T\}\mathbf{U}^T = \mathbf{U}\mathbf{R_s}\mathbf{U}^T = \mathbf{U}\mathbf{U}^T. \tag{4.6}$$

L'identification entre les équations (4.3) et (4.6) donne l'égalité $\mathbf{U}\mathbf{U}^T = \mathbf{I_p}$. La matrice $\mathbf{U}$ est donc orthogonale.

Le blanchiment des données ne permet évidemment pas d'assurer l'estimation des composantes indépendantes, mais simplifie celle-ci de moitié. Il est nécessaire d'estimer $p(p-1)/2$ inconnues de la matrice orthogonale $\mathbf{U}$ au lieu des $N \times p$ éléments de la matrice $\mathbf{A}$. L'estimation de la matrice $\mathbf{U}$ repose sur des fonctions objectif qui représentent une mesure d'indépendance.

Le blanchiment des données n'est pas systématique. Certains algorithmes d'ACI s'appliquent directement sur les données centrées comme l'algorithme Infomax de Bell et Sejnowski [10]. Le blanchiment est cependant recommandé puisqu'il permet d'accélérer la convergence des algorithmes [66]. La suite de cette section présente les principaux algorithmes d'ACI, JADE et FastICA, qui requièrent le blanchiment préalable des données.

4.3.5 Mesures d'indépendance et algorithmes

L'indépendance entre des composantes est estimée par quelques mesures caractéristiques et autour desquelles ont été développés différents algorithmes [20, 25, 63]. Certains ont eu un impact considérable dans la communauté de l'ACI par leur efficacité, leur généralité, et leur vitesse de convergence. Dans cette section, nous nous proposons de présenter ces mesures d'indépendance et les algorithmes associés utilisés dans les applications en spectroscopie Raman développées dans la suite du mémoire, à savoir FastICA et JADE.

4.3.5.1 Non-gaussianité et FastICA

Les sources à distribution gaussienne ne sont pas estimables par l'ACI si aucune autre hypothèse sur les signaux n'est faite[19]. Une mesure naturelle d'indépendance est donc la maximisation de la non-gaussianité. D'après le théorème "central limite" , une somme de variables aléatoires indépendantes et de même densité de probabilité tend vers une variable gaussienne lorsque le nombre de variables aléatoires tend vers l'infini. Ainsi, les composantes du vecteur $\mathbf{x}$ des mélanges possèdent une distribution plus proche de la distribution gaussienne que les composantes du vecteur $\mathbf{s}$ des sources.

La négentropie est un concept hérité de la théorie de l'information qui fournit une mesure de la non-gaussianité d'une variable aléatoire. Sa définition dépend de l'entropie que nous allons définir. Dans son contexte original, l'entropie mesure le degré d'incertitude d'une variable aléatoire. Plus une variable est aléatoire, non prédictible et non structurée, et plus son entropie est grande. Sous forme mathématique, l'entropie d'un vecteur aléatoire $\mathbf{s}$ de densité de probabilité $f(\mathbf{s})$ est définie par :

$$H(\mathbf{s}) = -\int f(\mathbf{s}) \log f(\mathbf{s}) d\mathbf{s}.$$

Cette quantité est maximisée par un vecteur gaussien parmi l'ensemble des vecteurs de même matrice de covariance.

La négentropie a été construite afin de fournir une quantité qui soit nulle pour une variable gaussienne et qui soit non-négative pour tout autre type de variable aléatoire. Elle est définie comme la différence entre l'entropie d'un vecteur gaussien $\mathbf{u}$ et l'entropie du vecteur considéré $\mathbf{s}$. Le vecteur gaussien utilisé dans la mesure de la négentropie est contraint à avoir une matrice de covariance identique à celle de $\mathbf{s}$. Sous forme mathématique, la négentropie est définie par :

$$J(\mathbf{s}) = H(\mathbf{u}) - H(\mathbf{s})$$

où $\mathbf{u}$ est un vecteur gaussien de même matrice de covariance que $\mathbf{s}$.

Dans sa forme actuelle, la négentropie n'est pas attractive pour l'estimation du modèle de l'ACI puisqu'elle dépend de la densité de probabilité du vecteur $\mathbf{s}$ qui est supposé inconnu et dont l'estimation

[19] Des hypothèses de non-stationnarité ou de corrélation temporelle des sources gaussiennes permettent d'estimer des mélanges de telles sources [19]

est particulièrement lourde. Une approximation de la négentropie a été développée dans [62] et s'écrit dans le cas monodimensionnel :

$$J(s_j) \approx c[E\{G(s_j)\} - E\{G(u_j)\}]^2$$

où G peut être toute fonction non-quadratique et c une constante.

Les composantes indépendantes sont obtenues par $\mathbf{Wz}$ où $\mathbf{z}$ est le vecteur des données blanchies et où la matrice $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_p]^T$ est estimée de manière à ce que ses lignes $\mathbf{w}_j^T$ maximisent la fonction objectif suivante :

$$J_G(\mathbf{W}) = \sum_{j=1}^{p} [E\{G(\mathbf{w}_j^T \mathbf{z})\} - E\{G(\nu)\}]^2 \tag{4.7}$$

sous la contrainte :

$$E\{(\mathbf{Wz})(\mathbf{Wz})^T\} = \mathbf{I_p}.$$

La variable ν est une variable gaussienne de même variance que $\mathbf{w}_j^T \mathbf{z}$.

Le choix de la fonction G doit obéir à certaines propriétés afin que l'estimateur $\mathbf{W}$ de la matrice $\mathbf{U}^T$ soit consistant, qu'il ait une variance asymptotique minimale et qu'il soit robuste [63]. Des fonctions ont été déduites de ces considérations et s'avèrent utilisables de manière générale :

$$\begin{aligned} G_1(s_j) &= \frac{1}{\alpha_1} \log\cosh(\alpha_1 s_j) \\ G_2(s_j) &= -\frac{1}{\alpha_2} \exp\left(-\frac{\alpha_2 s_j^2}{2}\right) \\ G_3(s_j) &= \frac{1}{4} s_j^4 \end{aligned}$$

où $1 \leq \alpha_1 \leq 2$ et $\alpha_2 \approx 1$ sont des constantes.

La minimisation de la fonction objectif (4.7) par la matrice $\mathbf{W}$ s'appuie sur un algorithme du point fixe. Chaque ligne de la matrice $\mathbf{W}$ est estimée selon une procédure itérative dont le schéma est le suivant :

$$\mathbf{w}_j = E\{\mathbf{z} g(\mathbf{w}_j^T \mathbf{z})\} - E\{g'(\mathbf{w}_j^T \mathbf{z})\} \mathbf{w}_j \tag{4.8}$$

$$\mathbf{w}_j = \mathbf{w}_j - \sum_{k=1}^{j-1} (\mathbf{w}_j^T \mathbf{w}_k) \mathbf{w}_k \tag{4.9}$$

$$\mathbf{w}_j = \frac{\mathbf{w}_j}{\sqrt{\mathbf{w}_j^T \mathbf{w}_j}} \tag{4.10}$$

où g et g' sont respectivement les dérivées de G et g. L'équation (4.8) estime une mise à jour de $\mathbf{w}_j$ qui va minimiser un peu plus la fonction objectif (4.7). L'équation (4.9) sert à rendre le vecteur $\mathbf{w}_j$ orthogonal aux lignes de $\mathbf{W}$ estimées aux étapes précédentes de l'algorithme. Enfin, l'équation (4.10) sert à assurer que les composantes indépendantes estimées soient de variance unité.

L'implémentation de cet algorithme garantit une estimation rapide d'une matrice composée de sources statistiquement indépendantes, la convergence ayant été prouvée dans [63], .

4.3.5.2 Cumulants d'ordre 4 et JADE

Fonctions caractéristiques : Une densité de probabilité est entièrement définie par la connaissance de l'une des deux fonctions caractéristiques $\phi_{\mathbf{s}}$ et $\psi_{\mathbf{s}}$ du vecteur aléatoire $\mathbf{s} \in \mathbb{R}^p$ définies par :

$$\phi_{\mathbf{s}}(\mathbf{u}) = E\{\exp(j\mathbf{u}^T\mathbf{s})\}$$

$$\psi_{\mathbf{s}}(\mathbf{u}) = \ln(\phi_s(\mathbf{u})).$$

La première fonction caractéristique $\phi_{\mathbf{s}}$ est utilisée pour définir les moments d'ordre r de $\mathbf{s}$ par :

$$E\{s_{\pi_1}s_{\pi_2}\dots s_{\pi_r}\} = (-j)^r \left.\frac{\partial^r \phi_{\mathbf{s}}(\mathbf{u})}{\partial u_{\pi_1}\partial u_{\pi_2}\dots\partial u_{\pi_r}}\right|_{\mathbf{u}=0}$$

avec $\pi_i \in \{1,\dots,p\}$ et $i \in \{1,\dots,r\}$.

La seconde fonction caractéristique $\psi_{\mathbf{s}}$ définit les cumulants d'ordre r de $\mathbf{s}$ par :

$$cum(s_{\pi_1}, s_{\pi_2}, \dots, s_{\pi_r}) = (-j)^r \left.\frac{\partial^r \psi_{\mathbf{s}}(\mathbf{u})}{\partial u_{\pi_1}\partial u_{\pi_2}\dots\partial u_{\pi_r}}\right|_{\mathbf{u}=0}.$$

Une densité de probabilité est entièrement définie par la connaissance soit de tous ses moments $E\{s_{\pi_1}s_{\pi_2}\dots s_{\pi_r}\}$, soit de tous ses cumulants $cum(s_{\pi_1}, s_{\pi_2}, \dots, s_{\pi_r})$, à tous les ordres r.

Cumulants : Les cumulants possèdent la propriété de multilinéarité. La multilinéarité permet d'exprimer les cumulants d'un vecteur $\mathbf{s} = \mathbf{Wz}$ en fonction des cumulants du vecteur $\mathbf{z}$ et de la transformation linéaire $\mathbf{W} = \{w_{ij} \mid (i,j) \in \{1,\dots,p\}^2\}$ par [25] :

$$cum(s_{\pi_1}, s_{\pi_2}, \dots, s_{\pi_r}) = \sum_{l_1,l_2,\dots,l_r} w_{\pi_1 l_1} w_{\pi_2 l_2} \dots w_{\pi_r l_r} cum(z_{l_1}, z_{l_2}, \dots, z_{l_r})$$

avec $(\pi_i, l_i) \in \{1,\dots,p\}^2$ et $i \in \{1,\dots,r\}$.

Les cumulants peuvent être exprimés sous forme de tenseurs [26]. Un tenseur peut être vu comme la généralisation de la notion de matrices. Les tenseurs de cumulants sont donc une généralisation de la notion de matrice de covariance. Les éléments de la diagonale principale des tenseurs de cumulants sont les cumulants marginaux du vecteur aléatoire considéré et les éléments non diagonaux sont nommés cumulants croisés. Un vecteur à composantes statistiquement mutuellement indépendantes offre des cumulants croisés nuls et des cumulants marginaux maximums :

$$cum(s_{\pi_1}, s_{\pi_2}, \dots, s_{\pi_r}) \neq 0 \text{ si et seulement si } \pi_1 = \pi_2 = \dots = \pi_r.$$

Une autre propriété intéressante est l'annulation des cumulants d'ordre supérieur à 2 de vecteurs gaussiens. Cette propriété est aussi appelée réjection gaussienne [20]. L'influence éventuelle du bruit additif supposé gaussien est éliminée par la considération simultanée de cette propriété et de la propriété d'additivité des cumulants d'une somme de deux vecteurs aléatoires.

L'estimation des cumulants se révèle être une alternative à l'estimation difficile de la densité de probabilité des sources. Cependant, il est impossible d'estimer tous les cumulants d'un vecteur aléatoire. C'est la raison pour laquelle l'estimation se limite à leur ordre 4 [89]. Une autre justification à l'utilisation des cumulants d'ordre 4 est l'annulation des cumulants d'ordre 3 pour des densités de probabilité symétriques.

Cumulants d'ordre 4 : Le tenseur des cumulants d'ordre 4 possède 4 dimensions. Chaque élément $cum(z_i, z_j, z_k, z_l)$, avec $(i, j, k, l) \in \{1, \ldots, p\}^4$, est un cumulant d'ordre 4 du vecteur de données $\mathbf{z}$. Le tenseur des cumulants d'ordre 4 du vecteur $\mathbf{z}$ sera noté $\mathbf{Cum_z}$ et chacun de ses éléments $Cum_{\mathbf{z}}(i, j, k, l)$ représentera le cumulant d'ordre 4, $cum(z_i, z_j, z_k, z_l)$. Il est donc naturel que certains auteurs aient cherché à généraliser les méthodes de décorrélation de variables. La décomposition en valeurs propres de la matrice de covariance assure sa diagonalisation, c'est-à-dire l'annulation de ses termes non diagonaux. Une généralisation de cette méthodologie a été proposée par Cardoso et Souloumiac [20] pour diagonaliser le tenseur $\mathbf{Cum_z}$ des cumulants d'ordre 4.

Tout comme une matrice est un opérateur linéaire de l'espace des vecteurs de dimensions p, le tenseur $\mathbf{Cum_z}$ des cumulants d'ordre 4 est un opérateur linéaire de l'espace des matrices de dimensions $p \times p$. La matrice $\mathbf{M} = \{m_{kl} \mid (k, l) \in \{1, \ldots, p\}^2\}$ donnée par cette transformation linéaire est définie par :

$$f_{ij}(\mathbf{M}) = \sum_{k=1}^{p} \sum_{l=1}^{p} m_{kl} Cum_{\mathbf{z}}(i, j, k, l).$$

Décomposition en valeurs propres : Comme tout opérateur linéaire symétrique, le tenseur des cumulants d'ordre 4 possède une décomposition en valeurs propres [66]. Dans le cas d'une matrice de covariance, la décomposition en valeurs propres est définie par des vecteurs propres et des valeurs propres. Dans le cas d'un tenseur d'ordre 4, elle peut être définie par p matrices propres $\mathbf{M}^q$ de $\mathbf{Cum_z}$ et p valeurs propres λ_q de $\mathbf{Cum_z}$, $q \in \{1, \ldots, p\}$, selon la décomposition suivante [66, chapitre 11] :

$$\mathbf{F}(\mathbf{M}^q) = \lambda_q \mathbf{M}^q$$

où $\mathbf{F} = \{f_{ij} \mid (i, j) \in \{1, \ldots, p\}^2\}$ est la matrice transformée de $\mathbf{M}$ par $\mathbf{Cum_z}$. Il a été montré dans [20] que les matrices propres $\mathbf{M}^q$, $q \in \{1, \ldots, p\}$ de $\mathbf{Cum_z}$ sont définies par :

$$\mathbf{M}^q = \mathbf{w}_q \mathbf{w}_q^T \text{ pour } q \in \{1, \ldots, p\}$$

où le vecteur $\mathbf{w}_q$ est une ligne de la matrice $\mathbf{W} = \mathbf{U}^T$ de l'équation (4.5).

Il est nécessaire d'estimer dans un premier temps les matrices propres $\mathbf{M}^q$, $q \in \{1, \ldots, p\}$, de $\mathbf{Cum_z}$. Dans ce but, les matrices $\mathbf{M}^q$ et $\mathbf{F}(\mathbf{M}^q)$ sont vectorisées[20] sous forme de vecteurs $\breve{\mathbf{m}}_q$ et $\breve{\mathbf{f}}(\breve{\mathbf{m}}_q)$. Deux nouvelles matrices $\breve{\mathbf{M}}$ et $\breve{\mathbf{F}}$ sont formées par $\breve{\mathbf{M}} = [\breve{\mathbf{m}}_1, \ldots, \breve{\mathbf{m}}_q, \ldots, \breve{\mathbf{m}}_p]$ $\breve{\mathbf{F}} = [\breve{\mathbf{f}}(\breve{\mathbf{m}}_1), \ldots, \breve{\mathbf{f}}(\breve{\mathbf{m}}_q) \ldots, \breve{\mathbf{f}}(\breve{\mathbf{m}}_p)]$. La matrice $\breve{\mathbf{F}}$ est traitée par une décomposition classique en valeurs propres dont la matrice estimée des

[20] vectoriser une matrice $\mathbf{M} \in \mathbb{R}^{p \times p}$ signifie que ses p^2 éléments sont considérés comme les éléments d'un vecteur $\mathbf{v} \in \mathbb{R}^{p^2}$

vecteurs propres n'est autre que la matrice $\breve{\mathbf{M}}$. Les vecteurs colonnes $\breve{\mathbf{m}}_q$ de $\breve{\mathbf{M}}$ sont réécrits sous forme de matrices par l'opération inverse de la vectorisation. Les matrices résultantes sont les matrices propres $\mathbf{M}^q$ recherchées de $\mathbf{Cum_z}$.

Les matrices $\mathbf{M}^q$ étant estimées, il reste à estimer la matrice $\mathbf{W}$. Les propriétés de multilinéarité, d'additivité et de réjection gaussienne permettent facilement de montrer que la matrice $\mathbf{W}$ diagonalise conjointement les p matrices $\mathbf{F}(\mathbf{M}^q)$, $q \in \{1,\ldots,p\}$ [20]. La matrice $\mathbf{W}$ peut donc être estimée de manière à rendre simultanément les matrices $\mathbf{WF}(\mathbf{M}^q)\mathbf{W}^T$ aussi diagonales que possible. Dans [20], les auteurs s'appuient sur ce principe pour définir la fonction objectif suivante à maximiser :

$$Q^{JADE}(\mathbf{W}) = \sum_{q=1}^{p} \left(\sqrt{\sum_{k=1}^{p} (\mathbf{W}\mathbf{M}^q\mathbf{W}^T)_{kk}^2} \right)^2 . \tag{4.11}$$

Diagonalisation conjointe : La maximisation de cette fonction, c'est-à-dire la diagonalisation simultanée des matrices $\mathbf{M}^q$, $q \in \{1,\ldots,p\}$, est réalisée par la technique de Jacobi étendue à plusieurs matrices [20]. Cette méthode s'appuie sur des rotations de Givens successives exécutées au sein de l'algorithme JADE (Joint Approximate Diagonalization of Eigen-matrices) [20]. Cet algorithme se décompose en 3 étapes. Tout d'abord, le tenseur des cumulants d'ordre 4, $\mathbf{Cum_z}$, des données blanchies $\mathbf{z}$ est calculé. Puis les matrices propres $\mathbf{M}^q$, $q \in \{1,\ldots,p\}$, de ce tenseur sont estimées. Finalement, l'ensemble formé par ces matrices propres est simultanément diagonalisé par une matrice $\mathbf{W}$.

L'algorithme JADE est très attractif car il est simple à implémenter et ne recquiert pas la gestion de paramètres d'optimisation. Il est nécessaire de lui préciser le nombre de sources à estimer. Pour des problèmes de SAS de faibles dimensions, son efficacité et sa rapidité de convergence sont des avantages indéniables. D'autres algorithmes ont été développés afin d'estimer des sources indépendantes à partir du modèle de l'ACI et quelques uns d'entre eux vont être brièvement introduit dans la section suivante.

4.3.5.3 Autres méthodes statistiques de SAS

SOBI : Des méthodes ont été développées pour exploiter la structure temporelle des sources. Lorsque les sources possèdent une cohérence temporelle, une identification du modèle de l'ACI est possible par des méthodes ne reposant que sur la manipulation d'outils statistiques d'ordre 2. Ces méthodes sont alors plus robustes que celles basées sur les statistiques d'ordre supérieur [12].

Ces méthodes s'appuient sur l'utilisation des matrices de covariances avec délai. La matrice de covariance d'un vecteur de sources centrées $\mathbf{s^c}(t)$ s'écrit $\mathbf{R_s} = E\{\mathbf{s^c}(t)\mathbf{s^c}(t)^T\}$. La matrice de covariance avec un délai τ est définie par :

$$\mathbf{R_s}(\tau) = E\{\mathbf{s^c}(t+\tau)\mathbf{s^c}(t)^T\}.$$

Des hypothèses suffisantes pour estimer complètement le modèle de l'ACI sont de supposer que les sources sont mutuellement décorrélées et que chacune possède une autocorrélation qui lui est propre et différente

des autres.

L'une des méthodes les plus populaires qui exploite ces propriétés est l'algorithme SOBI [12] qui signifie Second Order Blind Identification. Cette technique exploite la propriété suivante de la matrice $\mathbf{U}$ de l'équation (4.5). Les matrices de covariance avec délai $\mathbf{R_z}(\tau)$ des données blanchies $\mathbf{z}$ sont simultanément diagonalisées par $\mathbf{U}$. Les matrices de covariances $\mathbf{R_z}(\tau)$ pour plusieurs délais τ sont diagonalisées conjointement par une procédure de Jacobi. Cette dernière étape est rigoureusement identique à l'étape de diagonalisation de l'algorithme JADE. Le diagonaliseur ainsi estimé n'est rien d'autre que la matrice $\mathbf{U}$. Du fait de la diagonalisation de plusieurs matrices de covariance, cette méthode est beaucoup plus robuste que celles n'en exploitant qu'une seule.

Infomax : Un autre principe consiste à maximiser l'information transférée dans un réseau de neurones, d'où son nom *infomax* pour *information maximisation* en anglais [10]. Ce principe est équivalent à maximiser l'entropie des sorties non-linéaires du réseau. Cette méthode est équivalente aux méthodes basées sur le maximum de vraisemblance si les non-linéarité introduites dans le réseau pour calculer les sorties sont choisies égales aux fonctions de répartition des composantes indépendantes recherchées [66].

D'autres algorithmes existent et sont basés sur des notions d'indépendance ou des techniques d'optimisation différentes. Mais en pratique, les algorithmes décrits précédemment sont principalement utilisés de par leur efficacité, leur robustesse, leur simplicité de programmation et d'utilisation.

4.3.6 Applications

En pratique, les réalisations $\mathbf{x}(t)$ et $\mathbf{s}(t)$ des vecteurs aléatoires $\mathbf{x}$ et $\mathbf{s}$ sont supposées comme étant des réalisations de processus stationnaires et ergodiques connues sur une durée de temps suffisante. Un estimateur consistant de la matrice de covariance $\mathbf{R_x}$ de $\mathbf{x}(t)$ peut être calculé. Les signaux blanchis $\mathbf{z}(t)$ sont ainsi calculés. Des estimateurs consistants des cumulants d'ordre 4, $Cum_{\mathbf{z}}(i,j,k,l)$, avec $(i,j,k,l) \in \{1,\ldots,p\}^4$ des données blanchies $\mathbf{z}(t)$ peuvent être calculés. Ceci permet donc de calculer un estimé de la matrice de mélange $\mathbf{A}$ grâce à l'estimation de la matrice $\mathbf{V}$ définie à l'équation (4.4) et de la matrice $\mathbf{W} = \mathbf{U}^T$ en maximisant la fonction objectif définie par l'équation (4.11) ou en utilisant les équations (4.8)-(4.10).

De nombreux signaux réels sont modélisables par des réalisations de processus stationnaires et ergodiques. De nombreuses applications des techniques d'ACI sur des exemples réels ont ainsi pu être étudiées et les résultats cohérents ont permis de valider les méthodes d'ACI. Ces diverses techniques se sont révélées plus efficaces et robustes que prévu.

Les applications majeures sont faites dans le biomédical et les télécommunications, domaines dans lesquels les hypothèses de base de l'ACI sont facilement vérifiables.

Dans le domaine biomédical, de nombreux résultats concluants ont été publiés sur la séparation aveugle de données issues de l'électroencéphalographie (EEG) [86, 130, 69, 32, 68], de la magnétoencéphalographie (MEG) [131, 116, 68, 124], de l'électrocardiographie (ECG) [140, 30, 110, 22], de l'imagerie par résonance magnétique fonctionnelle (IRMf) [94, 15], la magnétoneurographie (MNG) [141, 42]. La décomposition de potentiels évoqués mesurés par l'EEG ou la MEG est également réalisée par l'ACI [133, 132, 124]. Grâce à l'ACI, l'identification des signaux d'intérêt est réalisable par des techniques d'enregistrement non-traumatiques, facilitant ainsi le diagnostic de maladies ou d'anomalies par les médecins.

Les techniques d'ACI ont été appliquées aux télécommunications [43, 111] et aux séries temporelles en économie [4, 73, 87].

L'extraction de caractéristiques d'images naturelles est un autre grand domaine où l'ACI est largement utilisée [11, 65], tout comme la déconvolution aveugle [96, 139].

4.4 Application de l'ACI à la spectroscopie Raman : le déparaffinage numérique

Après avoir présenté les techniques d'ACI, nous nous proposons d'étudier la possibilité d'appliquer ces techniques à l'analyse par spectroscopie Raman de tissus biologiques.

Les décès par cancer de la peau sont dûs pour les trois quarts aux mélanomes cutanés. Leur incidence augmente de 5% à 10% tous les ans, avec l'apparition de victimes de plus en plus jeunes à cause entre autres des surexpositions solaires. Les traitements sont peu efficaces dans la phase métastatique de la maladie. Seuls des diagnostics précoces permettent d'enrayer la maladie. Les dépistages sont réalisés sur des échantillons de peau qui sont ensuite stockés dans des tumorothèques pour des analyses futures. Les échantillons doivent donc se conserver plusieurs années sans altération de leur structure. Ce constat est valable pour tout type d'échantillons biologiques. Le paraffinage chimique des tissus est une pratique largement répandue en biomédical. L'analyse d'échantillons paraffinés nécessite l'élimination de la couche de paraffine pour accéder au tissu. Face aux limites des méthodes de déparaffinage chimique, nous nous proposons dans la suite de décrire une nouvelle méthode générale de déparaffinage mais numérique cette fois-ci. Le coeur de cette méthode repose sur le couplage de l'ACI à la spectroscopie Raman.

4.4.1 Paraffinage des échantillons biologiques et problèmes liés au déparaffinage classique

Afin d'être étudiés dans des conditions optimales et reproductibles, les échantillons biologiques doivent répondre à deux exigences majeures :

- Après analyse, les tissus étudiés par des spécialistes ont besoin d'être stockés dans une biblio-

thèque pour des analyses ultérieures. En cancérologie, les échantillons prélevés à un stade précoce de la maladie sur des patients sont stockés dans une tumorothèque afin d'être analysés de nombreuses années plus tard. Des corrélations sont ainsi étudiées entre les propriétés physiques, chimiques, spectroscopiques, ou autres, de l'échantillon et l'évolution de la maladie chez le patient, à savoir son décés, sa guérison ou une rechute. Les tissus doivent donc posséder une propriété de conservation.

- Les principales techniques d'analyse microscopiques et pathologiques, telles que les méthodes basées sur la spectroscopie Raman et les méthodes histochimiques et immunohistochimiques, s'appliquent sur de fines sections d'échantillon afin de transmettre la lumière. Or, la texture de l'échantillon doit être suffisamment consistante pour permettre une coupe aisée de l'échantillon en tranches fines.

Les échantillons ne répondent pas à ces exigences. Un conditionnement est capable de leur attribuer ces capacités. Il s'agit du paraffinage qui est un processus d'enveloppement des échantillons dans une couche de paraffine. La paraffine étant un excellent conservateur de tissus, la structure des échantillons n'est pas affectée et leur dégradation naturelle est stoppée. Les techniques de paraffinage d'échantillons biologiques sont parfaitement maîtrisées, simples à mettre en oeuvre et peu onéreuses. Ces avantages ont fait du processus de paraffinage une méthode très populaire de traitement des échantillons en biologie comme le prouvent ces nombreuses applications [45, 120, 41, 83, 39].

Les méthodes classiques d'analyse microscopiques et pathologiques citées précédemment ne s'appliquent pas sur les échantillons paraffinés puisque la paraffine empêche l'accession directe au spécimen à analyser. Un traitement chimique de l'échantillon permet d'éliminer la paraffine, cette étape s'appelle le déparaffinage ou le dégraissage. Les méthodes d'élimination les plus utilisées, telles que le xylène, Histoclear, HMAR[21] et Trilogy, consistent grossièrement au retrait de la paraffine et à la réhydratation de l'échantillon par trempage dans différents bains de solvants. Cette pratique est courante pour l'examen de tissues tumoraux issus par exemple du cancer du sein [52] et du cancer de la peau [45].

Une étude menée par Ó Faoláin a conclu à l'existence de trois principaux inconvénients lors de l'utilisation de ces méthodes :

- Le processus de déparaffinage est gourmand en réactifs chimiques et en temps. Plusieurs bains de solvants différents sont nécessaires pour nettoyer l'échantillon. Par exemple dans [41], l'agent de dégraissage est le xylène. Le protocole est particulièrement compliqué comme indiqué ci-après : les échantillons sont successivement plongés dans deux bains de xylène pendant 5 min, puis 4 min, deux bains d'éthanol pendant environ 3 min, puis 2 min, et ensuite dans un bain de Industrial Methylated Spirits à 95% pendant 1 min. Les échantillons sont finallement plongés dans un bain de xylène durant 18 heures.
  Ces méthodes peuvent aussi exiger des conditions expérimentales spécifiques, comme la technique HMAR qui impose de fortes températures et pressions aux échantillons.

[21] Heat-Mediated Antigen Retrieval

- Suite aux contacts répétés avec ces substances chimiques agressives et aux conditions expérimentales éprouvantes, la structure des échantillons peut être altérée. Ces méthodes vont à l'encontre même de l'objectif du paraffinage, c'est-à-dire conserver l'intégrité des échantillons au cours du temps et des différentes analyses.
- Les méthodes de dégraissage les plus populaires ne sont pas aussi efficaces que leur succès le sous-entend. Une couche résiduelle de paraffine perdure après l'étape de déparaffinage en certaines parties du tissus [41] et la présence de la paraffine peut fausser l'analyse du tissu.

Le développement d'une méthode efficace de dégraissage des échantillons paraffinés doit réaliser les opérations suivantes :

- Elle doit être rapide et simple à mettre en oeuvre.
- Elle ne doit pas altérer la structure biologique de l'échantillon à analyser.
- Le déparaffinage de l'échantillon doit être total.

4.4.2 Vers un déparaffinage numérique

Nous proposons maintenant de développer une procédure numérique de déparaffinage d'échantillons biologiques. Une méthode numérique peut possèder les deux premiers avantages recherchés, à savoir être rapide et simple, et ne pas dégrader l'échantillon. La dernière propriété énoncée et recherchée dépend quant à elle de l'efficacité de la méthode choisie. Le déparaffinage sera complet si la technique numérique est appropriée au modèle des données à analyser.

Le traitement du signal numérique agit sur des objets numériques. Une méthode d'analyse indirecte de l'échantillon paraffiné est donc nécessaire pour enregistrer des signaux numériques. Par *indirecte* nous entendons une méthode d'analyse qui s'applique sur l'échantillon paraffiné. Ainsi, l'établissement d'une procédure numérique de déparaffinage repose sur le choix d'une méthode d'analyse capable de sonder un échantillon paraffiné, et le choix d'un algorithme de traitement numérique des signaux enregistrés par la méthode d'analyse. Il est évident que le choix de la méthode d'analyse dictera partiellement le choix de la méthode de traitement numérique.

Les méthodes spectroscopiques allient l'analyse moléculaire non-destructive des échantillons, la forte sensibilité aux espèces chimiques présentes dans l'échantillon et la rapidité d'investigation. La spectroscopie Raman a été retenue pour sonder les échantillons aux vues de ses nombreux avantages sur les autres méthodes spectroscopiques décrits au paragraphe 1.3.2.3 à la page 23. Les pics Raman sont intenses et étroits en nombres d'onde. L'identification des bandes spectrales caractéristiques des différents constituants de l'échantillon sera facilitée.

La sensibilité de la spectroscopie Raman est duale lors de l'analyse d'un échantillon biologique paraffiné. Elle assure la détection de la présence de l'échantillon biologique situé sous la couche de paraffine. Mais la spectroscopie Raman est aussi très sensible à la présence de paraffine qui s'exprime sous forme de

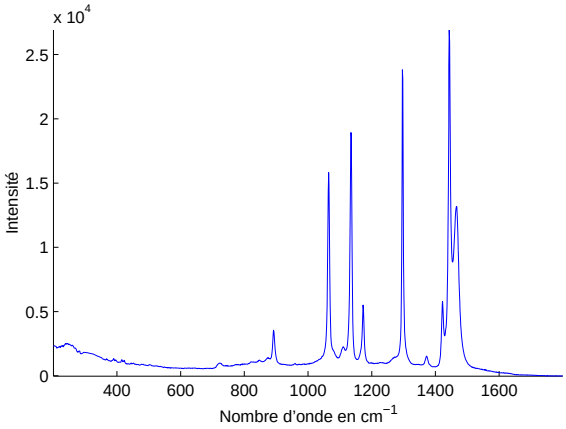


FIG. 4.1 – Spectre Raman de la paraffine seule

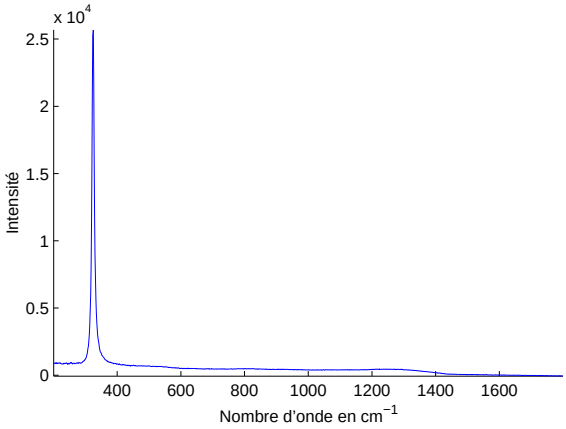


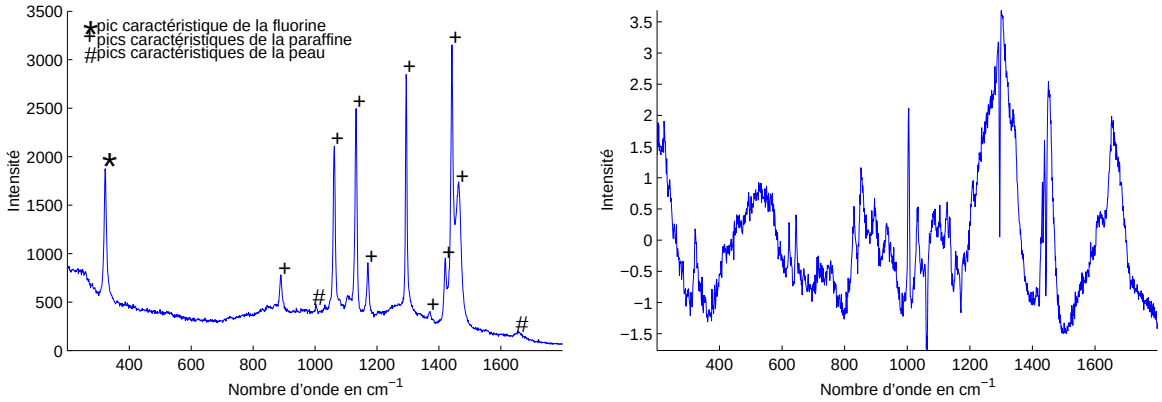
FIG. 4.2 – Spectre Raman de la fluorine seule

pics étroits et intenses, présentés sur la figure 4.1, qui masquent l'information vibrationnelle importante de l'échantillon sous-jacent. C'est cet effet de masque qui a jusqu'à maintenant empêché l'utilisation de la spectroscopie Raman sur des échantillons paraffinés. La seule étude menée avec succès sur des échantillons non chimiquement déparaffinés a été réalisée par Tfayli [126]. La spectroscopie infrarouge à transformée de Fourier a été utilisée avec succès pour la discrimination entre des nævi et des mélanomes sur des échantillons de peau paraffinés. Cette étude a été réalisée sans traitement numérique des spectres puisque la paraffine ne contribue pas aux bandes spectrales caractéristiques de la peau.

La paraffine n'est pas la seule espèce parasite des spectres Raman enregistrés sur un échantillon. L'acquisition des spectres n'est possible que si l'échantillon est fixé sur un support d'analyse comme expliqué au paragraphe 1.4.1.2 à la page 29. Ce support est choisi de façon à avoir l'activité Raman la plus faible et situé dans des bandes spectrales inactivées par l'échantillon analysé. Dans la suite de cette section, un support de fluorine, dont le spectre est présenté sur la figure 4.2, a été sélectionné.

Sans traitement adapté des spectres Raman enregistrés en différents points de l'échantillon biologique, il est impossible d'extraire les informations pertinentes reliées à l'échantillon situé sous la paraffine. Un spectre acquis sur un tissu paraffiné de peau humaine positionné sur un support en fluorine est visible sur la figure 4.3(a). Le pic labellisé par $*$ et centré à 325 cm^{-1} est caractéristique de la fluorine. Les pics de la paraffine sont repérés par les signes $+$ et sont centrés à 890 cm^{-1}, 1063 cm^{-1}, 1133 cm^{-1}, 1172 cm^{-1}, 1296 cm^{-1}, 1372 cm^{-1}, 1418 cm^{-1}, 1441 cm^{-1} et 1463 cm^{-1}. Sur ce spectre, il apparaît clairement que les pics caractéristiques de la peau, symbolisés par le signe $\#$ et situés à 1002 cm^{-1} et 1660 cm^{-1}, sont négligeables devant les pics de la paraffine et de la fluorine. Les autres informations sont totalement noyées dans le spectre et sont représentées par une estimation par ACI du spectre de la peau sur la figure 4.3(b). Une analyse visuelle s'avère donc limitée devant la faible intensité Raman de la peau.

Au lieu d'utiliser une méthode traditionnelle et inefficace de dégraissage des échantillons paraffinés, l'étude des propriétés physiques et statistiques des spectres Raman va nous guider vers la désignation



(a) (b)

FIG. 4.3 – (a) Exemple de spectre Raman enregistré sur un échantillon paraffiné de peau humaine sur support de fluorine, (b) Spectre Raman de la peau estimé par ACI

d'une méthode de traitement numérique du signal afin d'éliminer la paraffine et la fluorine des spectres Raman enregistrés.

4.4.3 Modélisation des spectres Raman

De par les lois physiques régissant les mécanismes de la spectroscopie Raman et étudiées au paragraphe 2.2.3 à la page 40, les spectres Raman sont modélisables par des mélanges linéaires et instantanés des spectres des composantes moléculaires de l'échantillon. Cette modélisation n'est cependant pas réaliste puisque des phénomènes parasites, énoncés au paragraphe 2.4.4 à la page 60 et dus à l'imperfection de l'instrumentation d'acquisition des spectres et à la nature biologique des échantillons, viennent s'ajouter aux spectres Raman idéaux des espèces moléculaires de l'échantillon. La matrice des données enregistrées $\mathbf{X}$ peut donc s'écrire :

$$\mathbf{X} = \mathbf{Z} + \mathbf{B} = \mathbf{Z^{utile}} + \mathbf{Z^{autre}} + \mathbf{B^1} + \mathbf{B^2}. \tag{4.12}$$

$\mathbf{B} = \mathbf{B^1} + \mathbf{B^2}$ est le sous-espace bruit. La partie $\mathbf{B^1}$ de $\mathbf{B}$ exprime la présence du courant noir, de la réponse spectrale du système, des rayons cosmiques, de la déviation en nombre d'onde, et est éliminée par les prétraitements présentés aux sections 2.3.4, page 48, et 2.4.4, page 60. La partie $\mathbf{B^2}$ de $\mathbf{B}$ modélise le fond de fluorescence exposé à la section 2.4.4.2, page 60. Ce sous-espace sera éliminé de l'analyse au paragraphe 4.4.5.1 par une procédure d'estimation et de soustraction de la ligne de base.

Le sous-espace $\mathbf{Z} = \mathbf{Z^{utile}} + \mathbf{Z^{autre}}$ se compose des informations vibrationnelles de l'échantillon analysé. Les influences Raman de la paraffine, de la fluorine, du tissu sous-jacent, mais aussi d'autres espèces moléculaires, qui peuvent être présentes dans l'échantillon, et d'un bruit de mesure à variations élevées se combinent pour former ce sous-espace. Seules nous intéressent les contributions de la paraffine, de la

fluorine et du tissu, et qui forment le sous-espace $\mathbf{Z}^{\mathbf{utile}}$. Nous voulons éliminer les autres contributions et le bruit à variations élevées qui composent le sous-espace $\mathbf{Z}^{\mathbf{autre}}$, comme nous le verrons à la section 4.4.5.4.

Le sous-espace $\mathbf{Z}^{\mathbf{utile}}$ est décomposable en trois sous-espaces :

$$\mathbf{Z}^{\mathbf{utile}} = \mathbf{Z}^{\mathbf{para}} + \mathbf{Z}^{\mathbf{fluo}} + \mathbf{Z}^{\mathbf{tissu}}. \tag{4.13}$$

avec $\mathbf{Z}^{\mathbf{utile}} = [\mathbf{z}_1^{\mathbf{utile}}, \ldots, \mathbf{z}_i^{\mathbf{utile}}, \ldots, \mathbf{z}_{N_{xy}}^{\mathbf{utile}}]^T \in \mathbb{R}^{N_{xy} \times N_\Lambda}$, où $\mathbf{z}_i^{\mathbf{utile}} = [z_{i1}^{utile}, \ldots, z_{i\Lambda}^{utile}, \ldots, z_{iN_\Lambda}^{utile}]^T \in \mathbb{R}^{N_\Lambda}$ est la partie signal d'un spectre expérimental. Les contributions spectrales respectives de la paraffine, de la fluorine et du tissu sous-jacent sont modélisées par les sous-espaces respectifs $\mathbf{Z}^{\mathbf{para}} = [\mathbf{z}_1^{\mathbf{para}}, \ldots, \mathbf{z}_i^{\mathbf{para}}, \ldots, \mathbf{z}_{N_{xy}}^{\mathbf{para}}]^T \in \mathbb{R}^{N_{xy} \times N_\Lambda}$, $\mathbf{Z}^{\mathbf{fluo}} = [\mathbf{z}_1^{\mathbf{fluo}}, \ldots, \mathbf{z}_i^{\mathbf{fluo}}, \ldots, \mathbf{z}_{N_{xy}}^{\mathbf{fluo}}]^T \in \mathbb{R}^{N_{xy} \times N_\Lambda}$ et $\mathbf{Z}^{\mathbf{tissu}} = [\mathbf{z}_1^{\mathbf{tissu}}, \ldots, \mathbf{z}_i^{\mathbf{tissu}}, \ldots, \mathbf{z}_{N_{xy}}^{\mathbf{tissu}}]^T \in \mathbb{R}^{N_{xy} \times N_\Lambda}$.

Les spectres $\mathbf{s}^{\mathbf{para}}$ de la paraffine, $\mathbf{s}^{\mathbf{fluo}}$ du support de fluorine et $\mathbf{s}^{\mathbf{tissu}}$ du tissu, de dimensions N_Λ, sont uniques et identiques d'un point d'acquisition à l'autre. Seule change l'intensité relative de chacun de ces spectres en relation directe avec la concentration de chacune de ses espèces dans l'échantillon à analyser. Les sous-espaces précédents s'écrivent ainsi sous la forme :

$$\mathbf{Z}^{\mathbf{para}} = \mathbf{a}^{\mathbf{para}}(\mathbf{s}^{\mathbf{para}})^T \tag{4.14}$$

$$\mathbf{Z}^{\mathbf{fluo}} = \mathbf{a}^{\mathbf{fluo}}\left(\mathbf{s}^{\mathbf{fluo}}\right)^T \tag{4.15}$$

$$\mathbf{Z}^{\mathbf{tissu}} = \mathbf{a}^{\mathbf{tissu}}\left(\mathbf{s}^{\mathbf{tissu}}\right)^T. \tag{4.16}$$

où les vecteurs colonnes $\mathbf{a}^{\mathbf{para}}$, $\mathbf{a}^{\mathbf{fluo}}$ et $\mathbf{a}^{\mathbf{tissu}}$, de dimensions N_{xy}, modélisent les profils de concentrations relatives de chaque espèce moléculaire présente dans l'échantillon. Le sous-espace signal $\mathbf{Z}^{\mathbf{utile}}$ peut ainsi s'écrire sous la forme :

$$\mathbf{Z}^{\mathbf{utile}} = \mathbf{a}^{\mathbf{para}}(\mathbf{s}^{\mathbf{para}})^T + \mathbf{a}^{\mathbf{fluo}}\left(\mathbf{s}^{\mathbf{fluo}}\right)^T + \mathbf{a}^{\mathbf{tissu}}\left(\mathbf{s}^{\mathbf{tissu}}\right)^T.$$

Cette décomposition de $\mathbf{Z}^{\mathbf{utile}}$ est estimable par un algorithme d'ACI grâce aux propriétés statistiques des spectres Raman des espèces chimiques pures d'un échantillon que nous allons exposer dans la section suivante.

4.4.4 Propriétés statistiques et ACI

Les équations (4.13)-(4.16) montrent qu'en un point de mesure i de l'échantillon, le spectre Raman enregistré $\mathbf{x}_i$ reflète la présence de la paraffine, de la fluorine et du tissu présent sous la couche de paraffine. Le spectre pur de la paraffine est présenté sur la figure 4.1. Ce spectre est composé de quelques pics Raman étroits et intenses centrés sur les nombres d'onde 890 cm^{-1}, 1063 cm^{-1}, 1133 cm^{-1}, 1172 cm^{-1}, 1296 cm^{-1}, 1372 cm^{-1}, 1418 cm^{-1}, 1441 cm^{-1} et 1463 cm^{-1}. Le reste du spectre est quasiment égal à zéro. Le spectre de la paraffine peut donc être qualifié de parcimonieux. Le spectre du support

de fluorine, présenté sur la figure 4.2, se compose uniquement d'un pic étroit et très intense centré en 325 cm^{-1}. Le reste du spectre est fixé à l'intensité nulle. Il peut donc être lui aussi modélisé comme parcimonieux. Les histogrammes de la distribution des intensités des spectres purs de la paraffine et de la fluorine sont représentés sur les figures 4.4(a) et 4.4(b). Ces histogrammes présentent une distribution d'intensités maximales voisines de zéros et justifient la propriété de parcimonie accordée à ces spectres. La parcimonie a été implicitement utilisée dans certains travaux sur l'ACI [10]. Les sources indépendantes recherchées étaient supposées à densité de probabilité surgaussienne, ce qui est caractéristique des sources parcimonieuses [8]. Les algorithmes d'ACI FastICA et JADE présentés respectivement aux sections 4.3.5.1 et 4.3.5.2 ont été développés dans le cadre général où les sources recherchées sont indépendantes, sans hypothèse sur leur densité de probabilité[22]. Ils sont donc tout à fait compétents pour estimer des sources parcimonieuses indépendantes. Une remarque importante est que les pics caractéristiques des spectres de ces espèces pures ne se recouvrent sur aucune bande spectrale. L'ensemble de ces arguments, la décorrélation et la parcimonie de ces spectres, permettent de supposer qu'ils sont statistiquement mutuellement indépendants entre eux.

Le tissu sous-jacent est généralement constitué de nombreuses espèces moléculaires différentes. Dans le cas d'un échantillon de peau paraffinée qui sera présenté en détail à la section , il est attendu d'estimer les spectres de la mélanine et de la kératine qui constituent en majorité l'épiderme de la peau. Cependant la mélanine possède une signature Raman beaucoup plus faible que la kératine. Les spectres de ces deux espèces étant de plus noyés dans les spectres de la paraffine et de la fluorine, l'intensité Raman de la mélanine devient négligeable et indétectable par des méthodes d'ACI. C'est pourquoi la présence de la mélanine n'est pas prise en compte dans la modélisation des spectres des échantillons de peau paraffinée.

Penser que le spectre Raman de la peau est identique au spectre Raman de la kératine est utopiste. La peau se compose d'une multitude d'espèces chimiques différentes. Dans le cas général, ces différentes espèces possèdent des signatures spectrales différentes qui activent presque toute la plage spectrale. L'indépendance statistique entre ces spectres ne peut donc pas être supposée. Pour surmonter ce problème, nous supposons que le tissu sous-jacent a une composition chimique suffisamment homogène pour qu'en chaque point de mesure la concentration de chaque espèce par rapport aux autres reste la même. Seule la concentration globale du tissu est autorisée à varier d'un point de mesure à un autre. Ainsi, en chaque point de mesure, l'ensemble du spectre du tissu est proportionnel à sa concentration, contrairement au cas d'un tissu de composition hétérogène où les bandes spectrales différentes varient d'un point de mesure à un autre en fonction de la composition locale du tissu. Il ne reste dans ce cas plus qu'à justifier l'indépendance entre le spectre du tissu et les spectres de la paraffine et de la fluorine.

Le spectre de la peau n'est pas parcimonieux et n'est pas composé de quelques pics étroits et intenses comme montré à la figure 4.3(b). L'histogramme de la distribution d'intensité de ce spectre estimé par l'ACI est représenté sur la figure 4.4(c) et illustre ce propos. La distribution du spectre de la peau

[22] D'autres algorithmes ne sont pas aussi généraux puisqu'ils imposent par exemple que les sources soient à densité de probabilité surgaussienne [10]

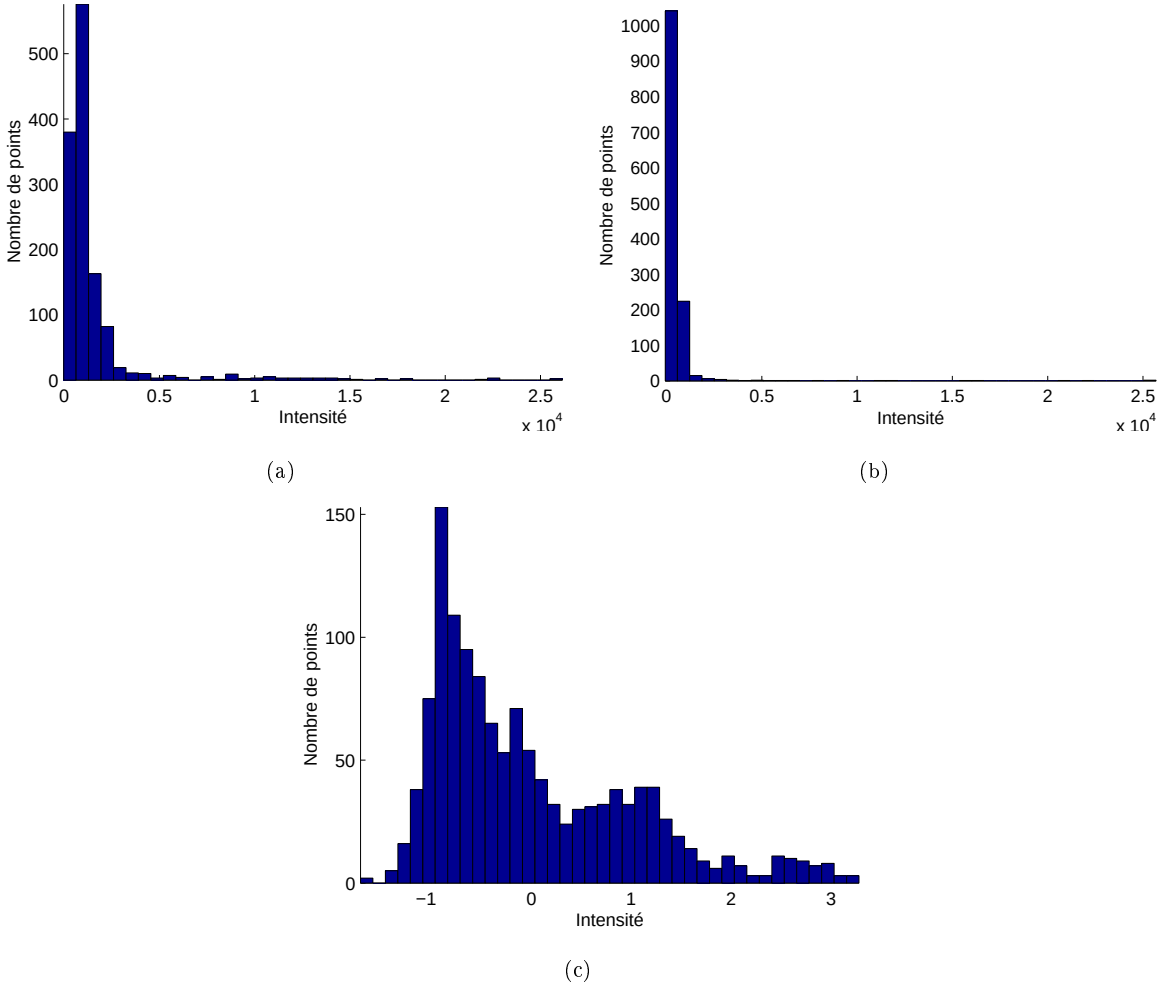


FIG. 4.4 – Histogrammes de la distribution des intensités (a) du spectre Raman de la paraffine pure, (b) du spectre Raman de la fluorine pure, (c) du spectre Raman estimé par ACI de la peau

s'étale plus largement que les spectres de la fluorine et de la paraffine. Les spectres Raman de tissus biologiques complexes ont une forme tout aussi complexe à analyser car de nombreux pics et bosses se superposent. Les intensités varient rapidement d'un nombre d'onde au suivant, conférant ainsi une forme non structurée aux spectres Raman de ces tissus. De plus, aucune réaction chimique entre la paraffine et le tissu d'une part, et entre l'ensemble (paraffine + tissu) et la fluorine d'autre part ne se produit. Les propriétés chimiques des constituants de l'échantillon sont donc indépendantes et suggèrent donc des propriétés vibrationnelles indépendantes donc des spectres indépendants. Nous posons ici l'hypothèse que la parcimonie des spectres Raman de la fluorine et de la paraffine, le caractère non structuré des intensités du spectre Raman du tissu biologique sous-jacent et l'indépendance des propriétés chimiques des composants de l'échantillon suffisent pour assurer l'indépendance statistique mutuelle de ces signaux.

L'hypothèse fondamentale d'indépendance statistique des sources sur laquelle repose l'ACI est suppo-

sée vérifiée pour les spectres Raman acquis sur des échantillons de tissus paraffinés. Les distributions des intensités des spectres sources ne sont évidemment pas gaussiennes. Dans les applications, le nombre de spectres enregistrés est toujours important et beaucoup plus grand que le nombre de sources à estimer. De plus, les concentrations de la paraffine, de la fluorine et du tissu sont variables d'un point de mesure à un autre. La matrice de mélange est donc supposée non-singulière.

Toutes les hypothèses requises pour la validation du modèle de l'ACI sont remplies. Les algorithmes d'ACI peuvent donc être appliqués sur le sous-espace $\mathbf{Z^{utile}}$ de jeux de spectres Raman $\mathbf{X}$ acquis sur des échantillons de tissus paraffinés et fixés sur des supports en fluorine. Il est évident que la parcimonie et la décorrélation entre les spectres de la paraffine et de la fluorine sont essentielles à l'applicabilité des méthodes d'ACI. Dans le cadre d'autres applications, si aucune espèce chimique ne possède un spectre parcimonieux et décorrélé des autres, alors les méthodes d'ACI ne seraient pas adaptées pour résoudre ce problème.

Avant de pouvoir appliquer l'ACI, il est nécessaire de soustraire les sous-espaces $\mathbf{B^1}$, $\mathbf{B^2}$ et $\mathbf{Z^{autre}}$ à la matrice des données $\mathbf{X}$ pour accéder au sous-espace $\mathbf{Z^{utile}}$. La section suivante va présenter des méthodes pour s'affranchir de ces sous-espaces.

4.4.5 Prétraitements

Comme indiqué au début de la section 4.4.3, le sous-espace $\mathbf{B^1}$ est éliminé par les prétraitements présentés aux sections 2.3.4, page 48, et 2.4.4, page 60. Le sous-espace $\mathbf{B^2}$ modélise le fond de fluorescence des spectres enregistrés. Il va être éliminé par une procédure d'estimation et de soustraction. Une non-linéarité n'a pas été insérée dans le modèle de l'équation (4.12) par soucis de simplification de l'étude. Il s'agit du non alignement des pics caractéristiques de la paraffine et de la fluorine d'un spectre enregistré à un autre. Cette non-linéarité va être compensée par un processus de recalage des pics Raman. Le sous-espace $\mathbf{Z^{autre}}$ peut ainsi être écartée de l'étude par une ACP appliquée sur les signaux recalés et corrigés de $\mathbf{B^1}$ et $\mathbf{B^2}$.

4.4.5.1 Élimination de la ligne de base

Le sous-espace bruit $\mathbf{B^2}$ résulte de la superposition d'une ligne de base $\mathbf{s_i^{ff}}$ au spectre $\mathbf{x}_i$ enregistrés au point de mesure i de l'échantillon. Cette ligne de base, également appelée fond de fluorescence, trouve son origine principalement dans l'émission de fluorescence par l'échantillon mais aussi dans le bruit d'instrumentation lié au spectromètre. Cet effet est classique lors de l'étude d'échantillons biologiques. La ligne de base se manifeste dans les spectres enregistrés par un signal à variation lente en fonction du nombre d'onde. Un exemple de cette ligne de base parasite est présenté sur la figure 4.5 où une estimation est indiquée en pointillés. Ce signal est décrit de façon plus approfondie au paragraphe 2.4.4.2 à la page 60.

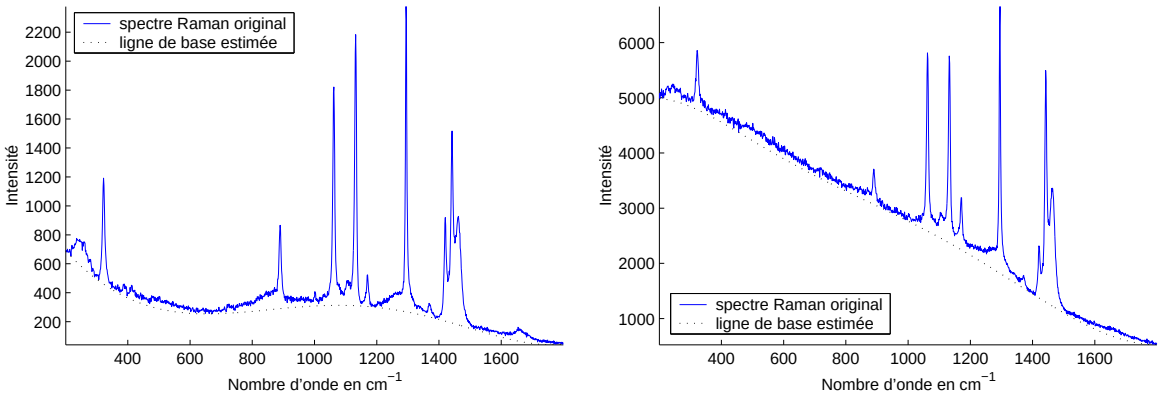


FIG. 4.5 – Exemples de spectres Raman enregistrés en deux points différents d'un échantillon de peau paraffinée fixé sur un support de fluorine et corrompus par des lignes de base d'intensités différentes

Caractéristiques de la ligne de base : La particularité de ce signal est qu'il n'a pas un comportement linéaire d'un spectre à un autre, c'est-à-dire que sa forme diffère d'un spectre à un autre comme montré sur la figure 4.5 qui représente des spectres Raman enregistrés en différents points d'un échantillon de peau paraffinée sur support de fluorine. Son intensité dépend principalement de la composition de l'échantillon en un point de mesure et possède une forte variance pour différents points d'acquisition. Sa modélisation ne peut donc pas se faire sous forme d'un sous-espace décomposable en un profil de concentration et un spectre constant d'un point d'acquisition à un autre, comme c'est le cas dans les équations (4.14)-(4.16).

Sa forme caractéristique d'une variation lente en fonction du nombre d'onde suggère sa modélisation par un polynôme. Pour chaque spectre enregistré, un polynôme doit être estimé indépendamment des polynômes estimés pour les autres spectres. La majorité des méthodes d'estimation des coefficients du polynôme s'appuie sur la minimisation de l'erreur de modélisation du spectre par le polynôme. Or en spectroscopie Raman, la forte intensité de quelques pics peut faire échouer l'estimation. Une méthode capable de s'affranchir de l'influence des pics Raman conduit à une estimation plus judicieuse de la ligne de base. C'est ce que propose Mazet et ses collaborateurs dans [92].

Modélisation : Chaque spectre enregistré peut s'écrire à partir de l'équation (4.12) sous la forme :

$$\mathbf{x}_i = \mathbf{z}_i^{\mathbf{utile}} + \mathbf{z}_i^{\mathbf{autre}} + \mathbf{b}_i^{\mathbf{1}} + \mathbf{b}_i^{\mathbf{2}}.$$

La modélisation du fond de fluorescence par un polynôme s'écrit mathématiquement sous la forme :

$$\mathbf{b}_i^{\mathbf{2}} = \mathbf{a}_i^T \mathbf{M}$$

où $\mathbf{a}_i$ et $\mathbf{M}$ sont de dimensions respectives $(L+1) \times 1$ et $(L+1) \times N_\Lambda$. Le nombre L représente l'ordre du polynôme. Le vecteur $\mathbf{a}_i$ définit le vecteur des coefficients du polynôme. La matrice $\mathbf{M}$ est la matrice

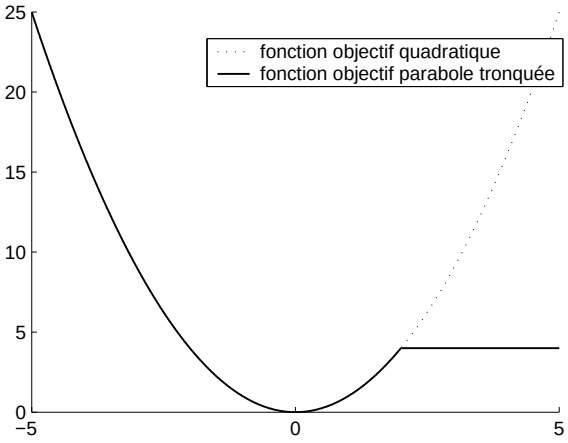


FIG. 4.6 – Comparaison entre les fonctions objectifs quadratique en trait pointillé et de forme parabole tronquée en trait plein

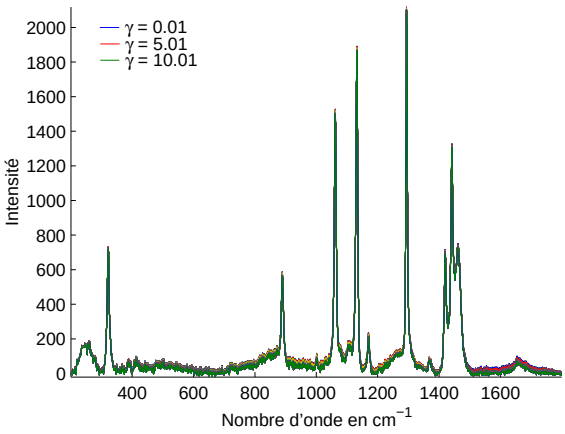


FIG. 4.7 – Spectre corrigé de sa ligne de base pour différentes valeurs du seuil γ

des nombres d'onde et est définie par :

$$\mathbf{M} = \begin{pmatrix} \bar{\nu}_1^0 & \cdots & \bar{\nu}_{N_\Lambda}^0 \\ \vdots & & \vdots \\ \bar{\nu}_1^L & \cdots & \bar{\nu}_{N_\Lambda}^L \end{pmatrix}.$$

Le but général est d'estimer le vecteur $\mathbf{a}_i$ qui va minimiser la fonction objectif suivante :

$$Q(\mathbf{a}_i) = \sum_{\Lambda=1}^{N_\Lambda} \varphi\left(b_{i\Lambda}^2 - \sum_{l=1}^{L+1} (a_{il} M_{l\Lambda}) \right)$$

où φ est une fonction qui va dicter la forme de la fonction objectif Q à minimiser.

Choix d'une fonction objectif tronquée : Le choix $\varphi(x) = x^2$ correspond à l'approche classique des moindres carrés. Elle est inefficace pour le traitement des spectres Raman car étant quadratique, les fortes valeurs du spectre $\mathbf{b_i^2}$ possèdent un fort coût. Or ces valeurs ne sont pas caractéristiques de la ligne de base et vont donc affecter fortement l'estimation du vecteur des coefficients $\mathbf{a}_i$. La fonction φ a été modeléee par Mazet dans [92] afin de ne pas accorder d'importance aux pics intenses du spectre $\mathbf{b_i^2}$ dans la procédure d'estimation du vecteur $\mathbf{a}_i$. Il a choisi une fonction φ qui est quadratique autour de zéro et constante au dessus d'un seuil γ à fixer par l'utilisateur. Cette fonction parabole tronquée est définie par :

$$\varphi(x) = \begin{cases} x^2 & \text{si } x < \gamma \\ \gamma^2 & \text{sinon.} \end{cases}$$

La figure 4.6 propose la comparaison entre une fonction objectif quadratique classique en trait pointillé et une fonction objectif de forme parabole tronquée en trait plein.

Grâce à cette fonction, toutes les valeurs des spectres ayant une intensité supérieure au seuil γ participent de la même manière à la fonction objectif. L'influence des pics est donc clairement limitée dans la procédure d'estimation du vecteur des coefficient du polynôme $\mathbf{a}_i$. La minimisation de Q n'étant pas aussi simple que celle des moindres carrés, la fonction objectif est minimisée par l'algorithme de minimisation semi-quadratique LEGEND [67].

En appliquant cette procédure sur tous les spectres Raman $\mathbf{x}_i$, $i \in \{1, \ldots, N_{xy}\}$, formant la matrice $\mathbf{X}$, le sous-espace $\mathbf{B^2}$ est estimé.

Choix de la valeur du seuil : Cet algorithme a été appliqué sur tous les spectres Raman utilisés dans la suite de cette application. Le seuil γ est à choisir en fonction du nombre et de l'intensité des pics Raman, et en fonction de l'écart-type σ_{bruit} du bruit. Dans [92], il est conseillé de choisir le seuil γ dans l'intervalle $[\sigma_{bruit}, 4\sigma_{bruit}]$. Ce paramètre a été ajusté visuellement après plusieurs essais mais n'influence que très faiblement l'estimation du fond de fluorescence dans nos applications puisque les spectres utilisés possèdent suffisament de points appartenant à la ligne de base. Pour illustrer ce propos, le spectre Raman de la figure 4.5(a) a été prétraité par cette procédure d'élimination de la ligne de base pour 100 valeurs différentes du seuil γ choisies dans l'intervalle $[0.01, 10.01]$. Les résultats sont visibles sur la figure 4.7. Les spectres ainsi obtenus sont quasiment égaux. Il est à noter qu'après la normalisation des spectres à une variance unité, comme présenté à la section 2.3.4.4, page 50, par l'équation (2.7), les variances des spectres corrigés calculées en chaque nombre d'onde n'excèdent pas 8×10^{-4}. Le choix de ce paramètre ne se révèle donc pas primordiale dans nos travaux..

Cependant, l'ordre du polynôme est un paramètre qui conditionne beaucoup plus l'efficacité de l'algorithme d'estimation de la ligne de base. Plusieurs essais sur de nombreux spectres Raman différents ont conclus sur un choix de l'ordre du polynôme égal à 5. Ce choix n'est réaliste que pour l'application qui va être traitée dans la suite. Pour un autre problème, ce nombre devrait être changé.

4.4.5.2 Alignement des pics

Un autre problème dans le traitement numérique des spectres Raman est le désalignement des pics Raman de l'échantillon analysé d'un spectre enregistré à un autre. Un exemple est donné sur la figure 4.8(a) où le pic Raman centré en 1063 cm^{-1} est dessiné pour des spectres acquis en différents points d'un échantillon de peau paraffinée fixée sur support de fluorine. Les sommets des pics sont clairement localisés en des nombres d'onde distincts d'un spectre à un autre. Cet effet est du à la résolution spectrale de l'appareil, à la sensibilité des détecteurs et aux aléas expérimentaux. Malgré la faible variation de la localisation des pics relativement à la longueur des spectres enregistrés, cet effet est très perturbateur pour les techniques d'ACI. Cette translation possible des maximums des pics se traduit par la présence de pics artéfactuels voisins sur les sources estimées par les techniques d'ACI. Il est nécessaire de réduire la variance de la localisation des pics afin d'appliquer les méthodes d'ACI dans des conditions optimales.

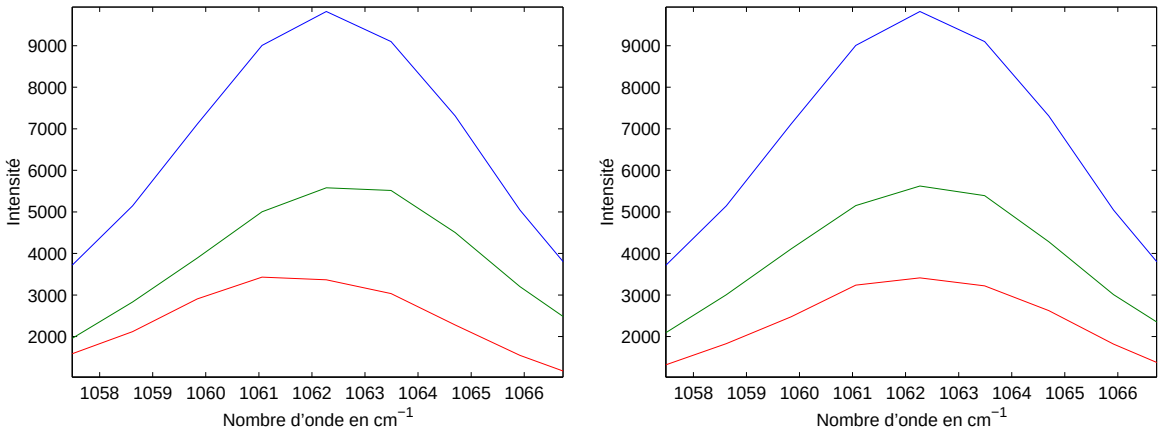


FIG. 4.8 – Exemple de (a) désalignement, et (b) réalignement, du maximum du pic centré en 1063 cm^{-1} sur des spectres Raman enregistrés sur des échantillons de peau paraffinée sur support de fluorine

Ce problème a été résolu par une procédure d'alignement des pics issue de la géophysique où elle est couramment utilisée [134]. Un spectre est choisi comme référence, c'est à dire que les positions de ses pics sont supposées correctes. Chaque spectre va être comparé à ce spectre référence. Chaque pic est considéré séparément. La transformée de Fourier discrète du pic est calculée pour la référence et pour le spectre à recaler. La fonction d'intercorrélation entre les pics est calculée dans le domaine fréquentiel. Des zéros sont insérés à la suite des coefficients des hautes fréquences. En calculant sa transformée de Fourier inverse, cette procédure a pour effet d'augmenter la fréquence d'échantillonnage du signal sans modifier son contenu spectral. Elle équivaut à une étape de suréchantillonnage de la fonction d'intercorrélation entre les pics de la référence et du spectre à recaler. L'abscisse pointant le maximum de la fonction d'intercorrélation suréchantillonnée est égal au décalage réel entre les deux pics au facteur de suréchantillonage près. Ce décalage est compensé par multiplication de la transformée de Fourier du pic à recaler avec la transformée de Fourier de la translation à appliquer. La transformée de Fourier inverse de ce résultat fournit un pic Raman centré au même nombre d'onde que le pic de la référence. Cette procédure est schématisée sur la figure 4.9.

Une étape supplémentaire consiste à traiter les bords des pics ainsi recalés afin d'harmoniser leurs intensités avec celle du spectre original et éviter des variations brusques entre le spectre original et le pic recalé. Pour achever ce but, le pic recalé est pondéré par une fenêtre de Hanning. Les bords du pic recalé sont ainsi négligés, tandis que le pic lui-même est mis en valeur. Une fenêtre de Hanning inversée, représentée sur la figure 4.10, pondère le pic original. Cette manipulation a l'effet inverse de la précédente, à savoir que les bords du pic sont conservés, tandis que le pic lui-même est annulé. La somme de ces deux signaux résultant donne le pic recalé et corrigé des effets de bords. Cette procédure est schématisée sur la figure 4.10.

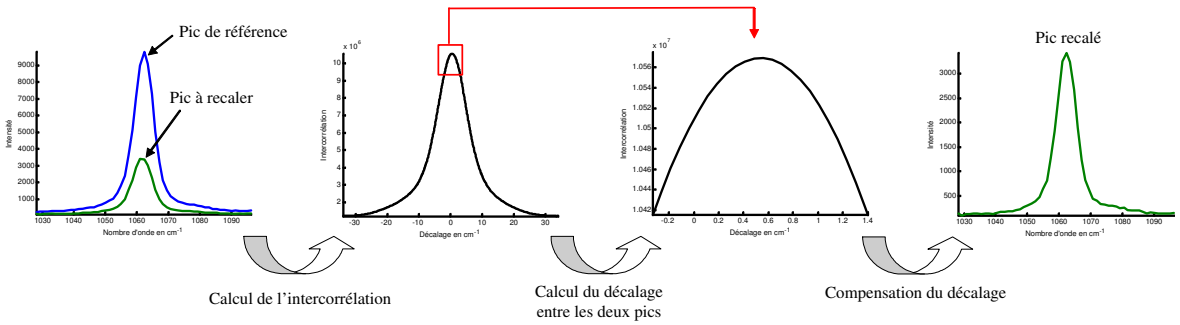


FIG. 4.9 – Schéma de principe du processus de recalage d'un pic

La figure 4.8(b) présente les résultats de la procédure complète d'alignement des pics Raman proposés sur la figure 4.8(a). Cette méthode est particulièrement efficace pour la tâche qui lui est dédiée. Elle a été appliquée sur les 8 pics caractéristiques de la fluorine et de la paraffine, c'est à dire aux bandes spectrales centrées allant de 325 cm^{-1} à 1441 cm^{-1}.

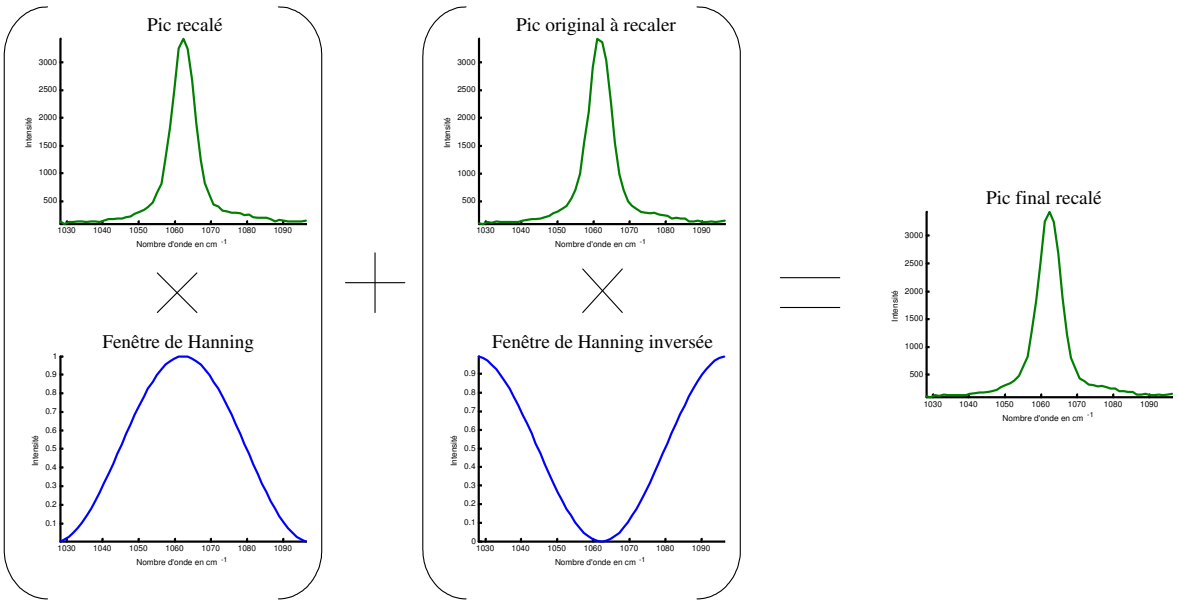


FIG. 4.10 – Schéma de principe de la procédure de correction des effets de bord

Ce prétraitement améliore les résultats obtenus par l'application des techniques d'ACI sur des spectres Raman. L'importance de cette procédure est illustrée dans l'annexe B où la comparaison est faite entre les sources estimées par l'application de l'ACI avec et sans ce prétraitement sur des spectres paraffinés de peau fixée sur un support de fluorine.

4.4.5.3 Centrage et réduction

À la suite des opérations précédentes, la matrice de données $\mathbf{X}$ est réduite au sous-espace signal $\mathbf{Z}$ de l'équation (4.12), c'est-à-dire $\mathbf{Z} = \mathbf{X} - \mathbf{B^1} - \mathbf{B^2}$.

De nombreuses méthodes statistiques sont développées dans le cadre de données centrées et réduites à une variance unité. Les méthodes d'ACI utilisées requièrent ce type de données. Chaque spectre constituant le sous-espace $\mathbf{Z}$ des données corrigées de la ligne de base et à pics recalés est donc centré et réduit. Le sous-espace résultant est noté $\bar{\mathbf{Z}}$. Cette étape a pour effet d'attribuer le même poids à chaque composante enregistrée.

Les équations (4.14), (4.15) et (4.16) sont transformées en :

$$\bar{\mathbf{Z}}^{\mathbf{para}} = \bar{\mathbf{a}}^{\mathbf{para}} \left(\bar{\mathbf{s}}^{\mathbf{para}}\right)^T \tag{4.17}$$

$$\bar{\mathbf{Z}}^{\mathbf{fluo}} = \bar{\mathbf{a}}^{\mathbf{fluo}} \left(\bar{\mathbf{s}}^{\mathbf{fluo}}\right)^T \tag{4.18}$$

$$\bar{\mathbf{Z}}^{\mathbf{tissu}} = \bar{\mathbf{a}}^{\mathbf{tissu}} \left(\bar{\mathbf{s}}^{\mathbf{tissu}}\right)^T \tag{4.19}$$

où $\bar{\mathbf{s}}^{\mathbf{para}}$, $\bar{\mathbf{s}}^{\mathbf{fluo}}$ et $\bar{\mathbf{s}}^{\mathbf{tissu}}$ sont les spectres centrés et réduits respectivement de la paraffine, de la fluorine et du tissu, et $\bar{\mathbf{a}}^{\mathbf{para}}$, $\bar{\mathbf{a}}^{\mathbf{fluo}}$ et $\bar{\mathbf{a}}^{\mathbf{tissu}}$ sont les concentrations relatives réduites associées.

4.4.5.4 Élimination de $\mathbf{Z^{autre}}$ par ACP

Les données à analyser se réduisent au sous-espace $\bar{\mathbf{Z}}$ à la suite des prétraitements précédents. Or $\bar{\mathbf{Z}} = \bar{\mathbf{Z}}^{\mathbf{utile}} + \bar{\mathbf{Z}}^{\mathbf{autre}}$ et les techniques d'ACI ne peuvent s'appliquer que sur $\bar{\mathbf{Z}}^{\mathbf{utile}}$ dont les sources sous-jacentes vérifient les hypothèses de l'ACI. L'élimination de $\bar{\mathbf{Z}}^{\mathbf{autre}}$ est simple car nous supposons que ce sous-espace est orthogonal à $\bar{\mathbf{Z}}^{\mathbf{utile}}$. Une ACP (voir section 2.3.5.2, page 52) appliquée sur les spectres de $\bar{\mathbf{Z}}$ permet de concentrer $\bar{\mathbf{Z}}^{\mathbf{utile}}$ sur les p premières composantes principales et de confiner $\bar{\mathbf{Z}}^{\mathbf{autre}}$ dans les $N_{xy} - p$ autres composantes.

4.4.6 Déparaffinage numérique

Les spectres sont maintenant nettoyés du fond de fluorescence et les pics sont recalés sur les mêmes nombres d'onde pour tous les spectres. La procédure de blanchiment de l'ACI a permis de s'affranchir du sous-espace $\bar{\mathbf{Z}}^{\mathbf{autre}}$. Le sous-espace signal centré et réduit $\bar{\mathbf{Z}}^{\mathbf{utile}}$ subsiste après ces opérations. Il est maintenant nécessaire de déparaffiner numériquement les spectres Raman par application de l'ACI pour faire apparaître le spectre Raman de l'échantillon situé sous la couche de paraffine.

L'application de l'ACI sur les spectres Raman acquis sur l'échantillon paraffiné permet d'estimer les spectres $\bar{\mathbf{s}}^{\mathbf{para}}$, $\bar{\mathbf{s}}^{\mathbf{fluo}}$ et $\bar{\mathbf{s}}^{\mathbf{tissu}}$ des espèces moléculaires présentes, à savoir respectivement la paraffine, la fluorine et le tissu, et leurs profils de concentrations relatives $\bar{\mathbf{a}}^{\mathbf{para}}$, $\bar{\mathbf{a}}^{\mathbf{fluo}}$ et $\bar{\mathbf{a}}^{\mathbf{tissu}}$ au sein de l'échantillon.

Les sous-espaces signaux associés à la paraffine et à la fluorine sont calculables respectivement par les équations (4.17) et (4.18). Le déparaffinage des spectres est donc simplement obtenu en soustrayant des spectres enregistrés l'influence du spectre de la paraffine. De plus, à cette étape, nous pouvons soustraire l'influence du spectre de la fluorine. Par abus de langage, le déparaffinage dénote l'élimination des spectres de la paraffine et de la fluorine. Le sous-espace $\bar{\mathbf{Z}}^{\mathbf{tissu}}$ informatif sur le tissu présent dans l'échantillon est estimé par :

$$\bar{\mathbf{Z}}^{\mathbf{tissu}} = \bar{\mathbf{Z}} - \bar{\mathbf{Z}}^{\mathbf{para}} - \bar{\mathbf{Z}}^{\mathbf{fluo}}.$$

Le déparaffinage numérique est achevé par l'estimation du sous-espace $\bar{\mathbf{Z}}^{\mathbf{tissu}}$.

La procédure de déparaffinage numérique qui vient d'être décrite s'appuie sur la combinaison d'une méthode d'analyse indirecte de l'échantillon qui est la spectroscopie Raman et d'une méthode de traitement statistique des signaux Raman qui est l'ACI. La complémentarité de ses deux techniques conduit à un procédé de déparaffinage simple et efficace qui a été décrit par les auteurs dans [49, 48]. Dans la suite de ce chapitre, nous allons prouver l'efficacité de cette méthodologie sur un exemple concret issu de la cancérologie. Il s'agit d'extraire le spectre de la peau à partir d'un échantillon paraffiné de cette peau, et d'en déduire la nature de ce tissu.

4.4.7 Application du déparaffinage numérique au diagnostic précoce de mélanomes

4.4.7.1 Mélanome cutané

Description : La peau est un organe stratifié composé de trois couches comme illustré sur la figure 4.11. La couche externe est l'épiderme qui est constituée principalement de deux types de cellules : les kératinocytes, qui produisent de la kératine, et les mélanocytes qui fabriquent la mélanine. La couche intermédiaire est le derme qui se compose de fibroblastes, de macrophages, de lymphocytes, de mastocytes et d'adipocytes. Ces cellules sécrètent du collagène et de l'élastine. La couche la plus interne est l'hypoderme, constitué de tissus adipeux et conjonctifs.

Les mélanomes trouvent leur origine dans la transformation cancéreuse des mélanocytes, cellules responsables de la pigmentation de la peau. Leur développement se fait en deux phases. La première est un grossissement latéral superficiel dans l'épiderme. Un dépistage précoce des mélanomes dans cette phase de développement permet une exérèse simple et rapide qui assure un taux de guérison élevé. La deuxième phase correspond à un développement vertical de la tumeur dans le derme. Cette phase est appelé la phase invasive. Une fois cette phase amorcée, la prolifération du cancer est rapide. Les traitements thérapeutiques deviennent plus lourds. Les chances de guérison sont beaucoup plus faibles que pour des traitements apportés au patient dans la phase de grossissement latéral.

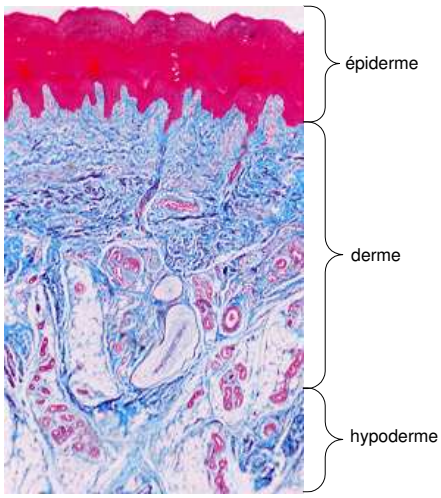


FIG. 4.11 – Coupe histologique transversale d'un échantillon de peau

Causes : Le mélanome cutané est la forme la plus sévère des cancers de la peau et contribue pour les trois quarts des morts par cancer de la peau. Cette transformation est principalement due à l'influence démontrée des rayons ultraviolets. Dans le monde, l'incidence des mélanomes augmente de 5% à 10% tous les ans avec l'apparition de plus de 100 000 nouveau cas dans le monde par an. Les causes en sont multiples : la multiplication des expositions solaires intenses de courte durée, l'altération de la couche d'ozone qui est la protection naturelle contre les rayonnements ultraviolets, le non enregistrement systématique de l'apparition de tout nouveau cas de mélanome. Cette dernière raison laisse d'ailleurs supposer une sous-estimation du nombre réel de cas de mélanomes.

Développement : Le mélanome se développe chez des adultes de plus en plus jeunes, entre 30 et 50 ans, et il est unanimement reconnu que les risques d'apparition de mélanomes sont grands lors de surexpositions solaires avant l'âge de 12 ans. Les traitements de la maladie disséminée sous forme de cancer métastatique sont peu efficaces. Seuls les diagnostics précoces et le dépistage sont fiables car ils permettent de détecter la maladie dans les prémices de son développement. Face à la menace que représente les mélanomes pour les jeunes générations, la volonté d'informer les populations des dangers représentés par les mélanomes et surtout par les expositions solaires a conduit certains pays à mener une politique d'enregistrement systématique de tout nouveau cas de mélanome comme en Belgique, en Argentine, en Australie ou en Nouvelle-Zélande.

Mélanome et nævus : Une difficulté dans le diagnostic précoce est la ressemblance visuelle entre les mélanomes et les nævi qui ne sont rien d'autre qu'un amas de cellules pigmentées et forment les grains de beauté. Les mélanomes ne se développent qu'à 30% à partir d'un nævus. Le développement de nouveaux outils non-invasifs pour le diagnostic précoce des mélanomes présente un intérêt crucial en pratique clinique. Ces nouvelles méthodes doivent être capables de discriminer entre un nævus bénin et un

mélanome malin. Les traitements thérapeutiques des mélanomes sont efficaces et mènent à la guérison des patients lorsqu'ils sont prescrits lors de la phase de grossissement latéral dans l'épiderme. Les méthodes de diagnostics précoces doivent donc être appliquées sur l'épiderme de la peau.

4.4.7.2 Études par spectroscopies vibrationnelles

Depuis quelques années, plusieurs études ont rapporté le potentiel des spectroscopies vibrationnelles pour caractériser et différencier des tissus cancéreux de tissus normaux. La spectroscopie infrarouge a été utilisée pour mesurer le degré clinicopathologique de lymphomes. Dans [2], les auteurs ont montré l'existence d'une corrélation entre le stade de développement de la tumeur et l'intensité de certaines régions du spectre infrarouge enregistré. Le potentiel de la spectroscopie infrarouge à diagnostiquer le cancer du colon a été démontré dans [3]. L'intensité des bandes spectrales de marqueurs moléculaires, tels que le phosphate et le rapport ARN/ADN, quantifie l'état pathologique de tissus de colons. De même les auteurs de [93] ont étudié les différences spectrales entre les spectres infrarouge de carcinomes baso-cellulaires et de tissus normaux. Le cancer du sein a été caractérisé grâce à la spectroscopie infrarouge appliqué sur des tissus sains et cancéreux [71]. La spectroscopie infrarouge a permis également de mettre en évidence des changements moléculaires entre mélanome malin et nævus bénin [53, 126].

La complémentarité entre la spectroscopie Raman et la spectroscopie infrarouge a conduit à l'exportation vers la spectrométrie Raman des recherches sur la caractérisation des cancers par spectroscopie infrarouge décrite précédemment. Mais depuis peu l'enjeu de ces études est devenu crucial grâce au développement de spectromètres Raman portatifs et de la construction de spectromètres Raman spécifiques à des études *in vivo*. La volonté d'instaurer des examens cliniques non-traumatiques et non-invasifs est le but ultime de ces recherches. Mais des recherches *ex vivo* préalables sont indispensables pour étudier les caractéristiques spectrales des tissus cancéreux et en déduire un diagnostic sûr.

La spectroscopie Raman a été utilisée de façon extensive à l'étude et la caractérisation de tissus cancéreux de toutes origines. Une altération de la structure des lipides, observée grâce aux variations des bandes spectrales définies par les cisaillements des CH_2 $(1420 - 1450\ cm^{-1})$ et la torsion des groupes $-(CH_2)_n-$ (autour de 1300 cm^{-1}), et des protéines, par les changements spectraux des bandes amide I $(1640 - 1680\ cm^{-1})$, amide III $(1220 - 1300\ cm^{-1})$ et l'élongation $C - C$ $(928 - 940\ cm^{-1})$, a été mise au jour dans des échantillons de carcinomes baso-cellulaires [45]. La recherche sur le cancer du sein a bénéficié de la spectroscopie Raman pour mettre au jour des différences de concentrations des protéines et du carbonate de calcium entre un tissu bénin et un tissu malin [52]. La spectroscopie Raman a été utilisée dans [85] pour diagnostiquer des cancers du col de l'utérus par traitement numérique de spectres Raman. Des études similaires ont été menées pour distinguer entre tissu normal et tissu cancéreux pour les cancers du poumon [61] et de la peau de type mélanomes malins [44].

4.4.7.3 Limites des spectroscopies vibrationnelles

Face à l'incidence mondiale des cancers, les études et les recherches se sont multipliées pour proposer des méthodes efficaces de diagnostic ou de traitement thérapeutique. Mais toute recherche nécessitant de la matière à étudier, des tumorothèques se sont développées partout dans le monde. Dans ces centres, des échantillons tissulaires cancéreux sont conservés à des fins d'analyses biochimiques et biomoléculaires. La conservation des tissus peut se faire selon deux méthodes de référence : la congélation ou le paraffinage. La congélation permet de garder intactes les propriétés des tissus, en particulier elle n'altère pas les acides nucléiques et les protéines. La technique d'inclusion dans la paraffine après fixation dans le formol permet une analyse histologique de bonne qualité mais une analyse moléculaire de faible qualité. Bien que la tendance actuelle est de favoriser la cryogénisation des échantillons tissulaires, de nombreuses banques possèdent un grand nombre d'échantillons paraffinés. Dans tous les articles de référence cités précédemment, les tissus analysés ont été conservés par l'une de ces techniques.

Avant d'être analysés, les échantillons paraffinés doivent être dégraissés par actions d'agents solvants divers sur l'échantillon. Comme expliqué dans le paragraphe 4.4.1, cette méthodologie altère la structure chimique de l'échantillon. Elle est pourtant largement répandue dans le monde, et à notre connaissance, seuls deux travaux ont été menés sur des tissus paraffinés non dégraissés et analysés par spectroscopie vibrationnelle. Dans [53] et [126], la spectroscopie infrarouge a été utilisée avec succès sur des sections paraffinées et non dégraissées de peau afin de discriminer entre des nævi et des mélanomes. Mais cette discrimination était basée sur des bandes vibratoires étroites où la paraffine n'a pas d'influence. L'utilisation de la spectroscopie Raman ne s'est jamais faite sur des échantillons paraffinés non dégraissés à des fins de diagnostic de cancers car la paraffine est très active en Raman, comme expliqué au paragraphe 4.4.2.

L'application de la méthode de déparaffinage décrite précédemment va permettre de mener une analyse complète sur des échantillons de peau non dégraissés et non altérés, et de s'affranchir des inconvénients liés à la présence de la paraffine.

4.4.7.4 Application du déparaffinage numérique au diagnostic précoce de mélanomes

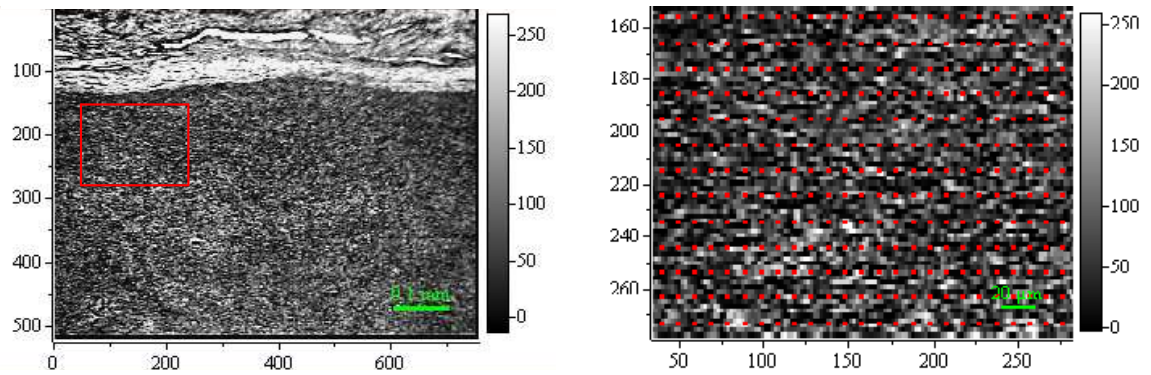
Considérations expérimentales : Comme expliqué précédemment, les mélanomes sont curables lorsqu'ils sont diagnostiqués dans leur première phase de développement latéral dans l'épiderme. Les études menées dans la suite de cette application ont été réalisées à partir du traitement numérique et de l'analyse visuelle de spectres Raman enregistrés sur des échantillons paraffinés d'épiderme de peau et fixés sur un support de fluorine.

Des sections transversales de 10 μm d'épaisseur ont été coupées sur des biopsies de mélanomes et de nævi fournies par le département de dermatologie du Centre Hospitalier Universitaire de Reims. La totalité des tissus analysés sont affectés par la prolifération des mélanocytes et ne présentent donc pas de

parties normales d'épiderme. Des images spectrales ont été collectées par un spectromètre Raman de type Labram (Dilor-Jobin Yvon, Lille, France) en mode point par point et par pas de 10 μm. La résolution spatiale des acquisitions est de 4 cm^{-1}. La source de lumière excitatrice est un laser titane-saphir de longueur d'onde 785 nm. En chaque point, le spectre a été enregistré en 1305 nombres d'onde différents couvrant une gamme spectrale s'étalant de 200 à 1800 cm^{-1}. La résolution spectrale est en moyenne de 1.22 cm^{-1}. Le temps d'acquisition de chaque spectre est de 30 secondes. Afin de limiter le bruit de mesure, deux enregistrements ont été effectués en chaque point. Le spectre moyen a été calculé pour chaque point de mesure. L'ensemble de ces spectres moyens forme un cube de données qui a été manipulé selon la méthode décrite au paragraphe 2.2.2 à la page 40 pour être mis sous une forme matricielle plus adaptée aux traitements numériques des signaux.

Au total, 3 mélanomes et 3 nævi ont été étudiés. L'étude suivante propose les résultats estimés à partir d'un mélanome et d'un nævus. L'annexe C présente l'ensemble des résultats sur les 3 mélanomes et les 3 nævi.

Étude d'un mélanome : Dans un premier temps, la procédure de déparaffinage numérique a été testée sur un échantillon paraffiné de mélanome fixé sur support en fluorine. L'image de l'échantillon à analyser est présentée dans son intégralité sur la figure 4.12(a). Le rectangle rouge encadre la zone de l'épiderme de l'échantillon qui a été soumise à l'analyse par spectroscopie Raman. Un zoom de cette zone est donné sur figure 4.12(b). Les points rouges de cette image représentent les points de l'échantillon où les spectres Raman ont été acquis. Le quadrillage de l'échantillon en 13 lignes et 25 colonnes a conduit à l'enregistrement d'un total de 325 spectres Raman selon 1305 nombres d'onde différents reflétant sa structure moléculaire. La matrice $\mathbf{X}$ des données à traiter est alors de dimensions 325×1305.

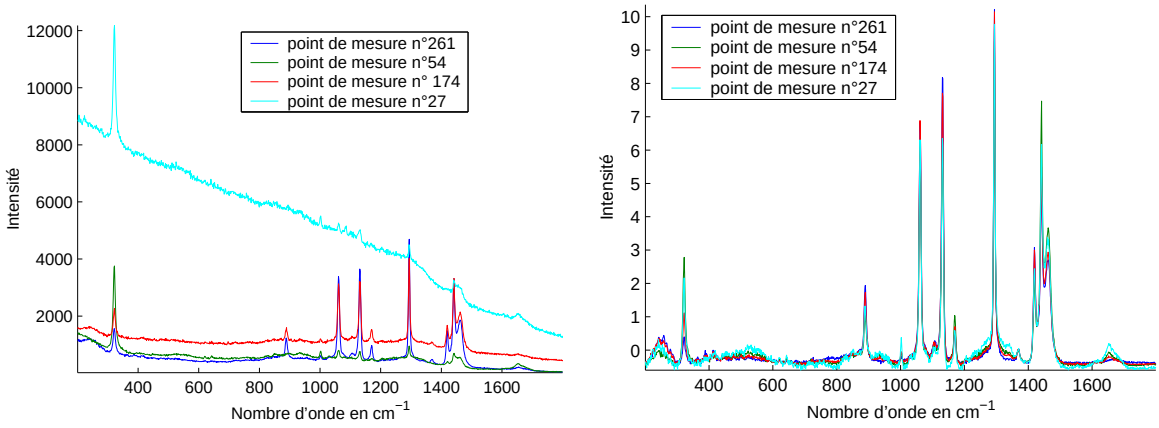


(a) (b)

FIG. 4.12 – Image de l'échantillon de mélanome paraffiné sur support de fluorine étudié (a) image complète de l'échantillon, (b) zoom sur la partie de l'épiderme analysée par spectroscopie Raman

Un sous-ensemble aléatoire de 4 spectres Raman appartenant à $\mathbf{X}$ est dessiné sur la figure 4.13(a).

Conformément à la figure, les spectres Raman bruts sont corrompus par la présence d'une ligne de base et expriment majoritairement la présence de la paraffine et de la fluorine. La présence de la peau est seulement visible dans les bandes centrées en 1002 cm^{-1} et 1660 cm^{-1}.



(a) (b)

FIG. 4.13 – Sous-ensemble de 4 spectres Raman choisis aléatoirement parmi l'ensemble des 325 spectres mesurés sur l'échantillon (a) spectres bruts, (b) spectres prétraités par l'élimination de la ligne de base, le recalage des pics, le centrage et la réduction des spectres

Les spectres bruts ont été corrigés et mis en forme grâce aux prétraitements décrits dans la partie 4.4.5 à la page 137. L'élimination de la ligne de base, le recalage des pics de la paraffine et de la fluorine, le centrage et la réduction des spectres résultants ont été appliqués aux spectres de la matrice $\mathbf{X}$ et forment la matrice des spectres prétraités, autrement dit le sous-espace $\bar{\mathbf{Z}}^{\mathbf{utile}}$. Les spectres de la figure 4.13(a) ainsi transformés sont visibles sur la figure 4.13(b).

▷ ***Séparation en 3 sources :*** Les techniques d'ACI peuvent maintenant être appliquées au jeu de données $\bar{\mathbf{Z}}^{\mathbf{utile}}$. Leur utilisation requiert la connaissance du nombre de sources p sous-jacentes au modèle. Il est évident que les échantillons analysés dans nos travaux sont constitués de trois espèces chimiques principales : la paraffine pour la conservation de l'échantillon, la fluorine pour fixer l'échantillon et le spécimen de peau à analyser. Ces trois espèces sont d'ailleurs à l'origine de la modélisation de la procédure de déparaffinage numérique présentée au paragraphe 4.4.3. Il nous est donc apparu légitime de supposer un modèle des données à 3 sources. L'algorithme JADE [20] a été appliqué sur le jeu de données $\bar{\mathbf{Z}}^{\mathbf{utile}}$.

Les trois sources estimées par cette méthode sont présentées sur la figure 4.14. La première source de la figure 4.14(a) présente clairement la présence de 6 pics, centrés aux nombres d'onde 890 cm^{-1}, 1063 cm^{-1}, 1133 cm^{-1}, 1296 cm^{-1}, 1372 cm^{-1}, 1418 cm^{-1} et labellisé par +, parmi les 9 pics caractéristiques de la paraffine. Les deux bandes spectrales caractéristiques de la peau et visibles à l'œil nu aux nombres d'ondes 1002 cm^{-1} et 1660 cm^{-1} sont modélisées dans cette même source mais avec une contribution négative. Dans la deuxième source de la figure 4.14(b) se mélangent des pics caractéristiques de la paraffine

avec des contributions positives, tels que les pics à 1063 cm^{-1}, 1172 cm^{-1}, 1296 cm^{-1}, 1441 cm^{-1} et 1463 cm^{-1}, et avec des influences négatives tels que les pics à 890 cm^{-1}, 1133 cm^{-1} et 1418 cm^{-1}, mais aussi des pics représentatifs de la peau en 1002 cm^{-1} et 1660 cm^{-1} mais cette fois-ci avec des maxima positifs. La troisième et dernière source de la figure 4.14(c) est constituée du pic de la fluorine centré en 325 cm^{-1} et trahit la présence de résidus de quelques pics de la paraffine. Les deux premières sources sont encore des mélanges linéaires des spectres de la paraffine et de la peau. La dernière source peut être considérée comme l'estimation du spectre de la fluorine. Cette modélisation à 3 sources n'est donc biologiquement pas judicieuse puisqu'une seule source est assimilable au spectre d'une espèce chimique pure, les spectres de la paraffine et de la fluorine devant être positifs et parcimonieux.

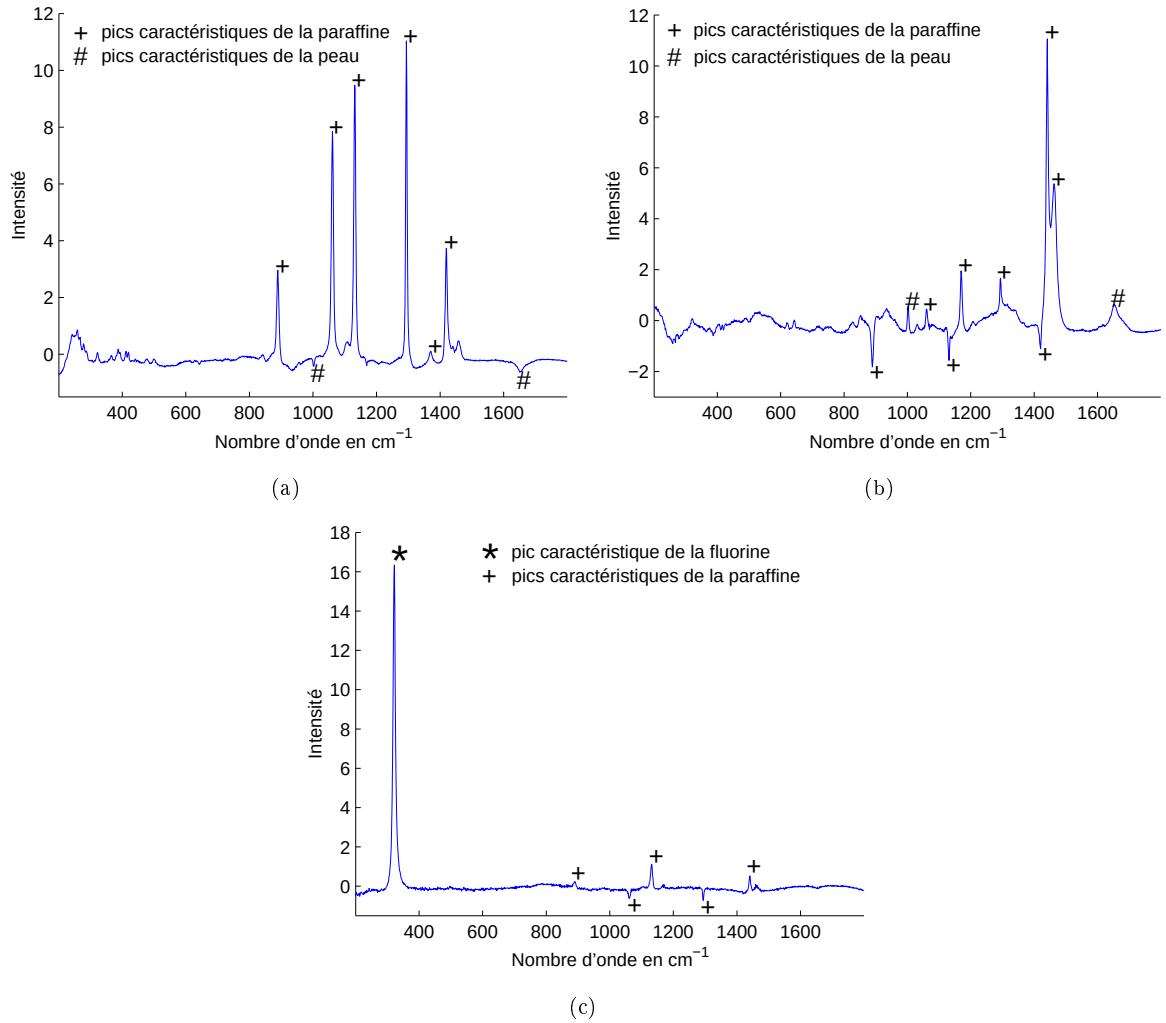


FIG. 4.14 – Sources estimées sur un mélanome par JADE pour un modèle à 3 sources : (a) première source, (b) deuxième source, (c) troisième source

▷ ***Séparation en 4 sources :*** De manière à séparer les spectres de la peau et de la paraffine, le nombre de sources supposé du modèle a été augmenté. Les résultats obtenus pour un modèle à 4 sources sont montrés sur la figure 4.15. La source 2 de la figure 4.15(b) ne reflète pratiquement plus la présence

de la peau. Les pics caractéristiques de la peau à 1002 cm^{-1} et 1660 cm^{-1} y sont de très faible intensité. Un problème est le mélange de pics de la paraffine mais avec des contributions positives ou négatives. Cependant, les pics tournés vers les intensités négatives sont de faible intensité en comparaison des pics orientés vers les intensités positives. Ce spectre peut donc être considéré, aux erreurs d'estimation près, au spectre des bandes spectrales de la paraffine centrées en 1172 cm^{-1}, 1441 cm^{-1} et 1463 cm^{-1}. La source 3 de la figure 4.15(c) peut toujours être assimilée au spectre Raman de la fluorine pure. Les sources 1 et 4 sont encore des mélanges linéaires des spectres de la paraffine et de le peau. Une amélioration notable est visible sur la source 4. Le spectre de la peau, bien qu'encore pollué par des pics rémanents de la paraffine, commence à émerger.

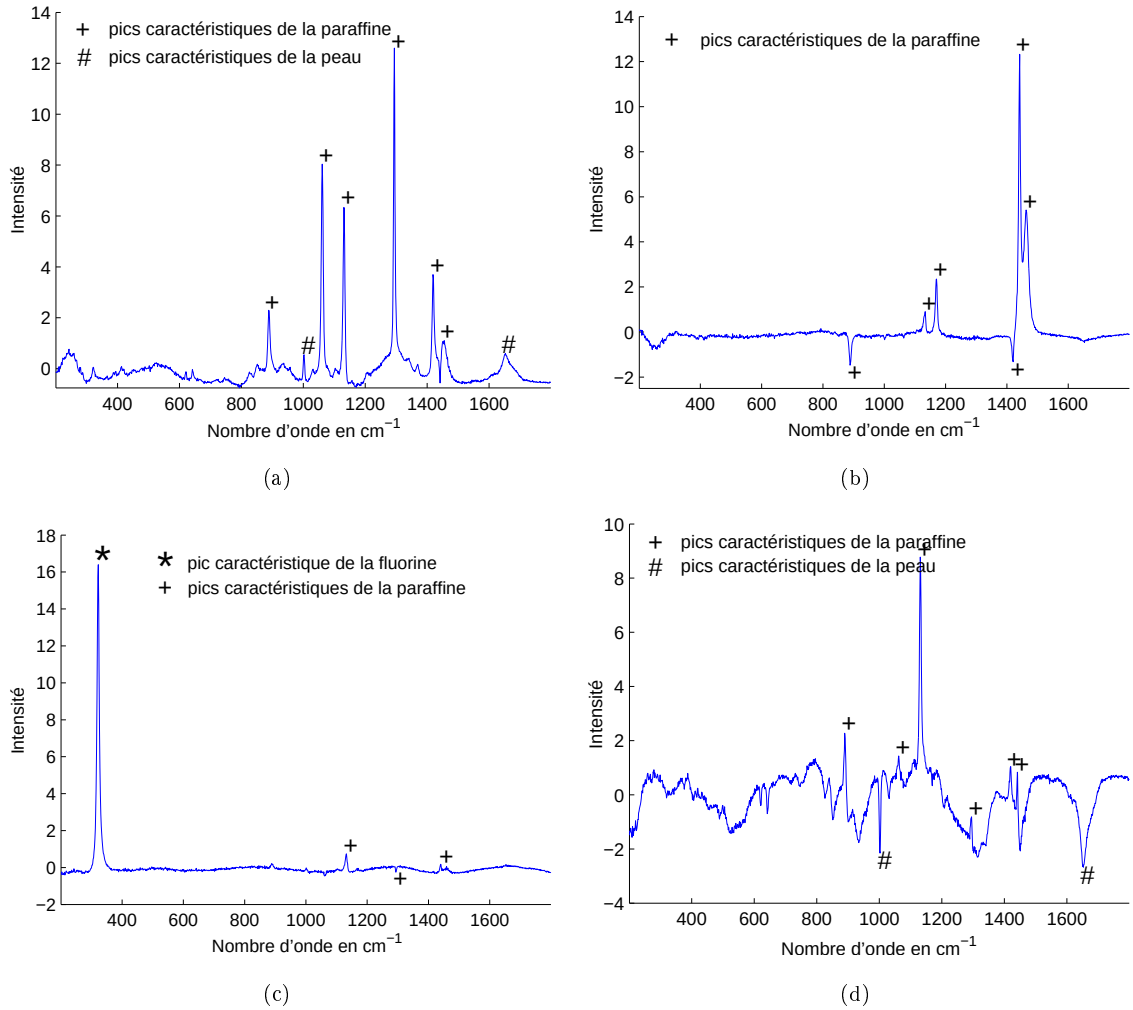


FIG. 4.15 – Sources estimées sur un mélanome par JADE pour un modèle à 4 sources : (a) première source, (b) deuxième source, (c) troisième source, (d) quatrième source

▷ ***Séparation en 5 sources :*** Le nombre de sources a donc été fixé à 5 afin de séparer les pics de la paraffine et le spectre de la peau. Les résultats de cette modélisation à 5 sources sont présentés sur la figure 4.16.

La source 1 correspond à la modélisation des pics de la paraffine centrés en 1063 cm^{-1}, 1296 cm^{-1} et 1372 cm^{-1}. Dans l'estimation de cette source, de petites imperfections sont visibles aux nombres d'onde 1133 cm^{-1}, 1372 cm^{-1}, 1418 cm^{-1}, 1441 cm^{-1} et 1463 cm^{-1}. Elles correspondent aux résidus des pics caractéristiques de la paraffine situés à ces mêmes nombres d'onde. L'estimation de sources ne pouvant pas être par principe parfaite, et ces pics parasites étant de très faible intensité, cette source est considérée comme une bonne estimation des bandes à 1063 cm^{-1} et 1296 cm^{-1} de la paraffine.

La source 2 est quasiment identique à la source 2 du modèle d'estimation à 4 sources de la figure 4.15(b). Cette source modélise quant à elle les bandes spectrales centrées en 1172 cm^{-1}, 1441 cm^{-1} et 1463 cm^{-1} de la paraffine. Les pics en 890 cm^{-1} et 1418 cm^{-1} de la paraffine présentent des résidus sur cette source. Une fois encore, nous considérons l'estimation de la source 2 comme une estimation réaliste des bandes à 1441 cm^{-1} et 1463 cm^{-1} de la paraffine puisque leur intensité est grande comparée à celle des bandes parasites.

La source 3 est assimilable à la modélisation des bandes spectrales de la paraffine centrées en 890 cm^{-1}, 1133 cm^{-1} et 1418 cm^{-1}. Aucune autre bande spectrale n'est présente dans cette source. Elle modélise donc parfaitement ces 3 bandes spectrales de la paraffine.

La source 4 présente un pic unique en 325 cm^{-1}. De légères interférences sont notables aux nombres d'onde 1133 cm^{-1} et 1441 cm^{-1}. Le spectre de la fluorine peut alors être considérée comme très bien estimé par la source 4.

La source 5 présente toutes les caractéristiques du spectre de la peau avec ses pics en 1002 cm^{-1} et 1660 cm^{-1}. D'autres bandes spectrales, jusqu'alors invisibles à l'œil nu, sont attribuables à la peau.

La région spectrale s'étalant de 1200 cm^{-1} à 1400 cm^{-1} est assignée aux bandes vibrationnelles des hydrates de carbone de la bande amide III. La région comprise entre 1440 cm^{-1} et 1500 cm^{-1} traduit la présence de protéines et de lipides.

Ce spectre présente cependant quelques anomalies qui se traduisent par la présence de pics caractéristiques de la paraffine et centrés sur les longueurs d'onde 1063 cm^{-1}, 1133 cm^{-1}, 1296 cm^{-1}, 1418 cm^{-1} et 1441 cm^{-1}. L'invasion de ces pics est faible comparée à l'intensité des bandes spectrales caractéristiques de la peau. Cette estimation du spectre de la peau est donc très bonne.

Comme expliqué à la section 4.4.4 à la page 134, bien que la mélanine soit une composante de l'épiderme de la peau, son spectre n'a pas été estimé par les méthodes d'ACI puisque dans la fenêtre spectrale d'enregistrement des spectres, sa puissance est trop faible comparée aux intensités Raman de la paraffine et de la fluorine. Nous parlons pour la source 5 du spectre de la peau et non du spectre de la kératine puisque la peau se compose d'autres espèces chimiques dont le spectre Raman n'est ni parcimonieux ni décorrélé du spectre de la kératine. La méthode de déparaffinage proposée ne permet d'estimer qu'une seule source non parcimonieuse et non décorrélée aux autres sources. La cinquième source représente donc un spectre moyen de la peau.

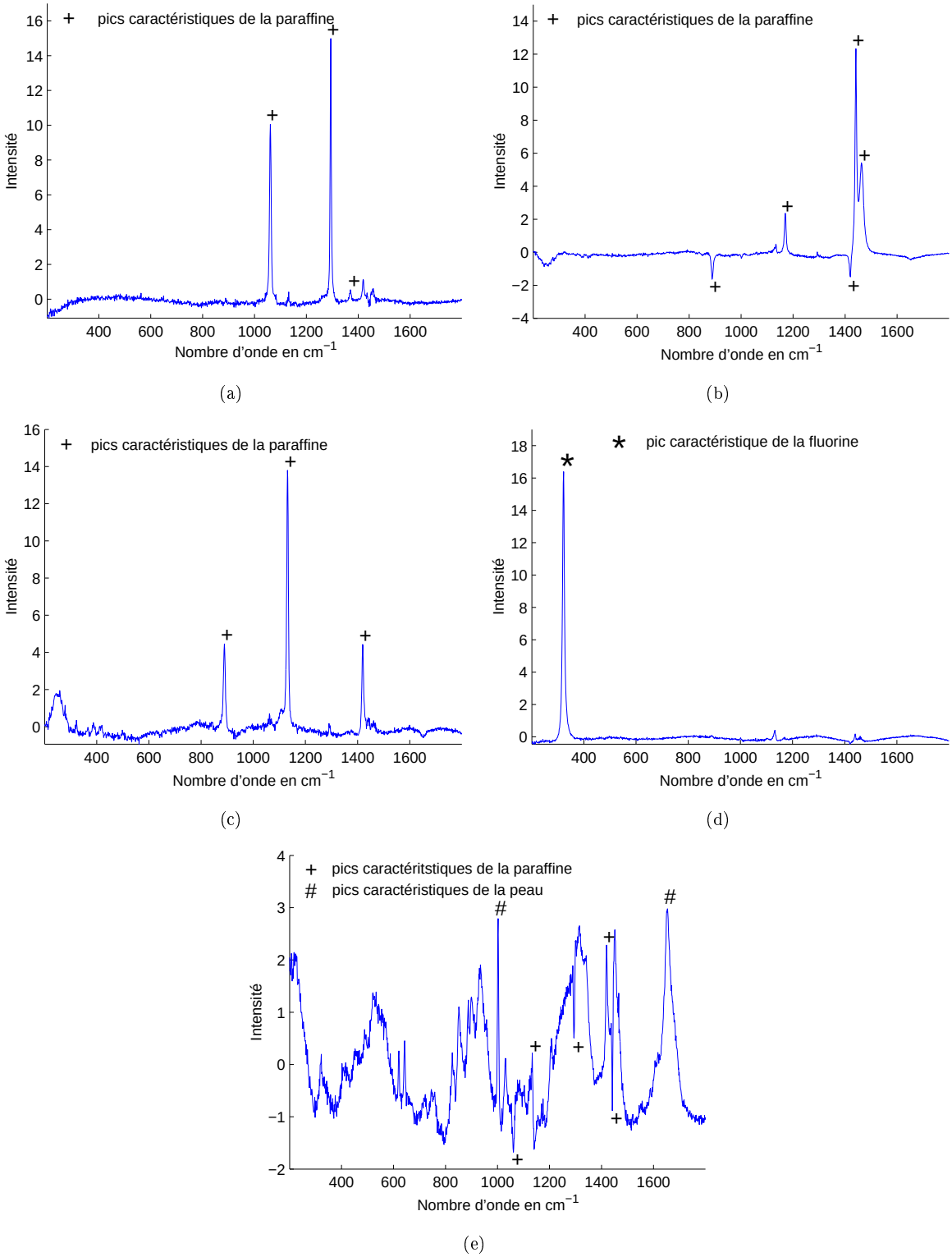


FIG. 4.16 – Sources estimées sur un mélanome par JADE pour un modèle à 5 sources : (a) première source, (b) deuxième source, (c) troisième source, (d) quatrième source, (e) cinquième source

Le sous-espace $\bar{\mathbf{Z}}^{\mathbf{para}} = \bar{\mathbf{a}}^{\mathbf{para}}(\bar{\mathbf{s}}^{\mathbf{para}})^T$ se modélise finalement par la contribution de trois sous-espaces indépendants $\bar{\mathbf{Z}}^{\mathbf{para_1}} = \bar{\mathbf{a}}^{\mathbf{para_1}}(\bar{\mathbf{s}}^{\mathbf{para_1}})^T$, $\bar{\mathbf{Z}}^{\mathbf{para_2}} = \bar{\mathbf{a}}^{\mathbf{para_2}}(\bar{\mathbf{s}}^{\mathbf{para_2}})^T$ et $\bar{\mathbf{Z}}^{\mathbf{para_3}} = \bar{\mathbf{a}}^{\mathbf{para_3}}(\bar{\mathbf{s}}^{\mathbf{para_3}})^T$:

$$\bar{\mathbf{Z}}^{\mathbf{para}} = \bar{\mathbf{Z}}^{\mathbf{para_1}} + \bar{\mathbf{Z}}^{\mathbf{para_2}} + \bar{\mathbf{Z}}^{\mathbf{para_3}}.$$

La fluorine s'exprime comme prédit par un sous-espace $\bar{\mathbf{Z}}^{\mathbf{fluo}} = \bar{\mathbf{a}}^{\mathbf{fluo}}\left(\bar{\mathbf{s}}^{\mathbf{fluo}}\right)^T$ et l'information utile du spécimen de peau se trouve résumée dans le sous-espace $\bar{\mathbf{Z}}^{\mathbf{tissu}} = \bar{\mathbf{a}}^{\mathbf{tissu}}\left(\bar{\mathbf{s}}^{\mathbf{tissu}}\right)^T$.

L'efficacité de la séparation des spectres des espèces chimiques pures est confirmée par plusieurs observations. Tout d'abord, les spectres estimés pour chaque espèce chimique sont nettoyés de l'influence des autres espèces. Chaque espèce est ainsi complètement isolée des autres. Ensuite, les spectres de la paraffine et de la fluorine sont parcimonieux et composés de quelques pics Raman étroits et intenses, comme nous l'avions remarqué sur les spectres de référence donnés sur les figures 4.1 et 4.2 à la page 132. Finalement, les colonnes $\bar{\mathbf{a}}_j$ de la matrice de mélange estimée $\bar{\mathbf{A}}$ et associées aux spectres sources $\bar{\mathbf{s}}_j$ sont composées exclusivement d'éléments positifs. Chaque colonne de $\bar{\mathbf{A}}$ représente le profil des concentrations relatives de chaque espèce chimique. Leur positivité est une évidence physique. La modélisation des données proposée par l'ACI est ainsi en total accord avec les contraintes physiques imposées aux spectres Raman et aux concentrations des espèces. Les résultats obtenus sont physiquement réalistes.

Pour résumer, la méthode de déparaffinage numérique développée par association de la spectroscopie Raman et des techniques d'ACI permet d'éliminer la paraffine et la fluorine de l'échantillon de peau. Le spectre de la peau est ainsi estimé avec une grande efficacité puisque les caractéristiques spectrales principales de la peau sont retrouvées. Il est à noter que contrairement à nos connaissances *a priori*, non pas 3 sources, mais 5 sources ont été estimées. Le spectre pur de la paraffine s'avère décomposable en 3 spectres différents. Chaque spectre modélise des bandes spectrales de la paraffine exclusivement présentes dans ce spectre. Nous pouvons en conclure que le spectre de la paraffine, contrairement aux modèles classiques, n'a pas un comportement linéaire en fonction du nombre d'onde. Il est composé de bandes spectrales qui fluctuent indépendamment les unes des autres. Ces conclusions sont disponibles dans [49, 48, 125]. Ces résultats ont été obtenus par l'application de l'algorithme JADE sur le jeu de données. Les mêmes sources sont estimées lorsque l'algorithme FastICA, décrit à la section 4.3.5.1, est employé. Ces résultats sont présentés dans l'annexe D. L'application de techniques de FMN sur ces données n'est pas efficace pour les séparer malgré les hypothèses de positivité sur les sources et les mélanges qui sont remplies par ces spectres, comme le montre l'annexe E. Il est légitime de se demander d'où vient cette différence de variation d'une bande à une autre. Elle correspond peut être à une interaction entre le tissu et la paraffine. Afin de répondre à cette question, plusieurs blocs de paraffine fixés sur de la fluorine ont été étudiés dans le paragraphe suivant.

Décomposition de la paraffine : Dans les mêmes conditions expérimentales que précédemment, des blocs de paraffine fixés sur un support de fluorine ont été analysés par spectroscopie Raman. Ces spectres ont ensuite été prétraités par les procédures décrites dans le paragraphe 4.4.5, puis traités par l'algorithme

JADE. Puisque le spécimen de peau n'est plus présent dans l'échantillon à analyser, un modèle à 4 sources est choisi, une source pour le spectre de la fluorine et trois sources pour représenter la totalité des bandes spectrales de la paraffine. Les sources estimées à partir de l'enregistrement à la surface d'un bloc de paraffine de 49×9 spectres Raman sont disponibles sur la figure 4.17.

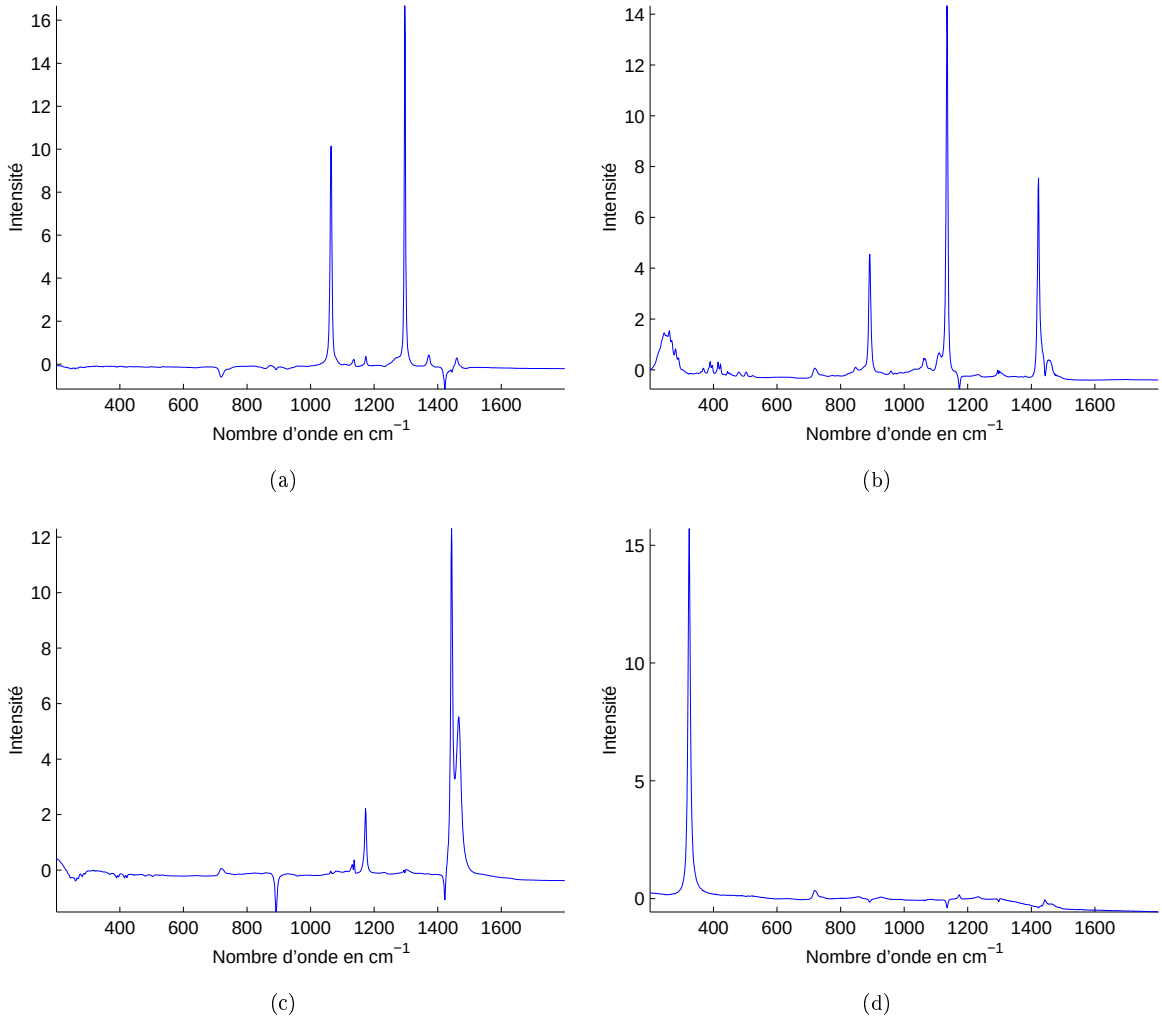


FIG. 4.17 – Sources estimées sur un bloc de paraffine pure par JADE pour un modèle à 4 sources : (a) première source, (b) deuxième source, (c) troisième source, (d) quatrième source

La première source estimée exhibe des pics centrés en 1063 cm^{-1} et 1296 cm^{-1}. La deuxième source possède des pics caractéristiques de la paraffine aux nombres d'onde 890 cm^{-1}, 1133 cm^{-1} et 1418 cm^{-1}. La troisième source modélise les pics centrés à 1172 cm^{-1}, 1441 cm^{-1} et 1463 cm^{-1}. La source 4 a un pic unique à 325 cm^{-1}. Les trois premières sources de la figure 4.17 modélise ainsi le spectre de la paraffine. La quatrième source représente le spectre de la fluorine. Ces résultats sont identiques à ceux estimés sur les spectres acquis sur un échantillon de peau paraffinée fixé sur un support de fluorine et présentés à la figure 4.16. La paraffine se modélise comme une somme de trois sources indépendantes pondérées par leurs coefficients de mélange qui varient d'un point d'acquisition à un autre. La paraffine doit être

représentée par un modèle linéaire à trois sources.

Une observation supplémentaire permet de conforter cette conclusion. Nous avons appliqué les techniques d'ACI sur différents blocs de paraffine et pour un nombre de sources supérieur à 4 (3 pour la paraffine et 1 pour la fluorine). Les sources estimées dépendent du jeu de données utilisé mais représentent toujours des sources de bruit et n'ont pas trouvé d'interprétation physique par les spécialistes. Ces observations donnent ainsi la preuve qu'un modèle composé de plus de 4 sources n'est pas pertinent, et donc que la paraffine se modélise par 3 sources indépendantes.

Nous allons maintenant tenter de donner à chaque source de la paraffine une interprétation physique. La source 1 est composée de bandes vibrationnelles traduisant l'extension des liaisons $C-C$ et la déformation des laisons CH_2. La source 2 se compose des énergies engendrées par la vibration dite de *wagging*, qui correspond à la vibration simultanée de même sens des angles $\widehat{C_1C_2H_1}$ et $\widehat{C_1C_2H_2}$ et qui est illustrée sur la figure 4.18, de l'étirement des liaisons $C-C$ et de la déformation des groupes CH_3. Les vibrations d'étirement des liaisons de valence $C-C$, de déformation des groupes CH_2 et de balancement dans le plan des groupes CH_2 constituent la source 3. L'ensemble des vibrations associées à la paraffine est récapitulé dans le tableau 4.1.

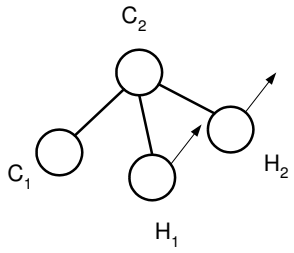


FIG. 4.18 – Illustration de la vibration dite de *wagging*

Nombre d'onde en cm^{-1}	Attribution
890	wagging des groupes CH_2
1063	étirement des liaisons $C-C$
1133	étirement des liaisons $C-C$
1172	étirement des liaisons $C-C$
1296	déformation des groupes CH_2
1418	déformation des groupes CH_3
1441	déformation des groupes CH_2
1463	balancement des groupes CH_2

TAB. 4.1 – Attributions des bandes spectrales de la paraffine

La séparation du spectre de la paraffine en trois sources ne suit aucune logique vibrationnelle. Dans chaque source, des vibrations liées aux groupes CH_2, CH_3 et aux liaisons $C-C$ se mélangent. Or il aurait été naturel de voir toutes les vibrations liées au groupe CH_2 regroupées dans une source, toutes les

vibrations liées au groupe CH_3 dans une autre source, et toutes les vibrations liées au liaisons $C-C$ dans une dernière source. Nous ne sommes pour le moment pas en mesure d'expliquer cette répartition des vibrations au sein des sources et d'en donner une interprétation physique réaliste. Plusieurs séparation des spectres Raman ont été appliquées sur trois blocs de paraffine différents et sur 6 échantillons de peau différents. Dans tous les cas, la même décomposition de la paraffine est obtenue. La paraffine est donc modélisable par trois sources différentes, comme les auteurs l'ont démontré dans [135]. Le déparaffinage numérique d'échantillon biologique a donc été prouvé comme efficace en pratique. Dans le paragraphe suivant, nous avons appliqué cette méthode sur des échantillons de mélanomes et de nævi afin de les discriminer .

Discrimination entre mélanome et nævus : La procédure de déparaffinage numérique est capable de restituer le spectre Raman de faible intensité d'un spécimen de peau noyé dans les spectres très énergétiques de la paraffine et de la fluorine. Face à la différence de puissance entre le spectre de la peau et les spectres de la paraffine et de la fluorine, la procédure d'estimation est suffisamment efficace pour faire apparaître l'ensemble des caractéristiques spectrales de la peau. Nous avons appliqué la méthode de déparaffinage numérique sur des spectres acquis sur des échantillons d'épidermes de mélanomes et de nævi afin d'étudier les différences spectrales entre ces deux types de tissus et pouvoir les discriminer.

▷ ***Application des techniques d'ACI :*** Par application de l'ACI sur les spectres Raman acquis à partir de l'épiderme d'un mélanome et d'un nævus, les sources estimées $\bar{\mathbf{s}}_1$, $\bar{\mathbf{s}}_2$ et $\bar{\mathbf{s}}_3$ sont attribuées dans les deux cas à la paraffine. Ces sources sont identiques dans les deux cas, comme le montrent les figures 4.16(a)-(c) et 4.19(a)-(c), confirmant une fois de plus la décomposition de la paraffine en trois sources quel que soit le type de tissu sous-jacent à analyser. Le support de fluorine se voit affecter la source $\bar{\mathbf{s}}_4$ qui est semblable dans les deux cas comme visible sur les figures 4.16(d) et 4.19(d). Seul le dernier spectre estimé $\bar{\mathbf{s}}_5$ présente des variations dans certaines régions spectrales qui traduisent les différences moléculaires entre un mélanome et un nævus, sur les figures 4.16(e) et 4.19(e). La figure 4.19 propose les sources estimées par l'application de JADE pour une modélisation à 5 sources du jeu de données spectrales de l'épiderme d'un nævus.

▷ ***Analyse :*** Des différences spectrales visibles à l'œil nu entre le spectre source de l'épiderme du nævus et l'épiderme du mélanome existent. Pour cela, nous avons représenté sur la figure 4.20 les deux sources superposées. Les spectres estimés à partir des épidermes de mélanome et de nævus ont été analysés par des biophysiciens. Les résultats ont été publiés par les auteurs dans [49, 48, 125] et sont les suivants :

- Le rapport des bandes du doublet de Fermi à 850 cm^{-1} et à 830 cm^{-1} diffère d'un type de tissu à un autre. Pour les mélanomes, il a été calculé comme égal en moyenne à 2.5, tandis que pour les nævi, sa valeur est de environ 1.6. Un tel changement nous informe sur l'état du cycle phénolique dans le résidu de tyrosine et le type de la bande moléculaire résultante (intra ou inter).
- Les variations de la structure secondaire sont marquées par une prédominance des vibrations de

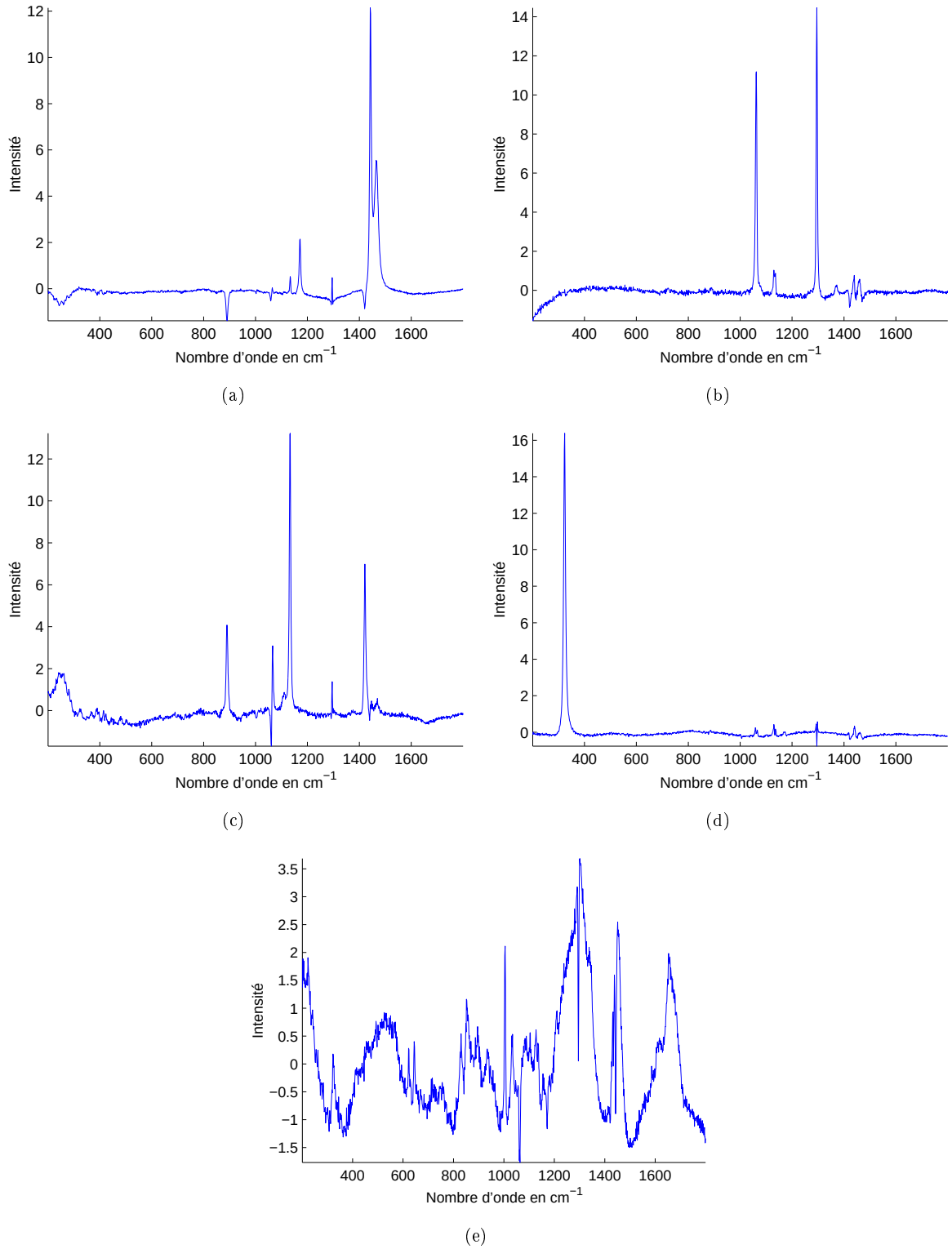


FIG. 4.19 – Sources estimées à partir d'épiderme de nævus par JADE pour un modèle à 5 sources : (a) première source, (b) deuxième source, (c) troisième source, (d) quatrième source, (e) cinquième source

l'hélice α à 1650 cm^{-1} dans la bande amide I du mélanome. Une information similaire peut être obtenue de la forte intensité de la bande à 934 cm^{-1} caractérisant l'extension des liaisons $C-C$ dans l'hélice α.

- La source du nævus présente un épaulement vers 1670 cm^{-1} qui révèle une contribution plus importante de la conformation périodique β.
- Les mêmes informations peuvent être obtenues par l'analyse des changements de la bande amide III dans la région $1220-1300$ cm^{-1}, et par l'intensité de la bande à 901 cm^{-1}.
- D'autres différences peuvent aussi être observées par exemple dans la bande à 1620 cm^{-1} qui est attribuée à l'ADN, ou dans la bande à 480 cm^{-1} qui est attribuée aux acides aminés aromatiques.
- D'autres changements dans le contenu résiduel en acides aminés des protéines peuvent être détectés dans les bandes à 620 cm^{-1}, 1003 cm^{-1}, 1033 cm^{-1} et 1610 cm^{-1} attribuées à la phénylalanine, et dans les bandes à 640 cm^{-1} et 1610 attribuées à la tyrosine.

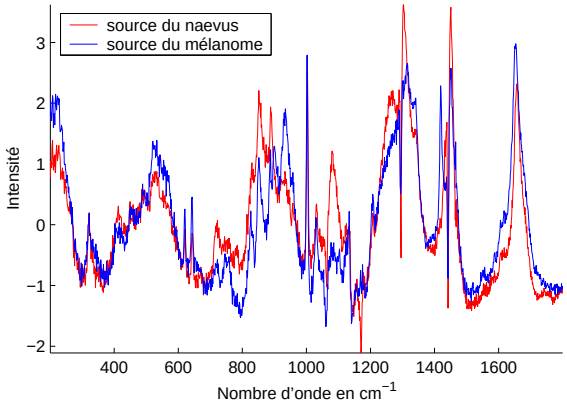


FIG. 4.20 – Comparaison entre les spectres de la peau estimés par JADE pour un mélanome (en bleu) et un nævus (en rouge)

Toutes ces différences spectrales définissent autant de descripteurs moléculaires du type de pathologie rencontré. De plus, les rapports d'intensités entre les bandes à 1650 cm^{-1} et 1450 cm^{-1} de la bande à amide I, à 1270 cm^{-1} et 1320 cm^{-1} de la bande à amide III, et à 930 cm^{-1} et 1000 cm^{-1} de la bande des extensions des liaisons $C-C$ dans les protéines sont différents d'un nævus à un mélanome. Les rapports d'intensités dans les bandes à amide I et à extension des liaisons $C-C$ dans les protéines ont été calculés à partir des spectres Raman estimés pour 3 nævi et pour 3 mélanomes et sont disponibles dans le tableau 4.2. Le rapport d'intensités des bandes à amide III n'a pu être déterminé puisqu'un pic rémanent de la paraffine subsiste à 1296 cm^{-1}.
Le rapport d'intensités des bandes à amide I présente une valeur plus faible pour les nævi que pour les mélanomes. Le rapport d'intensités des bandes vibrationnelles des liaisons $C-C$ dans les protéines est quant à lui obsolète puisque ses valeurs pour le mélanome 3 et le nævus 3 cassent la structure d'ordre instaurée par les deux premiers mélanomes et les deux premiers nævi.

Le rapport d'intensités entre les bandes à amide I s'avère être un indicateur du type de pathologie de l'échantillon de peau analysé. Mais les variations de cet indicateur sont faibles. Des méthodes de

	Amide I ($b_{1650\ cm^{-1}}/b_{1450\ cm^{-1}}$)	$C-C$ ($b_{930\ cm^{-1}}/b_{1000\ cm^{-1}}$)
Mélanome 1	1.155	0.684
Mélanome 2	0.811	0.756
Mélanome 3	0.79	1.33
Nævus 1	0.78	0.1265
Nævus 2	0.649	0.346
Nævus 3	0.548	1.01

TAB. 4.2 – Rapports d'intensités entre bandes spectrales pour 3 mélanomes et 3 nævi

classification doivent être développées pour exploiter non pas un simple rapport d'intensités de bandes mais l'ensemble des différences spectrales observées entre un mélanome et un nævus. Le spectre entier déparaffiné numériquement doit servir à définir la pathologie en présence.

4.5 Conclusion

La spectroscopie Raman est un outil puissant d'analyse de la structure et de la composition moléculaire d'un échantillon biologique. La superposition des informations vibrationnelles diffusées par les diverses espèces chimiques de l'échantillon rend son analyse difficile sans des traitements numériques appropriés. Cependant, aucune méthode n'a fait l'unanimité jusqu'à maintenant car les propriétés statistiques des spectres Raman ne sont pas exploitées à bon escient. L'Analyse en Composantes Indépendantes (ACI) propose une modélisation des données multidimensionnelles adaptée aux caractéristiques des spectres Raman. L'indépendance statistique des spectres des espèces chimiques pures de l'échantillon est suffisante pour assurer la séparation des spectres Raman. Des algorithmes simples et efficaces d'ACI ont été développés dans les années 90 et ont fait leurs preuves dans de nombreux domaines d'application. Une application innovante de l'ACI a été proposée dans ce chapitre.

Le déparaffinage numérique a été développé, à partir de l'association entre la spectrosopie Raman et l'ACI, pour proposer une alternative aux déparaffinages chimiques couramment utilisés pour préparer les échantillons biologiques à des analyses histopathologiques. Dans une première étape, les spectres Raman de l'échantillon paraffiné à analyser sont enregistrés. Ils sont inexploitables sans des traitements numériques adaptés puisque la paraffine est très active en Raman et masque les informations vibrationnelles du tissu sous-jacent. Le spectre de la paraffine et du tissu sous-jacent sont séparés par une ACI et l'analyse moléculaire du tissu est possible.

Cette méthode de déparaffinage numérique a été appliquée avec succès sur un échantillon paraffiné de peau dont le spectre a été extrait. Le spectre de la paraffine a été décomposé en trois sources indépendantes. Cette observation a été confirmée par l'étude de blocs de paraffine seule. Il en a été déduit que la paraffine possède des bandes spectrales à variations d'intensités indépendantes les unes des autres.

La méthode de déparaffinage présentée a également été testée afin d'effectuer la discrimination entre mélanomes et nævi. Les différences spectrales existantes entre les spectres de ces deux pathologies ont été mises à jour.

Conclusion et perspectives

Le travail effectué au cours de ce doctorat a permis de montrer l'importance du rôle du traitement du signal en spectroscopies optiques. L'efficacité des techniques de séparation de sources a été prouvée pour estimer les spectres et les concentrations d'espèces chimiques pures constituant des échantillons biologiques. L'analyse et le diagnostic des échantillons en sont facilités.

La Factorisation en Matrices Non-négatives (FMN) a séparé des spectres de fluorescence enregistrés à la surface de grains de blé et d'orge. La structure biologique ainsi révélée de ces grains a montré que l'acide férulique est un indicateur direct de la qualité meunière d'une farine. En spectroscopie Raman, l'Analyse en Composantes Indépendantes (ACI) a été appliquée avec succès au déparaffinage numérique d'échantillons d'épiderme de peau paraffinée fixée sur un support de fluorine. Bien que noyé parmi les pics intenses de la paraffine et de la fluorine, le spectre de la peau a été très bien estimé. La comparaison entre les spectres ainsi estimés de mélanomes et de nævi a mis en évidence l'existence de descripteurs moléculaires du type de pathologie étudiée. L'association de la spectroscopie Raman et de l'ACI s'est ainsi révélée un outil performant d'analyse et d'aide au diagnostic pour le cancer de la peau.

Le premier chapitre décrit les spectroscopies optiques qui sont l'absorption infrarouge, l'émission de fluorescence et la diffusion Raman. Les spectroscopies infrarouge et Raman sont complémentaires et mettent en jeu les transitions vibrationnelles et rotationnelles des molécules. Des règles de sélection, basées sur la polarisation et la polarisabilité des molécules étudiées, et des règles de symétries sur les molécules régissent l'activité ou l'inactivité Raman et infrarouge. La spectroscopie de fluorescence mesure les énergies électroniques des molécules. Des conditions sur le spin et les symétries de la molécule régissent l'émission de fluorescence. Ces effets sont mesurés par une chaîne d'acquisition, composée principalement d'un laser, de jeux de filtres, d'un spectromètre et d'un détecteur, qui fournit à l'utilisateur des spectres informatifs sur la nature microscopique des échantillons analysés. La puissance d'analyse des techniques spectroscopiques a popularisé leurs nombreuses applications en industrie, en environnemeent et en médecine.

Dans le deuxième chapitre, nous avons étudié les propriétés physiques des spectres Raman et de fluorescence pour en déduire une modélisation instantanée et linéaire commune. L'objectif de ces deux spectroscopies est d'estimer les spectres des espèces chimiques pures ainsi que leurs concentrations à par-

tir de spectres mesurés qui sont un mélange des spectres des espèces chimiques pures. Dans un premier temps, les imperfections de l'acquisition et les phénomènes parasites sont éliminés par des prétraitements spécifiques à chaque spectroscopie. Nous avons ensuite exploité les propriétés structurelles différentes des spectres Raman et de fluorescence pour décrire les techniques numériques couramment utilisées pour traiter ces spectres. L'Analyse en Composantes Principales et les méthodes par enveloppes sont particulièrement bien adaptées aux formes à variations lentes des spectres de fluorescence. Par contre la structure parcimonieuse des spectres Raman suggère l'utilisation de méthodes de type Band-Target Entropy Minimization (BTEM) ou SIMPLe to use Interactive Self-modeling Mixture Analysis (SIMPLISMA). Pour toutes ces techniques, les inconvénients liés au manque de précision, d'autonomie et de liberté nous ont conduit à proposer des algorithmes beaucoup plus généraux pour le traitement des spectres Raman et de fluorescence.

Dans le troisième chapitre, la modélisation linéaire et instantanée des spectres de fluorescence a été couplée aux propriétés naturelles de positivité des spectres et des concentrations des espèces chimiques pures composant un échantillon. Nous avons ensuite étudié la Factorisation en Matrices Non-négatives (FMN) comme une méthode générale de séparation des spectres de fluorescence. Nous avons décrit les algorithmes basés sur les minimisations de la distance euclidienne, de la divergence, de la distance euclidienne sous contrainte de parcimonie et de la divergence sous contrainte de localisation spatiale des sources. Les schémas d'optimisation retenus par ces méthodes s'expriment sous forme de règles de mises à jour multiplicatives des matrices des spectres purs et des concentrations. La simplicité de programmation de ces méthodes facilite leur application dans de nombreux domaines. Cependant, ces algorithmes sont localement convergents et l'application répétée de ces méthodes sur un jeu de données pour différentes conditions initiales est nécessaire afin d'assurer la convergence de l'algorithme vers une solution pertinente. Nous avons ensuite appliqué ces méthodes sur des spectres de fluorescence enregistrés sur des grains de céréales pour étudier la composition chimique de leurs structures biologiques. Les méthodes basées sur les minimisations de la distance euclidienne et de la divergence se sont montrées les plus efficaces grâce à leurs hypothèses non restrictives et par l'absence de paramètres à régler. Les sources estimées par ces méthodes à partir d'enregistrements de spectres de fluorescence sur un grain de blé ont été attribuées aux acides férulique libre, para-coumarique et férulique lié, qui sont les acides phénoliques les plus auto-fluorescents du grain de blé. Grâce à l'étude des profils de concentrations estimés, nous avons localisé l'acide férulique libre dans la couche à aleurone, l'acide para-coumarique dans la pliure du grain et l'acide férulique lié dans la couche à aleurone et dans la zone entourant la pliure du grain. Sur un grain d'orge, les spectres hybrides de la lignine, de l'acide férulique et de la cutine ont été estimés par les méthodes de FMN. Nous avons respectivement localisé chacune de ces espèces dans le péricarpe et la couche cireuse supérieure, dans la couche à aleurone et les glumelles, et dans la partie supérieure de la couche cireuse. L'étude structurelle de grains de céréales est ainsi facilitée par l'association de la spectroscopie de fluorescence, qui fournit des informations mélangées sur les transitions électroniques des espèces constitutives d'un échantillon, et les méthodes de FMN, qui s'appuient sur les propriétés physiques des spectres pour estimer des spectres sources et des concentrations réalistes. Cette étude a surtout permis

de montrer que l'acide férulique est un indicateur efficace de la contamination d'une farine par les sons puisqu'il est exclusivement concentré dans la couche à aleurone qui est connue pour être un indicateur de la qualité meunière d'une farine.

Dans le dernier chapitre, nous avons traité des spectres Raman provenant d'échantillons biologiques paraffinés par des méthodes d'Analyse en Composantes Indépendantes (ACI). Les problèmes liés au déparaffinage chimique, étape nécessaire à l'analyse des échantillons paraffinés, peuvent biaiser l'analyse d'un tissu. Afin de surmonter ces difficultés, nous avons proposé une méthode de déparaffinage numérique des échantillons. Cette méthode utilise l'enregistrement des informations vibrationnelles de l'échantillon paraffiné grâce à la spectroscopie Raman. De plus, l'échantillon est fixé sur un support de fluorine. La paraffine et la fluorine sont des espèces chimiques très actives en Raman et le spectre sous-jacent du tissu à analyser est noyé dans la masse d'informations provenant des deux autres espèces chimiques. L'utilisation d'une méthode de traitement du signal est nécessaire pour séparer les spectres de ces différentes espèces à partir des spectres Raman enregistrés sur des échantillons paraffinés. Dans ce but, nous avons étudié l'Analyse en Composantes Indépendantes (ACI), ses hypothèses, ses algorithmes et ses applications classiques. L'indépendance statistique mutuelle des sources recherchées par l'ACI est l'hypothèse fondamentale de ce concept. Or, les spectres Raman de la paraffine et de la fluorine ont été prouvés comme étant parcimonieux et décorrélés. Ces propriétés sont suffisantes pour supposer que ces deux spectres sont mutuellement indépendants. Le spectre du tissu sous-jacent a été supposé comme aléatoire et chimiquement indépendant de la fluorine et de la paraffine. Il est donc considéré comme indépendant de la fluorine et de la paraffine. Les différentes hypothèses nécessaires à l'utilisation des techniques d'ACI étant validées, l'ACI est applicable aux spectres Raman provenant d'échantillons paraffinés. Cependant, nous avons montré que les imperfections de l'acquisition et la nature des échantillons requiert l'ajout au modèle des signaux enregistrés de trois sous-espaces bruit et d'une non-linéarité devant être éliminés. Le premier bruit a pour origines le courant noir, la réponse non-linéaire du système, les rayons cosmiques et la déviation en nombre d'onde. Il est éliminé par des prétraitements spécifiques. Le deuxième exprime l'émission de fluorescence par l'échantillon qui s'ajoute au spectre Raman. Sa modélisation se fait par un polynôme qu'il faut estimer et soustraire du spectre Raman. Nous avons utilisé une procédure efficace d'estimation des coefficients de ce polynôme basée sur une fonction quadratique tronquée qui écarte de l'analyse les pics Raman intenses. La non-linéarité se traduit par des décalages en nombre d'onde des pics Raman entre différents spectres enregistrés. Pour l'éliminer, nous avons proposé une procédure de recalage employée en géophysique et qui s'appuie sur le suréchantillonnage et le calcul de l'intercorrélation entre un spectre de référence et un spectre à recaler. Le troisième sous-espace est composé du bruit blanc gaussien des mesures et des spectres Raman d'autres espèces chimiques. Ce sous-espace est supposé orthogonal au sous-espace utile composé des influences des spectres de la paraffine, de la fluorine et du tissu. Son élimination est facilement réalisée par une ACP appliquée sur les données nettoyées des deux premiers sous-espaces de bruit et de la non-linéarité, et préalablement centrées et réduites. Le sous-espace résultant est ainsi séparable par une ACI. Cette prodécure de déparaffinage numérique a été appliquée sur des échantillons de peau afin d'en estimer le spectre Raman. Nous avons montré que le spectre de la

paraffine se décompose en trois spectres indépendants, contrairement à sa modélisation classique. Nous avons confirmé ce résultat par une ACI appliquée sur des spectres enregistrés sur des blocs de paraffine seuls fixés sur un support de fluorine. L'efficacité de la procédure nous a amené à l'étendre à la discrimination précoce entre un mélanome et un nævus. Le déparaffinage numérique permet d'isoler les spectres d'épidermes de mélanome et de nævus. Leur analyse suggère l'existence de descripteurs moléculaires du type de pathologie étudiée. La discrimination entre un mélanome et un nævus est possible grâce à cette méthode de déparaffinage numérique.

Ce travail présente plusieurs perspectives.

Les différences spectrales entre le spectre Raman d'un mélanome et le spectre Raman d'un nævus ont permis de faire la discrimination entre ces deux pathologies. Afin d'automatiser l'identification du tissu tumoral analysé, des méthodes de classification restent à être appliquées aux spectres Raman estimés de la peau. Le diagnostic de la pathologie rencontrée sera ainsi facilitée pour le praticien.

Face à l'efficacité de la méthode de déparaffinage numérique, son application sur d'autres types de tissus paraffinés permettrait de valider cette approche et de généraliser son utilisation à tout type de tissu paraffiné analysé par spectroscopie Raman.

Les travaux présentés dans cette thèse sont basés sur le concept de la positivité des spectres et des concentrations des espèces chimiques pures. La FMN exploite fortement ces contraintes. Or les applications présentées des techniques d'ACI à des spectres Raman n'utilisent en aucun cas ces hypothèses. Nous avons testé des algorithmes d'ACI sous contraintes de positivité mais les résultats obtenus ne sont pas significativement différents de ceux estimés par les algorithmes classiques d'ACI. Il serait intéressant d'approfondir ce sujet et d'étudier l'apport possible de telles méthodes au déparaffinage numérique.

La séparation des spectres Raman d'échantillons de peau paraffinée par les méthodes d'ACI donne parfois des résultats difficilement interprétables. En effet, pour certaines acquisitions, un pic Raman n'a pas la même largeur d'un spectre à un autre à cause d'une dérive supposée de l'autofocus. Cette variation de largeur provoque une estimation faussée des spectres des espèces chimiques constitutives de l'échantillon. Il serait donc utile de développer des prétraitements capables d'harmoniser la largeur des pics Raman afin d'automatiser la procédure.

Annexes

Annexe A

Application de l'ACP sur des spectres Raman d'un mélanome

Dans le cadre du déparaffinage numérique, les spectres Raman de la paraffine, de la fluorine et du tissu sous-jacent doivent être estimés à partir de la matrice des données $\mathbf{X}$. A cette fin, avant d'appliquer des techniques de séparation de sources, il est indispensable de corriger les spectres acquis d'effets indésirables tels que la ligne de base, le décalage en fréquence des pics Raman et le bruit de mesure. Le sous-espace $\bar{\mathbf{Z}} = \bar{\mathbf{Z}}^{\mathbf{utile}} + \bar{\mathbf{Z}}^{\mathbf{autre}}$ résulte de l'élimination des sous-espaces bruit $\mathbf{B^1}$ et $\mathbf{B^2}$ du jeu de données original $\mathbf{X}$ défini par l'équation (4.12) à la page 133, du centrage et de la réduction de $\mathbf{Z}$. Les sources associées à la paraffine, à la fluorine et au mélanome sous-jacent sont regroupées dans le sous-espace $\bar{\mathbf{Z}}^{\mathbf{utile}}$. Comme expliqué à la section 4.4.5.4 à la page 143, l'ACP permet de s'affranchir du sous-espace $\bar{\mathbf{Z}}^{\mathbf{autre}}$ orthogonal à $\bar{\mathbf{Z}}^{\mathbf{utile}}$, mais aussi de faire un blanchiment des données, étape préliminaire à l'application de l'ACI comme expliqué à la section 4.3.4.2, page 121.

L'ACP estime des sources décorrélées entre elles (voir la section 2.3.5.2, page 52). Mais cette décorrélation n'est pas adaptée pour obtenir une bonne estimation des spectres des espèces chimiques présentes dans l'échantillon paraffiné de mélanome fixé sur un support en fluorine. Les courbes de la figure A.1 représentent les 5 premières sources décorrélées estimées par l'application de l'ACP sur les données $\bar{\mathbf{Z}}$.

La première composante principale sur la figure A.1(a) peut être vue comme un spectre moyen du jeu de données $\bar{Z}$. Les pics caractéristiques de cette composante reflètent la présence simultanée de la paraffine, de la fluorine et de la peau. La deuxième composante principale de la figure A.1(b) traduit la variance du jeu de données $\bar{Z}$ qui n'est pas modélisée dans la première composante principale. Tout comme cette dernière, sa forme révèle la présence de pics associés à la paraffine, à la fluorine et à la peau. Mais cette fois-ci, les pics ont des orientations différentes. Certains sont dirigés vers les intensités positives, d'autres vers les intensités négatives. Ces remarques sur la deuxième composante principale sont

valables pour les troisième, quatrième et cinquième composantes principales respectivement disponibles sur les figures A.1(c), A.1(d) et A.1(e).

Les formes des composantes principales ne sont pas physiquement interprétables comme des spectres Raman d'espèces chimiques pures. Les spectres de la paraffine, de la fluorine et de la peau ne sont pas séparés par l'ACP. Les composantes principales sont des mélanges linéaires de ces derniers.

Les profils de concentrations relatives estimés par l'ACP sont visibles sur la figure A.2. Le profil de la première composante principale est constitué de concentrations strictement positives à la vue de la figure A.2(a). Cette observation est en adéquation avec l'interprétation de cette composante comme un spectre moyen de $\bar{Z}$. Les profils dessinés sur les figures A.2(b), A.2(c), A.2(d) et A.2(e) exhibent des concentrations des composantes principales 2, 3, 4 et 5 alternativement positives et négatives.

Les coefficients qui composent les courbes de la figure A.2 ne sont pas physiquement interprétables comme des concentrations. L'application de l'ACP seule n'est pas suffisante pour proposer l'estimation physiquement réaliste des spectres de la paraffine, de la fluorine et de la peau. L'ACI va permettre de pallier ces problèmes en recherchant des composantes décorrélées, mais aussi indépendantes à l'ordre 4.

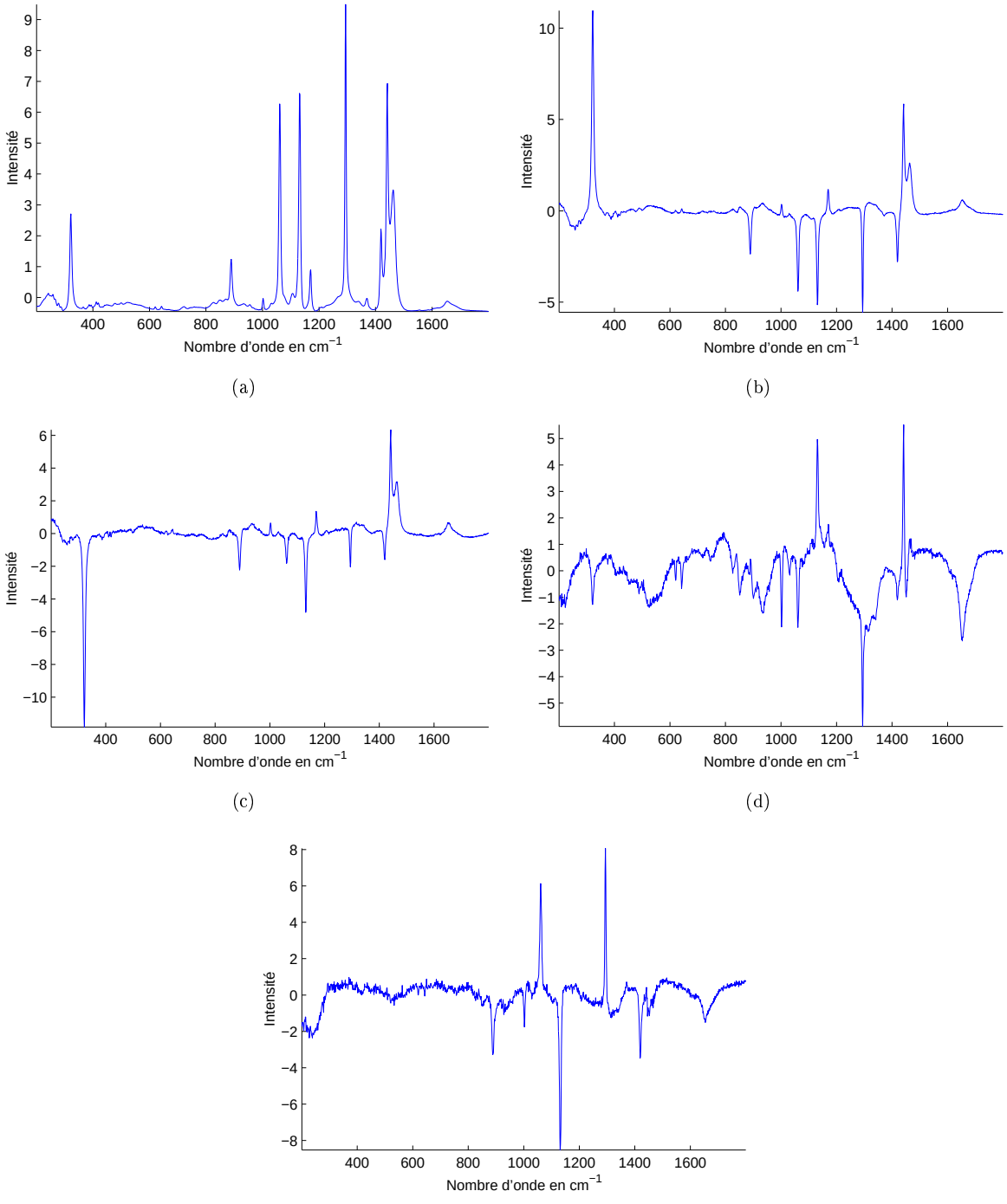


FIG. A.1 – Sources estimées sur un mélanome par ACP : (a) première source, (b) deuxième source, (c) troisième source, (d) quatrième source, (e) cinquième source

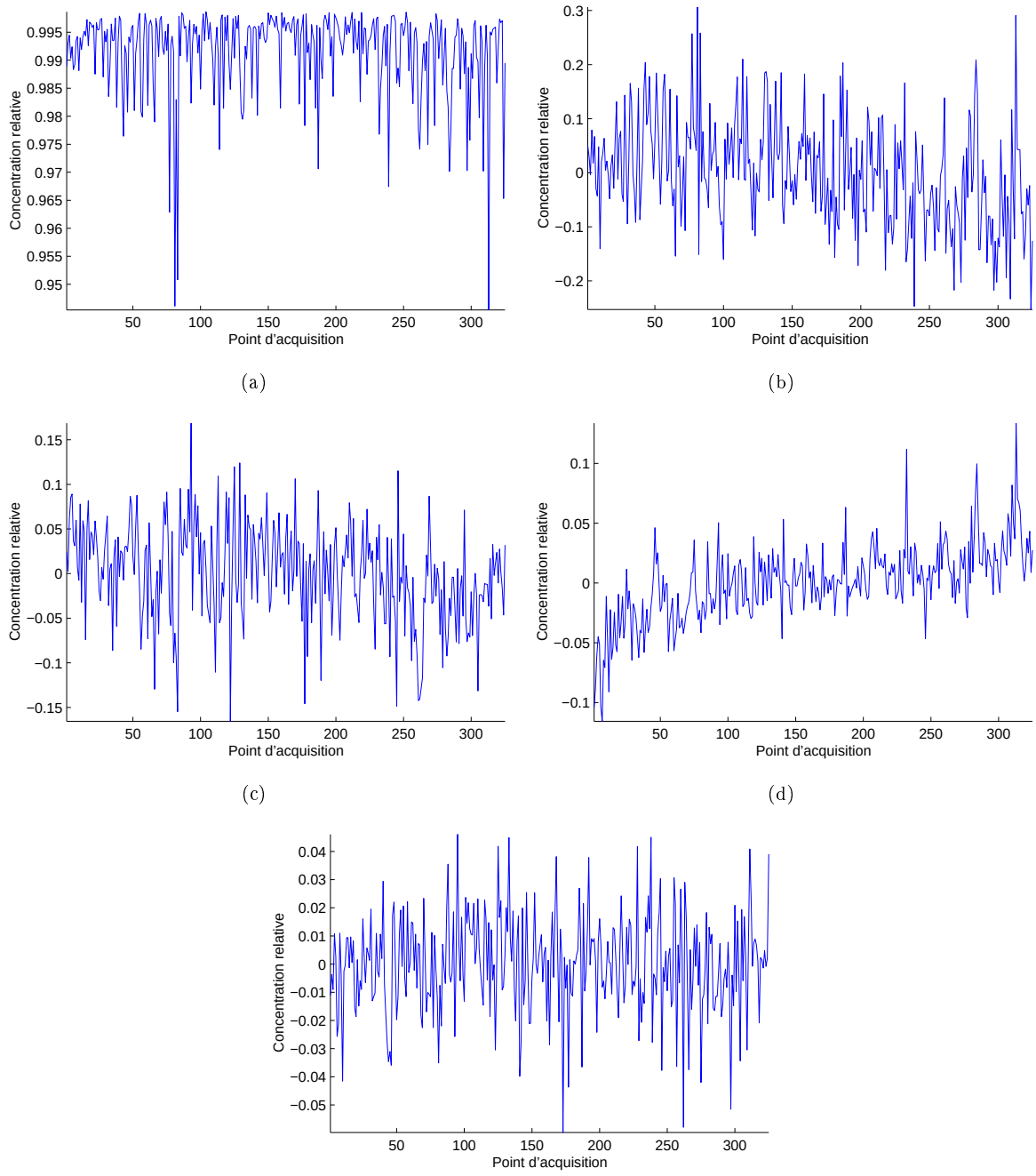


FIG. A.2 – Profils de concentrations estimés sur un mélanome par ACP : (a) premier profil, (b) deuxième profil, (c) troisième profil, (d) quatrième profil, (e) cinquième profil

Annexe B

Application de l'ACI sur des spectres Raman non recalés

L'application des techniques d'ACI sur le jeu de données $\mathbf{Z}$ est précédée par une procédure de recalage des pics Raman exposée à la section 4.4.5.2, page 140. L'utilité de cette étape est justifiée par le fait que des pics décalés pour différents spectres de la matrice $\mathbf{Z}$ entraînent l'apparition de pics artéfactuels dans les sources estimées par l'ACI. Ces pics artéfactuels s'expriment par des formes de dérivées de pics pour traduire la présence simultanée de deux pics voisins. Ces formes sont visibles sur la figure B.1 où les courbes rouges (respectivement bleues) sont les sources estimées par ACI après (respectivement sans) application de la procédure d'alignement des pics.

Le recalage des pics, préliminaire à une ACI, améliore la qualité des spectres estimés en supprimant l'apparition des pics artéfactuels.

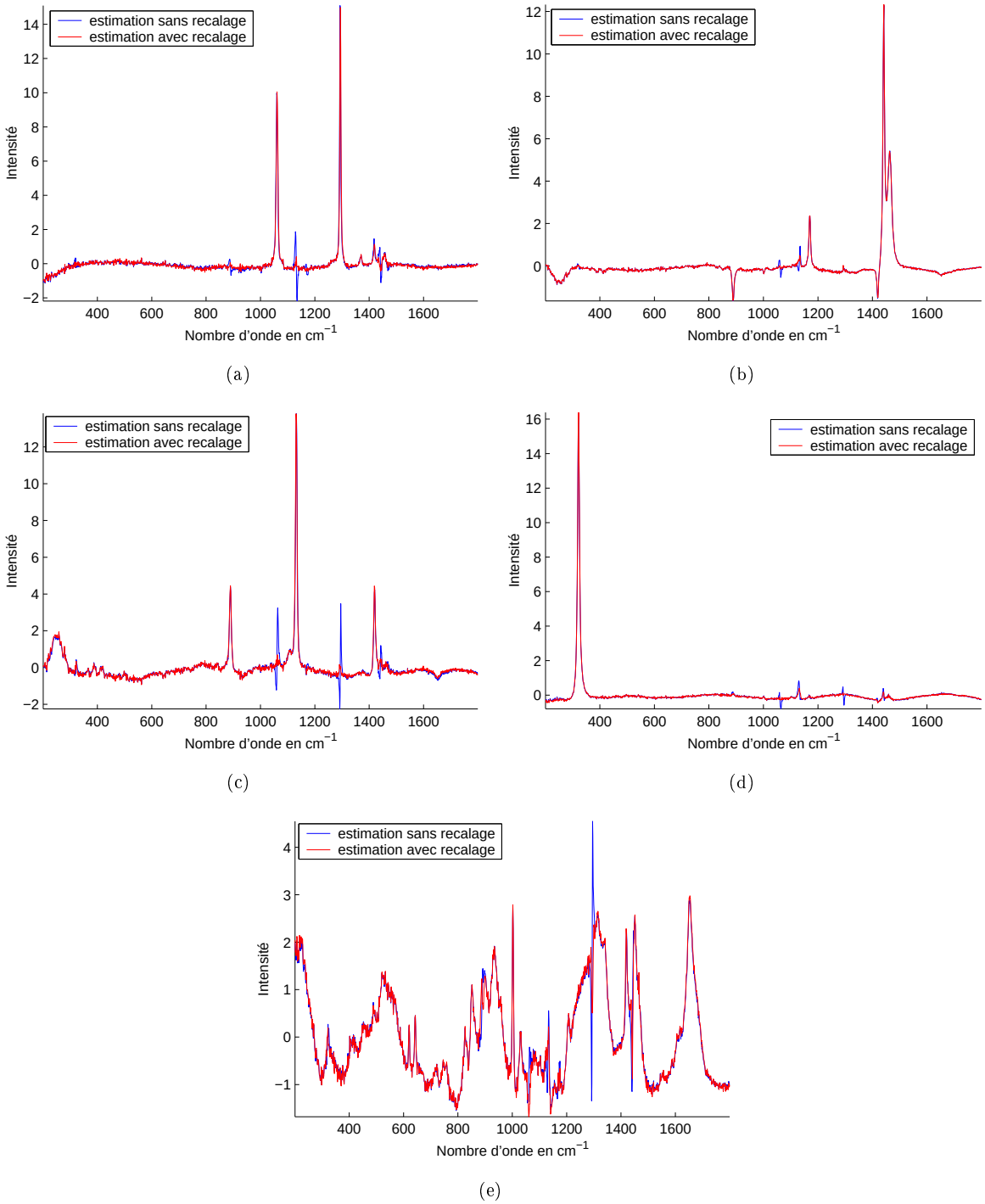


FIG. B.1 – Comparaison entres les sources estimées sur un mélanome par ACI sans (en bleu) ou avec (en rouge) procédure préalable de recalage des pics : (a) première source, (b) deuxième source, (c) troisième source, (d) quatrième source, (e) cinquième source profil

Annexe C

Application de l'ACI des spectres Raman de 3 mélanomes et 3 nævi

La méthodologie de déparaffinage numérique d'échantillons biologiques présentée à la section 4.4, page 129, a été appliquée sur des spectres Raman enregistrés sur des échantillons d'épidermes de 3 mélanomes et de 3 nævi.

Les figures C.1, C.2 et C.3 présentent les sources estimées sur les trois mélanomes.

Pour les mélanomes n°1 et n°2, les spectres sont enregistrés pour les nombres d'ondes allant de 200.413 cm^{-1} à 1799.58 cm^{-1}. Les 5 spectres estimés sont similaires pour ces deux mélanomes. Les 3 sources de la paraffine sont obtenues, le spectre de la fluorine est séparé, et le spectre du mélanome est estimé.

Les spectres du mélanome n°3 s'étendent de 650.474 cm^{-1} à 1820.42 cm^{-1}. Ainsi, la fluorine n'influence pas les spectres de ce mélanome puisque son spectre se compose d'un pic unique centré en 325 cm^{-1}. Le modèle à estimer se compose donc normalement de 3 sources de paraffine et une source du mélanome. Or pour un modèle à 4 sources, le spectre du mélanome n'est pas estimé. Seul un modèle à 5 sources conduit à son estimation. Une source traduit la présence du mélanome, 3 sources expliquent la paraffine, et la cinquième source ne correspond pas à une espèce chimique connue. Elle est interprétée comme une source artéfactuelle. Son origine provient probablement de la différence des largeurs des pics Raman, donc à un problème d'acquisition.

Les figures C.4, C.5 et C.6 présentent les sources estimées sur les trois nævi.

Pour chaque tissu, les mêmes sources sont estimées. Trois sources sont affectées à la décomposition du spectre de la paraffine, une source modélise le spectre de la fluorine, et une source estime le spectre du nævus.

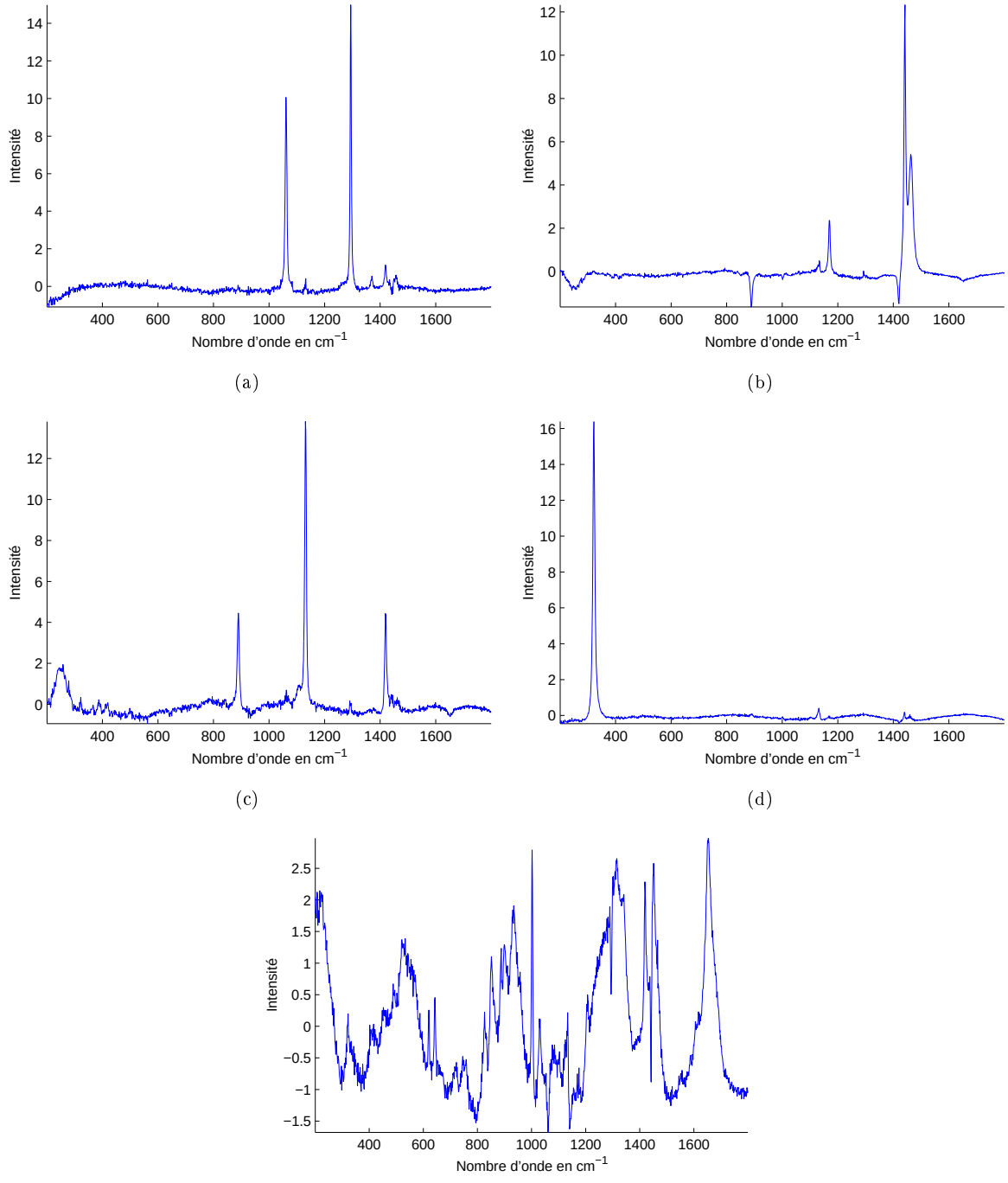


FIG. C.1 – Sources estimées sur le mélanome n°1 par JADE pour un modèle à 5 sources : (a) première source, (b) deuxième source, (c) troisième source, (d) quatrième source, (e) cinquième source

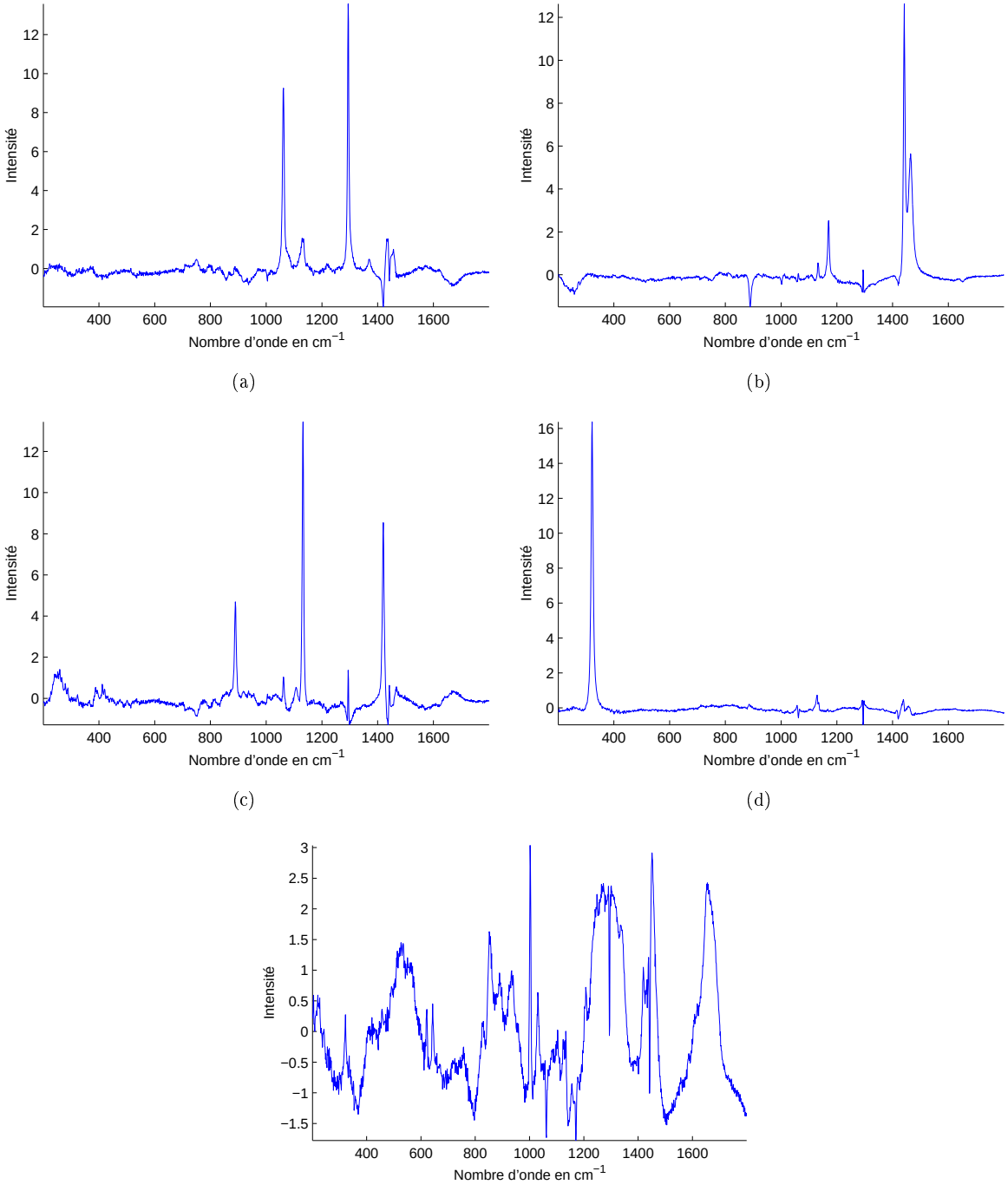


FIG. C.2 – Sources estimées sur le mélanome n°2 par JADE pour un modèle à 5 sources : (a) première source, (b) deuxième source, (c) troisième source, (d) quatrième source, (e) cinquième source

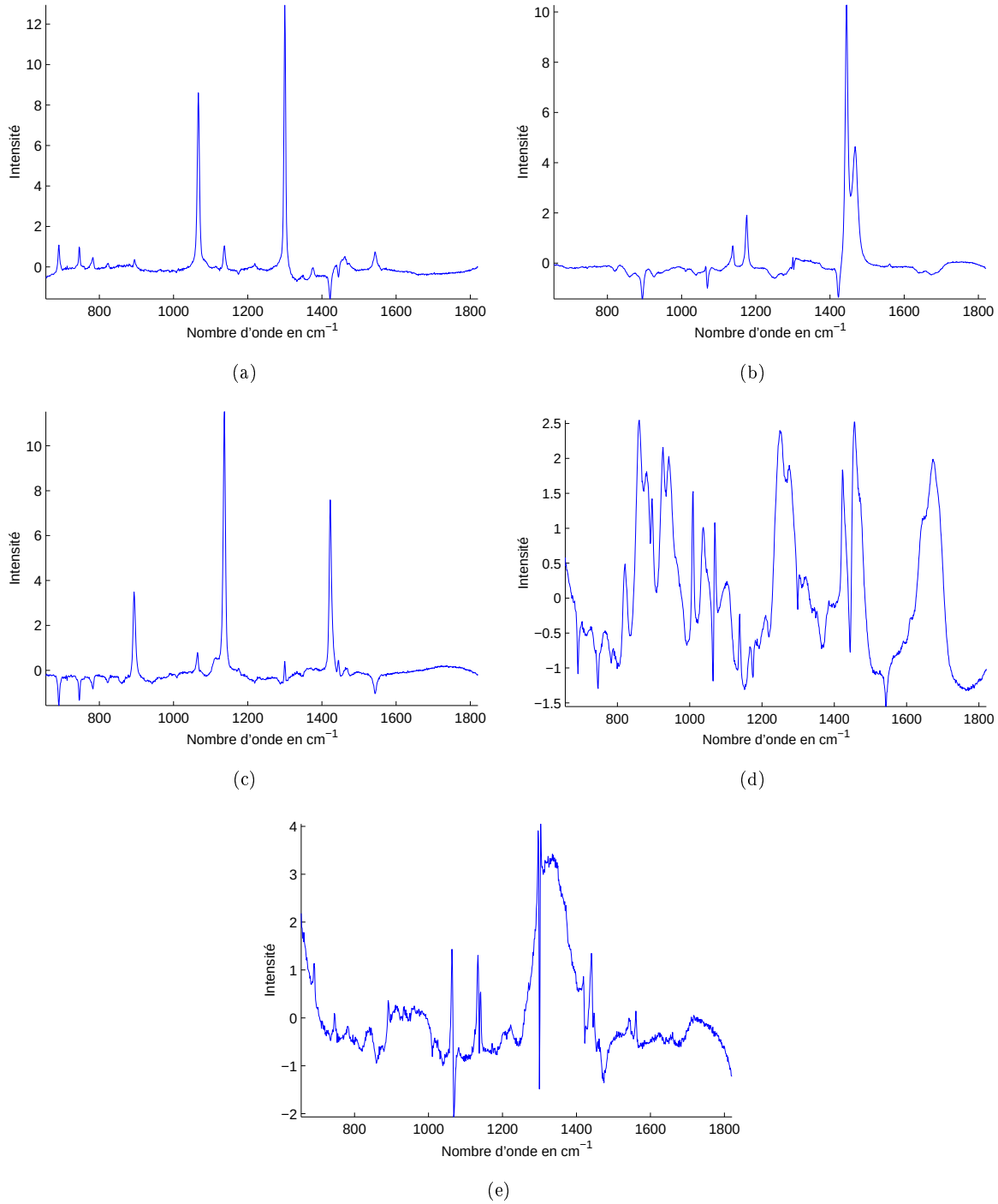


FIG. C.3 – Sources estimées sur le mélanome n°3 par JADE pour un modèle à 5 sources : (a) première source, (b) deuxième source, (c) troisième source, (d) quatrième source, (e) cinquième source

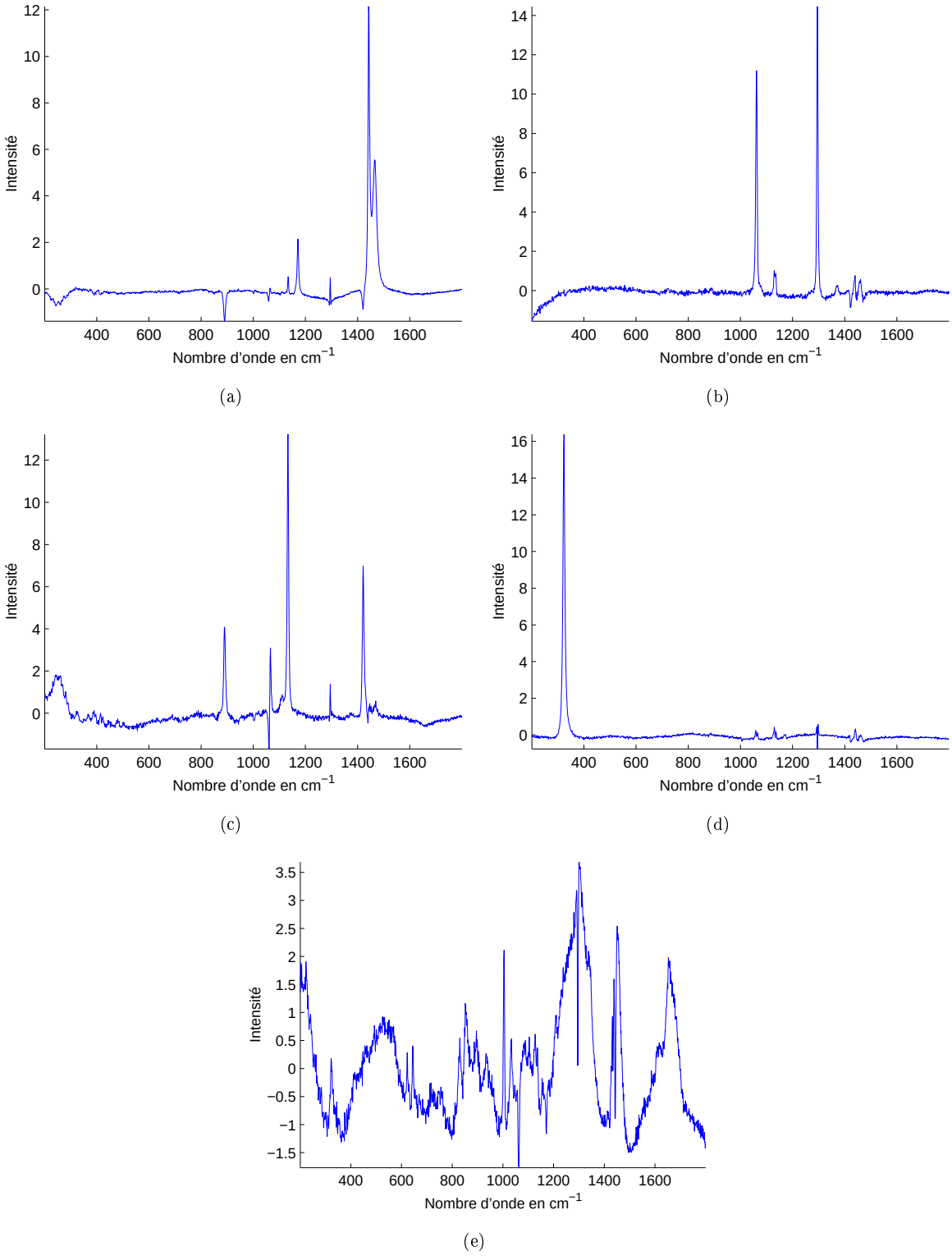


FIG. C.4 – Sources estimées sur le nævus n°1 par JADE pour un modèle à 5 sources : (a) première source, (b) deuxième source, (c) troisième source, (d) quatrième source, (e) cinquième source

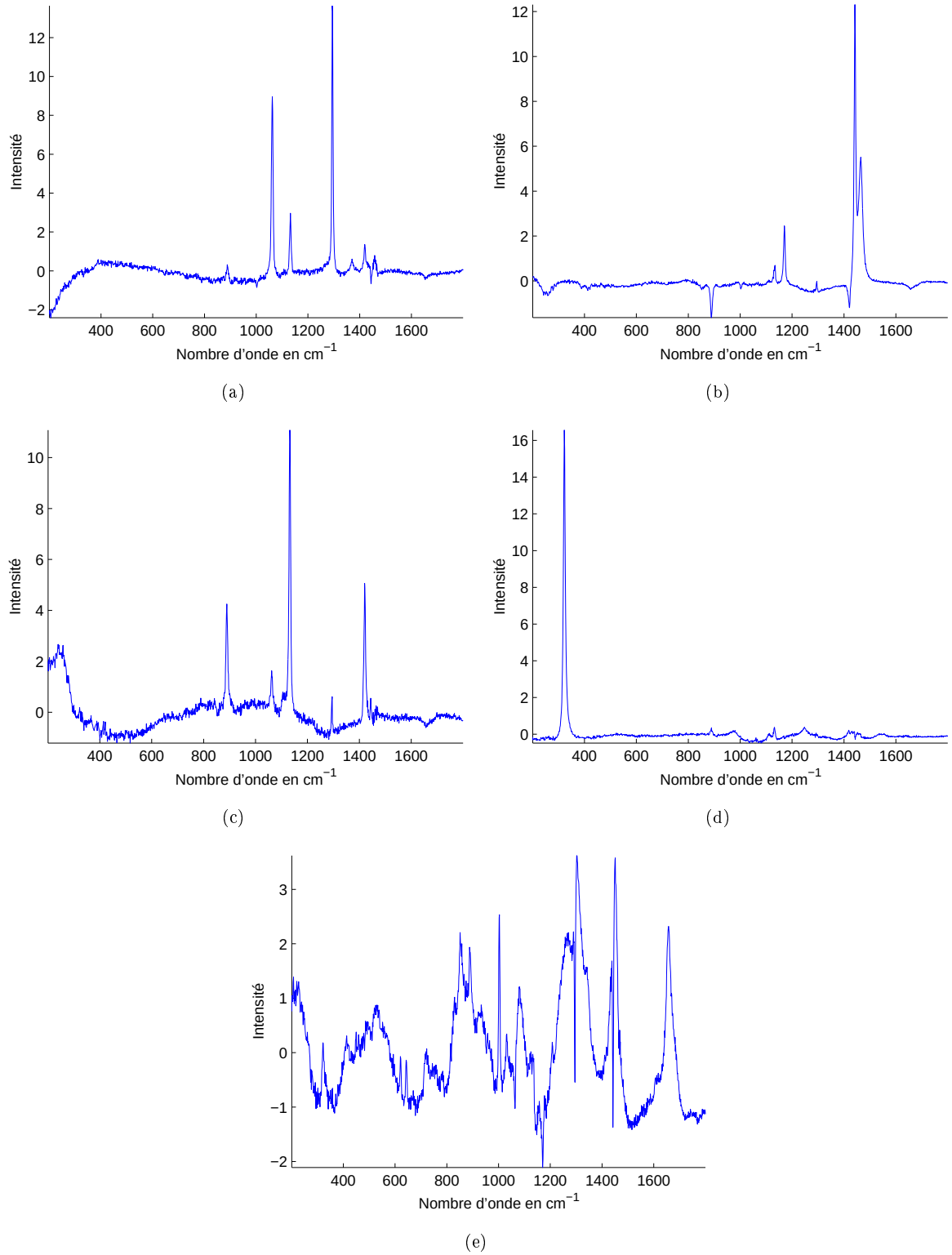


FIG. C.5 – Sources estimées sur le nævus n°2 par JADE pour un modèle à 5 sources : (a) première source, (b) deuxième source, (c) troisième source, (d) quatrième source, (e) cinquième source

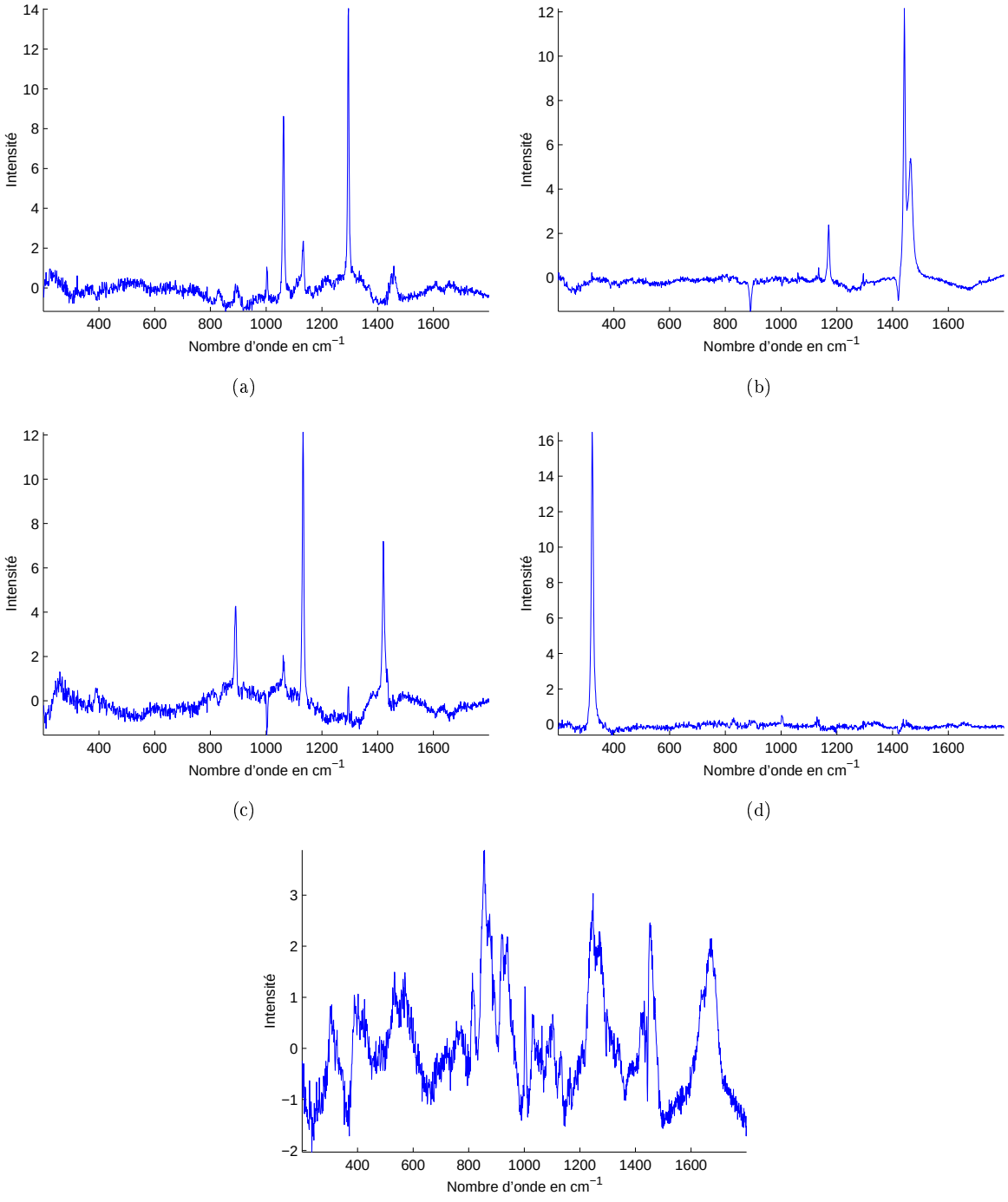


FIG. C.6 – Sources estimées sur le nævus n°3 par JADE pour un modèle à 5 sources : (a) première source, (b) deuxième source, (c) troisième source, (d) quatrième source, (e) cinquième source

Annexe D

Application de FastICA sur des spectres Raman de peau paraffinée

Les spectres Raman acquis sur un échantillon de peau paraffinée fixé sur un support de fluorine ont été traités par FastICA, algorithme d'ACI décrit à la section 4.3.5.1. Les 5 sources estimées sur le mélanome n°1 sont présentées sur la figure D.1 et sont similaires à celles estimées par JADE et présentées sur la figure 4.16 à la page 153 et sur la figure C.1 à la page 176. Quelque soit la non-linéarité choisie, l'algorithme converge vers des solutions équivalentes.

L'algorithme de diagonalisation maximale proposé par Comon dans [25] donne lui aussi des résultats similaires qui ne sont pas présentés.

La décomposition de la paraffine en trois sources est à nouveau confirmée et prouve l'indépendance réelle de ces sources puisque trois algorithmes différents basés sur trois mesures d'indépendance différentes fournissent les mêmes résultats.

Le spectre estimé de la peau présente les mêmes caractéristiques spectrales quelque soit la méthode d'ACI choisie. La discrimination entre un mélanome et un nævus reste possible et ne dépend pas de l'algorithme d'ACI.

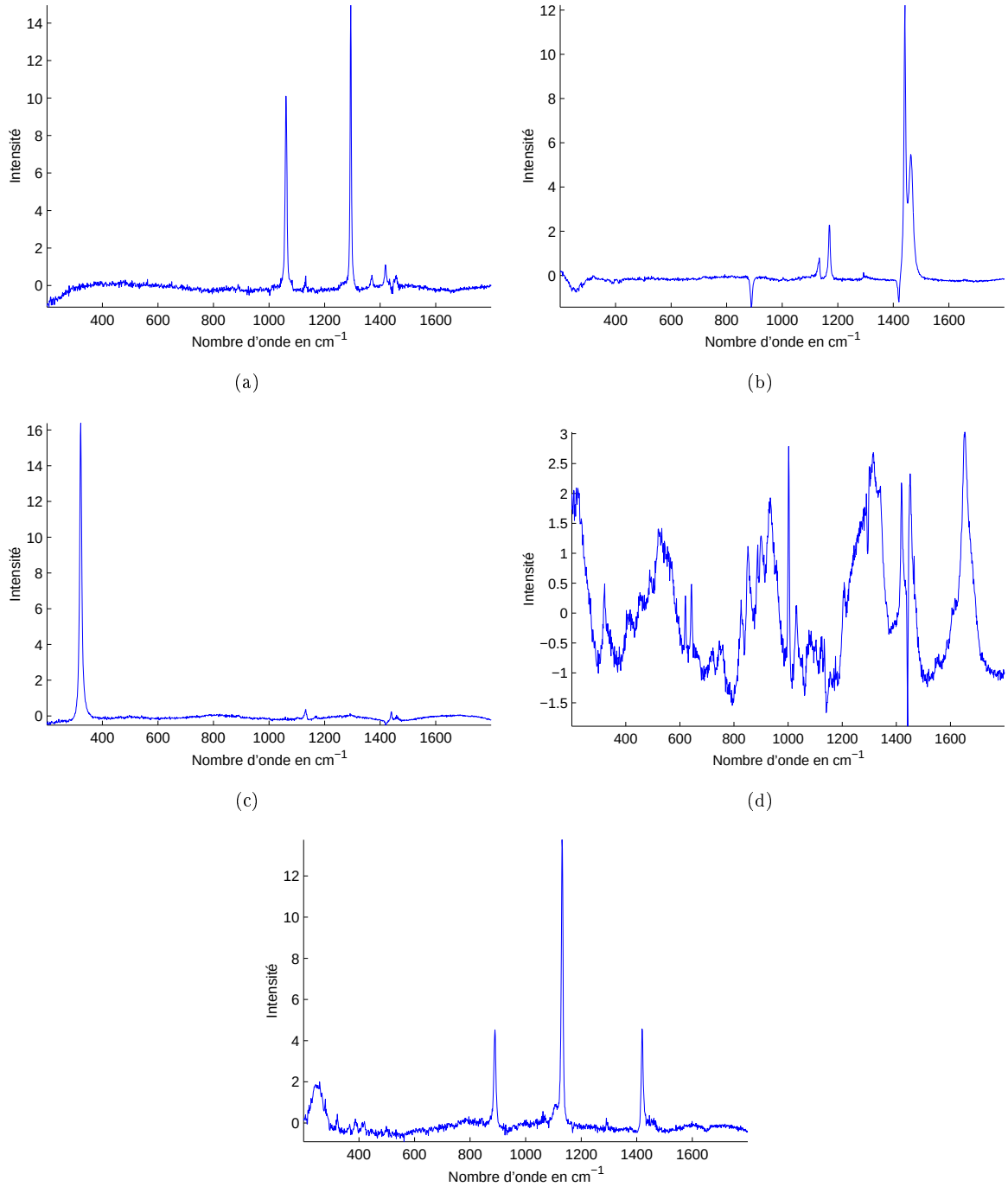


FIG. D.1 – Sources estimées sur un mélanome par FastICA : (a) première source, (b) deuxième source, (c) troisième source, (d) quatrième source, (e) cinquième source

Annexe E

Application de la FMN sur des spectres Raman de peau paraffinée

Les spectres Raman acquis sur un échantillon de peau paraffinée fixée sur un support de fluorine ont été séparés par les méthodes de FMN présentées à la section 3.4.2. La figure E.1 présente les sources estimées par l'algorithme de minimisation de la distance euclidienne pour un modèle à 5 sources. Les spectres Raman de la fluorine, de la paraffine et de la peau sont encore mélangés. Divers estimations ont été faites par les algorithmes de FMN pour plusieurs nombres de sources sous-jacentes. Aucun résultat pertinent n'a été estimé. L'hypothèse de positivité des spectres Raman et des concentrations des espèces n'est pas suffisante pour assurer une estimation efficace des spectres de la paraffine, de la fluorine et de la peau par la FMN.

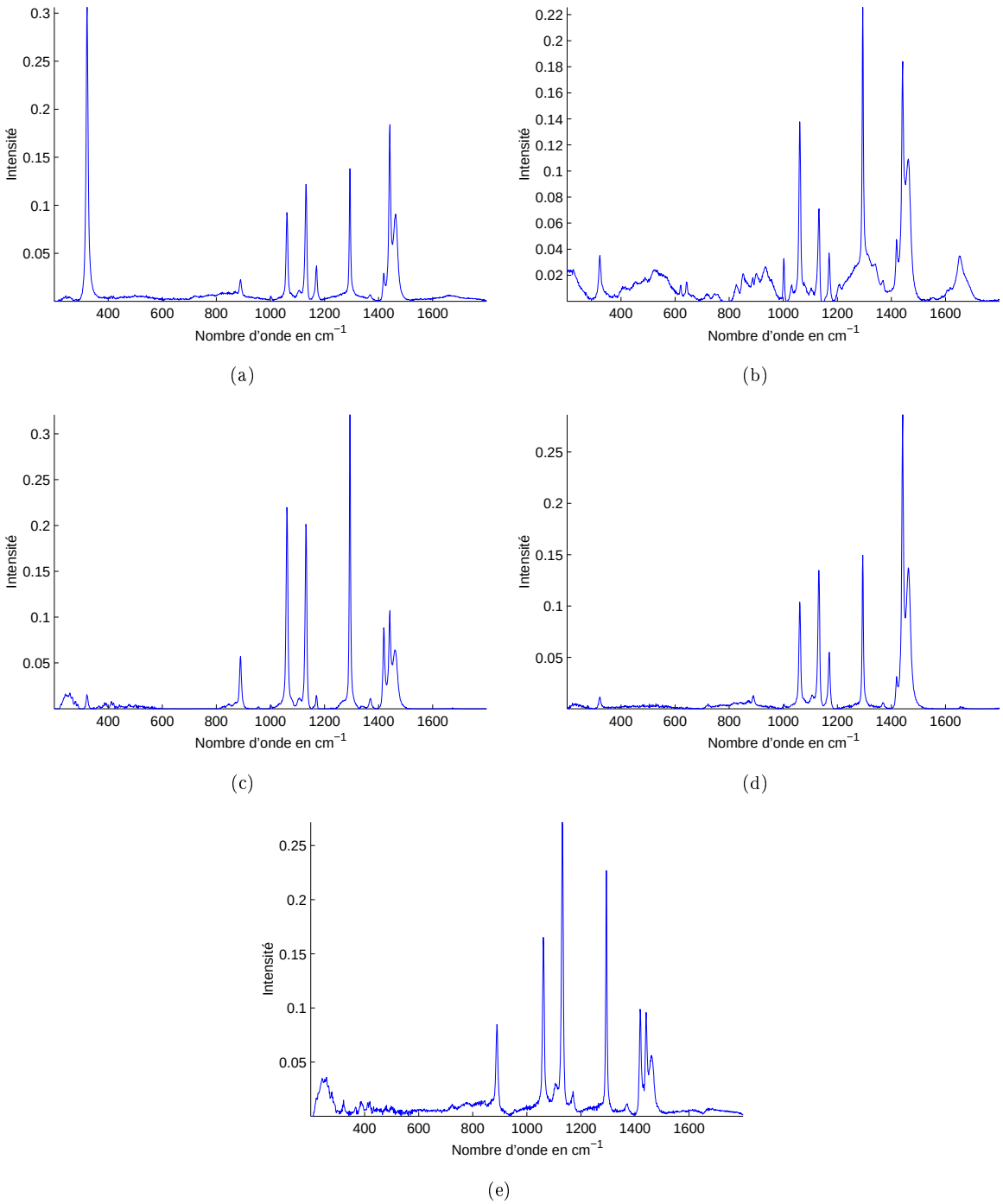
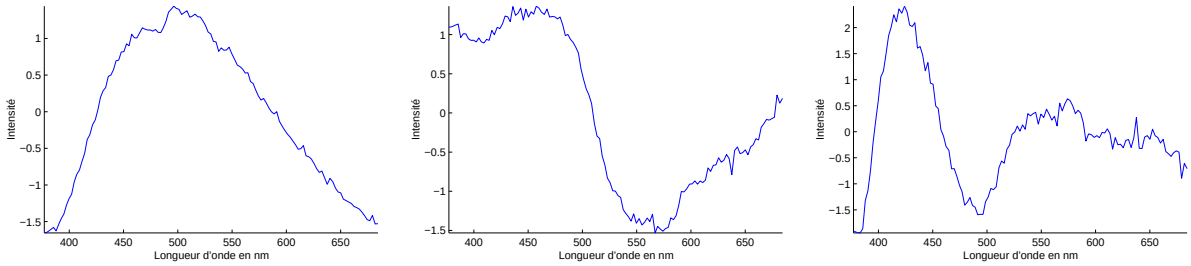


FIG. E.1 – Sources estimées sur des spectres Raman d'un échantillon paraffiné de peau fixé sur un support de fluorine par FMN pour un modèle à 5 sources : (a) première source, (b) deuxième source, (c) troisième source, (d) quatrième source, (e) cinquième source

Annexe F

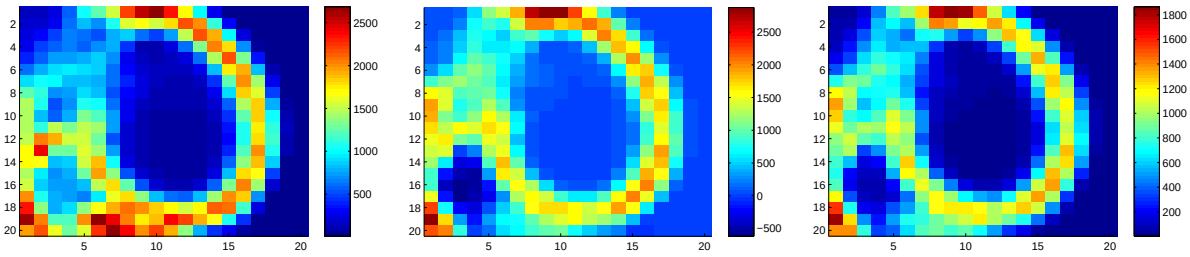
Application de l'ACI sur des spectres de fluorescence de grains de blé

Afin de montrer la spécificité de l'ACI à la spectroscopie Raman, les spectres de fluorescence acquis sur un grain de blé et présentés à la section 3.5 ont été séparés par l'ACI. Nous avons considéré dans un premier temps le problème direct $\mathbf{X} = \mathbf{AS}$.



(a) (b) (c)

FIG. F.1 – Sources estimées sur des spectres de fluorescence d'un grain de blé par JADE pour un modèle à 3 sources : (a) première source, (b) deuxième source, (c) troisième source



(a) (b) (c)

FIG. F.2 – Profils de concentrations estimées sur des spectres de fluorescence d'un grain de blé par JADE pour un modèle à 3 sources : (a) premier profil, (b) deuxième profil, (c) troisième profil

Les spectres des acides du grain de blé (les lignes de $\mathbf{S}$) sont donc supposés indépendants. Les sources et les profils de concentration estimés sont présentés respectivement sur les figures F.1 et F.2.

Les sources sont mal estimées puisqu'aucune ne ressemble aux spectres de référence des acides férulique libre, férulique lié et para-coumarique présentés sur la figure 3.3.

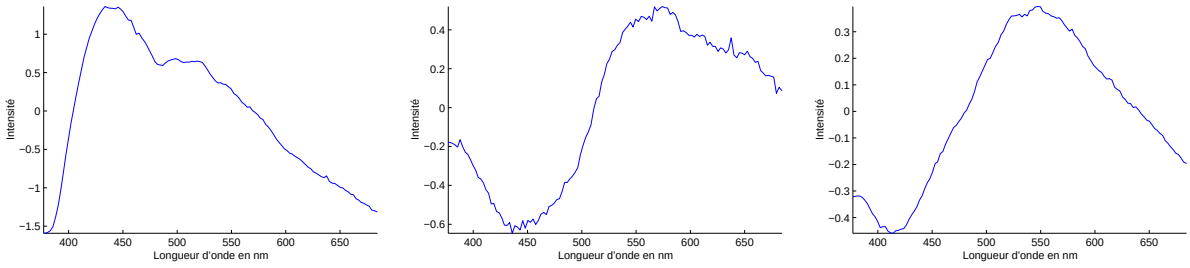
Les profils de concentration de chaque source sont fortement corrélés et ne traduisent pas la localisation spatiale de chaque espèce chimique dans une structure spécifique du grain de blé comme prouvé dans [113] et discuté au paragraphe **Discussion** à la page 105.

Les profils de concentrations réels des acides auto-fluorescents présents dans le grain de blé sont décorrélés d'après les études dans [113]. L'indépendance des profils de concentrations est donc plus probable que celle des spectres des acides. Nous avons donc étudié le problème transposé $\mathbf{X^T} = \mathbf{S^T A^T}$. Les résultats de l'application de JADE sur $\mathbf{X^T}$ sont proposés sur les figures F.3 et F.4.

Les profils de concentrations estimés montrent un partitionnement du grain de blé en trois régions indépendantes. La localisation des espèces estimées est obtenue.

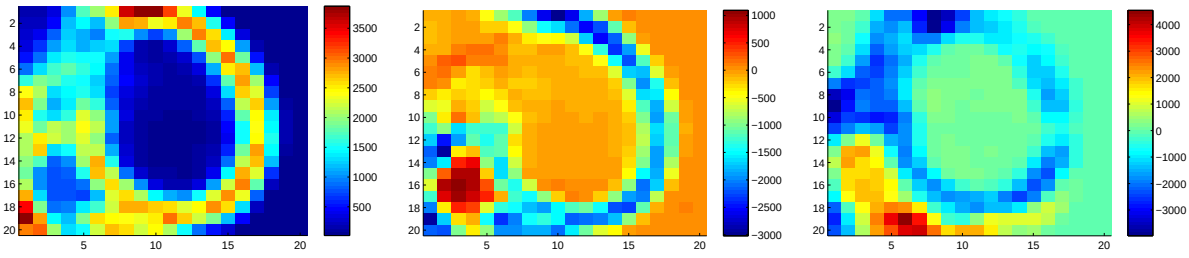
Les spectres estimés ne correspondent pas aux spectres réels des acides du grain de blé.

L'indépendance des spectres de fluorescence des acides phénoliques du grain de blé ou de leurs profils de concentrations ne sont pas des hypothèses satisfaisantes pour modéliser les spectres de fluorescence.



(a) (b) (c)

FIG. F.3 – Sources estimées sur des spectres de fluorescence d'un grain de blé par JADE pour un modèle transposé à 3 sources : (a) première source, (b) deuxième source, (c) troisième source



(a) (b) (c)

FIG. F.4 – Profils de concentrations estimées sur des spectres de fluorescence d'un grain de blé par JADE pour un modèle transposé à 3 sources : (a) premier profil, (b) deuxième profil, (c) troisième profil

Bibliographie

[1] Albani, J. R. *Absorption et fluorescence : principes et applications.* Tek & Doc, 2001.

[2] Andrus, P. G. L., and Strickland, R. D. Cancer grading by Fourier Transform Infrared spectroscopy. *Biospectroscopy 4*, 1 (1998), 37–46.

[3] Argov, S., Ramesh, J., Salman, A., Sinelnikov, I., Goldstein, J., Guterman, H., and Mordechai, S. Diagnostic potential of Fourier-transform infrared microspectroscopy and advanced computational methods in colon cancer patients. *Journal of Biomedical Optics 7*, 2 (2002), 248–254.

[4] Back, A. D., and Weigend, A. S. A first application of independent component analysis to extracting structure from stock returns. *International Journal of Neural Systems 8*, 4 (1997), 473–484.

[5] Bakker Shut, T. C., Witjes, M. J. H., Sterenborg, H. J. C. M., Speelman, O. C., Roodenburg, J. L. N., Marple, E. T., Bruining, H. A., and Puppels, G. J. In vivo detection of dysplastic tissue by Raman spectroscopy. *Analytical Chemistry 72* (2000), 6010–6018.

[6] Bandermann, F., Tausendfreund, I., Sasic, S., Ozaki, Y., Kleimann, M., Westerhuis, J., and Siesler, H. W. Fourier-transform Raman spectroscopic on-line-monitoring of the anionic dispersion block copolymerization of styrene and 1,3-butadiene. *Macromolecular Rapid Communications 22*, 9 (2001), 690–693.

[7] Barbillat, J., Bougeard, D., Buntinx, G., Delhaye, M., Dhamelincourt, P., and Fillaux, F. Spectrométrie Raman. *Techniques de l'Ingénieur, traité Analyse et Caractérisation P2 865* (1999), 1–31.

[8] Bartlett, M. S., Movellan, J. R., and Sejnowski, T. J. Face recognition by Independent Component Analysis. *IEEE Transactions on Neural Networks 13*, 6 (2002), 1450–1464.

[9] Beljebbar, A., Morjani, H., Sockalingum, G. D., and Manfait, M. Rapid identification of the multidrug resistance in the human leukemic cells by near infrared Fourier transform Raman microspectroscopy. *Proceedings of Infrared Spectroscopy : New tool in medecine 3257* (1998), 61–65.

[10] Bell, A. J., and Sejnowski, T. J. An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation 7*, 6 (1995), 1129–1159.

[11] Bell, A. J., and Sejnowski, T. J. The "independent components" of natural scenes are edge filters. *Vision Research 37*, 23 (1997), 3327–3338.

[12] Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., and Moulines, E. A Blind Source Separation technique using second-order statistics. *IEEE Transactions on Signal Processing 45*, 2 (1997), 434–444.

[13] Bermond, O., and Cardoso, J.-F. Méthodes de séparation de sources dans le cas sous-déterminé. *Proceedings of GRETSI'99 – 17e colloque GRETSI sur le traitement du signal et des images* (Vannes, France, 1999), 749–752.

[14] Bertrand, D., and Scotter, C. N. G. Application of multivariate analyses to NIR spectra of gelatinized starch. *Applied Spectroscopy 46* (1992), 1420–1425.

[15] Biswal, B. B., and Ulmer, J. L. Blind source separation of multiple signal sources of fMRI data sets using independent component analysis. *Journal of Computer Assisted Tomography 23*, 2 (1999), 265–271.

[16] Booksh, K. S., Muroski, A. R., and Myrick, M. L. Single measurement excitation/emission matrix spectrofluorometer for determination of hydrocarbons in ocean water. 2. Calibration and quantitation of naphtalene and styrene. *Analytical Chemistry 68* (1996), 3539–3544.

[17] Boughriet, A., Figueiredo, R. S., Laureyns, J., and Recourt, P. Identification of newly generated iron phases in recent anoxic sediments : ^{57}Fe Moessbauer and microRaman spectroscopic studies. *Journal-Chemical Society Faraday Transactions 93*, 17 (1997), 3209–3215.

[18] Brookner, C., Utzinger, U., Follen, M., Richards-Kortum, R., Cox, D., and Atkinson, E. N. Effects of biographical variables on cervical fluorescence emission spectra. *Journal of Biomedical Optics 8*, 3 (2003), 479–483.

[19] Cardoso, J.-F. Dependence, correlation and gaussianity in independent component analysis. *Journal of Machine Learning Research 4* (2003), 1177–1203.

[20] Cardoso, J.-F., and Souloumiac, A. Blind beamforming for non-Gaussian signals. *IEE ProceedingsF 140*, 46 (1993), 362–370.

[21] Caspers, P. J., Lucassen, G. W., Wolthuis, R., Bruining, H. A., and Puppels, G. J. *In vitro* and *in vivo* Raman spectroscopy of human skin. *Biospectroscopy 4*, S5 (1998), S31–S39.

[22] Castells, F., Rieta, J., Millet, J., and Zarzoso, V. Spatiotemporal blind source separation approach to atrial activity estimation in atrial tachyarrhythmias. *IEEE Transactions on Biomedical Engineering 52*, 2 (2005), 258–267.

[23] Chang, S. K., Dawood, M. Y., Staerkel, G., Utzinger, U., Atkinson, E. N., Richards-Kortum, R. R., and Follen, M. Fluorescence spectroscopy for cervical precancer detection : Is there variance across the menstrual cycle? *Journal of Biomedical Optics 7*, 4 (2002), 595–602.

[24] Chew, W., Widjaja, E., and Garland, M. Band-Target Entropy Minimization (BTEM) : An advanced method for recovering unknown pure component spectra. Application to the FTIR spectra of unstable organometallic mixtures. *Organometallics 21* (2002), 1982–1990.

[25] Comon, P. Independent component analysis–a new concept? *Signal Processing 36* (1994), 287–314.

[26] Comon, P. Tensor decompositions : state of the art and applications. *IMA Conference on Mathematics in Signal Processing* (Warwick, UK, 2000).

[27] Corana, A., Marchesi, M., Maini, C., and Ridella, S. Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm. *ACM Transactions on Mathematical Software 13*, 3 (1987), 262–280.

[28] Dalibart, M., and Servant, L. Spectroscopie dans l'infrarouge. *Techniques de l'Ingénieur, traité Analyse et Caractérisation P2 845* (2000), 1–26.

[29] de Groot, P. J., Postma, G. J., Melssen, W. J., Buydens, L. M. C., Deckert, V., and Zenobi, R. Application of principal component analysis to detect outliers and spectral deviations in near-field surface-enhanced Raman spectra. *Analytica Chimica Acta 446* (2001), 71–83.

[30] De Lathauwer, L., Moor, B. D., and Vandewalle, J. Fetal electrocardiogram extraction by blind source subspace separation. *IEEE Transactions on Biomedical Engineering 47*, 5 (2000), 567–572.

[31] de Oliveira Nunes, L., Martin, A. A., Jr., L. S., and Zampieri, M. FT-Raman spectroscopy study for skin cancer diagnosis. *Spectroscopy : An International Journal 17*, 2/3 (2003), 597–602.

[32] Delorme, A., Makeig, S., and Sejnowski, T. Automatic artifact rejection for EEG data using high-order statistics and independent component analysis. *Proceedings of ICA 2001 – the 3rd International Conference on Independent Component Analysis and Blind Source Separation* (San Diego, California, USA, 2001), 457–462.

[33] Dempster, M. M., Laird, N. M., and Jain, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society 39* (1977), 1–38.

[34] Depczynski, U., Jetter, K., Molts, K., and Niemöller, A. The fast wavelet transform on compact intervals as a tool in chemometrics : I. Mathematical background. *Chemometrics and Intelligent Laboratory Systems 39*, 1 (1997), 19–27.

[35] Dexter, J. E., and Matsuo, R. R. Effect of semolina extraction rate on semolina characteristics and spaghetti quality. *Cereal Chemistry 56* (1978), 841–852.

[36] Diamantaras, K. I., and Kung, S. Y. *Principal component neural networks : Theory and applications.* Wiley, 1996.

[37] Donoho, D., and Stodden, V. When does Non-negative Matrix Factorization give a correct decomposition into parts? *Advances in Neural Information Processing 16* (2004).

[38] Drezek, R., Sokolov, K., Utzinger, U., Boiko, I., Malpica, A., Follen, M., and Richards-Kortum, R. Understanding the contributions of NADH and collagen to cervical tissue fluorescence spectra : Modeling, measurements, and implications. *Journal of Biomedical Optics 6*, 4 (2001), 385–396.

[39] Endl, E., Kausch, I., Baack, M., Knippers, R., Gerdes, J., and Scholzen, T. The expression of Ki-67, MCM3, and p27 defines distinct subsets of proliferating, resting, and differentiated cells. *Journal of Pathology 195* (2001), 457–462.

[40] Eudes, D. Caractérisation tissulaire par micro-spectroscopie Raman et analyses statistiques multivariées : application au cancer du côlon. Mémoire ingénieur, CNAM Centre Régional de Champagne-Ardenne, 2003.

[41] Faoláin, E. O., Hunter, M. B., Byrne, J. M., Kelehan, P., Lambkin, H. A., Byrne, H. J., and Lyng, F. M. Raman spectroscopic evaluation of efficacy of current paraffin wax section dewaxing agents. *Journal of Histochemistry & Cytochemistry 53*, 1 (2005), 121–129.

[42] Farina, D., Févotte, C., Doncarli, C., and Merletti, R. Blind separation of linear instantaneous mixtures of nonstationary surface myoelectric signals. *IEEE Transactions on Biomedical Engineering 51*, 9 (2004), 1555–1567.

[43] Feng, M., and Kammayer, K.-D. Application of source separation algorithms for mobile communications environment. *Proceedings of ICA 1999 – the 1st International Conference on Independent Component Analysis and Blind Source Separation* (Aussois, France, 1999), 431–436.

[44] Gniadecka, M., Philipsen, P. A., Sigurdsson, S., Wessel, S., Nielsen, O. F., Christensen, S. H., Hercogova, J., Rossen, K., Thomsen, H. K., Gniadecka, R., Hansen, L. K., and Wulf, H. C. Melanoma diagnosis by Raman spectroscopy and neural networks : structure alterations in proteins and lipids in intact cancer tissue. *Journal of Investigative Dermatology 122*, 2 (2004), 443–449.

[45] Gniadecka, M., Wulf, H. C., Nymark-Mortensen, N., Faurskov-Nielsen, O., and Christensen, D. H. Diagnosis of basal cell carcinoma by Raman spectroscopy. *Journal of Raman Spectroscopy 28* (1997), 125–129.

[46] Gobinet, C., Elhafid, A., Vrabie, V., Huez, R., and Nuzillard, D. About importance of positivity constraint for source separation in fluorescence spectroscopy. *Proceedings of EUSIPCO 2005 – the 13th European Signal Processing Conference* (Antalya, Turquie, 2005).

[47] Gobinet, C., Perrin, E., and Huez, R. Application of Non-negative Matrix Factorization to fluorescence spectroscopy. *Proceedings of EUSIPCO 2004 – the 12th European Signal Processing Conference* (Vienne, Autriche, 2004).

[48] Gobinet, C., Tfayli, A., Piot, O., Vrabie, V., and Huez, R. Independent Component Analysis and Raman spectroscopy on paraffinised non dewaxed cutaneous biopsies : A promising methodology for melanoma early diagnosis. *Proceedings of BPC 2005 – the First International Workshop on Biosignal Processing and Classification* (Barcelone, Espagne, 2005), 19–26.

[49] Gobinet, C., Tfayli, A., Piot, O., Vrabie, V., and Huez, R. A method of digital deparaffining based on Raman spectroscopy and Independent Component Analysis - Application to melanoma early diagnosis. *IFMBE Proceedings, Vol. 11, Proceedings of the 3rd European Medical & Biological Engineering Conference - EMBEC´05* (Prague, République Tchèque, 2005), 3688–3693.

[50] Goehner, R. Background subtract subroutine for spectral data. *Analytical Chemistry 50* (1978), 1223–1225.

[51] Guimet, F., Boqué, R., and Ferré, J. Cluster analysis applied to the exploratory analysis of commercial spanish olive oils by means of excitation-emission fluorescence spectroscopy. *Journal of Agricultural and Food Chemistry 52* (2004), 6673–6679.

[52] Haka, A., Shafer-Peltier, K., Fitzmaurice, M., Crowe, J., Dasari, R., and Feld, M. Identifying microcalcifications in benign and malignant breast lesions by probing differences in their chemical composition using Raman spectroscopy. *Cancer Research 62* (2002), 5375–5380.

[53] Hammody, Z., Sahu, R. K., Mordechai, S., Cagnano, E., and Argov, S. Characterization of malignant melanoma using vibrational spectroscopy. *The Scientific World Journal 5* (2005), 173–182.

[54] Hata, T. R., Scholz, T. A., Ermakov, I. V., McClane, R. W., Khachik, F., Gellermann, W., and Pershing, L. K. Non-invasive Raman spectroscopic detection of carotenoids in human skin. *Journal of Investivative Dermatology 115*, 3 (2000), 441–448.

[55] Henry, R. C. Current factor analysis receptor models are ill-posed. *Atmospheric Environment 21* (1987), 1815–1820.

[56] Herbert, S., Riaublanc, A., Bouchet, B., Gallant, K. J., and Dufour, E. Fluorescence spectroscopy investigation of acid- or rennet-induced coagulation of milk. *Journal of Dairy Science 82* (1999), 2056–2062.

[57] Hopke, P. K. Target transformation factor analysis as an aerosol mass apportionment method : a review and sensitivity study. *Atmospheric Environment 22* (1988), 1777–1792.

[58] Hoyer, P. O. Non-negative sparse coding. *Neural Networks for Signal Processing XII* (2002), 557–565.

[59] Hoyer, P. O. Non-negative Matrix Factorization with sparseness constraints. *Journal of Machine Learning Research 5* (2004), 1457–1469.

[60] Hérault, J., Jutten, C., and Ans, B. Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. In *Proceedings of GRETSI'85 – 10^e colloque GRETSI sur le traitement du signal et des images* (Nice, France, 1985), pp. 1017–1022.

[61] Huang, Z., McWilliams, A., Lui, H., McLean, D. I., Lam, S., and Zeng, H. Near-infrared Raman spectroscopy for optical diagnosis of lung cancer. *International Journal of Cancer 107* (2003), 1047–1052.

[62] Hyvärinen, A. New approximations of differential entropy for Independent Component Analysis and projection pursuit. *Advances in Neural Information Processing Systems 10* (1998), 273–279.

[63] Hyvärinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks 10*, 3 (1999), 626–634.

[64] Hyvärinen, A. Survey on Independent Component Analysis. *Neural Computing Surveys 2* (1999), 94–128.

[65] Hyvärinen, A., and Hoyer, P. O. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation 12*, 7 (2000), 1705–1720.

[66] Hyvärinen, A., Karhunen, J., and Oja, E. *Independent Component Analysis.* Wiley, New York, 2001.

[67] Idier, J. Convex half-quadratic criteria and interacting auxiliary variables for image restoration. *IEEE Transactions on Image Processing 10* (2001), 1001–1009.

[68] James, C., and Gibson, O. Temporally constrained ICA : an application to artifact rejection in electromagnetic brain signal analysis. *IEEE Transactions on Biomedical Engineering 50*, 9 (2003), 1108–1116.

[69] Jung, T.-P., Humphries, C., Lee, T.-W., Makeig, S., McKeown, M. J., Iragui, V., and Sejnowski, T. J. Extended ICA removes artifacts from electroencephalographic recordings. *Advances in Neural Information Processing Systems 10* (1998), 894–900.

[70] Jutten, C., and Herault, J. Blind separation of sources, Part I : An adaptive algorithm based on neuromimetic architecture. *Signal Processing 24* (1991), 1–10.

[71] Kang, K. A., Chance, B., Zhao, S., Srinivasan, S., Patterson, E., and Troupin, R. Breast tumor characterization using near-infrared spectroscopy. *Proceedings of SPIE 1888* (1993), 487–499.

[72] Kendall, M. G., and Stuart, A. *The advanced theory of statistics, volume 1 - Distribution theory.* Charles Griffin and Company Ltd, London, 1963.

[73] Kiviluoto, K., and Oja, E. Independent component analysis for parallel financial time series. *Proceedings of ICONIP'98 – the 5^{th} International Conference on Neural Information Processing 2* (1998), 895–898.

[74] Lawton, W. H., and Sylvestre, E. A. Self modeling curve resolution. *Technometrics 13*, 3 (1971), 617–633.

[75] Leblanc, L., and Dufour, E. Monitoring the identity of bacteria using their intrinsic fluorescence. *FEMS Microbiology Letters 211* (2002), 147–153.

[76] Lee, D. D., and Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature 401* (1999), 788–791.

[77] Lee, D. D., and Seung, H. S. Algorithms for non-negative matrix factorization. *Advances in Neural and Information Processing Systems 13* (2001), 556–562.

[78] Lee, E., Chan, C. K., and Paatero, P. Application of Positive Matrix Factorization in source apportionment of particulate pollutants in Hong Kong. *Atmospheric Environment 33* (1999), 3201–3212.

[79] Lee, J. S., Lee, D. D., Choi, S., and Lee, D. S. Application of non-negative matrix factorization to dynamic positron emission tomography. *Proceedings of ICA 2001 – the 3rd International Conference on Independent Component Analysis and Signal Separation* (2001), 629–632.

[80] Li, S. Z., Hou, X. W., Zhang, H. J., and Cheng, Q. S. Learning spatially localized, parts-based representation. *Computer Vision and Pattern Recognition 2001 1* (2001), I–207–I–212.

[81] Lieber, C. A., and Mahadevan-Jansen, A. Automated method for subtraction of fluorescence from biological Raman spectra. *Applied Spectroscopy 57*, 11 (2003), 1363–1367.

[82] Lin, W.-C., Toms, S. A., Johnson, M., Duco Jansen, E., and Mahadevan-Jansen, A. In vivo brain tumor demarcation using optical spectroscopy. *Photochemistry and Photobiology 73*, 4 (2001), 396–402.

[83] Lowry, A., Wilcox, D., Masson, E. A., and Williams, P. E. Immunohistochemical methods for semiquantitative analysis of collagen content in human peripherical nerve. *Journal of Anatomy 191* (1997), 367–374.

[84] Luypaert, J., Heuerding, S., de Jong, S., and Massart, D. An evaluation of direct orthogonal signal correction and other preprocessing methods for the classification of clinical study lots of a dermatological cream. *Journal of Pharmaceutical and Biomedical Analysis 30* (2002), 453–466.

[85] Mahadevan-Jansen, A., Follen Mitchell, M., Ramanujam, N., Malpica, A., Thomsen, S., Utzinger, U., and Richards-Kortum, R. Near-infrared Raman spectroscopy for in vitro detection of cervical precancers. *Photochemistry and Photobiology 68*, 1 (1998), 123–132.

[86] Makeig, S., Bell, A. J., Jung, T.-P., and Sejnowski, T. J. Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems 8* (1996), 145–151.

[87] Malaroiu, S., Kiviluoto, K., and Oja, E. Time series prediction with independent component analysis. *Proceedings of the International Conference on Advanced Investment Technologies* (1999).

[88] Manoharan, R., Baraga, J. J., Feld, M. S., and Rava, R. P. Quantitative histochemical analysis of human artery using Raman spectroscopy. *Journal of Photochemistry and Photobiology B–Biology 16* (1992), 211–233.

[89] Mansour, A., Barros, A. K., and Ohnishi, N. Blind Separation of Sources : Methods, assumptions and applications. *IEICE Transactions on Fundamentals E83-A*, 8 (2000), 1498–1512.

[90] Maquelin, K., Choo-Smith, L.-P., van Vreeswijk, T., Endtz, H. P., Smith, B., Bennett, R., Bruining, H. A., and Puppels, G. J. Raman spectroscopic method for identification of clinically relevant microorganisms growing on solid culture medium. *Analytical Chemistry 72*, 1 (2000), 12–19.

[91] Mauchien, P. Spectrofluorimétrie moléculaire et spectrométrie de fluorescence atomique. *Techniques de l'Ingénieur, traité Analyse chimique et Caractérisation PE2 835* (1990), 1–13.

[92] Mazet, V., Carteret, C., Brie, D., Idier, J., and Humbert, B. Background removal from spectra by designing and minimising a non-quadratic cost function. *Chemometrics and Intelligent Laboratory Systems 76*, 2 (2005), 121–133.

[93] McIntosh, L. M., Jackson, M., Mantsch, H. H., Stranc, M. F., Pilavdzic, D., and Crowson, A. N. Infrared spectra of basal cell carcinomas are distinct from non-tumor-bearing skin components. *Journal of Investigative Dermatology 112*, 6 (1999), 951–956.

[94] McKeown, M. J., Makeig, S., Brown, G. G., Jung, T. P., Kindermann, S. S., Bell, A. J., and Sejnowski, T. J. Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping 6* (1998), 160–188.

[95] Muroski, A. R., Booksh, K. S., and Myrick, M. L. Single-measurement excitation/emission matrix spectrofluorometer for determination of hydrocarbons in ocean water. 1. Instrumentation and background correction. *Analytical Chemistry 68* (1996), 3534–3538.

[96] Nguyen Thi, H.-L., and Jutten, C. Blind source separation for convolutive mixtures. *Signal Processing 45* (1995), 209–229.

[97] Ohta, N. Estimating absorption bands of component dyes by means of principal component analysis. *Analytical Chemistry 45* (1973), 553–557.

[98] Paatero, P. Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems 37* (1997), 23–35.

[99] Paatero, P., and Tapper, U. Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics 5* (1994), 111–126.

[100] Paatero, P., Tapper, U., Aalto, P., and Kulmala, M. Matrix factorization methods for analysing diffusion battery data. *Journal of Aerosol Sciences 22*, S1 (1991), S273–S276.

[101] Palmer, G. M., Zhu, C., Breslin, T. M., Xu, F., Gilchrist, K. W., and Ramanujam, N. Comparison of multiexcitation fluorescence and diffuse reflectance spectroscopy for the diagnosis of breast cancer (march 2003). *IEEE Transactions on Biomedical Engineering 50*, 11 (2003), 1233–1242.

[102] Papadopoulos, A. J., Zhadin, N. N., Steinberg, M. L., and Alfano, R. R. Fluorescence spectroscopy of normal, SV40-transformed human keratinocytes, and carcinoma cells. *Cancer Biochemistry Biophysics 17* (1999), 13–23.

[103] Paterson, K. G., Sagady, J. L., Hooper, D. L., Bertman, S. B., Carroll, M. A., and Shepson, P. B. Analysis of air quality data using Positive Matrix Factorization. *Environmental Science & Technology 33*, 4 (1999), 635–641.

[104] Petit, A. Théorie des spectres atomiques. *Techniques de l'Ingénieur, traité Analyse et Caractérisation P2 655* (1999), 1–22.

[105] Piot, O. *Caractérisation par microspectroscopie Raman des espèces moléculaires responsables de la cohésion des grains de blé tendre*. PhD thesis, U.F.R. de Pharmacie, Université de Reims Champagne-Ardenne, 2000.

[106] Pradhan, A., Pal, P., Durocher, G., Villeneuve, L., Balassy, A., Babai, F., Gaboury, L., and Blanchard, L. Steady state and time-resolved fluorescence properties of metastatic and non-metastatic malignant cells from different species. *Journal of Photochemistry and Photobiology 31* (1995), 101–112.

[107] Pussayanawin, V., Wetzel, D. L., and Fulcher, R. G. Fluorescence detection and measurement of ferulic acid in wheat milling fractions by microscopy and HPLC. *Journal of Agricultural and Food Chemistry 36* (1988), 515–520.

[108] Ramadan, Z., Song, X. H., and Hopke, P. K. Identification of sources of Phoenix aerosol by Positive Matrix Factorization. *Journal of Air & Waste Management Association 50*, 8 (2000), 1308–1320.

[109] Ren Ong, L., Widjaja, E., Stanforth, R., and Garland, M. Fourier transform Raman spectral reconstruction of inorganic lead mixtures using a novel Band-Target Entropy Minimization (BTEM) method. *Jounal of Raman Spectroscopy 34* (2003), 282–289.

[110] Rieta, J., Castells, F., Sanchez, C., and Zarzoso, V. Atrial activity extraction for atrial fibrillation analysis using blind source separation. *IEEE Transactions on Biomedical Engineering 51*, 7 (2004), 1176–1186.

[111] Ristaniemi, T., and Joutsensalo, J. On the performance of blind source separation in CDMA downlink. *Proceedings of ICA 1999 – the 1st International Conference on Independent Component Analysis and Blind Source Separation* (Aussois, France, 1999), 437–441.

[112] Roscoe, B. A., and Hopke, P. K. Comparison of weighted and unweighted target transformation rotations in factor analysis. *Computers & Chemistry 5* (1981), 1–7.

[113] Saadi, A., Lempereur, I., Sharonov, S., Autran, J. C., and Manfait, M. Spatial distribution of phenolic materials in durum wheat grain as probed by confocal fluorescence spectral imaging. *Journal of Cereal Science 28*, 2 (1998), 107–114.

[114] Sajda, P., Du, S., Brown, T. R., Stoyanova, R., Shungu, D. C., Mao, X., and Parra, L. C. Nonnegative Matrix Factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain. *IEEE Transactions on Medical Imaging 23*, 12 (2004), 1453–1465.

[115] Saltiel, J., Sears, D. F., Choi, J.-O., Sun, Y.-P., and Eaker, D. W. Fluorescence, fluorescence-excitation, and ultraviolet absorption spectra of trans-1-(2-Naphthyl)-2-phenylethene conformers. *Journal of Physical Chemistry 98*, 1 (1994), 35–46.

[116] Sander, T. H., Lueschow, A., Curio, G., and Trahms, L. Removal of alpha-wave artefacts in MEG data by independent component analysis. *Proceedings of Biomag2000* (2000), 857–860.

[117] Sasaki, K., Kawata, S., and Minami, S. Constrained nonlinear method for estimating component spectra from multicomponent mixture. *Applied Optics 22* (1983), 3599–3603.

[118] Shafer-Peltier, K. E., Haka, A. S., Motz, J. T., Fitzmaurice, M., Dasari, R. R., and Feld, M. S. Model-based biological Raman spectral imaging. *Journal of Cellular Biochemistry Supplement 39* (2002), 125–137.

[119] Shen, J., and Israël, G. W. A receptor model using a specific non-negative transformation technique for ambient aerosol. *Atmospheric Environment 23*, 10 (1989), 2289–2298.

[120] Sigurdsson, S., Philipsen, P., Hanson, L., Larsen, J., Gniadecka, M., and Wulf, H. Detection of skin cancer by classification of Raman spectra. *IEEE Transactions on Biomedical engineering 51* (2004), 1784–1793.

[121] Smaragdis, P., and Brown, J. C. Non-negative Matrix Factorization for polyphonic music transcription. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (2003), 177–180.

[122] Stanimirovic, O., Boelens, H., Mank, A., Hoefsloot, H., and Smilde, A. Profiling of liquid crystal displays with Raman spectroscopy : Preprocessing of spectra. *Applied Spectroscopy 59*, 3 (2005), 267–274.

[123] Sun, Y.-P., Bennett, G., Johnston, K. P., and Fox, M. A. Quantitative resolution of dual fluorescence spectra in molecules forming twisted intramolecular charge-transfer states. Toward establishment of molecular probes for medium effects in supercritical fluids and mixtures. *Analytical Chemistry 64* (1992), 1763–1768.

[124] Tang, A. C., Phung, D., Pearlmutter, B. A., and Christner, R. Localization of independent components from magnetoencephalography. *Proceedings of ICA 2000 – the 2nd International Conference on Independent Component Analysis and Blind Source Separation* (Helsinki, Finlande, 2000), 387–392.

[125] Tfayli, A., Gobinet, C., Vrabie, V., Piot, O., and Huez, R. Melanoma characterisation on paraffinised biopsies with raman spectroscopy and independent component analysis. *Proceedings of Dataspec 2005 – the 1st International Workshop on Data Analysis and Biospectroscopy* (Reims, France, 2005).

[126] Tfayli, A., Piot, O., Durlach, A., Bernard, P., and Manfait, M. Discriminating nevus and melanoma on paraffin-embedded skin biopsies using FTIR microspectroscopy. *Biochimica et Biophysica Acta (BBA) - General Subjects 1724* (2005), 262–269.

[127] Theis, F. J., Stadlthanner, K., and Tanaka, T. First results on uniqueness of sparse Non-negative Matrix Factorization. *Proceedings of EUSIPCO 2005 – the 13th European Signal Processing Conference* (Antalya, Turquie, 2005).

[128] Valeur, B. *Invitation à la fluorescence.* Éditions de Boeck Université, 2004.

[129] van de Poll, S. W. E., Romer, T. J., Puppels, G. J., and van der Laarse, A. Raman spectroscopy of atherosclerosis. *Journal of Cardiovascular Risk 9*, 5 (2002), 255–261.

[130] Vigário, R. Extraction of ocular artifacts from EEG using independent component analysis. *Electroencephalography and Clinical Neurophysiology 103*, 3 (1997), 395–404.

[131] Vigário, R., Jousmäki, V., Hämäläinen, M., Hari, R., and Oja, E. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. *Advances in Neural Information Processing Systems 10* (1998), 229–235.

[132] Vigário, R., Särelä, J., Jousmäki, V., and Oja, E. Independent component analysis in decomposition of auditory and somatosensory evoked fields. *Proceedings of ICA 1999 – the 1st International Conference on Independent Component Analysis and Blind Source Separation* (Aussois, France, 1999), 167–172.

[133] Vigário, R., Särelä, J., and Oja, E. Independent component analysis in wave decomposition of auditory evoked fields. *Proceedings of ICANN'98 – the 8th International Conference on Artificial Neural Networks* (1998), 287–292.

[134] Vrabie, V. *Statistiques d'ordre supérieur : applications en géophysique et électrotechnique.* PhD thesis, Institut National Polytechnique de Grenoble, 2003.

[135] Vrabie, V., Huez, R., Gobinet, C., Piot, O., Tfayli, A., and Manfait, M. On the modelling of paraffin through Raman spectroscopy. *Proceedings of MCBMS'06 – the 6th IFAC Symposium on Modelling and Control in Biomedical Systems* (Reims, France, 2006). Accepté.

[136] Wentzell, P. D., Nair, S. S., and Guy, R. D. Three-way analysis of fluorescence spectra of polycyclic aromatic hydrocarbons with quenching by nitromethane. *Analytical Chemistry 73* (2001), 1408–1415.

[137] Widjaja, E., Li, C., and Garland, M. Semi-batch homogeneous catalytic in-situ spectroscopic data. FTIR spectral reconstructions using Band-Target Entropy Minimization (BTEM) without spectral preconditioning. *Organometallics 21* (2002), 1991–1997.

[138] Windig, W., and Guilment, J. Interactive self-modeling mixture analysis. *Analytical Chemistry 63* (1991), 1425–1432.

[139] Yellin, D., and Weinstein, E. Multichannel signal separation : methods and analysis. *IEEE Transactions on Signal Processing 44* (1996), 106–118.

[140] Zarzoso, V., Nandi, A. K., and Bacharakis, E. Maternal and foetal ECG separation using blind source separation methods. *IMA Journal of Mathematics Applied in Medicine and Biology 14*, 3 (1997), 207–225.

[141] Ziehe, A., Müller, K.-R., Nolte, G., Mackert, B.-M., and Curio, G. Artifact reduction in magnetoneurography based on time-delayed second-order correlations. *IEEE Transactions on Biomedical Engineering 47*, 1 (2000), 75–87.

Résumé

Les techniques de spectroscopies optiques fournissent des informations moléculaires sur un échantillon. Leur séparation est nécessaire pour identifier les espèces présentes et analyser leur distribution spatiale.
La modélisation des données et l'analyse de leurs propriétés nous permettent de désigner des techniques de séparation de sources propres à chaque spectroscopie.
La positivité et la forme à variations lentes des spectres suggèrent l'utilisation de la Factorisation en Matrices Non-négatives (FMN) pour séparer les informations. Son efficacité est illustrée sur des grains de blé et d'orge dont les acides phénoliques principaux sont identifiés par la FMN. L'acide férulique, constituant principal de la couche à aleurone qui est indicatrice de la qualité d'une farine, peut être utilisé comme un indicateur de la contamination d'une farine par les sons.
Ensuite, une méthode de déparaffinage numérique basée sur l'association de la spectroscopie Raman et de l'Analyse en Composantes Indépendantes (ACI) est proposée. La propriété d'indépendance chimique entre les espèces constitutives d'un échantillon et les propriétés de parcimonie et de décorrélation de leurs spectres Raman suffisent à admettre leur indépendance mutuelle. L'application de l'ACI permet d'estimer les spectres de la paraffine, de la fluorine et du tissu sous-jacent. Lors de l'application de cette procédure sur des échantillons paraffinés de peau, la paraffine est prouvée comme modélisable par trois sources indépendantes, et les spectres d'épidermes de mélanome et de nævus sont isolés. La mise en évidence de discripteurs moléculaires rend possible la discrimination entre ces deux types de pathologies.

Mots-clés: Séparation Aveugle de Sources, Analyse en Composantes Indépendantes, Factorisation en Matrices Non-négatives, spectroscopie Raman, spectroscopie de fluorescence, grains de céréales, diagnostic précoce du cancer de la peau

Abstract

Optical spectroscopies give molecular informations of a sample. The separation of these informations is required in order to identify pure present chemical species and analyze their spatial distribution.
The modelization of spectroscopic data and the analysis of their properties lead to choose source separation techniques well suited for each spectroscopy.
Positivity and smooth shape of fluorescence spectra suggest the use of Non-negative Matrix Factorization (NMF) to separate spectra. Its efficacy is illustrated on wheat and barley grains. Their major phenolic acids are identified by NMF. Ferulic acid is the major component of the aleuron layer. This layer is an indicator of the quality of a flour. This acid can be used as an indicator of the bran contamination of a flour.
Then, a method to numerically dewax samples is proposed. It is based on the association of Raman spectroscopy with Independent Component Analysis (ICA). Chemical independence of pure species and sparsity and decorrelation of their Raman spectra are sufficient to assume their mutual independence. The application of ICA leads to the estimation of spectra of paraffin, fluorine and the underlying tissue. This dewaxing process is applied on paraffinised skin tissues. Paraffin must be modeled by three independent sources and spectra of melanoma and nevi epidermis can be isolated by this numerical dewaxing process. The discrimination between these two pathologies is possible thanks to molecular descriptors that has been found.

Keywords: Blind Source Separation, Independent Component Analysis, Non-negative Matrix Factorization, Raman spectroscopy, fluorescence spectroscopy, cereal grains, early diagnosis of skin cancer