De l'importance des traitements préalables à l'application de l'Analyse en Composantes Indépendantes en spectroscopie Raman

Cyril Gobinet[1], Valeriu Vrabie[2], Éric Perrin[2], Régis Huez[2]

[1]Unité MéDIAN, UMR CNRS 6142, Université de Reims Champagne-Ardenne
51 rue Cognacq-Jay, 51096 Reims Cedex

[2]CReSTIC, Université de Reims Champagne-Ardenne
Chaussée du Port, 51000 Châlons-en-Champagne

cyril.gobinet@univ-reims.fr, valeriu.vrabie@univ-reims.fr
eric.perrin@univ-reims.fr, regis.huez@univ-reims.fr

Résumé – La spectroscopie Raman est un outil puissant d'analyse de la composition moléculaire d'échantillons biologiques. La complexité de certains tissus rend l'analyse des spectres tributaire de méthodes d'extraction des informations pertinentes. L'Analyse en Composantes Indépendantes est montrée comme un outil efficace de séparation des spectres à condition que des prétraitements adaptés et développés dans ce papier soient appliqués afin de linéariser le modèle génératif des spectres déformés par l'instrumentation.

Abstract – Raman spectroscopy is a powerful tool to analyze the molecular composition of biological samples. Some tissues are complex and require the use of numerical methods in order to extract useful information. Independent Component Analysis is shown as an efficient Raman spectra separation tool with the proviso that preprocessing steps are applied to linearize the generative model of Raman spectra distorted by the instrumentation. Such preprocessing steps are developed in this article.

1 Introduction

La spectroscopie Raman est une technique vibrationnelle d'analyse moléculaire basée sur l'interaction inélastique entre un laser et la matière à analyser. En chaque point de mesure, l'instrumentation enregistre un spectre Raman qui est un vecteur $\mathbf{I}_k = [\ldots, I_k(\bar{\nu}), \ldots]$ dont chaque élément $I_k(\bar{\nu})$ représente l'intensité de la lumière diffusée au nombre d'onde $\bar{\nu}$. L'intensité diffusée est proportionnelle au nombre de molécules vibrant en ce nombre d'onde. Une propriété intéressante de la spectroscopie Raman est que chaque molécule possède un spectre Raman unique qui représente sa signature spectrale. L'analyse du spectre d'un échantillon permet donc d'identifier les différentes molécules qui y sont présentes. La simplicité d'utilisation et la puissance d'analyse de cette méthode en ont fait son succès qui se traduit par de multiples applications notamment dans le biomédical [1].

Un spectre Raman est composé d'intensités positives ou nulles et peut être vu comme une superposition de pics assimilables à des gaussiennes ou des lorentziennes. L'étude d'échantillons biologiques complexes peut s'avérer difficile à cause des recouvrements de pics. En particulier, les signatures Raman de certaines espèces chimiques peuvent cacher celles d'autres espèces d'intérêt. Des exemples de tels spectres Raman acquis sur une biopsie paraffinée de peau humaine, qui est le support d'étude de cet article, sont proposés sur la figure 1. La signature Raman intense de la paraffine cache les informations vibrationnelles de la peau et empêche le diagnostic de cancers de la peau.

Une méthode efficace d'analyse a été proposée dans [2]. La modélisation des spectres Raman acquis à la surface d'un échantillon biologique comme un mélange linéaire et instantané des signatures spectrales des espèces moléculaires présentes dans cet échantillon et la forme générale des spectres Raman (pics parcimonieux non superposés des différentes espèces) y ont été montrées suffisantes pour l'application des méthodes de séparation de sources, en particulier de l'Analyse en Composantes Indépendantes (ACI) [3], sur ces types de signaux. Dans [4], les spectres acquis à la surface d'une petite coupe paraffinée de peau humaine (280×280 µm^2) ont été prouvés constitués de 3 sources de paraffine et d'une source de peau représentées sur la figure 2. Des experts ont mis en évidence dans le spectre estimé de la peau plusieurs pics et bandes spectrales caractéristiques de ses divers constituants (ADN, protéines, ...) [2], ce qui rend possible la discrimination des mélanomes et confirme l'efficacité de la méthode.

Expérimentalement, aux signaux Raman purs s'ajoutent des effets perturbateurs tels que le fond de fluorescence, le décalage en nombre d'onde des pics Raman et l'hétérogénéité des largeurs de pics. Les spectres Raman bruts ne respectent donc plus le modèle linéaire et instantané de l'ACI dont l'application aboutit à l'estimation de sources aberrantes comme le prouve la figure 3 qui montre l'application de l'ACI sur les spectres Raman bruts acquis sur une grande biopsie paraffinée de peau humaine. Il est évident que ces sources sont complètement différentes de

celles présentées sur la figure 2.

Dans le présent article, nous nous intéressons au développement de prétraitements adaptés à la correction des effets perturbateurs précités (afin de retrouver le modèle linéaire et instantané de l'ACI) et à leurs effets sur les estimations de l'ACI appliquée sur des spectres Raman de peau humaine paraffinée.

2 Description des données et du modèle

L'étude est faite sur 952 spectres Raman acquis sur une grande coupe paraffinée (10×10 mm^2) de peau humaine dans l'intervalle spectral allant de 657 à 1821 cm^{-1}, soit 990 échantillons par spectre. Le jeu de données est donc une matrice $\mathbf{I} = [\ldots, \mathbf{I}_k^T, \ldots]^T \in \mathbb{R}^{N \times N_{\bar{\nu}}}$ où $.^T$ est l'opérateur de transposition, $N = 952$ est le nombre de spectres, et $N_{\bar{\nu}} = 990$ est le nombre de nombres d'onde d'acquisition. La figure 1 représente deux spectres (en traits pleins) acquis en deux points différents de l'échantillon.

Le modèle a été prouvé linéaire et instantané (après l'application de prétraitements adaptés qui vont être décrits dans la suite) [2], et constitué de 4 sources dans une étude préliminaire sur des données similaires [4] acquises sur une petite coupe paraffinée (280×280 µm^2) : 3 sources de paraffine et 1 source de peau représentées sur la figure 2. La parcimonie et la non-superposition des différents pics de la paraffine mènent à l'annulation des cumulants croisés d'ordres supérieurs des 4 sources [2], prouvant leur indépendance. L'ACI peut alors être utilisée pour extraire ces sources. Nous avons utilisé parmi les techniques d'ACI l'algorithme JADE (Joint Approximate Diagonalization of Eigenmatrices) [5] mais l'application d'autres algorithmes tels que MD (Maximal Diagonality) [6] ou FastICA [3] mène aux mêmes résultats.

3 La ligne de base

Les échantillons biologiques sont connus pour émettre de la fluorescence qui est un effet en compétition avec la diffusion Raman. Ce fond de fluorescence s'exprime sous la forme d'une ligne de base différente d'un spectre acquis à un autre et est représenté en pointillés sur la figure 1. Au point d'acquisition k, le fond de fluorescence $\mathbf{F}_k$ se modélise par une fonction polynômiale $\mathbf{a}_k^T \mathbf{\Lambda}$ d'ordre L dont les coefficients diffèrent d'un point d'acquisition à un autre. Le vecteur $\mathbf{a}_k = [\ldots, a_{k_i}, \ldots]^T \in \mathbb{R}^{L+1}$ définit le vecteur des coefficients du polynôme et la matrice $\mathbf{\Lambda} = [\ldots, \mathbf{\Lambda}_i, \ldots] \in \mathbb{R}^{(L+1)\times N_{\bar{\nu}}}$ est la transposée de la matrice de Vandermonde des nombres d'onde [7].

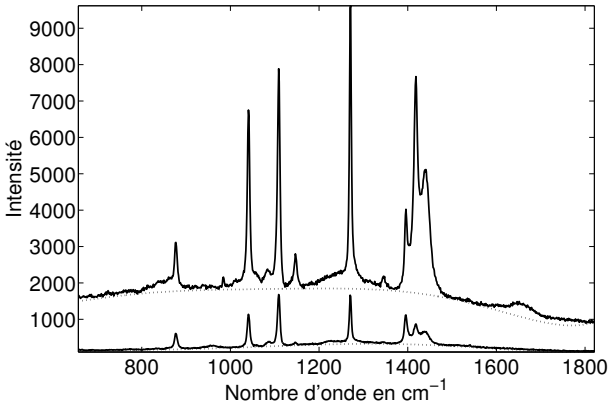


FIG. 1 – Exemples de 2 spectres Raman (continus) et leurs lignes de base associées (pointillés) acquis en 2 points différents d'une grande coupe paraffinée de peau

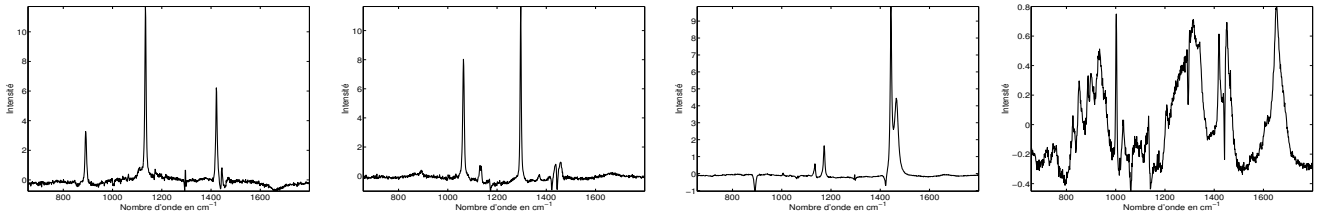
Dans cet article, la méthode de Mazet [7], basée sur la minimisation d'une fonction coût quadratique tronquée $Q(\mathbf{a}_k) = \sum_{i=1}^{N_{\bar{\nu}}} \phi(\mathbf{I}_k(\bar{\nu}_i) - \mathbf{a}_k^T \mathbf{\Lambda}_i)$ où $\phi(x) = x^2$ si $x < \gamma$ et $\phi(x) = \gamma^2$ sinon, a été utilisée puisqu'elle permet de s'affranchir des pics Raman de forte intensité et d'incorporer le seuil γ lié à la puissance du bruit. Une fois estimée, cette ligne de base $\mathbf{F}_k$ est soustraite au spectre brut $\mathbf{I}_k$ [8].

L'algorithme JADE [5] d'ACI a été appliqué sur les données brutes et sur les données corrigées de leur ligne de base. Les sources estimées dans ces deux cas sont respectivement montrées sur les figures 3 et 4. Une nette amélioration est visible puisque les composantes à variation lente (dues à la ligne de base) visibles sur chaque source estimée de la figure 3 (en particulier la figure 3(d)) sont absentes sur les sources de la figure 4 estimées à partir des données corrigées de leur ligne de base.

4 Le décalage des pics en nombre d'onde

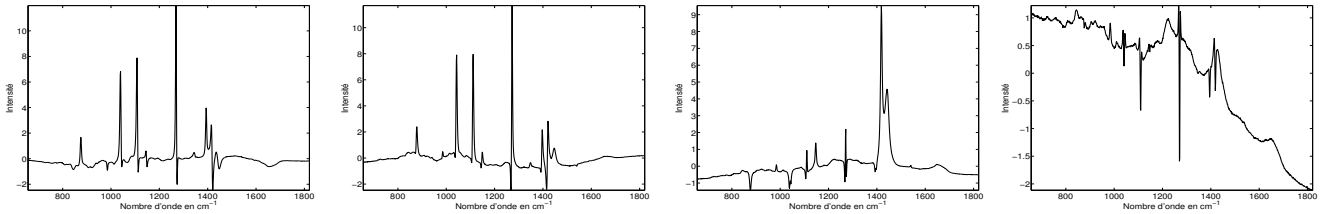
La résolution spectrale de l'ordre de 4 cm^{-1} introduit une incertitude sur la position exacte des maxima des pics les plus intenses de la paraffine. Cette incertitude se manifeste par l'estimation de 2 sources (figures 4(b) et 4(a)) qui ne diffèrent que par les positions voisines de leurs pics principaux ([1039.5, 1107.7, 1268.9] et [1043, 1111.2, 1272.5] cm^{-1} respectivement).

Une procédure d'alignement des pics a été appliquée aux spectres corrigés de leur ligne de base afin d'éliminer cet effet [8]. Chaque pic de la paraffine est délimité par son support en nombres d'onde. La procédure étant identique pour tous les pics de la paraffine, nous nous limiterons à l'expliquer dans le cas d'un pic de support $\mathcal{D}$. Un sous-spectre de référence $\mathbf{I}_{Ref}(\mathcal{D})$ doit être choisi afin d'aligner les autres sous-spectres sur son maximum. Ce sous-spectre est celui qui a la distance euclidienne globale avec tous les autres sous-spectres la plus petite. La procédure étant identique pour tous les sous-spectres, nous allons donc nous limiter au recalage d'un seul sous-spectre $\mathbf{I}_k(\mathcal{D})$. La fonction d'intercorrélation $\mathbf{C}_{Ref,k}$ entre le sous-spectre de référence $\mathbf{I}_{Ref}(\mathcal{D})$ et le sous-spectre $\mathbf{I}_k(\mathcal{D})$ est calculée. Le maximum de cette intercorrélation se situe au décalage en nombre d'onde nécessaire pour aligner le pic avec sa référence. Afin d'obtenir des décalages inférieurs à un échantillon, ces sous-spectres pourront être sur-échantillonnés. Finalement $\mathbf{I}_k(\mathcal{D})$ est recalé en utilisant la transformée de Fourier et se transforme en $\tilde{\mathbf{I}}_k(\mathcal{D})$. Le recalage peut introduire des discontinuités entre $\mathbf{I}_k(\mathcal{D})$ et $\tilde{\mathbf{I}}_k(\mathcal{D})$ sur les bords de $\mathcal{D}$. Une pondération de chaque



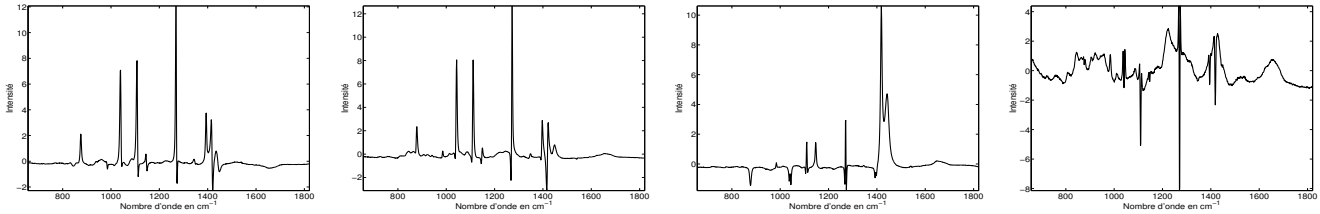
(a) Paraffine (b) Paraffine (c) Paraffine (d) Peau

FIG. 2 – Sources estimées par ACI sur une petite coupe paraffinée de peau dans [4]



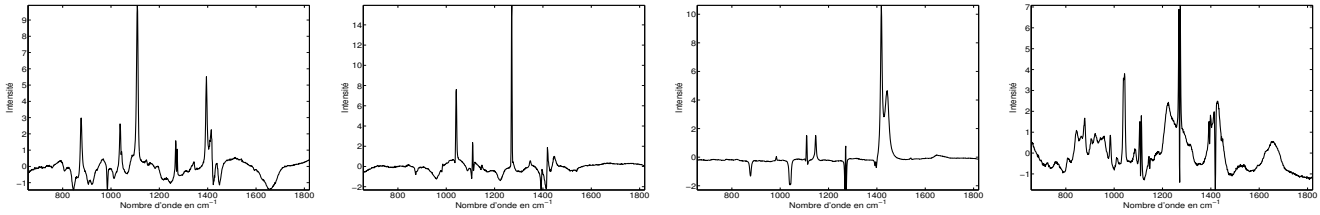
(a) Paraffine (b) Paraffine (c) Paraffine (d) Peau

FIG. 3 – Sources estimées par JADE sur les données brutes



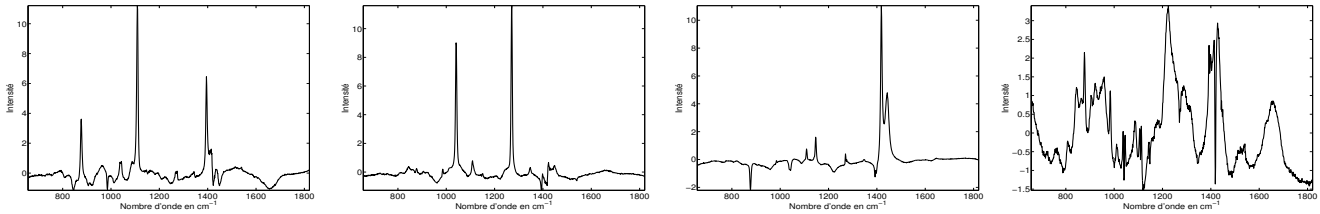
(a) Paraffine (b) Paraffine (c) Paraffine (d) Peau

FIG. 4 – Sources estimées par JADE sur les données corrigées de leur ligne de base



(a) Paraffine (b) Paraffine (c) Paraffine (d) Peau

FIG. 5 – Sources estimées par JADE sur les données alignées en nombre d'onde



(a) Paraffine (b) Paraffine (c) Paraffine (d) Peau

FIG. 6 – Sources estimées par JADE sur les données uniformisées en largeur de pics

signal respectivement par une fenêtre de Hanning $\mathbf{H}_h$ et par son complémentaire à 1 permet de s'affranchir de ces effets : $\bar{\mathbf{I}}_k(\mathcal{D}) = (\bar{\mathbf{I}}_k(\mathcal{D}) \times \mathbf{H}_h) + (\mathbf{I}_k(\mathcal{D}) \times (1 - \mathbf{H}_h))$.

La figure 5 présente les spectres estimés par JADE sur ces données doublement corrigées. L'estimation s'est encore une fois améliorée avec 4 sources distinctes contrairement à la figure 4.

5 L'hétérogénéité des largeurs des pics

L'inclinaison, les imperfections de découpe et l'altération de l'échantillon sont autant de facteurs qui peuvent faire varier la largeur des pics Raman en des points d'acquisition spatialement éloignés. Ces variations biaisent les estimations par ACI comme le montrent les figures 5(b) et 5(d) qui présentent, au voisinage du nombre d'onde 1271.4 cm^{-1}, respectivement un pic Raman et la forme d'une dérivée seconde d'un pic Raman. Cette dernière forme est caractéristique de l'hétérogénéité de la largeur du pic.

Une procédure d'uniformisation de la largeur des pics Raman caractéristiques de la paraffine a été appliquée aux spectres Raman corrigés de leur ligne de base et de leur décalage des pics en nombre d'onde. Comme précédemment, la méthode est présentée brièvement pour un pic de paraffine et un spectre acquis. Étant plus simple d'agrandir la largeur d'un pic que de la diminuer par la procédure présentée, le sous-spectre de référence $\bar{\mathbf{I}}_{Ref}(\mathcal{D})$ est calculé comme la moyenne des 20% de sous-spectres de plus grande largeur afin d'éliminer divers pics aberrants . Puis le noyau de convolution $\mathbf{H} = [h_2, h_1, 1, h_1, h_2]$ est estimé par minimisation de la distance euclidienne entre $\bar{\mathbf{I}}_{Ref}(\mathcal{D})$ et le sous-spectre transformé $\breve{\mathbf{I}}_k(\mathcal{D}) = \bar{\mathbf{I}}_k(\mathcal{D}) * \mathbf{H}$ où $*$ symbolise le produit de convolution. Finalement, comme dans la section précédente, les effets de bords sont traités grâce à des fenêtres de Hanning par : $\hat{\mathbf{I}}_k(\mathcal{D}) = (\breve{\mathbf{I}}_k(\mathcal{D}) \times \mathbf{H}_h) + (\bar{\mathbf{I}}_k(\mathcal{D}) \times (1 - \mathbf{H}_h))$

La figure 6 présente les spectres estimés par JADE sur ces données triplement corrigées. Ces 4 sources sont spectroscopiquement interprétables. Les figures 6(a), 6(b) et 6(c) restituent les informations spectrales de la paraffine [4], tandis que la figure 6(d) correspond pratiquement parfaitement au spectre Raman de la peau [2].

6 Discussion

L'analyse successive des figures 3, 4, 5 et 6 montre l'efficacité des traitements appliqués aux spectres avant l'utilisation de l'ACI. Alors que sur les figures 3, 4 et 5, les sources estimées ne représentent aucune signature spectrale des entités présentes dans l'échantillon, celles de la figure 6 sont facilement identifiables aux sources de la figure 2 obtenues lors d'une étude précédente [4] validée par des experts en spectroscopie Raman.

Un autre sujet de discussion concerne la positivité. Il est vrai qu'une propriété des spectres Raman est leur positivité. On pourrait donc s'attendre à estimer des sources positives. Cependant, les spectres sont centrés et normalisés à une variance unité avant l'application de JADE. Les sources estimées seront donc forcément centrées.

Cependant, la matrice de mélange estimée par ACI (en même temps que la matrice des sources et qui définit les profils de concentrations associés [2, 4, 8]) n'est composée que d'éléments positifs. Cette observation est en accord avec la positivité des concentrations des espèces constitutives de l'échantillon analysé auxquelles sont assimilés les coefficients de mélange.

7 Conclusion

Une ACI peut s'avérer mal adaptée lorsqu'elle est appliquée sur des spectres Raman bruts. Des déformations d'origine matérielle cassent le modèle linéaire et instantané de l'ACI. Afin de linéariser le modèle génératif des spectres Raman, des prétraitements efficaces et adaptés à la ligne de base, au décalage en nombre d'onde des pics Raman intenses et à la déformation de la largeur de ces pics ont été proposés. Leur application sur des spectres Raman acquis à la surface d'un échantillon paraffiné de peau humaine a été prouvée comme préalablement indispensable à l'application de l'ACI afin d'estimer efficacement des informations cachées par la signature Raman intense de la paraffine.

Références

[1] E.E. Lawson, B.W. Barry, A.C. Williams et H.G.M. Edwards. *Biomedical applications of Raman spectroscopy.* Journal of Raman Spectroscopy, vol. 28, n°2-3, p. 111–117, 1997.

[2] V. Vrabie, C. Gobinet, O. Piot, A. Tfayli, P. Bernard, R. Huez et M. Manfait. *Independent Component Analysis of Raman spectra : Application on paraffin-embedded skin biopsies.* Biomedical Signal Processing and Control, vol. 2, n°1, p. 40–50, 2007.

[3] A. Hyvarinen, J. Karhunen et E. Oja. *Independent Component Analysis.* Wiley, 2001.

[4] V. Vrabie, R. Huez, C. Gobinet, O. Piot, A. Tfayli et M. Manfait. *On the modelling of paraffin through Raman spectroscopy.* 6th IFAC Symposium on Modelling and Control in Biomedical Systems (MCBMS'06). Reims, France, 20–22 septembre 2006.

[5] J.-F. Cardoso et A. Souloumiac. *Blind beamforming for non-Gaussian signals.* IEE Proceedings-F, vol. 140, n°6, p. 362–370, 1993.

[6] P. Comon. *Independent component analysis, a new concept ?* Signal Processing, vol. 36, p. 287–314, 1994.

[7] V. Mazet, C. Carteret, D. Brie, J. Idier et B. Humbert. *Background removal from spectra by designing and minimising a non-quadratic cost function.* Chemometrics and Intelligent Laboratory Systems, vol. 76, n°2, p. 121–133, 2005.

[8] C. Gobinet. *Application de techniques de séparation de sources à la spectroscopie Raman et à la spectroscopie de fluorescence.* Thèse de l'URCA, Reims, France, mars 2006.