Décomposition d'une séquence de spectres avec modèle markovien et algorithme RJMCMC à deux variables de dimension

Vincent MAZET[1], Sylvain FAISAN[1], Lionel POISSON[2], Marc-André GAVEAU[2], Jean-Michel MESTDAGH[2]

[1]ICube, Université de Strasbourg, CNRS ; 300 boulevard Sébastien Brant, BP 10413, 67412 Illkirch, France
[2]Laboratoire Francis Perrin, CEA, IRAMIS, CNRS ; Service des Photons Atomes et Molécules, 91191 Gif-sur-Yvette, France.
{vincent.mazet,faisan}@unistra.fr, {prenom.nom}@cea.fr

Résumé – L'objectif de ce travail est de décomposer une séquence temporelle de spectres de photoélectrons en raies dont on estime les paramètres (position, amplitude, largeur). Comme les raies évoluent doucement à travers les données, la décomposition est effectuée conjointement sur toute la séquence et les raies sont regroupées en trajectoires. Dans un contexte bayésien, cette évolution douce est modélisée par un a priori markovien sur les paramètres des raies. De plus, le nombre de raies et de trajectoires étant inconnues, l'algorithme RJMCMC est choisi pour échantillonner la loi a posteriori. L'une des originalités de notre travail est d'avoir deux variables de dimension : cela impose d'introduire de nouveaux mouvements. Enfin, une procédure originale a été mise au point pour accélérer la convergence de l'algorithme. Les résultats obtenus à partir de données réelles sont satisfaisants et permettent de confirmer l'opinion des experts.

Abstract – This work aims at decomposing a temporal sequence of photoelectron spectra in peaks whose parameters (position, amplitude, width) are estimated. As the peaks move slowly through the data, the decomposition is performed jointly over the whole sequence and the peaks are grouped into tracks. In a Bayesian framework, this slow evolution is modeled by a Markovian prior on the peak parameters. In addition, the number of peaks and tracks being unknown, the RJMCMC algorithm is chosen to sample the posterior. One of the novelties of our work is to deal with two variables of dimension: this requires the introduction of new moves. Finally, an original procedure was proposed to accelerate the convergence of the algorithm. Results obtained on real data illustrate the performance of the method.

1 Introduction

Un photoélectron est un électron émis d'un échantillon de matière suite à l'absorption d'une radiation électromagnétique. La distribution de ces électrons en fonction de leur énergie est appelée spectre de photoélectrons ; il est généralement modélisé comme une somme de raies superposée sur un continuum. Une séquence de spectres de photoélectrons est un ensemble de spectres acquis à différents instant de l'expérience [11]. L'échantillonnage temporel des acquisitions est suffisamment fin pour considérer que les raies évoluent lentement, c'est-à-dire que leurs paramètres varient peu entre deux spectres consécutifs. En outre, le nombre de raies peut varier à travers la séquence car les raies peuvent apparaître ou disparaître. L'objectif de ce travail est à la fois d'estimer le nombre de raies et leurs paramètres (centres, amplitudes et largeurs) et de les suivre à travers la séquence.

À notre connaissance, cette problématique n'a jamais été abordée bien qu'un grand nombre de travaux proposent des solutions à des problèmes similaires, comme par exemple en spectroscopie Raman [4], en temps-fréquence [2, 10], en sismique-réflexion [7], ou en suivi de cibles [1, 9]. Cela dit, ces travaux ne peuvent pas être directement utilisés pour notre problème. De même, notre problème n'est pas un problème de séparation de source [15] ou de démelange spectral [8] car les positions et formes des raies varient à travers la séquence. Or, la décomposition d'un *unique* spectre a été largement étudiée dans le passé. En particulier, les méthodes bayésiennes couplées à un algorithme MCMC (*Monte Carlo Markov Chain*) sont des techniques très performantes [3, 6, 12].

Une décomposition séquentielle, dans laquelle les spectres sont décomposés indépendamment les uns des autres, est inadaptée [7, 13] pour deux raisons : elle peut aboutir à des décompositions très différentes de spectres pourtant contigus, en contradiction avec l'hypothèse d'évolution lente des raies ; elle ne permet pas de suivre une raie à travers les données, nécessitant alors un post-traitement. Au contraire, une décomposition conjointe, où les spectres sont décomposés en même temps, permet de prendre en compte l'évolution lente des raies si leur évolution est régularisée, fournissant ainsi des résultats cohérents. Elle permet également de classer les raies et donc de les suivre au cours de la séquence.

Nous avons récemment proposé une telle approche dans [13, 14], mais cet article apporte des modifications majeures : le nombre de raies est inconnu et nécessite d'utiliser l'algorithme RJMCMC, la plupart des hyperparamètres sont estimés, de nouvelles lois a priori sont proposées, nous mettons en avant et discutons brièvement du fait que le modèle a maintenant deux variables de dimensions, et enfin le nombre de spectres traités est plus important.

Ce travail a été soutenu par l'INSII dans le cadre du PEPS Rupture SpectroDec.

2 Modèle bayésien

Chaque spectre $\boldsymbol{y}_s$ ($s \in \{1,\dots,S\}$) est modélisé comme une somme de raies $\boldsymbol{x}_s$ et d'un terme de bruit $\boldsymbol{v}_s$: $\boldsymbol{y}_s = \boldsymbol{x}_s + \boldsymbol{v}_s$ [6, 11]. Afin de suivre et de régulariser l'évolution des raies au cours de la séquence, les raies sont regroupées en trajectoires : une trajectoire est composée d'au plus une raie par spectre et s'étend sur des spectres consécutifs ; on note K le nombre de trajectoires. La k-ème trajectoire apparaît au spectre b_k et se termine au spectre b_k+l_k-1 où l_k représente le nombre de raies de la trajectoire. Comme les raies sont modélisées par une gaussienne, le n-ème élément de $(\boldsymbol{x}_s)$, noté $(\boldsymbol{x}_s)_n$, s'écrit :

$$(\boldsymbol{x}_s)_n = \sum_{k=1}^{K}\sum_{m=1}^{l_k} a_{k,m}\exp\left(-\frac{(n-c_{k,m})^2}{2w_{k,m}^2}\right)\delta_{b_k+m-1,s} \tag{1}$$

avec $n \in \{1,\dots,N\}$, N représente la taille d'un spectre. La m-ème raie de la trajectoire k est paramétrée par son centre $c_{k,m}$, son amplitude $a_{k,m}$ et sa largeur $w_{k,m}$; sa présence dans le spectre $\boldsymbol{y}_s$ est codée par le symbole de Kronecker $\delta_{b_k+m-1,s}$ qui vaut 1 si $b_k + m - 1 = s$, 0 sinon. Le nombre total de raies est $M = \sum_{k=1}^{K} l_k$. M et K définissent la dimension du problème : les trajectoires sont modélisées par $2K$ paramètres ($\boldsymbol{b}$ et $\boldsymbol{l}$, qui définissent la « structure » du modèle), les raies par $3M$ paramètres ($\boldsymbol{c}$, $\boldsymbol{a}$, $\boldsymbol{w}$), et il y a 4 hyperparamètres (définis dans la suite). Ainsi, pour M et K fixes, le problème est de dimension $3M + 2K + 4$.

A priori joint sur M, K, $\boldsymbol{l}$ et $\boldsymbol{b}$ L'a priori $p(M,K,\boldsymbol{l},\boldsymbol{b})$ doit favoriser les solutions avec le plus petit nombre de raies et de trajectoires. La procédure traditionnelle consiste à définir la loi a priori ainsi : $p(M,K,\boldsymbol{l},\boldsymbol{b}) = p(M,K)p(\boldsymbol{l},\boldsymbol{b}|M,K)$. Or, $p(\boldsymbol{l},\boldsymbol{b}|M,K)$ a une influence très variable suivant les valeurs de M et K. Par exemple, si $M = SK$, il n'y a qu'une configuration (toutes les traces commencent au spectre 1 et finissent au spectre S) et elle est de probabilité 1. Au contraire, si $M \approx SK/2$, alors le nombre de configurations devient important lorsque M augmente, conduisant ainsi à une probabilité extrêmement faible. Ce phénomène a donc tendance à privilégier de grandes trajectoires et il est alors très difficile de définir $p(M,K)$ pour respecter notre connaissance a priori. Aussi, nous définissons l'a priori joint $p(M,K,l,b)$ comme :

$$p(M,K,\boldsymbol{l},\boldsymbol{b}) \propto \xi^{M+K}\,\mathbb{1}_{\mathcal{X}}(M,K,\boldsymbol{l},\boldsymbol{b}) \tag{2}$$

où $\mathbb{1}$ est la fonction indicatrice et $\mathcal{X}$ l'ensemble des valeurs possibles du quadruplet $(M,K,\boldsymbol{l},\boldsymbol{b})$. En effet, pour que $p(M,K,\boldsymbol{b},\boldsymbol{l})$ soit intégrable, le nombre de raies est borné par une valeur maximale $K_{\max}$. Les contraintes $K \in \{1,\dots,K_{\max}\}$ et $M \in \{K,\dots,SK\}$ définissent l'ensemble des espaces de dimensions constantes et sont schématisées sur la figure 1. Il existe également des contraintes sur $\boldsymbol{l}$ et $\boldsymbol{b}$: $\boldsymbol{l}$ et $\boldsymbol{b}$ sont des vecteurs de taille K ; les composantes de $\boldsymbol{l}$ et $\boldsymbol{b}$ prennent leurs valeurs dans $\{1,\dots,S\}$; $\forall k, b_k + l_k - 1 \leq S$; et $M = \sum_{k=1}^{K} l_k$. Notez que pour M et K fixé, le modèle proposé revient à considérer $p(\boldsymbol{l},\boldsymbol{b}|M,K)$ uniforme donc à ne privilégier aucune configuration.

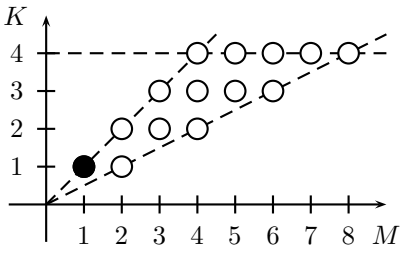

FIGURE 1 – Valeurs possibles de M et K pour $K_{\max} = 4$ et $S = 2$. Chaque cercle représente l'espace des solutions pour le couple (M,K) ; le cercle noir correspond à l'initialisation.

Paramètres de raies Nous supposons que les positions des raies $\boldsymbol{c}_k = (c_{k,1},\dots,c_{k,l_k})$ de la trajectoire k évoluent lentement à travers la séquence. Elles sont par ailleurs considérées mutuellement et conditionnellement indépendantes sachant $r_{\boldsymbol{c}}$ et $\boldsymbol{l}$. Comme $p(\boldsymbol{c}_k|r_{\boldsymbol{c}},\boldsymbol{l}) = p(\boldsymbol{c}_k|r_{\boldsymbol{c}},l_k)$, il vient :

$$p(\boldsymbol{c}|r_{\boldsymbol{c}},\boldsymbol{l}) = \prod_{k=1}^{K} p(\boldsymbol{c}_k|r_{\boldsymbol{c}},l_k). \tag{3}$$

Enfin, les positions sont contraintes à être dans l'intervalle $\mathcal{C} = [1,N]$. Par conséquent, les positions de la trajectoire k sont modélisées comme un champ de Markov gaussien 1D, la position de la première raie étant distribuée suivant une loi uniforme, d'où, pour $k \in \{1,\dots,K\}$:

$$\begin{aligned} p(\boldsymbol{c}_k|r_{\boldsymbol{c}},l_k) &= p(c_{k,1}) \times \prod_{m=2}^{l_k} p(c_{k,m}|c_{k,m-1},r_{\boldsymbol{c}},l_k) \\ &= \frac{1}{N-1} \times \frac{1}{(2\pi r_{\boldsymbol{c}})^{\frac{l_k-1}{2}}}\exp\left(-\frac{1}{2r_{\boldsymbol{c}}}\|\boldsymbol{D}\boldsymbol{c}_k\|^2\right)\mathbb{1}_{\mathcal{C}}(\boldsymbol{c}_k) \end{aligned} \tag{4}$$

où $\boldsymbol{D}$ définit une dérivée discrète d'ordre un (pour privilégier les évolutions droites) et $r_{\boldsymbol{c}}$ permet de régler la force de l'a priori. Comme $r_{\boldsymbol{c}}$ est petit, l'effet de la troncature sur $\mathcal{C}$ est négligeable et l'a priori peut être approximé par l'équation (4). De la même manière, les a priori sur les amplitudes $\boldsymbol{a}_k$ et largeurs $\boldsymbol{w}_k$ ($\forall k$) sont similaires, les hyperparamètres associés étant respectivement $r_{\boldsymbol{a}}$ et $r_{\boldsymbol{w}}$.

Autres lois a priori Le bruit est supposé blanc gaussien de moyenne nulle et de variance $r_{\boldsymbol{v}}$: $\forall s,n, (\boldsymbol{v}_s)_n|r_{\boldsymbol{v}} \sim \mathcal{N}(0,r_{\boldsymbol{v}})$. Des lois uniformes sur $\mathbb{R}^+$ sont choisies pour les hyperparamètres $r_{\boldsymbol{c}}$, $r_{\boldsymbol{a}}$, $r_{\boldsymbol{w}}$ et $r_{\boldsymbol{v}}$ (bien qu'elles soient impropres, la loi a posteriori sera propre).

3 Algorithme RJMCMC

La loi a posteriori est échantillonnée avec l'algorithme RJMCMC (*reversible jump Monte Carlo Markov chain*) [5] car l'exploration de la loi a posteriori est extrêmement difficile et ceci pour quatre raisons : la loi présente de nombreux minima locaux ; l'espace des solutions est de très grande dimension ; sa dimension est également inconnue ; sa dimension dépend

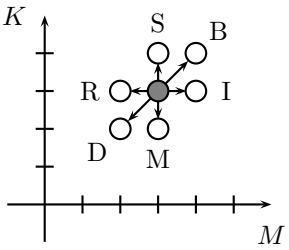

FIGURE 2 – Représentation des mouvements dans l'espace (M, K) à partir du cercle gris.

en fait de deux variables (M et K). Cette dernière raison est une particularité de notre problème qui, à notre connaissance, n'a jamais été traitée dans le cadre de l'algorithme RJMCMC. Ainsi, contrairement à l'algorithme RJMCMC traditionnel où un couple de mouvements suffit pour changer de dimension (généralement *birth* et *death*), il nous faut deux couples de mouvements de base. De plus, pour réduire le temps de calcul, nous utilisons également d'autres mouvements qui sont des combinaisons des mouvements de base (cf. figure 2). À ceux-ci, il faut enfin ajouter le mouvement d'échantillonnage de l'a posteriori à dimension constante pour aboutir finalement aux huit mouvements suivants :

- ajout (B : *birth*) et suppression (D : *death*) d'une trajectoire d'une raie (changement de dimension de (M, K) à $(M + 1, K + 1)$ ou $(M - 1, K - 1)$ respectivement) ;
- scission (S : *split*) d'une trajectoire en deux et fusion (M : *merge*) de deux trajectoires contigues en une unique trajectoire (saut de (M, K) à $(M, K + 1)$ ou $(M, K - 1)$) ;
- ajout (I : *increase*) et retrait (R : *reduce*) d'une raie à l'une des extrémités d'une trajectoire (saut de (M, K) à $(M + 1, K)$ ou $(M - 1, K)$) ;
- « ré-affectation » des raies (L : *labelling*), ce mouvement consiste à échanger des raies entre deux trajectoires ; il est équivalent à une combinaison de fusions et scissions mais ne change pas la dimension du problème car M et K ne sont pas modifiés ;
- mise à jour des paramètres inconnus $\boldsymbol{c}$, $\boldsymbol{a}$, $\boldsymbol{w}$, $r_{\boldsymbol{c}}$, $r_{\boldsymbol{a}}$, $r_{\boldsymbol{w}}$, $r_{\boldsymbol{v}}$ à dimension constante (U : *update*). Ce mouvement est effectué à l'aide d'un échantillonneur de Gibbs utilisant des simulations directes et des algorithmes de Metropolis-Hastings [16].

L'algorithme est initialisé avec une seule trajectoire contenant une unique raie et une estimation du maximum a posteriori est obtenue en récupérant l'échantillon généré le plus probable. Plusieurs critères de convergence existent dans la littérature, mais ils donnent souvent des résultats différents. Comme en plus l'espace des solutions est énorme (sa dimension est typiquement plus grande que 1000), il est difficile de déterminer la convergence. C'est pourquoi nous adoptons une approche conservative en fixant le nombre d'itérations manuellement.

Lois candidates mixtes Pour accélérer le temps de calcul, nous utilisons des lois candidates mixtes, c'est-à-dire que les candidats sont générés aléatoirement suivant une distribution uniforme ou une distribution définie à partir du modèle et/ou des données (*data/model-driven proposal*). En effet, une loi candidate uniforme permet d'explorer tout l'espace des solutions mais peut être très longue avant de proposer un bon candidat. En revanche, une distribution définie à partir du modèle et/ou des données a plus de chance de proposer un bon candidat, mais peut paradoxalement faire baisser le rapport d'acceptation. En effet, comme ce rapport est proportionnel à la probabilité du mouvement dual, il est difficile de sortir d'une configuration peu probable car le mouvement dual a peu de chance d'être proposé.

Ajout de perturbations Malgré nos efforts pour construire des mouvements et des propositions efficaces, l'algorithme n'explore pas toujours convenablement l'espace. La raison principale est que les mouvements ci-dessus sont locaux dans l'espace (M, K) et ne sont pas toujours suffisants : il serait donc nécessaire de mettre en œuvre des mouvements transdimensionnels qui puissent faire des sauts importants au dessus de configurations peu probables. Cependant, proposer de tels mouvements est difficile pour deux raisons. La première est que le nombre de variables à simuler est très important et peut être variable. La seconde est qu'il existe des dépendances très fortes entre raies d'une même trajectoire.

C'est pourquoi nous proposons une stratégie basée sur l'observation suivante : les problèmes de convergence n'apparaissent que très rarement à dimension constante. Au contraire, ils sont dus à une mauvaise estimation de la structure du modèle (variables $\boldsymbol{b}$ et $\boldsymbol{l}$ définissant la position des trajectoires dans les données), ce qui arrive lorsque les trajectoires sont proches les unes des autres. Notre stratégie consiste à considérer l'estimation $\boldsymbol{\theta}^{(1)}$ obtenue par l'algorithme RJMCMC comme étant un minimum local. On propose alors plusieurs états perturbés de $\boldsymbol{\theta}^{(1)}$ soit en reliant deux trajectoires éloignées de plusieurs spectres, ou en supprimant une partie de trajectoire (pas seulement sur un spectre) proche d'une autre, ou encore en ajoutant ou supprimant des raies à l'extrémité d'une trajectoire. Chaque état $\boldsymbol{\theta}_i^{(2)}$ obtenu ainsi permet d'initialiser un échantillonneur de Gibbs pour vérifier si la perturbation permet de s'échapper du potentiel minimum local. Si une meilleure estimation que $\boldsymbol{\theta}^{(1)}$ est obtenue, elle est utilisée comme initialisation d'un nouvel algorithme RJMCMC. Dans le cas contraire, la procédure est terminée et on conserve $\boldsymbol{\theta}^{(1)}$ comme estimation finale. On évite par ailleurs de générer des états $\boldsymbol{\theta}_i^{(2)}$ dont la structure (variables $\boldsymbol{b}$ et $\boldsymbol{l}$) a déjà été testée sur d'autres états : de cette manière, on s'assure que la procédure converge.

4 Résultats et conclusion

La méthode proposée a été appliquée sur la séquence de spectres de photoélectrons étudiée dans [11] constituée de 44 spectres (sur 3,47 ps) de $N = 181$ échantillons (de 0,02 eV à 2,49 eV) : cf. figure 4. La méthode a été codée en Matlab (le code et les données sont disponibles à l'adresse miv.u-strasbg.fr/mazet/jointdec). L'algorithme RJMCMC seul (sans perturbation) nécessite environ 4 h de calcul sur un poste de

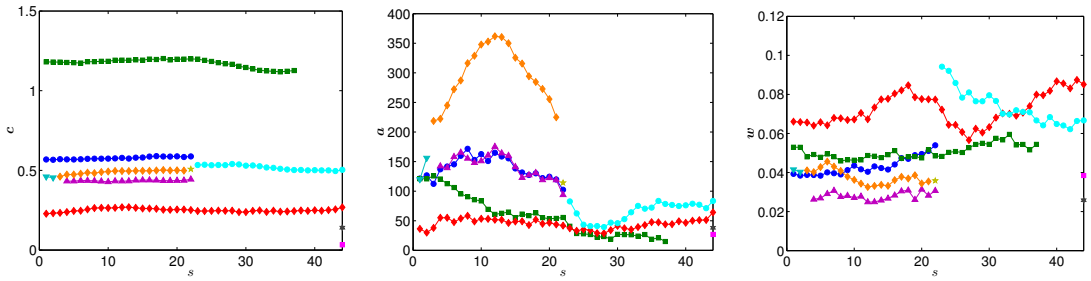

FIGURE 3 – Centres $\boldsymbol{c}$, amplitudes $\boldsymbol{a}$ et largeurs $\boldsymbol{w}$ des raies estimées. Chaque couleur correspond à une trajectoire.

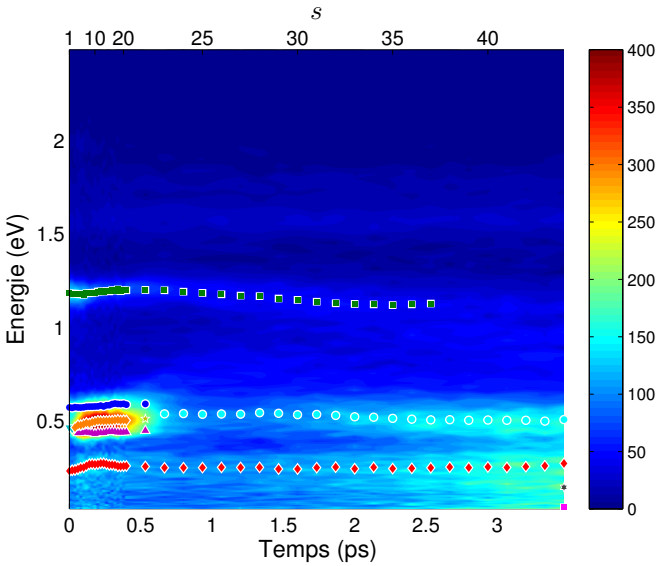

FIGURE 4 – Séquence réelle de spectres de photoélectrons : la distribution des électrons est tracée en fonction de l'énergie (n, axe vertical) et du temps (s, axe horizontal). Les points correspondent à la position estimée des raies.

travail classique, ce à quoi il faut ajouter le temps de calcul de chaque perturbation (qui dépend du nombre et de la configuration des trajectoires estimées) qui est d'environ 30 s. Les trajectoires des raies estimées sont représentées figures 3 et 4. Le nombre de trajectoires évolue dans le temps : la trajectoire à 1,2 eV disparaît après 2,5 ps et plusieurs raies évoluent autour de 0,5 eV et 0,3 eV. Ces résultats confirment quantitativement les observations des experts effectuées dans [11], à savoir que la dynamique présentée est le résultat de plusieurs processus simultanés, et que l'énergie (*i.e.* le centre) des raies varie dans le temps, conduisant à la disparition de raies. Enfin, notons qu'à 0,5 eV plusieurs raies apparaissent vers 0,2 ps dont la structure et l'intensité évolue très rapidement : aussi, les trajectoires estimées jusqu'à 0,5 ps sont remplacées par une unique trajectoire au-delà.

Références

[1] Y. BAR-SHALOM, F. DAUM et J. HUANG: The probabilistic data association filter. *IEEE Control Systems Magazine*, 29, 2009.

[2] M. DAVY, S.J. GODSILL et J. IDIER : Bayesian analysis of polyphonic western tonal music. *Journal of the Acoustical Society of America*, 119(4):2498–2517, 2006.

[3] R. FISCHER et V. DOSE : Analysis of mixtures in physical spectra. *Bayesian methods*, pages 145–154, 2001.

[4] C. GOBINET, V. VRABIE, M. MANFAIT et O. PIOT : Preprocessing methods of Raman spectra for source extraction on biomedical samples : Application on paraffin-embedded skin biopsies. *IEEE Transactions on Biomedical Engineering*, 56(5):1371–1382, 2009.

[5] P.J.GREEN: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 1995.

[6] S. GULAM RAZUL, W.J. FITZGERALD et C. ANDRIEU : Bayesian model selection and parameter estimation of nuclear emission spectra using RJMCMC. *Nuclear Instruments and Methods in Physics Research A*, 497:492–510, 2003.

[7] J. IDIER et Y. GOUSSARD : Markov modeling for Bayesian restoration of two-dimensional layered structures. *IEEE Transactions on Information Theory*, 39(4):1356–1373, 1993.

[8] N. KESHAVA et J.F. MUSTARD : Spectral unmixing. *IEEE Signal Processing Magazine*, 19(1):44–57, 2002.

[9] R.P.S. MAHLER : Multitarget Bayes filtering via first-order multitarget moments. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4), 2003.

[10] S.G. MALLAT et Z. ZHANG : Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[11] A. MASSON, L. POISSON, M.-A. GAVEAU, B. SOEP, J.-M. MESTDAGH, V. MAZET et F. SPIEGELMAN : Dynamics of highly excited barium atoms deposited on large argon clusters. I. General trends. *J. Chem. Phys.*, 133, 2010.

[12] V. MAZET : *Développement de méthodes de traitement de signaux spectroscopiques : estimation de la ligne de base et du spectre de raies*. Thèse de doctorat, Université Henri Poincaré, Nancy 1, 2005.

[13] V. MAZET : Joint Bayesian decomposition of a spectroscopic signal sequence. *IEEE Signal Processing Letters*, 2011.

[14] V. MAZET, S. FAISAN, A. MASSON, M.-A. GAVEAU, L. POISSON et J.-M. MESTDAGH : Joint Bayesian decomposition of a spectroscopic signal sequence with RJMCMC. *In IEEE Workshop on Statistical Signal Processing*, Ann Arbor, USA, 2012.

[15] H.-L. NGUYEN THI et C. JUTTEN : Blind source separation for convolutive mixtures. *Signal Processing*, 45(2):209–229, 1995.

[16] C. ROBERT et G. CASELLA : *Monte Carlo statistical methods*. Springer, 2e édition, 2004.