

Comment la chimie analytique peut contribuer à l'étude des images astronomiques

Danielle NUZILLARD¹,

¹Laboratoire d'Automatique et de Micro-électronique
Université de Reims Champagne-Ardenne, Moulin de la Housse, BP 1039, F-51687 Reims Cedex 2, France
Danielle.Nuzillard@univ-reims.fr

Résumé – De nombreux travaux théoriques, validés par des réalisations pertinentes, ont contribué à l'essor des méthodes de séparation aveugles de sources (SAS). Un aperçu de la diversité des applications illustre les possibilités offertes en la matière. Deux applications spécifiques ayant trait à l'analyse spectroscopique par RMN et à l'astrophysique, caractérisées par des contraintes communes, sont présentées pour mettre en évidence les problèmes de modélisation, du choix de l'algorithme et les transformations associées indispensables pour justifier les hypothèses de la SAS.

Abstract – Many theoretical works, giving rise to relevant achievements, contributed to the development of the blind source separation methods (BSS). An outline of the diversity of the applications illustrates the offered possibilities. Two specific applications related to the NMR spectroscopy and astrophysics, that share features, are presented. They highlight the problems of specific mixture modeling, of the algorithm selection and of data transformation requirements so that the hypothesis of BSS can be fulfilled.

1 Introduction

Le concept d'Analyse en Composantes Indépendantes (ACI) et de Séparation Aveugle de Sources (SAS) a été introduit par Ch Jutten et J. Héroult [20], et développé par P. Comon [5]. Sa finalité est de rechercher dans des données, la structure la plus indépendante possible. L'essor de ces méthodes a été favorisé par la création en 1999 du symposium international ICA : [9], [10], [11], [12]. Les serveurs et pages : [14], [15], [16], [20], [18], [19] fournissent de nombreux points d'entrée ainsi que les ouvrages [4], [13] qui regroupent des informations détaillées sur les algorithmes et les modèles de base.

Les premières applications ont émergé dans des domaines aussi divers que la neurobiologie, la parole, les télécommunications [7], [8], le traitement d'antenne. L'instrumentation en général a mis en oeuvre ces méthodes pour la prospection sismique, les vibrations des machines, les applications biomédicales (EEG, MEG, ECG, EMG, ...), l'analyse physico-chimique. D'autres travaux ont été développés en astrophysique [23], [1], [6], pour la décomposition en bases naturelles de signaux sonores ou d'images, pour le traitement des textes ou le multimédia. L'image est source de nombreuses applications : segmentation d'images texturées, filigrane, traitement des réflexions. L'imagerie multi et hyperspectrale en télédétection [21], pour la cartographie des ressources, en agriculture ou dans le domaine du vivant (biologie, médecine, ...) est un thème prometteur ne serait-ce que pour ses besoins en réduction de dimensionnalité.

Dans toutes ces situations, il est indispensable que les données puissent se ramener à des mélanges qui vérifient les conditions d'application de la SAS. Les deux types d'application qui suivent, valident ces méthodes sur des problèmes réels et illus-

trant comment il est possible de tenir compte du contexte physique dans la recherche d'une représentation des sources. En outre, les acquis des applications en *Analyse spectrale de mélange de petites molécules par résonance magnétique nucléaire (RMN)* [26], [22], [28] ont été transposés dans celle concernant *l'Identification des composantes d'une galaxie* [23].

La section 2 regroupe les considérations communes aux deux applications : la problématique, les caractéristiques communes des données, le choix de l'algorithme et les développements qui en découlent. Les sections 3 et 4 sont plus spécifiques à chacune des applications. La section 5 tire les conclusions et présente des perspectives.

2 Liens entre les applications

2.1 Problématique

De manière générale, plusieurs sources émettent N signaux $S = (S_1, \dots, S_N)^T$ qui sont reçus après leur propagation en M mélanges linéaires bruités $X = (X_1, \dots, X_M)^T$ sur un réseau de capteurs, où $M > N$ et A est la matrice de mélange. L'objectif est de remonter aux signaux émis. La matrice A est identifiée à un facteur scalaire près et à une permutation près [5], [20].

2.2 Caractéristiques des données

Pour chacune des applications suivantes, les données sont modélisables sous la forme de mélanges physiques instantanés bruités : les sources inconnues recherchées ont une origine physique et leurs coefficients de mélange, inconnus eux aussi, sont des proportions donc des coefficients positifs [26], [29]. Les sources (1D ou 2D) sont des signaux réels donc bruités,

inconnus certes, mais dont le contenu intrinsèque est temporellement ou spatialement corrélé; par ailleurs les mélanges présentent une information spectrale localisée. Cette description conditionne le choix de l'algorithme.

2.3 Choix de l'algorithme

Les méthodes à l'ordre 2 prennent en compte la corrélation spatiale ou temporelle des données [2]. La diagonalisation conjointe des matrices de corrélation fournit des solutions acceptables pour les signaux bruités. Aussi un algorithme du type SOBI [2] est tout à fait adapté à la situation.

Par ailleurs, pour les applications traitées, la contrainte d'orthogonalité des sources a été relaxée au profit de celle de positivité des coefficients des mélanges et des sources quand celle-ci est vérifiée [26].

La nature même des données a suggéré de nouveaux développements avec l'algorithme f -SOBI [28], dérivé de SOBI. L'algorithme f -SOBI utilise les matrices de corrélation des données pour séparer dans l'espace de Fourier de celles-ci. Ce nouvel algorithme est tout particulièrement indiqué pour sa rapidité et sa robustesse lorsque l'information spectrale est concentrée dans peu de raies alors que l'information temporelle couvre l'ensemble de l'enregistrement comme c'est le cas pour les deux types applications présentées.

Le traitement des spectres de raies superposés au bruit de capteur, par les méthodes basées sur les statistiques d'ordre supérieur, conduit à des résultats peu satisfaisants car le bruit fausse l'évaluation des propriétés statistiques des signaux d'intérêt.

3 Analyse spectrale de mélanges

3.1 Contexte

3.1.1 Description des données

La RMN est une technique d'analyse physico-chimique qui permet d'établir des structures moléculaires. Un spectromètre de RMN fournit des données brutes qui sont des signaux temporels à 1, 2 ou 3 dimensions. Les spectres sont produits par transformation de Fourier des données temporelles. L'axe des fréquences reflète l'environnement électronique des noyaux atomiques de l'échantillon étudié et constitue une source d'information structurale de première importance.

La RMN du ^{13}C , est particulièrement adaptée à la situation, sachant que les spectres produits sont formés de raies fines en nombre limité. Par contre, le rapport signal sur bruit est faible. Par comparaison la RMN du proton ^1H fournit des spectres aux raies larges concentrées sur un domaine spectral réduit. L'exploitation de données de la RMN du ^1H est toutefois facilitée par l'introduction d'une seconde dimension temporelle en RMN 2D. Un spectre 2D comporte deux axes de fréquence. Un pic spectral est produit lorsque deux noyaux présentent un couplage magnétique, couplage dont la détection constitue une source d'information structurale. La répartition de l'informa-

tion sur une surface rend moins probable les superpositions de pics, permettant de valider plus aisément qu'en RMN 1D l'hypothèse d'indépendance des sources.

3.1.2 Validation des hypothèses

Les signaux temporels fournis par le spectromètre sont modélisables par une somme de fonctions sinusoïdales amorties à laquelle s'ajoute un bruit gaussien de capteur. Les raies spectrales sont modélisables par une somme de lorentziennes et l'application de la SAS à l'analyse de mélanges par RMN a été conditionnée par le respect des contraintes de linéarité et d'indépendance statistique des sources.

Il y a linéarité si le spectre d'un mélange est réellement la combinaison des spectres des produits purs avec comme coefficients leurs concentrations. Cela n'est vrai en toute rigueur qu'à dilution infinie. Si l'écart à la situation idéale est important, il faut faire appel à un schéma d'analyse des données plus complexe.

L'indépendance statistique des sources implique l'orthogonalité de leurs signaux. Plus les lignes spectrales sont étroites, plus les signaux ont une forte probabilité d'être indépendants.

3.2 Illustration des problèmes de modélisation

Les applications concernent l'aide à la détermination de la structure :

- d'une molécule,
- des molécules entrant dans la composition d'un seul mélange,
- des molécules entrant dans la composition de plusieurs mélanges.

Chacune d'elle soulève des problèmes particuliers illustrés ci-dessous qui n'auraient pas pu être résolus sans les méthodes de séparation de sources. Des difficultés supplémentaires liées au non respect de l'indépendance stastistique sont également mises en évidence.

3.2.1 Structure d'une molécule

L'objectif est l'identification des groupes CH_n , n étant le nombre d'atomes d'hydrogène portés par un carbone. L'échantillon est excité par une séquence d'impulsion reliée à un paramètre d'angle θ dont la valeur n'est pas connue avec précision. Cette séquence conditionne l'intensité des signaux émis en réponse par les groupes CH_n , l'intensité des raies spectrales est modulée par la valeur des angles θ et dépend de n . De plus, la linéarité des mélanges n'est pas forcément respectée à cause d'instabilités instrumentales.

3.2.2 Mélanges de produits

Composition d'un seul mélange physique. Il existe des situations pour lesquelles il n'est pas possible d'obtenir les composants purs à l'état liquide. C'est le cas du glucose pour lequel il se produit une réaction d'isomérisation de l' α -glucose en β -glucose dès qu'il est dissout dans l'eau (D_2O) modifiant ainsi

les concentrations des produits. Celles-ci entraînent un glissement des fréquences des réponses et introduisent des erreurs de non-linéarité dans l'évaluation des sources.

Composition de plusieurs mélanges. Si n substances sont présentes dans m mélanges linéairement indépendants, les spectres obtenus sont idéalement des combinaisons linéaires des spectres des substances isolées lorsqu'il n'y a pas d'interactions entre les molécules, donc pas de modification des fréquences des signaux. L'information est étendue dans un plan (RMN 2D) pour s'affranchir des possibles recouvrements des spectres. Le traitement des non-linéarités fait l'objet de pré et post traitements spécifiques.

3.3 Résolution des problèmes de modélisation

La notion même de mélange a été considérée de trois manières différentes :

- le spectre d'une substance organique unique est décomposable par exemple en sous-spectres des groupes CH, CH₂, CH₃,
- à partir d'un seul mélange de n substances, il est possible de produire m spectres ($m > n$) où les composantes vont répondre spécifiquement à un paramètre extérieur ajustable,
- n substances sont présentes dans m mélanges.

Les erreurs de modélisation introduites par l'instrumentation ou par le non respect par le mélange expérimental des hypothèses de la SAS influent directement sur la qualité des résultats, ainsi il est nécessaire de développer des pré-traitements pour y remédier.

D'abord, l'**indépendance statistique** est améliorée en produisant des spectres à deux dimensions. La résolution de la spectroscopie 1D étant supérieure à celle 2D [27], celle-ci est moins sujette aux instabilités instrumentales.

Les **instabilités fréquentielles** liées soit à des instabilités instrumentales, soit à des glissements de fréquence des spectres des mélanges, sont prises en compte. Pour cela, il ne faut pas considérer les signaux fréquentiels eux-mêmes, mais l'intégrale du signal sur des fenêtres bien définies. Pour chaque mélange, les valeurs des l'intégrales fournissent un vecteur de données réduites [22]. L'ensemble de ces vecteurs est séparé. Un post-traitement permet de reconstruire des spectres 2D pertinents.

La **positivité** des signaux et des mélanges remplace la contrainte d'orthogonalité s'il existe un sous-domaine spectral pour lequel les sources sont effectivement indépendantes. [26]. La positivité a été utilisée par de nombreux auteurs, en particulier dans [30], dans le cadre de l'imagerie par RMN.

Enfin le **spécialiste**, dont le rôle est déterminant pour le choix de la technique expérimentale à mettre en oeuvre, **valide** les résultats soit par comparaison avec les spectres purs, soit par des considérations physico-chimiques. Dans ces exemples, il est remarquable que la technique permette à partir d'un seul échantillon d'obtenir un ensemble de signaux de mélanges adapté au traitement mathématique des séparations.

4 Astronomie

4.1 Contexte

En astronomie, l'analyse des images a pour but d'extraire des structures caractéristiques qui peuvent aider les astronomes à améliorer la connaissance des corps célestes. Cette application a pour objet de tester la SAS en tant qu'outil exploratoire sur un jeu d'images multi-spectrales d'une galaxie. Les images multi-spectrales sont produites grâce à des détecteurs à transfert de charges et à des filtres colorés de caractéristiques spectrales bien définies.

4.2 Préparation des données

Souvent l'intensité du pixel provient d'un mélange. On admet d'une part que ce mélange est linéaire, et d'autre part que chaque image provient de sources physiquement indépendantes. Le bruit de détecteur est faible tandis que celui de photon est dominant dans la région de la galaxie. Une transformation d'Anscombe stabilise sa variance [23]. La valeur du fond est estimée et soustraite. Après cette préparation, la linéarité du mélange est quelque peu perturbée, réduisant la dynamique des données et en particulier le rôle du noyau de la galaxie.

4.3 Traitements

Divers algorithmes ont été testés pour extraire des sources des images pré-traitées. Les meilleurs résultats [23] de qualité équivalente sont fournis par f-SOBI (séparation dans le domaine des fréquences spatiales) avec 16 matrices de corrélation, SOBI-2D (séparation dans le domaine des coordonnées spatiales) avec 8 matrices et fastICA avec $g(y) = \tan(y)$. Des tests complémentaires sur des données simulées bruitées semblables aux données célestes ont confirmé la pertinence et la robustesse de f-SOBI [3].

Les techniques de séparation ont été appliquées à la galaxie de Seyfert 3C120 observée par le télescope spatial Hubble. Elles ont permis de mettre en évidence trois sources significatives : la région centrale de la galaxie, les régions ionisées entourant le noyau, les anneaux correspondant à la fonction d'étalement de la caméra pour la longueur d'onde de la raie H_{α} . Ce modèle a été validé par des considérations astrophysiques.

Pour améliorer la décomposition, il faut disposer de plus d'images, mais la SAS peut-être considérée comme un outil exploratoire intéressant qui peut suggérer une nouvelle approche analytique. Son usage est bien adapté à l'extraction d'information à partir de grandes bases de données.

5 Conclusion et Perspectives

Ces deux applications illustrent les possibilités offertes par les méthodes de SAS. Pour l'analyse spectrale, les pré-traitements sont presque indispensables pour transformer les données pour vérifier les hypothèses de la SAS. Par compte, en astrophysique, la réduction du bruit a été faite au détriment de l'exactitude du mélange linéaire. Néanmoins, l'approche reste va-

lable dans ce cadre où la SAS se positionne comme un outil d'analyse exploratoire. Il est probable que la qualité des sources serait améliorée si l'on disposait de plus de données et que ces techniques sont performantes sur des données de types multi/hyper spectrales.

Il n'existe pas de modèle universellement réaliste pour séparer des signaux. Il est nécessaire d'étendre ces méthodes de SAS à des modèles plus complexes et plus réalistes dans le cadre d'applications bien spécifiques en collaboration avec des spécialistes des autres disciplines pour inclure plus des contraintes structurelles ou même des informations marginales sur les sources. Par opposition la compression des données, l'objectif est de minimiser l'information mutuelle conduisant à une représentation mathématique sans aucune signification physique.

La SAS est un outil supplémentaire qui doit faire partie intégrante du traitement du signal. Il existe des algorithmes performants, il est maintenant indispensable de développer les étapes de pré-traitement (filtrage, débruitage, transformations, ...) et de post-traitement pour que les données remplissent les conditions de séparabilité et livrent des résultats exploitables par les spécialistes. Par ailleurs, les mélanges non-linéaires, dépendants du temps ou convolutifs se rencontrent fréquemment dans des contextes réels. Leur traitement nécessite l'exploration de voies nouvelles de recherche.

Références

- [1] C. Baccigalupi, L. Bedini, et al. *Neural networks and separation of background and foregrounds in astrophysical sky maps*, Mon. Monthly Notices of the Royal Astronomical Society, 318(3), pp. 769-780, 2000.
- [2] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. *A blind source separation technique using second-order statistics*, IEEE Trans. SP, vol. 45, pp. 434-444, 1997.
- [3] A. Bijaoui, D. Nuzillard. *Smoothing and Adaptive Denoising for Blind Source Separation*, ICA'01 San Diego USA, pp. 96-101, 9-12 dec 2001.
- [4] A. Cichocki, S. Amari. *Adaptive Blind Signal and Image Processing*, John Wiley & Sons, 2002.
- [5] P. Comon. *Independent component analysis, A new concept?*, Signal Processing, vol. 36, pp. 287-314, 1994.
- [6] J. Delabrouille, J.-F. Cardoso and G. Patanchon. *Multi-detector multi-component spectral matching and applications for CMB data analysis*, soumis à Mon. Monthly Notices of the Royal Astronomical Society, 2003. Disponible à <http://arxiv.org/abs/astro-ph/0211504>
- [7] N. Delfosse and P. Loubaton. *Adaptive blind separation of independent sources: a deflation approach*, Signal Processing, vol. 45, pp. 59-83, 1995.
- [8] Y. Deville and J. Damour and N. Charkani, *Multi-Tag Radio-Frequency Identification Systems Based on New Source Separation Neural Networks*, Neurocomputing, vol. 49, pp. 369-388, 2002.
- [9] ICA'99, ICA and BSS, INPG, St Martin d'Hères, France, pp. 261-266, january 11-15, 1999.
- [10] ICA'00, ICA and BSS, University of Technology, pp. 33-38, Helsinki June, 2000.
- [11] ICA'01, 3rd Int. Workshop on ICA and BSS, ucsd, San Diego USA, 9-12 dec. 2001.
- [12] ICA'03 4th Int. Symposium on ICA and BSS, <http://ica2003.jp/> Nara Japon, April 1-4, 2003.
- [13] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*, John Wiley & Sons, 2001.
- [14] <http://www.bsp.brain.riken.go.jp/ICALAB/>
- [15] <http://www.cis.hut.fi/projects/ica/>
- [16] http://www.cnl.salk.edu/~tewon/ica_cnl.html
- [17] http://www.lis.inpg.fr/demos/sep_sourc/ICAdemo/
- [18] <http://www.sccn.ucsd.edu/~scott/ica.html>
- [19] <http://www.tsi.enst.fr/icacentral/>
- [20] Ch. Jutten, J. Héroult. *Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture*, Signal Processing, vol. 24, pp. 1-10, 1991.
- [21] M. Lennon, G. Mercier, M.C. Mouchot, L. Hubert-Moy. *Independent component analysis as a tool for the dimensionality reduction and the representation of hyperspectral images*, in IGARSS, 2001.
- [22] D. Nuzillard. *Adaptation de SOBI à des données fréquentielles*, GRETSI'99, pp. 745-748, 13-17 sept. 1999.
- [23] D. Nuzillard, A. Bijaoui. *Blind Source Separation and Analysis of multispectral Astronomical Images*, Astronomy & Astrophysic, Sup. Series, vol. 147, pp. 129-138, 2000.
- [24] D. Nuzillard, S. Bourg, J.-M. Nuzillard. *Model-free Analysis of Mixtures by NMR*, Journal of Magnetic Resonance, vol. 133, pp. 358-363, 1998.
- [25] D. Nuzillard, J.-M. Nuzillard, *Application de la séparation aveugle de sources à la RMN*, Hermès-Paris, C2I'98, pp. 435-442, 18-19 nov. 1998.
- [26] D. Nuzillard, J.-M. Nuzillard. *BSS applied to non-orthogonal signals*, ICA'99, Aussois, France, pp. 25-30, 11-15 janvier 1999.
- [27] D. Nuzillard, J.-M. Nuzillard. *Application of Blind Source Separation to 1-D and 2-D NMR Spectroscopy*, IEEE Signal Processing Letter, vol. 5, pp. 209-211, aug. 1998.
- [28] D. Nuzillard, J.M. Nuzillard. *Second Order Blind Source Separation*, Signal Processing, vol. 83(3), pp. 627-631, march 2003.
- [29] M. D. Plumbey. *Conditions for nonnegative independent component analysis*, IEEE Signal Processing Letters, vol. 9, no. 6, pp. 177-180, june 2002.
- [30] *Recovery of constituent spectra in 3D chemical shift imaging using non-negative matrix factorization*, P. Sadja, S. Du, et al. ICA'03 Nara Japon, April 1-4, 2003.