Séparation de Sources tenant compte de la Corrélation Temporelle ou Spatiale

Danielle NUZILLARD

Laboratoire d'Automatique et de Microélectronique
UFR Sciences Exactes et Naturelles, Moulin de la Housse, BP 1039, 51687 Reims Cedex 02, France
Danielle.Nuzillard@univ-reims.fr

Résumé – Cette communication traite de la recherche de sources qui tiennent compte de l'agencement temporel ou spatial des données. La séparation a lieu soit dans l'espace direct, soit dans l'espace de Fourier. Les corrélations associées sont prises dans l'espace où elles sont le plus pertinentes. Les données peuvent être réelles ou complexes. Il est possible d'introduire des contraintes de positivité sur les mélanges et/ou sur les sources. Les algorithmes de séparations sont testés sur des données simulées pour évaluer leur qualité.

Abstract – This communication deals with the search of sources which take into account the temporal or spatial organisation. The separation takes place as well as in the direct space and in the Fourier space. The associated correlations are calculated in the space the one they are the more pertinent. Data is real or complex. It is possible to introduce a positivity contraint on both the mixtures and the sources. The algorithms are tested on simulated data to allow an evaluation quality.

1 Introduction.

La séparation de sources consiste à rechercher une description élémentaire d'un phénomène physique. Une classe d'algorithmes de séparation de sources utilise les statistiques du second ordre pour fournir des sources non corrélées. Si les densité de probabilité des sources sont gaussiennes, celles-ci sont alors indépendantes. Si non, une voie possible est de rechercher des sources les plus indépendantes possibles. L'information mutuelle peut-être utilisée comme une mesure de dépendance des i sources notées S_i. L'information mutuelle d'un ensemble de I sources $\{S_1, ..., S_i, ..., S_I\}$, notée IMS est définie comme il suit :

$$IMS = \sum_{n_1,\ldots,n_I} p(n_1,\ldots,n_I) \log_2 \frac{p(n_1,\ldots,n_I)}{\prod_{i=1,I} p_{S_i}(n_i)} \quad (1)$$

qui est la divergence de Kullback-Leibler entre la loi de probabilité conjointe $p(n_1,\ldots,n_I)$ et la loi de probabilité obtenue à partir des probabilités marginales $p_{S_i}(n_i)$. Si les sources sont indépendantes, la probabilité conjointe est le produit des probabilités marginales et cette divergence est égale à 0. La propriété inverse est aussi vraie.

En supposant que le vecteur X des phénomènes physiques observés s'écrit :

$$X = AS + N \quad (2)$$

où A est la matrice de mélange, S le vecteur des sources et N celui du bruit additif et en négligeant le bruit, il vient :

$$S \approx BX. \quad (3)$$

où B est la matrice pseudo-inverse de A. L'entropie de l'ensemble des sources $E(S_1,\ldots,S_I)$ est l'entropie de l'ensemble des mélanges $E(X_1,\ldots,X_N)$ plus le logarithme du Jacobien de la transformation :

$$E(S_1,\ldots,S_I) = E(X_1,\ldots,X_N) + log_2|\det B| \quad (4)$$

d'où :

$$IMS = \sum_i E(S_i) - E(X_1,\ldots,X_N) - log_2|\det B| \quad (5)$$

Pour obtenir les sources, il n'est pas utile de recourir aux lois de probabilités conjointes, il suffit de maximiser la fonction [5] :

$$C = -\sum_i E(S_i) + log_2|\det B| \quad (6)$$

C est une fonction de contraste dont le cadre a été formalisé par P. Comon [5] et qui dépend de l'évaluation $E(S_i)$ de l'entropie de chaque source i.

L'équation (4), et par conséquence les équations (5) et (6) sont vérifiées seulement comme étant la limite pour un nombre élevé de pixels. Ce nombre doit être tel qu'il permette d'obtenir une fonction de densité de probabilité (FDP) expérimentale avec un très petit intervalle d'échantillonnage. Or la densité de probabilité des sources n'est jamais connue précisément. Le critère à optimiser est toujours une approximation. Il est possible de rechercher l'indépendance statistique en introduisant les statistiques d'ordre supérieur grâce aux fonctions seuils d'un réseau de neurones ou à l'optimisation de fonctions vérifiant les propriétés des fonctions de contraste. Celles-ci ne tiennent pas compte de l'agencement spatial or temporel des données. Or, cette organisation peut jouer un rôle très important car on ne sépare pas un bruit vérifiant ayant des propriétés statistiques mais des signaux ayant une certaine cohérence. En prenant en compte cette corrélation intrinsèque des données, un algorithme au second ordre peut

fournir de meilleures performances qu'un algorithme basé sur les statistiques d'ordre supérieur.

2 Corrélation temporelle ou spatiale.

SOBI est un algorithme du second ordre [1] très performants sur des signaux physiques. La séparation a lieu dans l'espace direct i.e. l'espace des données. La matrice issue d'un décalage nul est diagonalisée de façon à obtenir une matrice orthogonale. Celle-ci est appliquée sur les données. Ensuite la matrice de séparation et la matrice de mélange sont obtenues à une constante et à une permutation près. Pour cela, le critère utilisé consiste à maximiser la corrélation intrinsèque des signaux et minimiser la corrélation extrinsèque entre les signaux via un algorithme de diagonalisation conjointe [4] de p matrices de variance covariance des signaux et des signaux décalés.

Cet algorithme a fourni d'excellentes séparations après un pré-traitement des données en analyse chimique par Résonance Magnétique Nucléaire [7]. Une variante, notée *f-SOBI*, de l'algorithme a été ensuite aménagée pour séparer des données dans l'espace de Fourier [8]. Devant la pertinence des résultats, une autre variante a été développée, puis testée avec succès comme outil exploratoire en astronomie [9] en analyse d'images multispectrales (versions *f-SOBI* 2-*D* et *SOBI* 2-*D*). Pour le spectrographe l'espace direct est celui des fréquences et l'espace de Fourier est le temps. Pour le traiteur d'images, l'espace direct est celui de l'image, l'espace de Fourier est celui des fréquences spatiales.

2.1 Une Dimension.

2.1.1 Espace de Fourier.

La transformée de Fourier est une opération linéaire, la matrice de séparation peut donc être obtenue à partir de l'espace direct ou de Fourier et s'appliquer dans n'importe quel espace. La séparation dans l'espace de Fourier est pertinente dans les cas répertoriés ci-dessous.

- Les données temporelles nécessitent des pré et/ou post traitements dans l'espace de Fourier,
- Le nombre d'échantillons pertinents F en fréquence est faible devant le nombre d'échantillons T de l'espace temporel,
- Le nombre d'échantillons F est suffisamment faible devant le nombre p des matrices de variance-covariance et introduit un biais dans l'estimation de ces dernières,
- Les données ne sont pas corrélées dans l'espace direct mais le sont dans leur espace de Fourier, ou ont une corrélation plus importante dans cet espace que dans le direct. En particulier, cette approche se révèle utile pour des signaux à énergie à très basse fréquence qui nécessitent la prise en compte de corrélation à longue distance car les moments d'ordre élevé ne sont pas définis.

2.1.2 Coefficients de corrélation de *f-SOBI*.

Deux spectres a et b sont échantillonés des intervalles de fréquences équidistants Δf. Les échantillons sont indicés par i, avec $0 \leq i < F$. Les échantillons correspondants dans le domaine temporel sont notés x_j and y_j :

$$x_j = \sum_{i=0}^{F-1} a_i w^{ij} \quad \text{et} \quad y_j = \sum_{i=0}^{F-1} b_i w^{ij} \quad (0 \leq j < \infty) \quad (7)$$

où $w = \exp(2i\pi/F)$. La fonction de corrélation $R^f_{xy}(\tau)$ entre deux signaux x et y est définie par :

$$R^f_{xy}(k\Delta t) = \lim_{T\to\infty} \frac{1}{T} \sum_{j=0}^{T-1} x_j . y^*_{j-k} \quad (8)$$

où Δt est l'intervalle d'échantillonnage : $\Delta f.\Delta t = 1$, et k l'indice qui définit le décalage temporel $\tau = k.\Delta t$. L'exposant f réfère à f-SOBI. Compte tenu de la cyclicité de la TF, il vient $x_{j+F} = x_j$ et $y_{j+F} = y_j$, l'expression se réduit à :

$$R^f_{xy}(k\Delta t) = \frac{1}{F} \sum_{j=0}^{F-1} x_j . y^*_{j-k}. \quad (9)$$

En introduisant l'équation (7) dans (9), il vient :

$$R^f_{xy}(k\Delta t) = \frac{1}{F} \sum_{j=0}^{F-1} a_j . b^*_j . w^{jk} \quad (10)$$

qui est la relation bien connue de la fonction de corrélation dans l'espace de Fourier des données [8]. Cette expression est utilisée dans *f-SOBI*.

2.2 Deux Dimensions.

A partir d'une image, un vecteur signal peut être construit par concaténation des pixels des lignes ou des colonnes. L'organisation spatiale n'est alors prise en compte que dans une seule direction. Si les effets de bord des images sont négligeables parce que le fond est régulier c'est suffisant. Dans le cas général, et compte tenu du sens physique de la corrélation, il est nécessaire de considérer l'image dans sa globalité, i.e. dans les deux dimensions. Le choix des corrélations spatiales entre deux images dans les différentes directions : horizontale, verticale, diagonale, a été rendu paramétrable. De même que pour les signaux (1-*D*), certaines images compte-tenu de leur information spatiale ont intérêt à être traitées dans l'espace de Fourier. Les coefficients des matrices de corrélation ont les expressions suivantes.

Dans l'Espace de Fourier (*f-SOBI* 2*D*) :

$$R^f_{xy}(k_1, k_2) = \frac{1}{F_1 F_2} \sum_{j_1=0}^{F_1-1} \sum_{j_2=0}^{F_2-1} a_{j_1 j_2} . b^*_{j_1 j_2} . w_1^{j_1 k_1} w_2^{j_2 k_2} \quad (11)$$

où k_1 et k_2 sont les décalages de fréquences spatiales, F_1 et F_2 sont les nombres de ligne et de colonne, j_1 et j_2 sont les indices des pixels des images a et b, $w_1 = exp(2i\pi/F_1)$ et $w_2 = exp(2i\pi/F_2)$.

Dans l'Espace direct ($SOBI$ $2D$) :

$$R_{xy}(j_1, j_2) = \frac{1}{F_1 F_2} \sum_{k_1=0}^{F_1-1} \sum_{k_2=0}^{F_2-1} x_{k_1 k_2} . x^*_{k_1-j_1, k_2-j_2} \tag{12}$$

où j_1 et j_2 sont les décalages de l'image.

De même que pour les signaux, il est possible de choisir l'espace des corrélations des images le mieux approprié pour l'application à traiter.

2.3 Signaux réels ou complexes.

Dans SOBI, le critère suivant est minimisé :

$$\mathcal{C}(\mathcal{M}_{\mathcal{R}}, V) = \sum_{j=1}^{J} off(V^H . R_X(j) . V) \tag{13}$$

où *off* est l'ensemble des termes hors diagonale et où $V = \{v_1, v_2, ..., v_N\}$ est la base qui minimise le critère. Cette base s'obtient grâce à une extension de l'algorithme de Jacobi, en recherchant étape par étape une matrice unitaire V qui est appliquée à tous les éléments repérés par le même indice [4]. Le vecteur v associé à la plus grande valeur propre λ minimise le critère. Les termes de la matrice de rotation complexe vérifient les relations :

$$\cos\theta = \sqrt{\frac{1+v(1)}{2}}; \tag{14}$$

$$\sin\theta e^{-i\varphi} = \frac{v(2) - iv(3)}{2\cos\theta}; \tag{15}$$

$$\sin\theta e^{+i\varphi} = \text{conjugué}(\sin\theta e^{-i\varphi}) \tag{16}$$

Pour traiter des données réelles, il suffit de choisir la matrice V réelle. Alors, les variantes algorithmiques qui fournissent des signaux réels utilisent :

$$\cos\theta = \sqrt{\frac{1+v(1)}{2}}, \quad \sin\theta = \frac{v(2)}{2\cos\theta} \tag{17}$$

2.4 Contrainte de Positivité.

En analyse par Résonance Magnétique Nuclaire, les signaux traités sont des mélanges positifs de spectres de raies positives. En analyse d'images du ciel, les photons proviennent de différents objets célestes. Ils sont reçus sur le CCD à travers un jeu de filtres de couleur. Ils se comportent en première approximation comme un mélange linéaire positif de contributions positives. Pour ces deux applications, la contrainte de positivité a été introduite au détriment de la contrainte d'orthogonalité puisque celle-ci n'est pas vérifiée.

L'algorithme de séparation aveugle de sources (SAS) est appliqué, suivi d'une procédure dite ALS : 'Alternated Least Squares' [6]. Si $X = \hat{A}\hat{S}$, $\hat{A}$ (resp. $\hat{S}$) peuvent être évaluées à partir de $\hat{S}$ (resp. $\hat{A}$) et de X au sens des moindres carrés par les relations :

$$\hat{S} = ({}^t\hat{A}.\hat{A})^{-1} . {}^t\hat{A}.X, \tag{18}$$

$$\hat{A} = X . {}^t\hat{S} . (\hat{S} . {}^t\hat{S})^{-1}. \tag{19}$$

Les vraies sources sont obtenues par l'algorithme ALS grâce au processus itératif suivant :

1. Mise à zéro des parties négatives de $\hat{S}$,
2. Estimation de $\hat{A}$,
3. Mise à zéro des coefficients négatifs de $\hat{A}$,
4. Estimation de $\hat{S}$,
5. Retour au point 1 si la convergence n'est pas assurée.

Le processus est arrêté quand aucune évolution du résultat n'est perceptible. Par expérience personnelle la poursuite du processus dégrade les résultats et une ou deux itérations sont suffisantes.

3 Qualité des sources restaurées.

3.1 Indice de Gini.

La qualité des séparations a été évaluée à partir d'un jeu de données simulées. Les vrais coefficients de mélange $\overline{a}_{ij}$ sont donc connus. Un critère basé sur l'estimation de la concentration de l'énergie en fonction des sources originales semble pertinent [2]. En négligeant le bruit, il vient :

$$X = \overline{A}\,\overline{S} \tag{20}$$

où $\overline{A}$ est la vraie matrice de mélange et $\overline{S}$ le vrai vecteur source. Alors :

$$S = B\overline{A}\,\overline{S} \tag{21}$$

Si la séparation est parfaite, la matrice $D = B\overline{A}$ est le produit d'une matrice diagonale et d'une matrice de permutation. Si non, les coefficients sont dispersés. Les sources restaurées étant une combinaison des sources originales, chaque source restaurée S_j s'écrit :

$$S_j = \sum_{j'} d_{jj'} \overline{S}_{j'} \tag{22}$$

Les termes d'énergie $e_{j'} = d^2_{jj'}$ sont calculés et triés par valeur croissante. Pour une restauration parfaite, seul le dernier terme n'est pas égale à 0, pour n'importe quelle permutation entre les sources restaurées et les sources originales. Si $r_{j'}$ est le rang de la source $e_{j'}$, la concentration de l'énergie est obtenue par l'indice de Gini :

$$G_j = \frac{1}{n-1}[2\frac{\sum_{j'} e_{j'} r_{j'}}{\sum_{j'} e_{j'}} - (n+1)] \tag{23}$$

$G_j = 0$ si tous les $e_{j'}$ sont égaux, $G_j = 1$ si tous les $e_{j'}$ sont nuls excepté une valeur. Alors G_j donne une information valable sur la restauration des sources. La restauration est évaluée par :

$$G = \sum_j G_j \tag{24}$$

G fournit une information objective sur les sources qui est confirmée par la sélection visuelle.

3.2 Contraste.

L'information mutuelle des sources (IMS) est minimale lorsque le contraste C est maximum. Ces deux grandeurs font intervenir la densité de probabilité dont l'évaluation dépend du pas de codage de l'histogramme des niveaux des sources.

– Si le pas est trop petit, le nombre de pixels par cellule est trop faible et l'estimation n'est pas valable.
– Si le pas est trop grand, la FDP est lissée et elle n'est pas sensible aux caractéristiques non gaussiennes.

Alors la détermination de l'IMS [9] est réalisée en quatre étapes :

1. Pour chaque source i on détermine la valeur moyenne m_i et la déviation standard σ_i après un rejet à $3-\sigma_i$. Dans cet algorithme, on calcule de manière itérative les paramètres et on rejette les valeurs qui sont en dehors de l'intervalle $[m_i - 3\sigma_i, m_i + 3\sigma_i]$. Après quelques itérations (4 à 5) l'algorithme converge. Dans le cas de vraie distribution gaussienne, la moyenne obtenue est correcte, tandis que le biais σ_i est de l'ordre de 2%. Si la distribution n'est pas gaussienne, ces paramètres définissent un noyau gaussien de la FDP, et on aura mesuré plus de valeurs en dehors de l'intervalle $[m_i - 3\sigma_i, m_i + 3\sigma_i]$ que pour une FDP à noyau gaussien.
2. L'histogramme $H_i(k)$ de la source i est déterminé avec une taille de cellule égale à cette déviation σ_i. On peut évaluer l'entropie de la source E_i par :

$$E_i = -\sum_k \frac{H_i(k)}{N} \log_2 \frac{H_i(k)}{N} \qquad (25)$$

   où N est le nombre de pixels. Cette information est indépendante de la taille de la cellule pour des valeurs plus petites que σ_i, pour un grand nombre de pixels. Un faible biais est introduit pour un petit nombre de pixels par cellule. Un compromis possible est de choisir une taille de cellule égale à σ_i.
3. L'équation du contraste (6) est vérifiée seulement comme étant la limite pour un nombre élevé de pixels. Sa valeur est maximale lorsque les sources sont indépendantes.

3.3 Simulations.

Le jeu d'images de test a une texture semblable à celle du ciel. Des motifs de forme gaussienne ont été simulés aléatoirement et mélangés au hasard avec des coefficients positifs [2]. Pour divers algorithmes de SAS, les quantités C et G ont été calculées. Les meilleures séparations sont obtenues avec FastICA-dt [10] (méthode de déflation, fonction tanh(y)), et f-SOBI. Les autres options de FastICA ne fournissent pas de résultats corrects. On remarque qu'avec JADE [3] et fastICA st (méthode symétrique, fonction tanh(y)), C indique une bonne séparation alors que G ne le confirme pas.

C prend seulement en compte la FDP du pixel alors que G, indice qui concorde avec l'examen visuel, considère l'image dans sa globalité. Les algorithmes basés sur la FDP, tels que KL, JADE et FastICA, sont favorisés par C tandis que SOBI et ses variantes sont pénalisés.

La contrainte de positivité se marie très bien avec SOBI et fournit des spectres positifs et des images à valeurs positives avec des coefficients de mélanges positifs. La positivité dégrade les résultats avec fastICA et JADE pour ces images dans lesquelles la corrélation spatiale est très forte. Cette procédure qui tient compte d'une donnée physique supplémentaire ne se limite pas aux algorithmes que j'ai utilisés.

SAS	G	C	SAS +	G	C
Sources	4.	30.09			
Mélanges	2.53	17.88			
KL	2.60	17.88			
FastICA-dt	3.99	28.84	FastICA-dt+	2.58	23.57
FastICA-st	3.49	29.00	FastICA-st+	3.17	23.94
JADE	3.10	26.99	JADE+	2.80	28.98
SOBI-1D	1.44	21.47	SOBI-1D+	3.24	20.58
SOBI-2D	1.49	22.03	SOBI-2D+	2.85	23.12
f-SOBI-1D	3.97	26.96	f-SOBI-1D+	4.00	30.03
f-SOBI-2D	3.96	26.71	f-SOBI-2D+	3.99	28.93

Références

[1] A. Belouchrani, K. Abed-Meraim, J-F. Cardoso, E. Moulines, *A blind Source Separation Technique Using Second-Order Statistics*, IEEE Trans. On Signal Processing, vol. 45, no. 2, pp. 464-443, feb. 1997.

[2] A. Bijaoui, D. Nuzillard, *Multispectral Analysis Blind Source Separation and Mutual Information,* soumis IEEEIP en fvrier 2001.

[3] J.F. Cardoso, A. Souloumiac, *Blind Beamforming for non-Gaussian signals*, IEE Proceedings-F, 40(6), pp. 362-370, 1993.

[4] J.F. Cardoso and A. Souloumiac, *Jacobi angles for simultaneous diagonalization*, SIAM J. Mat. Anal. Appl., vol.17 pp. 161-164, 1996.

[5] P. Comon, *Independent component analysis, A new concept ?*, Signal Processing, vol. 36, pp. 287-314, 1994.

[6] L.C.M. Van Gorkom and T.M. Hancewiecz, *Analysis of DOSY and GPC-NMR experiments on polymers by multivariate curve resolution,* J. Magnetic Resonance, vol. 130, pp. 125-130, 1998.

[7] D. Nuzillard, S. Bourg, J.-M. Nuzillard, *Model-free Analysis of Mixtures by NMR*, Journal of Magnetic Resonance, vol. 133, pp. 358-363, aug. 1998.

[8] D. Nuzillard, *Adaptation de SOBI à des données fréquentielles*, GRETSI'99, pp. 745-748, Vannes, 1999.

[9] D. Nuzillard, A. Bijaoui, *Blind Source Separation and Analysis of multispectral Astronomical Images*, Astronomy & Astrophysic, Sup. Series, vol. 147, pp. 129-138, 2000.

[10] A. Hyvärinen and E. Oja, E., "A fast fixed-point algorithm for independent component analysis," *Neural Computation,* vol. 9, pp. 1483-1492, 1997.