SEPARATION OF NON ORTHOGONAL SPECTRAL DATA

Danielle Nuzillard

L.A.M. Université de Reims Champagne-Ardenne,
Moulin de la Housse BP 1039, 51687 Reims Cedex 2, France.
e-mail: danielle.nuzillard@univ-reims.fr

ABSTRACT

Independant component analysis relies on the statistical independance of the sources, a constraint that is not always fulfilled when dealing with particular real life problems. The positivity of mixing coefficients and of spectral source data, imposed by physical reasons, is also a strong constraint that permits to improve the solution of source separation problems. The necessity of dealing with spectral data led first to adapt the Second-Order Blind Identification (SOBI) algorithm to frequency domain data sets. The SOBI algorithm is not able to retrieve non-orthogonal sources from mixtures. However, it may produce solutions that are close to reality. ATheir refinement through the Alternated Least Squares procedure (ALS) introduces the positivity constraint and improves greatly the quality of the separation.

1. INTRODUCTION

Independant component analysis attempts to find a linear decomposition of observed data (the mixtures) that minimizes the statistical dependence between components (the sources). The sources can be retrieved even though only a set of their linear combinations is available [4] [2] [1] [3]. The problem is well-defined only when the sources are statistically independent. This primary constraint imposes at least the sources orthogonality. The basis of source separation are exposed in section 2. In applications the sources are not always orthogonal. The solution produced by ICA algorithms is then necessarily inaccurate and alternatives must be investigated in order to retrieve the true sources.

Non-orthogonal source signals can be handled in two different ways. Either some identified spectral regions present non-overlapping signals and standard ICA techniques can be applied to them to find the mixing matrix, or new constraint are imposed to solve the problem. For physical reasons, and especially in what concerns spectroscopic analysis, both spectral data and mixing coefficients are positive. Spectra represent an energy absorption versus a wavelength variable and does not vary in sign. Mixing coefficients represent sample concentrations and are therefore inherently positive amounts. Other scientific fields may benefit from this approach, as far as positivity constraint applies.

Source separation in the frequency domain is appropriate [7] when spectra are directly produced by experimental recording devices or when signals require pre and/or post-processing in the frequency-domain, as shown in section 3. These processings include successive FT and IFT steps. The FT being a linear operation, the separating matrix can be applied on time-domain signals as well on spectra. Separation matrix can be obtained from an adaptation of the Second Order Blind Identification (SOBI) algorithm, so that the required correlation matrices are computed directly from spectra, as explained in section 4.

The introduction of positivity constraint in the separation of frequency data is explained in section 5. Approximate orthogonal spectral sources obtained by ICA methods are refined using an iterative procedure named Alternated Least Squares (ALS). The efficiency of such a decomposition, regarding the influence of the noise level, is evaluated on simulated data in section 6. It illustrates the interest of non-orthogonal source separation when performed in the frequency domain, and constrained by positivity considerations.

2. BLIND SOURCE SEPARATION

The aim of Blind Source Separation is to retrieve n sources from m linear combinations ($m \geq n$) provided by sensors. The mixed signal can be written as:

$$X = Y + N = A.S + N, \qquad (1)$$

where X are the detected signals, Y are the mixed signals, A is the mixing matrix, S are the source signals and N is the noise introduced by the sensors.

I thank Dr. J.-M. Nuzillard for constant encouragements.

Resolution techniques require statistically independent signals and white sensor noise. Two signals are uncorrelated if their second order cross cumulant (or scalar product) is zero. This does not mean these two signals are independent. The independence is proved to an order k if cross cumulants up to order k are null. This explains why the used separation criteria rely generally on higher-order statistic considerations. One of the methods based on second order only statistics is the Principal Components Analysis (PCA). Its goal is to find orthogonal principal directions and to reduce the dimension of the data set when possible. Nevertheless, PCA does not solve the source separation problem because it lacks a criterion for statistical independence. Alternatively, introducing the hypothesis of temporally correlated source signals, an efficient separation is allowed with a second order only algorithm (SOBI) [1], as shown below.

The source separation problem is under-determined even in the absence of detector noise, since:

$$x_i(t) = \sum_{i=1}^{n} \frac{a_{i,j}}{\alpha_j}.\alpha_j.s_j(t). \tag{2}$$

As a consequence the power of the sources can be considered as normalized to unity without loss of generality.

The first step of the separation is the search of a whitening matrix that transforms the original set of m signals into a reduced basis of n orthogonal signals named the whitened data set. The correlation matrix of any vector signal $Z(t)$ is defined by:

$$R_Z(\tau) = E\left(z(t).z^*(t+\tau)\right). \tag{3}$$

If the noise is white, uncorrelated with itself and with the signals, then:

$$R_X(\tau) = R_Y(\tau). \tag{4}$$

Therefore, the correlation matrix ($\tau = 0$) of the whitened mixture is the identity. The whitening matrix W is obtained by diagonalization of $R_X(0)$. For any τ value:

$$R_{WY}(\tau) = W.R_Y(\tau).W^H = \underbrace{(W.A)}_{U}.R_S(\tau).\underbrace{(W.A)^H}_{U^H} \tag{5}$$

When $\tau = 0$, this relation shows that $W.A$ is a unitary matrix U because $R_S(0) = I_n$. For other τ values $R_S(\tau)$ is still a diagonal matrix and therefore U is the unitary matrix that diagonalizes $R_{WY}(\tau)$. Since the matrix Y is not available because of the sensor noise, only estimates of the whitened matrix $\hat{W}$ of W and $\hat{U}$ of U are available. The matrix of the whitened sensor data verifies $R_{\hat{W}X} = I_n$. The whole separation process may be summarized as follows:

- The matrix $\hat{W}$ is estimated by diagonalization of $R_X(0)$,
- A set of $R_{\hat{W}X}(\tau)$ is calculated,
- $\hat{U}$ is evaluated so that it jointly diagonalizes this set,
- The mixing matrix is retrieved using:

$$\hat{A} = \hat{W}^{\#}.\hat{U}, \tag{6}$$

- The source signals are estimated by:

$$\hat{S} = \hat{A}^H.R_X^{-1}.X. \tag{7}$$

3. SEPARATION IN THE FREQUENCY DOMAIN

Separation of frequency data is worth of interest when spectral data are directly available or when signals require pre and/or post processing in frequency domain [6]. Preprocessing is applied to correct artifacts or to extract pertinent data. One should take into account a particular case where the location of the spectral lines may vary during the recording process [5]. Such fluctuations caused by the experimental setup and do not depend on the sample to be analyzed. An IFT is performed to reconstruct time domain signals. A suitable algorithm is finally applied to separate sources. The visualization of spectral sources required an other FT step. Theses successive forward and backward Fourier transformation steps are useless if spectra can be directly separated. Candidates for an efficient separation in the frequency domain are data for which frequency content is low compared to the number of temporal samples as we will see below. The FT being a linear operation, the separating matrix can be applied on temporal data or on spectra as well. The computation of the mixing matrix by a technique related to the SOBI algorithm requires the knowledge of the covariance matrices. These ones are directly computed from the spectral data [7]. Thus, a new SOBI algorithm named f-SOBI dealing with spectral data was developed.

A collection of amplitudes in the frequency domain corresponds to a sum of sine functions in time domain. A rebuilt signal is written:

$$A(l.\Delta t) = \sum_k a_k.\exp\left(i2\pi.\frac{kf_e}{T}.l.\Delta t\right), \tag{8}$$

$$A(l.\Delta t) = \sum_k a_k.w^{kl}, \tag{9}$$

where:

- $w = \exp\left(\frac{i2\pi}{T}\right)$,
- f_e is the sampling frequency,
- T is the number of samples,
- $\frac{k f_e}{T}$ are the frequency lines whose amplitudes are a_k,
- Δt is sampling period.

The correlation function between two vectors A and B is expressed as:

$$R_{AB}(\tau) = \frac{1}{L} \sum_l A(l.\Delta t) . B^*(l.\Delta t - \tau), \qquad (10)$$

where $\tau = n.\Delta t$, then:

$$R_{AB}(\tau) = \frac{1}{L} \sum_l \left(\sum_k a_k . w^{kl} \right) . \left(\sum_{k'} b^*_{k'} . w^{-k'.(l-n)} \right), \qquad (11)$$

finally:

$$R_{AB}(\tau) = \sum_k a_k b^*_k w^{kn}. \qquad (12)$$

It is the IFT of the term by term product of the vectors A and B^*.

The m mixture vectors provide m^2 correlation functions represented by $p \times m^2$ matrices where p is the number of τ values to consider. Each correlation matrix being hermitian, there are only $pm(m+1)/2$ correlation coefficients to compute.

When the spectral content is low compared to the number of samples, the numbers p of delays and of spectral lines k are small, it is not necessary to resort to a FT algorithm to obtain the $R_{AB}(n)$ values. A direct calculation can be undertaken.

The rest of the separation is then achieved using the native SOBI algorithm renamed from here t-SOBI. The matrix $R_x(n=0)$ is diagonalized to compute the whitening matrix, which is then applied on spectra to yield whitened data. The separation and mixing matrices are obtained within scaling factors and within a permutation through joint diagonalisation of the other $(n \neq 0)$ covariance matrices. The separation matrix is computed on the fly by multiplying all the transformation matrices applied to the original mixtures that yield the sources.

This extention can be viewed as an alternative choice for the cross correlation we want to reduce: t-SOBI takes into account the correlation at short distances in the time (or direct) space, while in f-SOBI the correlations in the Fourier transform space are considered. The f-SOBI was developed for NMR spectrograms displaying narrow peaks. The spectral correlation rapidly decreases, while it is always strong enough

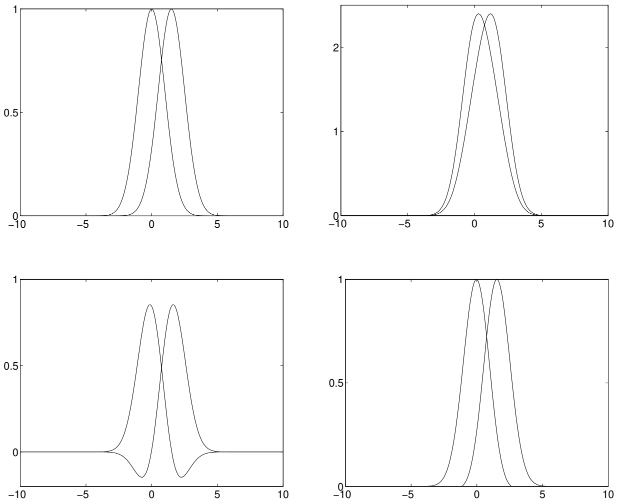

Figure 1: Top left: simulated sources, right: mixtures; bottom left: orthogonal separation, right: positive separation

in the Fourier space. With spectrograms whith large lines this argument is not valuable and spectral data fit more adequately the requirements of t-SOBI.

4. SEMI-BLIND SEPARATION

A possible way of dealing with non-orthogonal sources consists in searching non-overlapping sub-spectra in the frequency representation of the signal [6]. The selected regions must be chosen so that they include signals from all sources. The selection itself requires some degree of data interpretation, underlying on prior knowledge on the data themselves. The source separation process, applied on the extracted spectral regions, yields the separating matrix. Clearly, even though this matrix is evaluated on partial data, it applies to the entire data set. Its application to the original mixture data leads to desired source signals. This processing can be performed as well on temporal data as spectra since FT is a linear operation. The overall process can be resumed as follows:

1. Record time-domain data,
2. FT,
3. Select independent areas to build pseudo-spectra,
4. Perform IFT to obtain pseudo-data in time domain,
5. Apply a separation algorithm to pseudo-data,

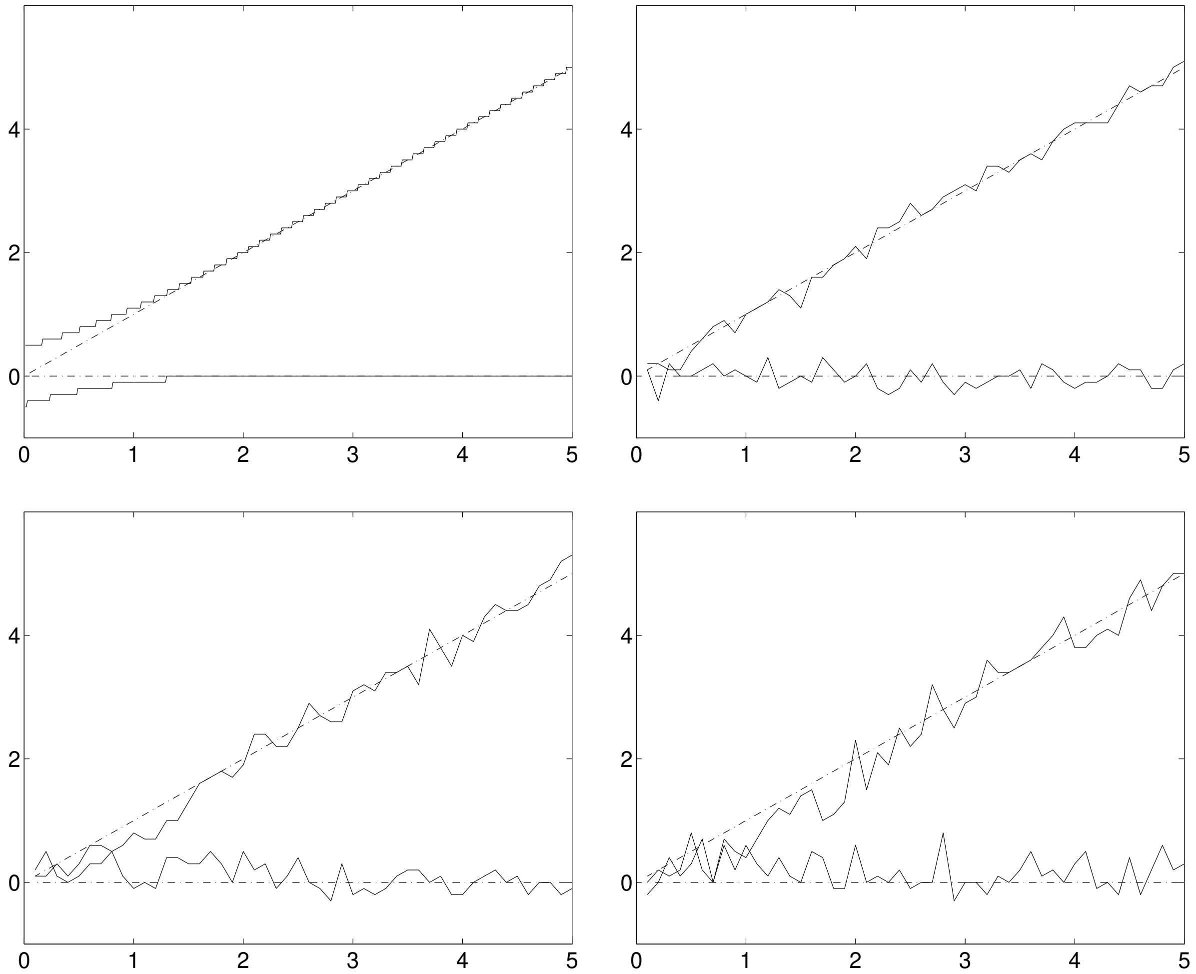

Figure 2: Comparison between true (dot-dashed line) and estimated (solid line) central positions, depending on the distance (on x axis) of two gaussian sources and according to the signal noise ratio (SNR). The first one is always centered on zero while the second one is moving along the x axis. Top left: without any noise, right: 5% SNR; bottom left: 10% SNR, right: 20% SNR

6. Calculate the separating matrix,
7. Apply the separating matrix on initial data to obtain the sources.

Steps 4, 5 and 6 can be more efficiently replace by direct separation in frequency domain as explained in part 3.

5. POSITIVE DECOMPOSITION

The preceding approach is feasible only if orthogonal sub-spectra can be identified. A more versatile approach consists in relaxing the orthogonality assumption. Positivity of source data and mixing coefficients has proved to be a useful alternative constraint. Many physical applications provide positive signals. Mixing coefficients are related to proportions or substance concentrations and therefore are positive. Under these assumptions, the separation of the responses into individual components is possible. A method, the Multi-Components Resolution (MCR) analysis [8], involves three steps:

1. Abstract factor analysis, conceptually identical to whitening,
2. Real factor analysis, that involves a fourth order statistical data analysis,
3. Search of a solution that follows the positivity requirement.

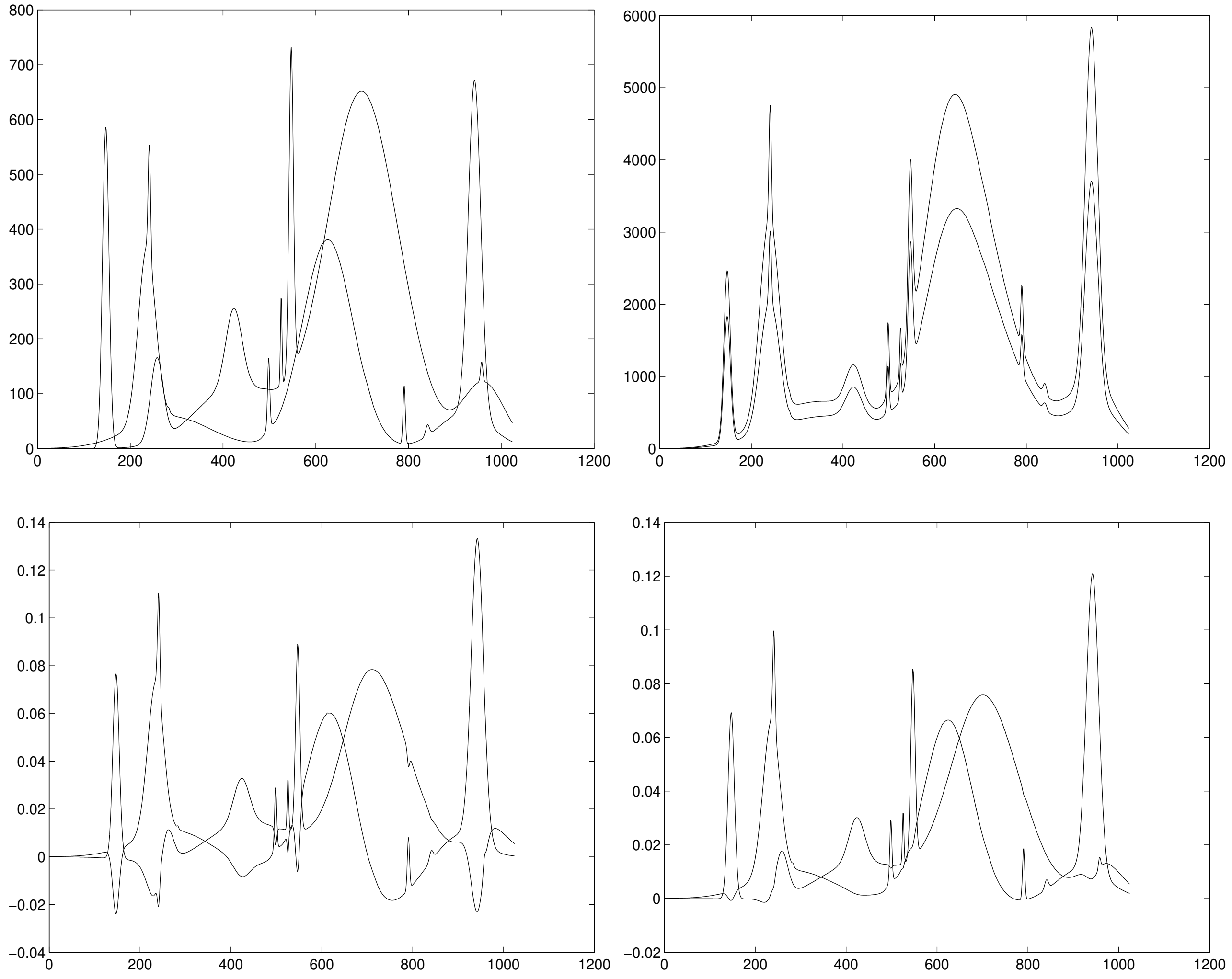

Figure 3: Top left: simulated spectra, right: mixtures of spectra; bottom left: orthogonal demixing, right: demixing when positive constraint is considered.

The first two steps produce solution sources that are orthogonal and cannot be the true ones. The third step is named ALS, Alternated Least Squares. It is a projection on convex set (POCS) algorithm, justifying its convergence. Assuming that $X = \hat{A}\hat{S}$, $\hat{A}$ (resp. $\hat{S}$) can be evaluated from $\hat{S}$ (resp. $\hat{A}$) and X:

$$\hat{S} = ({}^t\hat{A}.\hat{A})^{-1}.{}^t\hat{A}.X, \qquad (13)$$

$$\hat{A} = X.{}^t\hat{S}.(\hat{S}.{}^t\hat{S})^{-1}, \qquad (14)$$

in the least-squares sense. True sources are obtained by ALS following an iterative process:

1. Set to zero the negative parts of $\hat{S}$,

2. Estimation of $\hat{A}$,

3. Set to zero the negative coefficients of $\hat{A}$,

4. Estimation of $\hat{S}$,

5. Return to 1 if convergence is not achieved.

The process is stopped when no evolution of the result is perceptible.

The first two steps of the MCR analysis may be replaced by a SOBI analysis if time correlated data mixtures are available and if the Fourier transform of the source signals provide positive spectra.

6. RESOLUTION POWER OF THE POSITIVE DECOMPOSITION

We check the resolution power of the positive decomposition on simulated data. The orthogonal separation can yield sources with negative parts. Then, original sources are retrieved by taking into accompt positivity constraint as shown in figure 1. For a set of tests, two gaussian sources were simulated, their standard deviation σ is 1. The mixing proportions [1 2, 2 1] provide two mixtures. The first source is centered on 0 while the second is moving along the x axis from position 0.2 to 5. For each gaussian source, the true and the estimated central positions are compared. They are drawn in figure 2 where the x axis is the distance between two gaussian sources. Without any noise, the measured error becomes negligible when the distance between two centers is higher than 1.5. For each mixture with added noise, hundred separation trials are performed. The estimated positions are averaged. A bias is visible when the distance between centers is less than 2. This sets the limits of the positive separation capability for such signals.

A more realistic example is provided using source spectra randomly generated. The standard deviation and the position of the amplitude of gaussian peaks are calculated as $1+\exp(ax)$ where a is constant and x is uniformly drawn on the interval [0 1]. The position of the center of each peak is uniformly drawn over the frequency range. The elements of the mixing matrix are integers drawn on the [0 1 0] interval. Each source consists in a sum of fifteen such gaussian peaks. Orthogonal decomposition naturely shows negative parts. The positive decomposition succeeds to retrieve spectra similar to initial randomly drawn sources as shown in figure 3. Many trials following this scheme were undertaken and confirm the feasability of our approach to positive decomposition.

When no prior knowledge is available, one cannot extract orthogonal sub-spectra, the application of the positivity constraint provides an alternative to allow a good separation.

7. CONCLUSION

This contribution enlightens the interest of positive decomposition for applications compatible with positivity hypothesis. This is especially the case of spectroscopy. The separation is performed in the frequency domain from which the correlation matrices are calculated and then supplied to a second order algorithm. It is sufficient since it gives an orthogonal decomposition which provides a sufficiently good starting point for proceeding to the positive decomposition. Then, an alternated least square algorithm including positivity constraint is applied. Positive sources are retrieved for simulated data as well for real spectra. Illustration borrowed from NMR spectroscopy will be presented during the workshop.

8. REFERENCES

[1] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, 'A blind source separation technique using second-ordre statistics,' *IEEE Trans. SP,* vol. 45, pp. 434-444, 1997.

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

[3] A. Hyvärinen, 'Fast and robust ifxed-point algorithms for independant component analysis,' *IEEE Transactions on Neural Networks,* 10(3) pp. 626-634, 1999.

[4] C. Jutten, J. Hérault, 'Blind separation of sources, part I: An adaptative algorithm based on neuromimetic architecture,' *Signal Processing,* 24:1-10, 1991.

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

[6] D. Nuzillard, J.-M. Nuzillard, 'BSS applied to non-orthogonal signals,' *1st Workshop ICA'99,* Aussois, France, pp. 25-30, 1999.

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

[8] L. C. M. Van Gorkom and T. M. Hancewiecz, 'Analysis of DOSY and GPC-NMR experiments on polymers by multivariate curve resolution,' *J. Magn. Reson.,* vol. 130, pp. 125-130, 1998.