Séparation aveugle sous-déterminée de sources audio par la méthode EMD (Empirical Mode Decomposition)

Abdeldjalil AISSA-EL-BEY, Karim ABED-MERAIM, Yves GRENIER

Département Traitement du Signal et des Images,
ENST-Paris, 46 rue Barrault 75634, Paris Cedex 13, France
{elbey, abed, grenier}@tsi.enst.fr

Résumé – Dans le cadre de la séparation aveugle de sources, nous montrons dans cet article comment effectuer la séparation de mélanges instantanés de sources audio en utilisant une méthode basée sur l'algorithme de Décomposition Modale Empirique (ou EMD, pour Empirical Mode Decomposition). Cette approche nous permet, en particulier de traiter le cas sous déterminé (c'est à dire le cas où l'on a moins de capteurs que de sources). L'approche EMD se base sur le fait que les signaux audio (et particulièrement les signaux musicaux) peuvent être bien modélisés localement par une somme de signaux périodiques. Ces signaux seront donc décomposés en utilisant l'algorithme EMD et recombinés par classification suivant leurs directions spatiales regroupant ainsi les composantes de chacune des sources. Nous présenterons quelques résultats de simulation qui permettent d'évaluer les performances de cette nouvelle méthode.

Abstract – This paper introduces new algorithm for the blind separation of audio sources using Empirical Mode Decomposition (EMD). Indeed, audio signals and, in particular, musical signals can be well approximated by a sum of damped sinusoidal (modal) components. Based on this representation, we propose a two steps approach consisting of a signal analysis (extraction of the modal components) using EMD followed by a signal synthesis (pairing of the components belonging to the same source) using vector clustering. For the signal analysis, a major advantage of the proposed method resides in its ability to separate more sources than sensors. Simulation results are given to assess the performance of the proposed algorithm.

1 Introduction

La séparation aveugle de sources est un problème qui consiste à retrouver des signaux statistiquement indépendants à partir de leurs mélanges (observations) et cela sans connaissance *a priori* de la structure des mélanges ou des signaux sources.
La séparation de sources intervient dans des applications diverses [1] telles que la localisation et la poursuite de cibles en radar et sonar, la séparation de locuteurs (problème dit de "cocktail party"), la détection et la séparation dans les systèmes de communication à accès multiple, l'analyse en composantes indépendantes de signaux biomédicaux (e.g., EEG ou ECG), etc. Ce problème a été intensément étudié dans la littérature et beaucoup de solutions efficaces ont déjà été proposées [1].
Néanmoins, le cas sous-déterminé où le nombre de sources est supérieur à celui des capteurs (observations) reste relativement peu traité et sa résolution est l'un des problèmes ouverts de la séparation aveugle de sources. Dans le cas de signaux non-stationnaires (incluant en particulier les signaux audio), certaines solutions utilisant la transformée temps-fréquence des observations existent pour le cas sous-déterminé [4, 6]. Nous proposons ici une approche alternative utilisant la décomposition modale empirique EMD (Empirical Mode Decomposition) des signaux observés [2, 3]. Cette technique, récemment proposée, permet la décomposition des différents modes d'un signal supposé *localement périodique*, sans qu'il soit nécessairement harmonique au sens de Fourier. Les signaux audio et plus particulièrement les signaux musicaux peuvent être modélisés par une somme de sinusoïdes amorties [7]. C'est cette propriété que nous proposons d'exploiter ici pour la séparation de sources audio par le biais de la décomposition EMD.

2 Formalisation du problème

Le modèle de séparation aveugle de sources suppose l'existence de N signaux indépendants $s_1(t), \ldots, s_N(t)$ et M observations $x_1(t), \ldots, x_M(t)$ qui représentent les mélanges. Ces mélanges sont supposés instantanés, i.e.

$$x_i(t) = \sum_{j=1}^{N} a_{ij} s_j(t) \qquad i = 1, \ldots, M \tag{1}$$

Ceci peut être représenté par l'équation de mélange :

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{2}$$

où $\mathbf{s}(t) \stackrel{\text{def}}{=} [s_1(t), \ldots, s_N(t)]^T$ est le vecteur colonne de dimension $N \times 1$ qui regroupe les signaux sources, le vecteur $\mathbf{x}(t)$ regroupe de la même manière les M signaux observés, et $\mathbf{A} \stackrel{\text{def}}{=} [\mathbf{a}_1, \ldots, \mathbf{a}_N]$ est la matrice de mélange de taille $M \times N$ où $\mathbf{a}_i = [a_{1i}, \ldots, a_{Mi}]^T$ contient les coefficients du mélange. Nous supposerons que pour tout couple (i, j) les vecteurs $\mathbf{a}_i$ et $\mathbf{a}_j$ sont linéairement indépendants. Les signaux sources sont supposés décomposables en somme de composantes modales $c_i^j(t)$, i.e :

$$s_i(t) = \sum_{j=1}^{l_i} c_i^j(t) \qquad t = 0, \ldots, T-1 \tag{3}$$

L'hypothèse d'indépendance des sources est remplacée ici par l'hypothèse de quasi-orthogonalité des composantes modales, i.e.

$$\frac{\langle c_i^j | c_{i'}^{j'} \rangle}{\|c_i^j\| \|c_{i'}^{j'}\|} \approx 0 \quad \text{pour} \quad (i,j) \neq (i',j') \tag{4}$$

où

$$\langle c_i^j | c_{i'}^{j'} \rangle \stackrel{\text{def}}{=} \sum_{t=0}^{T-1} c_i^j(t) c_{i'}^{j'}(t)^* \quad (5)$$

et

$$\|c_i^j\|^2 = \langle c_i^j | c_i^j \rangle \quad (6)$$

L'hypothèse (4) peut être parfois restrictive. Ainsi nous proposons dans la section 3.4 une méthode permettant de s'affranchir de cette contrainte.

3 La séparation de sources utilisant l'approche EMD

En se basant sur le modèle précédent, on propose une approche en deux étapes qui sont :

- *Étape d'analyse* : Dans cette étape on applique une décomposition modale sur chaque sortie de capteur afin d'en extraire toutes les composantes harmoniques ou pseudo-harmoniques. Cette décomposition consiste en l'algorithme EMD introduit en [2, 3].

- *Étape de synthèse* : Dans cette étape nous regroupons ensemble les composantes modales correspondant au même signal source afin de reconstituer le signal d'origine. Ceci est fait par une méthode de classification basée sur la direction spatiale des composantes que l'on estime par corrélation de celles-ci avec le signal d'antenne observé.

3.1 Analyse des signaux utilisant l'EMD

Une nouvelle technique non-linéaire, appelée EMD a été récemment introduite par N.E. Huang et al. pour représenter les signaux non-stationnaires [2]. Dans l'algorithme EMD le signal non stationnaire est considéré à l'échelle de ses oscillations locales. Plus précisément, la décomposition d'un signal $z(t)$ par l'EMD se résume comme suit :

1. Identification de tous les extrema de $z(t)$.
2. Interpolation entre les minima (resp. maxima) du signal, conduisant à une enveloppe $e_{min}(t)$ (resp. $e_{max}(t)$).
3. Calcul de la moyenne $m(t) = (e_{min}(t) + e_{max}(t))/2$.
4. Extraction du détail $d(t) = z(t) - m(t)$.
5. Itération sur le résidu $m(t)$ jusqu'à l'obtention d'un résidu final d'énergie quasi-nulle.

En appliquant alors l'algorithme EMD sur les x_i qui s'écrivent $x_i(t) = \sum_{j=1}^{N} a_{ij} s_j(t) = \sum_{j=1}^{N} \sum_{k=1}^{l_j} a_{ij} c_j^k(t)$, on obtient des estimées $\widehat{c}_j^k(t)$ des composantes $c_j^k(t)$.

3.2 Synthèse des signaux utilisant la classification vectorielle

Pour la synthèse des signaux sources on observe que l'hypothèse de quasi-orthogonalité nous permet d'avoir :

$$\frac{\langle \mathbf{x} | c_i^j \rangle}{\|c_i^j\|^2} \stackrel{\text{def}}{=} \frac{1}{\|c_i^j\|^2} \begin{bmatrix} \langle x_1 | c_i^j \rangle \\ \vdots \\ \langle x_M | c_i^j \rangle \end{bmatrix} \approx \mathbf{a}_i \quad (7)$$

où $\mathbf{a}_i$ représente la i^{eme} colonne de $\mathbf{A}$. nous pouvons alors associer chaque composante $\widehat{c}_j^k$ à une direction spatiale (vecteur colonne de $\mathbf{A}$) que l'on estime par

$$\widehat{\mathbf{a}}_j^k = \frac{\langle \mathbf{x} | \widehat{c}_j^k \rangle}{\|\widehat{c}_j^k\|^2}$$

Deux composantes du même signal source étant associées au même vecteur colonne de $\mathbf{A}$, nous proposons de regrouper ces composantes par classification sur les vecteurs $\widehat{\mathbf{a}}_j^k$ selon N classes. Pour cela, nous avons utilisé l'algorithme k-means [9] qui se résume comme suit :

1. Initialement, sélectionner N centroïdes arbitrairement dans l'ensemble des vecteurs.
2. Assigner chaque vecteur à la classe dont le centroïde est le plus proche au sens de la distance Euclidienne.
3. Calculer les centroïdes de chaque classe.
4. Répéter l'étape 2 et 3 jusqu'à ce qu'il n'y ait aucun changement de vecteurs entre les classes.

Finalement, on pourra reconstruire les sources initiales à une constante près en additionnant les différentes composantes d'une même classe.

3.3 Association et sélection

Remarquons, qu'en appliquant l'approche décrite précédemment (analyse plus synthèse) sur toutes les sorties d'antenne $x_1(t), \cdots, x_M(t)$, nous obtenons M estimées de chacune des sources. Nous avons observé que la qualité d'estimation des sources varie d'un capteur à un autre, et ceci dépend fortement des coefficients du mélange, en particulier, du rapport signal à interférence (RSI) de la source désirée. En conséquence, nous proposons une méthode aveugle de sélection pour choisir la meilleure des M estimées d'une source donnée. Pour cet effet, nous avons besoin tout d'abord de regrouper les estimées d'une même source ensemble. Ceci est réalisé par corrélation i.e. un signal est associé aux $(M-1)$ signaux (les signaux issus des $(M-1)$ autres capteurs) qui lui sont les plus corrélés. Le facteur de corrélation de deux signaux s_1 et s_2 est calculé par $\frac{\langle s_1 | s_2 \rangle}{\|s_1\| \|s_2\|}$. Une fois l'association des sources effectuée, on propose de sélectionner la source estimée qui a l'énergie maximale, i.e.

$$\widehat{s}_i(t) = \max_j \{E_i^j = \|\widehat{s}_i^j(t)\|^2, \quad j = 1, \cdots, M\} \quad (8)$$

où E_i^j représente l'énergie de la i^{eme} source obtenue à partir du j^{eme} capteur.

3.4 Projection sous-espace

Afin de relâcher la contrainte (4), on suppose qu'une composante $c_j^k(t)$ peut être présente dans plusieurs sources. Ceci est le cas pour certains signaux musicaux tels que ceux traités dans [5]. Pour simplifier, nous supposons ici qu'une composante appartient au plus à deux sources. Supposons donc que la composante $c_j^k(t)$ est présente dans les sources $s_{j_1}(t)$ et $s_{j_2}(t)$ avec les amplitudes α_{j_1} et α_{j_2} respectivement. Il s'en suit que la direction spatiale associée à cette composante estimée par (7) est donnée par :

$$\widehat{\mathbf{a}}_j^k \approx \alpha_{j_1} \mathbf{a}_{j_1} + \alpha_{j_2} \mathbf{a}_{j_2}. \quad (9)$$

Il s'agit maintenant de trouver les indices j_1 et j_2 des deux sources associées à cette composante, ainsi que les amplitudes α_{j_1} et α_{j_2}. Pour ce faire, on propose une approche basée sur la projection en sous-espace. Supposons que la matrice de mélange $\mathbf{A}$ est connue et vérifie les conditions $M > 2$ et tout triplet de vecteurs colonnes de $\mathbf{A}$ sont linéairement indépendants. Observons alors :

$$\mathbf{P}_{\widetilde{\mathbf{A}}}^{\perp}\mathbf{a}_j^k = 0$$

si est seulement si $\widetilde{\mathbf{A}} = [\mathbf{a}_{j_1}\ \mathbf{a}_{j_2}]$ où $\mathbf{P}_{\widetilde{\mathbf{A}}}^{\perp}$ représente la matrice de projection orthogonale sur l'orthogonal du sous-espace image de $\widetilde{\mathbf{A}}$, i.e.

$$\mathbf{P}_{\widetilde{\mathbf{A}}}^{\perp} = \mathbf{I} - \widetilde{\mathbf{A}}\left(\widetilde{\mathbf{A}}^H\widetilde{\mathbf{A}}\right)^{-1}\widetilde{\mathbf{A}}^H. \qquad (10$$

En pratique, en prenant en compte le bruit, on détecte les colonnes j_1 et j_2 en minimisant :

$$(j_1, j_2) = \underset{(l,m)}{\arg\min}\left\{\|\mathbf{P}_{\widetilde{\mathbf{A}}}^{\perp}\widehat{\mathbf{a}}_j^k\| \mid \widetilde{\mathbf{A}} = [\mathbf{a}_l\ \mathbf{a}_m]\right\}$$

Une fois $\widetilde{\mathbf{A}}$ trouvée, on estime les pondérations α_{j_1} et α_{j_2} par

$$\begin{bmatrix} \alpha_{j_1} \\ \alpha_{j_2} \end{bmatrix} = \widetilde{\mathbf{A}}^{\#}\widehat{\mathbf{a}}_j^k \qquad (11$$

où $\widetilde{\mathbf{A}}^{\#}$ représente la pseudo-inverse de $\widetilde{\mathbf{A}}$. Dans cette article nous avons traité toutes les composantes comme étant associées à deux signaux sources. Si jamais une composante n'est présente que dans une seule source, un des deux coefficients estimés dans (11) devrait être zéro ou proche de zéro.
Dans ce qui précède la matrice de mélange $\mathbf{A}$ est supposée connue, il faut donc l'estimer avant d'appliquer la projection par sous-espace. Nous nous proposons d'estimer les colonnes de $\mathbf{A}$ comme étant les centroïdes moyennés (i.e. comme chaque source est estimée M fois, et que chaque sources correspond à une classe et à un centroïde, alors, on moyenne les centroïdes des M classes d'une source donnée) des N classes obtenues dans l'étape de synthèse de notre algorithme. Cette approche suppose implicitement que la majorité des composantes n'appartiennent qu'à une seule source et par conséquent que les vecteurs $\widehat{\mathbf{a}}_j^k$ représentent dans leurs majorité une des colonnes de la matrice $\mathbf{A}$.

Remarques :

- <u>*Cas sur-déterminé*</u> : Dans ce cas, on peut effectuer la séparation par une inversion de la matrice de mélange $\mathbf{A}$. Cette dernière est estimée par l'approche décrite précédemment.
- <u>*Estimation du nombre de sources*</u> : C'est une tâche difficile dans le cas sous-déterminé. Il existe certaines solutions utilisant soit des approches tensorielles [8] soit d'autres utilisant des techniques de classifications avec estimation conjointe du nombre de classes [9]. Cependant, ces méthodes sont très sensibles au bruit, à la dynamique des signaux sources et au conditionnement de la matrice de mélange. Dans cet article on suppose que le nombre de sources est connu (ou correctement estimé).

4 Simulations

Nous présentons ici quelques résultats de simulation pour illustrer l'exécution de notre algorithme de séparation de sources aveugle. Pour cela, nous considérons une antenne linéaire uniforme (ALU) de capteurs avec $M = 3$ capteurs recevant $N = 4$ signaux sources audio (à l'exception de la troisième expérience où N varie dans l'interval [2,6]). Selon le modèle ALU, la matrice $\mathbf{A}$ est engendrée avec des angles d'arrivés des sources aléatoires. La taille des observations $T = 10000$ (les signaux sources sont échantillonnés à une fréquence de 44.1 khz). Les signaux observés sont corrompus par un bruit blanc additif de covariance $\sigma^2\mathbf{I}$ (σ^2 étant la puissance de bruit). La qualité de séparation est mesurée par l'erreur quadratique moyenne normalisée (EQMN) des sources estimées pour 100 tirages aléatoires du bruit.

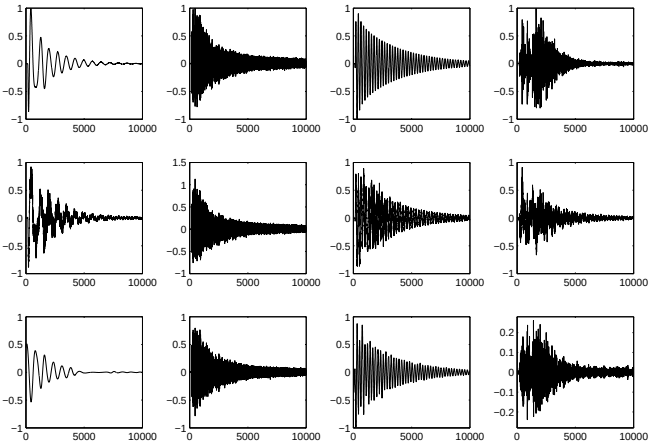

FIG. 1 – Exemple de séparation de 4 signaux audio pour 3 capteurs : signaux originaux (ligne du haut) ; signaux estimés par pseudo-inversion quand la matrice de mélange est connue (ligne du milieu) ; signaux estimés par EMD (ligne du bas).

La figure 1 représente les résultats obtenus en utilisant notre algorithme pour les 4 sources audio représentées par la première ligne de la figure. La deuxième ligne montre le résultat obtenu en utilisant la pseudo inverse de la matrice de mélange $\mathbf{A}$ pour estimer les sources originales en connaissant $\mathbf{A}$ avec exactitude, et la dernière donne les estimées des sources en utilisant notre algorithme. La figure 2 représente la variation de l'erreur quadratique moyenne normalisée (EQMN) des estimées des signaux sources en fonction du rapport signal à bruit (RSB). Sur cette figure nous comparons les résultats obtenus par une sélection basée sur un critère énergétique (qui consiste à sélectionner l'estimée la plus énergétique parmi les M estimées da la même source), une sélection optimale au sens de l'EQMN (celle-ci est donnée seulement à titre comparatif car elle suppose la connaissance des signaux sources) et les résultats obtenus par pseudo-inversion de la matrice $\mathbf{A}$. La figure 3 représente la variation de l'erreur quadratique moyenne normalisée des estimées des signaux sources en fonction du nombre de sources qui varie de 2 à 6. Dans les cas ou $N = 2$ et $N = 3$ (cas sur-déterminé) la séparation est effectuée par pseudo-inversion de l'estimée de la matrice de mélange $\mathbf{A}$. La figure 4 représente la variation de l'EQMN des estimées des signaux sources en fonction du RSB. Sur cette figure nous comparons les résultats obtenus par l'utilisation de la décomposition EMD uniquement et celles obtenues avec l'utilisation de

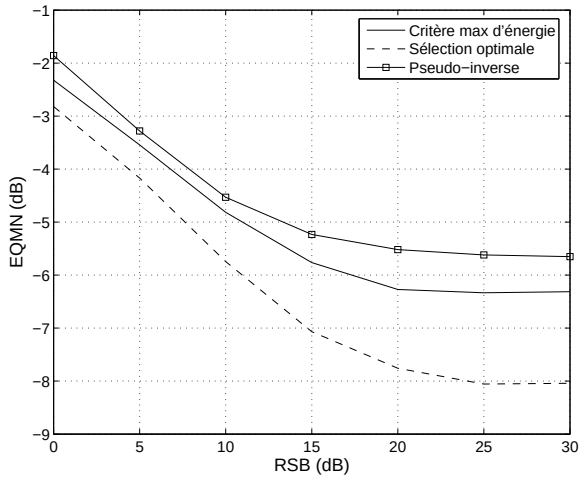

FIG. 2 – Performances de l'algorithme de séparation de 4 sources audio pour 3 capteurs : les courbes représentent la valeur moyenne de l'erreur quadratique de l'estimation des signaux sources en fonction du rapport signal à bruit.

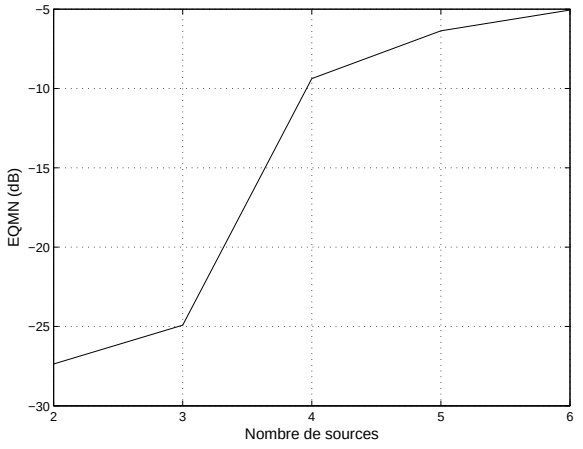

FIG. 3 – Performance de l'algorithme de séparation pour 3 capteurs : la courbe représente la valeur moyenne de l'erreur quadratique de l'estimation des signaux sources en fonction de leur nombre.

la décomposition EMD conjuguée à la projection sous-espace. On peut observer un gain à moyen et fort RSB mais une légère dégradation à faible RSB pour la méthode EMD avec projection sous-espace. Cette dégradation est due au fait que l'on associe systématiquement toute composante à deux sources est que l'estimation des amplitudes par (11) est fortement bruitée à faible RSB. Ainsi, si une composante n'appartient effectivement qu'à une seule source s_{j_1} le coefficient de la deuxième source α_{j_2} devrait être nul. En présence de bruit l'équation (11) donne une valeur non nulle de ce coefficient qui dégrade l'estimation des sources.

5 Conclusion

Dans cet article nous présentons une nouvelle méthode de séparation aveugle de sources audio utilisant l'algorithme EMD.

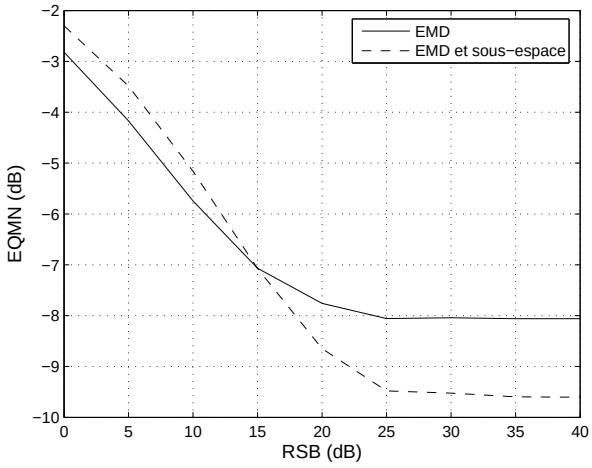

FIG. 4 – Performances des l'algorithmes de séparation de 4 sources audio pour 3 capteurs : les courbes représentent la valeur moyenne de l'erreur quadratique de l'estimation des signaux sources en fonction du rapport signal à bruit pour l'algorithme EMD classique ainsi que celle pour l'EMD avec projection sous-espace.

L'avantage majeur de la méthode proposée est sa capacité à séparer plus de sources que de capteurs, et dans ce cas elle produit une meilleure qualité de séparation que ce que l'on obtient par pseudo-inversion de la matrice de mélange (même si elle est connue).

Références

[1] A.K. Nandi (editor), "Blind estimation using higher-order statistics." *Kluwer Acadimic Publishers*, Boston 1999.

[2] N.E Huang, Z. Shen, S.R. Long, M.L. Wu, H.H. Shih, Q. Zheng, N.C. Yen, C.C. Trung and H.H. Liu, "The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary times series analysis", *Proc. Roy. Soc.* London A, Vol. 454, pp. 903-995, 1998.

[3] P. Flandrin, G. Rilling and P. Goncalvès, "Empirical mode decomposition as a filter bank", *IEEE SPL*, 2004.

[4] L. Nguyen, A. Belouchrani, K. Abed-Meraim and B. Boashash, "Separating more sources than sensors using time-frequency distributions." in *Proc. ISSPA*, Malysia, 2001.

[5] J. Rosier, Y. Grenier, "Unsupervised Classification Techniques for Multipitch Estimation", in *116th Convention of the Audio Engineering Society*, Berlin, 2004.

[6] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals : demixing n sources from 2 mixtures," in *ICASSP*, Turkey, 2000.

[7] R. Boyer and K. Abed-Meraim, "Audio modeling based on delayed sinusoids." *IEEE-Tr-SAP*, Mars 2004.

[8] L. De Lathauwer, B. Moor, J. Vandewalle, "ICA techniques for more sources than sensors", *Higher-oder statistic Proc. of the IEEE Sig. Proc. Workshop*, 1999.

[9] I.E. Frank and R. Todeschini, "The data analysis handbook", *Elsevier, Sci. Pub. Co.*, 1994.