De la séparation de sources à l'analyse en composantes indépendantes, et au-delà

Christian Jutten, Pierre Comon

▶ **To cite this version:**

Christian Jutten, Pierre Comon. De la séparation de sources à l'analyse en composantes indépendantes, et au-delà. 2023. hal-04106245v1

HAL Id: hal-04106245

https://hal.science/hal-04106245v1

Preprint submitted on 25 May 2023 (v1), last revised 7 Jul 2023 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

De la séparation de sources à l'analyse en composantes indépendantes, et au-delà

Christian Jutten and Pierre Comon *

25 mai 2023

Table des matières

*GIPSA-lab, 11, rue des Mathématiques, BP 46, 38402 Saint Martin d'Hères Cedex, France. email: christian.jutten@grenoble-inp.fr, pierre.comon@gipsa-lab.fr

Résumé

La séparation de sources, apparue dans les années 80, a été largement étudiée en France, avec des contributions remarquables de plusieurs pionniers, le soutien du GdR ISIS et des ses directeurs. Cet article retrace quelques étapes de cette histoire.

Abstract— Source separation, which appeared in the 80's, has been widely investigated in France, with outstanding contributions of a few pioneers, support of GdR ISIS and its directors. This paper traces a few steps of this story.

Dans cet article, on se limitera sauf mention contraire au corps des réels par simplicité, mais la plupart des résultats restent vrais dans le corps des complexes, souvent au prix de complications d'écriture et de calcul.

1 Un problème inspiré de la biologie

Tout a commencé à Grenoble dans le laboratoire de Traitement d'Images et de Reconnaissance de Formes (LTIRF) où Jeanny Hérault avait initié des travaux de recherche sur la modélisation du neurone et des réseaux de neurones. L' objectif était de mieux comprendre les principes de fonctionnement du cerveau pour

traiter l'information. Mais, à la fin des années 70, il était difficile de présenter ces travaux atypiques dans les conférences de traitement du signal. Aussi, en 1982, Jeanny Hérault, avec Bernard Ans et Christian Jutten, a créé les journées interdisciplinaires "Neurosciences et Sciences de l'Ingénieur", pour rassembler les chercheurs intéressés par la modélisation et la simulation du cerveau. Ces premières journées ont réuni à Grenoble une soixantaine de personnes venues de toute la France, et de toutes disciplines : neurosciences, physique, mathématiques, informatique et traitement du signal.

1.1 Décodage du mouvement humain

Durant ces journées, Ch. Jutten et J. Hérault et B. Ans ont discuté longuement avec Jean-Pierre Roll et Jean-Claude Gilhodes, qui travaillaient sur le décodage du mouvement chez les vertébrés, dans un laboratoire de neurosciences de Marseille. Lorsque nous bougeons une articulation, des capteurs "fusoriaux", localisés sur les tendons du muscle qui actionne l'articulation, mesurent l'étirement et la vitesse d'étirement du muscle. Ces informations, dites de proprioception, sont transmises au système nerveux central par deux types de fibres nerveuses : les fibres primaires et secondaires. Pour un mouvement d'une articulation à différentes vitesses constantes, on mesure sur ces fibres les signaux illustrés à la figure 1.

La figure 1.A regroupe trois tracés : en bas, le mouvement de l'articulation à vitesse constante entre deux positions (de 90 à 100 degrés) ; au milieu, le train d'impulsions (spikes) mesurés sur les fibres primaires ; en haut, la fréquence instantanée des impulsions (i.e., l'inverse de l'intervalle de temps entre deux impulsions successives) appelée fréquencegramme. Les tracés des parties B et C sont similaires, mais correspondent à des vitesses de mouvement plus grandes. Sur les fréquencegrammes, on observe que les positions initiales et finales de l'articulation sont codées avec des fréquences instantanées différentes mais constantes, croissantes avec la position angulaire de l'articulation. Pendant le mouvement (rampe), la fréquence instantanée est augmentée, proportionnellement à la vitesse du mouvement. On obtient des tracés similaires avec les fibres secondaires. Cependant, les fibres secondaires (dynamiques) sont plus sensibles à la vitesse de l'articulation que les fibres primaires.

1.2 Modélisation

En négligeant le pic initial (au début du mouvement), on observe que la fréquence instantanée sur chaque type de fibres est un mélange pondéré de l'étirement, $p(t)$, et de la vitesse d'étirement, $v(t)$, du muscle. La fréquence, $f_I(t)$, des fibres primaires (statiques) est plus sensible à l'étirement, alors que la fréquence , $f_II(t)$, des fibres secondaires (dynamiques) est plus sensible à la vitesse d'étirement. Ch. Jutten et J. Hérault et B. Ans ont alors proposé le modèle linéaire très simple suivant :

$$\begin{cases} f_I(t) &= a_{11}p(t) + a_{12}v(t) \\ f_{II}(t) &= a_{21}p(t) + a_{22}v(t) \end{cases}$$

Dans cette équation, on mesure seulement les fréquences $f_I(t)$ et $f_{II}(t)$, les autres quantités sont inconnues. Les a_{ij} modélisent les

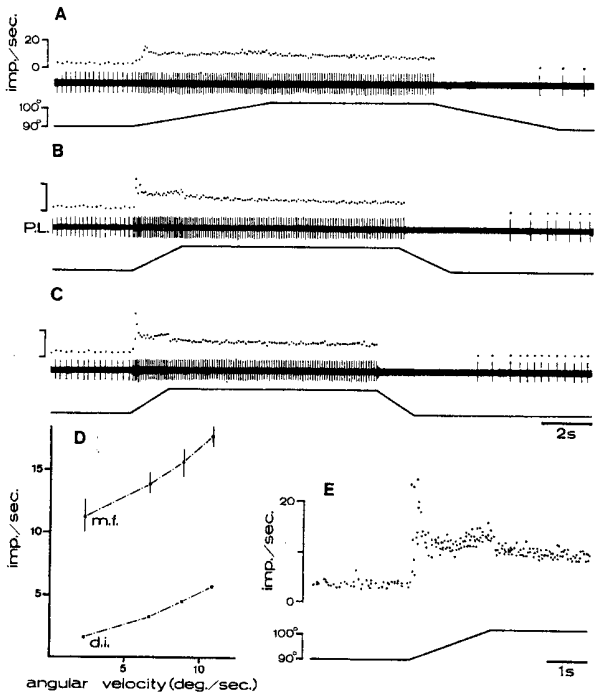


FIGURE 1 – Mesures sur les fibres primaires. (A, B et C) Réponses pour un mouvement avec trois vitesses angulaires différentes. (D) Fréquence instantanée en fonction de la vitesse angulaire de l'articulation. (E) Superposition de plusieurs réponses pour une même vitesse angulaire. Repris de Roll [131].

gains inconnus des fibres à l'étirement ($p(t)$) et à la vitesse d'étirement ($v(t)$). Puisque les fibres primaires sont plus sensibles à l'étirement qu'à la vitesse d'étirement, on a l'hypothèse :

$$a_{11} > a_{21} \text{ et } a_{22} > a_{12}. \tag{1}$$

En notant $\boldsymbol{x}(t) = [f_I(t), f_{II}(t)]^\mathsf{T}$, $\boldsymbol{A}$ la matrice 2×2 de coefficients a_{ij}, et $\boldsymbol{s}(t) = [p(t), v(t)]^\mathsf{T}$, (1) peut s'écrire sous la forme compacte :

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

où $\boldsymbol{A}$ et $\boldsymbol{s}(t)$ sont inconnus. Dans ces conditions, la question était la suivante : est-il possible d'estimer les signaux sources $\boldsymbol{s}(t)$ à partir de la seule mesure de $\boldsymbol{x}(t)$? Et si oui, comment ?

1.3 Première solution

Compte tenu de l'équation (1), on peut déduire que la matrice $\boldsymbol{A}$ est inversible. Ainsi l'estimation des caractéristiques du mouvement pourrait se faire en estimant une matrice $\boldsymbol{B}$, inverse de $\boldsymbol{A}$, ce qui est justifié en l'absence de bruit [47] :

$$\hat{\boldsymbol{s}}(t) = \boldsymbol{B}\boldsymbol{x}(t). \tag{3}$$

La première solution proposée par Ch. Jutten et J. Hérault en 1985 s'inspire de réseaux de neurones à inhibitions latérales récurrentes, et son schéma de principe est présenté à la figure 2, dans le cas le plus simple de 2 mélanges de 2 sources. Dans ce schéma, les connexions récurrentes sont représentées par des

flèches qui vont des sorties du réseau vers les entrées, avec les poids $-c_{12}$ et $-c_{21}$, le signe moins indiquant l'inhibition.

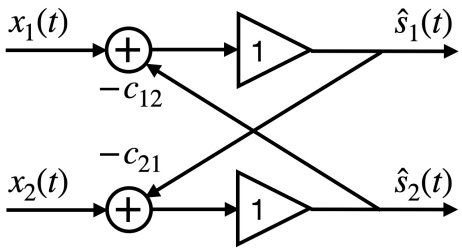


FIGURE 2 – Réseau de neurones à inhibitions latérales récurrentes pour résoudre le problème de séparation de sources. Les sorties, $\hat{s}_1(t)$ et $\hat{s}_e(t)$, sont des estimations des sources ajustées par adaptation des coefficients c_{12} et c_{21}.

Les sorties du réseau sont les estimations des sources inconnues et leurs équations s'écrivent :

$$\begin{cases} \hat{s}_1(t) &= x_1(t) - c_{12}\hat{s}_2(t) \\ \hat{s}_2(t) &= x_2(t) - c_{21}\hat{s}_1(t) \end{cases}$$

que l'on peut écrire également :

$$\begin{cases} \hat{s}_1(t) &= \frac{x_1(t)-c_{12}x_2(t)}{1-c_{12}c_{21}} \\ \hat{s}_2(t) &= \frac{x_2(t)-c_{21}x_1(t)}{1-c_{12}c_{21}} \end{cases}$$

L'idée est d'adapter les coefficients $-c_{12}$ et $-c_{21}$ de sorte que les sorties (les sources estimées) deviennent statistiquement indépendantes. Si on impose aux sorties d'être simplement décorrélées, on a les équations d'adaptation suivantes pour les coefficients :

$$c_{ij}(t+1) = c_{ij}(t) + \mu \hat{s}_i(t)\hat{s}_j(t),\ \forall i,j \in \{1,2\}. \qquad (4)$$

Lorsque les sorties – supposées centrées – sont décorrélées, l'espérance du terme $\hat{s}_i(t)\hat{s}_j(t)$ s'annule et les coefficients ne varient plus. Cette idée n'a pas été retenue car elle induit des variations des coefficients $-c_{12}$ et $-c_{21}$ identiques, en raison de la symétrie du produit. De plus, la décorrélation n'est pas suffisante pour garantir l'indépendance. Intuitivement, Ch. Jutten et J. Hérault ont cherché à casser la symétrie de la règle d'adaptation, en utilisant des fonctions non linéaires impaires différentes des sorties estimées :

$$c_{ij}(t+1) = c_{ij}(t) + \mu f(\hat{s}_i(t))\, g(\hat{s}_j(t)),\ \forall i,j \in \{1,2\}. \qquad (5)$$

Avec cette règle, on voit que les corrections apportées aux poids $-c_{12}$ et $-c_{21}$ sont différentes, et intuitivement, on comprend que les produits $f(\hat{s}_i(t))\, g(\hat{s}_j(t))$ font intervenir des statistiques d'ordre supérieur à deux, de façon à approcher l'indépendance. En développant les fonctions $f(.)$ et $g(.)$ en séries de Taylor, on comprend le lien de (l'espérance de) ce terme avec des cumulants croisés.

Cet algorithme intuitif était très simple. Il marchait bien (en général, malgré des cas pathologiques), avec une convergence rapide vers la décorrélation et une convergence plus lente vers la solution correspondant à l'indépendance, atteinte seulement en moyenne. Il se généralise aussi facilement à une dimension quelconque.

Cependant, dans les années 80, la puissance des ordinateurs était limitée (fréquence d'horloge de l'ordre de 10 MHz, mémoire vive de base de l'ordre de 256 kO), et l'exécution de cet algorithme (même pour 2 sources et 2 capteurs) demandait plusieurs dizaines de minutes jusqu'à la convergence. Aussi, en 1985, Ch. Jutten avait conçu une petite machine électronique [91], à base de transistors à effet de champ et d'amplificateurs opérationnels (Fig. 3), qui implantait l'algorithme de séparation de sources pour 2 sources et 2 capteurs. Les sources étaient soit des signaux provenant de générateurs de signaux, soit des signaux audio : on pouvait visualiser la convergence soit sur un oscilloscope, soit écouter les signaux sonores mélangés et séparés sur un haut-parleur. La convergence prenait de l'ordre de la seconde, et avait été ralentie pour faciliter sa visualisation. Plus tard, en 1989, E. Vittoz et X. Arreguit de l'EPFL avaient conçu un circuit intégré CMOS qui implantait cet algorithme de séparation de sources [158].

D'un point de vue théorique, cet algorithme était très différent des approches classiques de minimisation d'un critère de coût (fréquemment, une erreur quadratique) par des méthodes de gradient, et Jutten et Hérault n'avaient pas de résultats théoriques sur sa convergence. P. Comon a étudié en détail cet algorithme [48]. Il a montré notamment que cet algorithme cherche les zéros d'une fonction (et non le minimum d'une fonction de coût), et qu'il correspond à un algorithme d'itération stochastique de Robbins-Monro [130].

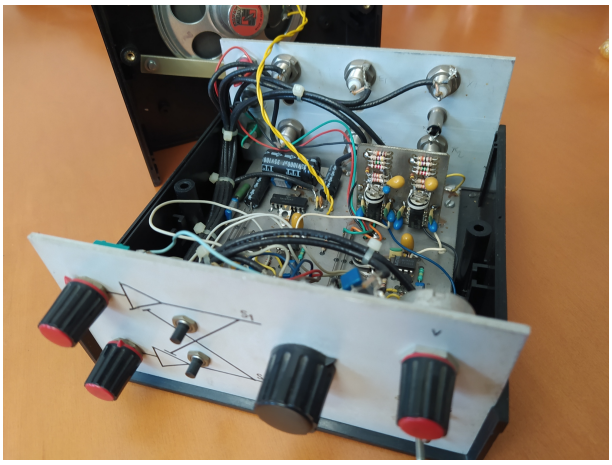


FIGURE 3 – Implantation électronique d'un algorithme de séparation de sources (2 sources et 2 capteurs). Cet appareil, réalisé en 1985, utilise des transistors à effet de champ et des amplificateurs opérationnels.

Ces résultats ont été présentées au GRETSI en 1985 [83], et à Neural Networks for Computing (qui est ensuite devenu NIPS) en 1986 [82], puis un peu plus tard dans le journal *Signal Processing* [92, 48] avec les résultats théoriques de Comon sur la convergence de l'algorithme. Mais, il restait beaucoup à comprendre !

2 Contexte scientifique des années 80

En 1982, quand on s'intéressait aux modèles mathématiques du cerveau, on connaissait le modèle de McCullogh et Pitts [107], le perceptron de Rosenblatt (1957) [132], le Neocognitron de Fukushima (1980) [72]. Les travaux de Kohonen [96] sur les cartes auto-organisatrices et les modèles de Hopfield [85] venaient juste de paraître. La naissance du perceptron multi-couches et de la rétropropagation du gradient [136] était à venir (1986). Dans la communauté réseaux de neurones artificiels, nos travaux ont retenus l'attention de quelques chercheurs, mais ont été occultés par ces autres avancées.

Dans la communauté signal, l'hypothèse de gaussianité des signaux et du bruit était dominante de sorte que les statistiques d'ordre deux (SOS) étaient privilégiées et la décorrélation même confondue avec l'indépendance statistique. Par ailleurs, dans ce contexte de SOS, des travaux avaient montré que le problème que nous abordions n'était pas soluble [15][47, ch.1].

L'accueil au GRETSI 85 a donc été empreint de surprise voire de scepticisme ! Il faut dire que le titre du papier [83] était long et peu explicite ! Les appelations Analyse en Composantes Indépendentes (ACI) et (en anglais) Independent Component Analysis (ICA), n'interviendront que plus tard, en 1987 [90], avant d'être formalisés rigoureusement par P. Comon dans le colloque HOS de Chamrousse en 1991 [37], puis dans *Signal Processing* en 1994 [40].

3 Quelques notes historiques

3.1 Quelques pionniers

À la suite du GRETSI 85, Laurent Kopp, intrigué par le papier [83], avait convaincu Thomson Sintra d'embaucher Pierre Comon en 1988 pour comprendre pourquoi et comment notre algorithme marchait. En parallèle, J.-L. Lacoume encadrait deux thèses au CEPHAG, d'abord celle de P. Ruiz qui a mis en évidence en 1988 l'apport des cumulants [135], puis celle de M. Gaeta qui a proposé d'approximer la vraisemblance par un développement de Gram-Charlier [73]. En 1987, Jean-François Cardoso, après une visite à Grenoble, s'est aussi intéressé au problème. P. Comon et J-F. Cardoso ont été des pionniers dans la compréhension théorique de l'ICA et dans la proposition de nouvelles idées : rôle des cumulants [34] [33], décomposition tensorielle [23, 40], diagonalisation conjointe [25, 12], concept d'équivariance [24]... Nous en donnons un aperçu dans la section 5.

3.2 Soutien en France et en Europe

Au sein du Groupe de Recherche Coordonnées "Systèmes Adaptatifs en Robotique et Traitement du Signal et Automatique" (GRECO SARTA), créé en 1984 pour 4 ans, une réflexion sur le traitement du signal et des images aboutit en 1988 à la création du GdR TDSI (Traitement du signal et de l'image), dirigé par Claude Gueguen puis Odile Macchi, qui deviendra plus tard le GdR ISIS.

De 1988 à 1993, le groupe de travail 9 (GT 9) "Ordres Supérieurs", est très animé autour de la non-gaussianité, des statistiques d'ordre supérieur à deux (HOS pour Higher Order Statistics), et de la séparation de sources. Sous la houlette de Jean-François Cardoso, de 1990 et 1997, de nombreuses journées sont organisées chaque année, avec des exposés tutoriaux et de doctorants qui sont très nombreux à s'ouvrir à ce domaine. Ce GT a ensuite été animé par Eric Moreau avec une forte activité jusqu'au milieu des années 2000. Par ailleurs Philippe Loubaton a animé le GT "multivariable" sur la déconvolution aveugle multivariée de 1996 à 1997, avec l'organisation de plusieurs journées sur la séparation de sources dans des mélanges convolutifs.

Enfin entre 1991 et 1995, un groupe de travail ATHOS (Advanced Topics in High Order Statistics), financé par la Commission Européenne et conduit par Pierre Comon, a contribué à promouvoir les HOS, la séparation de sources et l'ICA dans la communauté signal en Europe.

3.3 Workshops

L'intérêt pour la non gaussianité s'est concrétisé en fin des années 80, avec un premier colloque sur les statistiques d'ordres supérieurs à deux (HOS pour High Order Statistics) à Vail (USA) en 1989, puis le second à Chamrousse en 1991. Durant ces workshops bisannuels HOS, de nombreuses contributions sur l'ICA sont présentées.

Dans les conférences sur les réseaux de neurones, le thème existe mais reste un peu marginal, jusqu'au milieu des années 90 avec les algorithmes Infomax (1995) de Bell et Sejnowsky [11] et FastICA (1999) d'Hyvärinen [88]. Ce dernier algorithme a bénéficié d'une certaine popularité grâce à une publicité extensive, mais présente de gros inconvénients souvent passés sous silence [165] [166] [163].

Les acronymes BSS et ICA tardent à rentrer dans les EDICS d'IEEE SP Society : ce sera fait seulement en 2003. Entretemps, le premier workshop ICA est organisé à Aussois en janvier 1999 par Jean-François Cardoso, Philippe Loubaton et Christian Jutten. Dans cette petite station des Alpes, 130 scientfiques venus du monde entier viennent passer une semaine. Ce workshop aura lieu environ tous les deux ans, jusqu'en 2019 : Helsinky (Finlande) en 2000, San Diego (USA) en 2001, Nara (Japon) en 2003, Granada (Espagne) en 2004, Southampton (UK) en 2005, Charleston (USA) en 2006, Londres en 2007, Paraty (Brésil) en 2009, Saint-Malo en 2010, ..., 2017 à Grenoble. Dans le cadre de ces workshops, on peut noter l'organisation d'une compétition internationale de séparation de signaux audio, lancée à la suite d'un projet (soutenu par le GDR ISIS) de Rémi Gribonval, Cédric Févotte et Emmanuel Vincent [156].

4 Quelques résultats remarquables

Idées fondamentales Un résultat de G. Darmois [49] [47, p.330] fournit une réponse précise sur la séparabilité fondée sur l'indépendance statistique. Ce théorème peut être énoncé comme suit :

> **Théorème 1** *Soient s_n $n = 1, \ldots, N$, N variables aléatoires statistiquement indépendantes, et deux com-*

binaisons linéaires de ces variables :

$$x_1 = \sum_n a_n s_n \quad \text{et} \quad x_2 = \sum_n a'_n s_n.$$

Si x_1 et x_2 sont statistiquement indépendantes, alors les variables s_n pour lesquelles $a_n\, a'_n \neq 0$ sont nécessairement gaussiennes.

Une preuve simple est donnée dans [47, sec. 9.2.4]. Une preuve plus compliquée peut être trouvée dans [93, ch. 3] [69, sec. XV.8] [129, ch. 5] sous des hypothèses plus générales.

Soit $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$, un vecteur aléatoire égal au produit d'une matrice inversible (inconnue) $\boldsymbol{A}$, et d'un vecteur aléatoire (inconnu) $\boldsymbol{s}$ dont les composantes (les sources) sont des variables aléatoires mutuellement indépendantes (*i.e.* dans leur ensemble). Si $\boldsymbol{x}$ correspond au vecteur d'observations, on peut chercher une matrice de séparation $\boldsymbol{B}$ de manière à ce que $\boldsymbol{y} = \boldsymbol{B}\,\boldsymbol{x}$ soit une estimation de $\boldsymbol{s}$ (Fig. 4). En d'autres termes, $\boldsymbol{y}$ est relié aux sources par une matrice globale, $\boldsymbol{G} = \boldsymbol{B}\boldsymbol{A}$, et $\boldsymbol{y} = \boldsymbol{G}\boldsymbol{s}$. Comme nous allons le voir, il est utile d'utiliser le théorème suivant, attribué conjointement à Marcinkiewicz et Dugué [63, 93, 40] :

Théorème 2 *Soient $\boldsymbol{s}$ et $\boldsymbol{y}$ deux vecteurs aléatoires tels que $\boldsymbol{y} = \boldsymbol{G}\,\boldsymbol{s}$, où $\boldsymbol{G}$ est une matrice éventuellement rectangulaire, avec plus de colonnes que de lignes. Si les composantes de $\boldsymbol{s}$ sont mutuellement indépendantes, et si celles de $\boldsymbol{y}$ sont indépendantes deux à deux, alors s_k est gaussienne dès que $\boldsymbol{G}$ a deux composantes non nulles dans la même colonne.*

En effet, ce théorème 2 permet de se ramener à l'indépendence de paires de variables (y_i, y_j), et donc d'utiliser le théorème 1. Il permet aussi de comprendre que l'indétermination est réduite à une permutation-échelle seulement pour les composantes non gaussiennes [40], puisque chaque colonne de $\boldsymbol{G}$ ne contient qu'une seule composante non nulle.

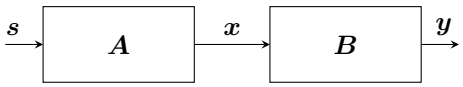


FIGURE 4 – Modélisation : seul le vecteur $\boldsymbol{x}$ est observé, $\boldsymbol{s}$ et $\boldsymbol{A}$ sont inconnues ; $\boldsymbol{y}$ est une estimation des sources, et $\boldsymbol{B}$ une matrice séparatrice.

Partitionnons $\boldsymbol{s}$ en deux parties, une partie gaussienne $\boldsymbol{s}_g$, et une partie non gaussienne, $\boldsymbol{s}_h$. Le vecteur observé peut donc s'écrire $\boldsymbol{x} = \boldsymbol{A}_h\,\boldsymbol{s}_h + \boldsymbol{A}_g\,\boldsymbol{s}_g$. Alors on peut établir le corollaire suivant [93, ch. 10] :

Corollaire 3 *La matrice $\boldsymbol{A}_h$ est identifiable à une permutation-échelle près, mais la matrice $\boldsymbol{A}_g$ n'est pas unique.*

Autrement dit, l'estimation de $\boldsymbol{s}$ à partir de $\boldsymbol{x}$ est impossible [1] pour les composantes de $\boldsymbol{s}$ qui sont à la fois gaussiennes indépendantes, et blanches stationnaires (iid).

Ce résultat conduit à deux approches fondamentalement différentes :

1. Si les sources sont des variables aléatoires **iid et non gaussiennes**, la solution conduit à l'ICA et requiert l'utilisation de statistiques d'ordre supérieur à deux. La prise en compte des liens entre échantillons successifs n'est pas nécessaire, mais la limitation est de ne pas pouvoir séparer des signaux gaussiens.
2. Si les sources sont des variables aléatoires **non iid et gaussiennes**, les sources doivent être soit des signaux colorés (échantillons successifs **dépendants**) [12] soit non stationnaires (échantillons successifs **non identiquement distribués**) [120]. L'avantage est la résolution avec des statistiques d'ordre deux, la limitation est de nécessiter des conditions supplémentaires sur l'autocorrélation (signaux colorés) ou le profil de variances (signaux non stationnaires) des sources.

5 Mélanges instantanés de sources non gaussiennes

À partir de ces résultats, Comon et Cardoso ont proposé des cadres théoriques rigoureux à l'ACI (Analyse en Composantes Indépendantes), qui soulignent les limites d'identifiabilité et le rôle du blanchiment à l'ordre 2 de $\boldsymbol{x}$. Dans la suite on notera P le nombre de capteurs et N le nombre de sources, de sorte que :

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} + bruit \tag{6}$$

où $\boldsymbol{A}$ est de dimension $P \times N$. Si $P < N$, le mélange est qualifié de "sous-déterminé", et de "sur-déterminé" dans le cas contraire, $P > N$. En présence de bruit, on doit généralement supposer $P \leq N$.

5.1 Blanchiment spatial

Dans cette section on présente le blanchiment spatial dans le cas complexe car cela ne complique pas les notations. L'indépendance statistique entre les composantes s_n implique en premier lieu leur décorrélation (et l'inverse n'est vrai que si les s_n sont gaussiennes). Une première idée est de décorréler les observations x_i par une transformation de *blanchiment spatial* $\tilde{\boldsymbol{x}} = \boldsymbol{W}\,\boldsymbol{x}$, souvent qualifiée de *standardisation* [40, section 2.2] [106]. La méthode la plus robuste consiste à utiliser l'Analyse en Composantes Principales (ACP) de $\boldsymbol{x}$ comme suit.

Soit $\boldsymbol{X}$ la matrice de dimensions $P \times T$ contenant les T observations. On calcule sa Décomposition en Valeurs Singulières (SVD) : $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H$. Les colonnes de $\boldsymbol{V}^H$ contiennent des vecteurs $\tilde{\boldsymbol{x}}$ dont la covariance est identité, et sont donc les versions standardisées de $\boldsymbol{x}$. Plus précisément, on a les relations $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}\boldsymbol{U}^H$ et $\boldsymbol{B} = \boldsymbol{Q}\boldsymbol{W}$, où $\boldsymbol{Q}$ sera unitaire.

Cette apparente facilité cache en réalité un problème difficile, sur lequel nous reviendrons dans la section 5.5. En effet, la standardisation n'est possible que si la matrice de covariance de $\boldsymbol{x}$ est bien conditionnée. Si ce n'est pas le cas, on peut projeter $\boldsymbol{x}$ sur les espaces singuliers dominants, ce qui réduira la dimension de $\boldsymbol{x}$; mais cela impose de choisir un seuil (sur les valeurs singulières) à partir duquel on néglige certaines données.

1. Cependant, si le nombre de sources n'excède pas le nombre de capteurs, et si une seule source est gaussienne, l'identification de $\boldsymbol{A}_h$ permettra quand même d'estimer la source gaussienne, par élimination.

Après standardisation, les choses sont théoriquement plus simples, puisqu'il suffit de chercher une matrice de séparation $\boldsymbol{Q}$ qui laissera la covariance de $\tilde{\boldsymbol{x}}$ inchangée ; cette matrice devra donc être orthogonale ($\boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I}$), et par définition bien conditionnée.

5.2 Information mutuelle

Dans [37, 40], Comon propose de minimiser l'information mutuelle des composantes de $\hat{\boldsymbol{s}}$, tout en permettant au modèle (2) d'être perturbé par un bruit additif de distribution inconnue.

L'idée est de chercher une métrique mesurant l'écart entre la distribution conjointe des sorties y_p et le produit de leurs distributions marginales. Si on adopte la divergence de Kullback $K(\cdot,\cdot)$ comme métrique, on obtient l'information mutuelle :

$$K(p_{\boldsymbol{y}}, \prod_p p_{y_p}) = \int_{\boldsymbol{u}} p_{\boldsymbol{y}}(\boldsymbol{u}) \log \left[\frac{p_{\boldsymbol{y}}(\boldsymbol{u})}{\prod_i p_{y_i}(u_i)} \right] d\boldsymbol{u} = I(p_{\boldsymbol{y}}). \tag{7}$$

Cette dernière est toujours négative, et nulle si et seulement si $p_{\boldsymbol{y}}(\boldsymbol{u}) = \prod_i p_{y_i}(u_i)$ presque partout. De plus, elle est invariante par changement d'échelle, c'est-à-dire par un changement $(\boldsymbol{y}, \boldsymbol{u}) \to (\boldsymbol{\Lambda}\boldsymbol{y}, \boldsymbol{\Lambda}\boldsymbol{u})$ où $\boldsymbol{\Lambda}$ est une matrice diagonale régulière. C'est une condition souhaitable [2] puisqu'on sait que la solution recherchée n'est déterminée qu'à un facteur d'échelle près.

Ensuite, on peut la décomposer en trois parties :

$$I(p_{\boldsymbol{y}}) = I(g_{\boldsymbol{y}}) + J(p_{\boldsymbol{y}}) - \sum_n J(p_{y_n}), \tag{8}$$

où $J(p_{\boldsymbol{y}})$ désigne la néguentropie de $\boldsymbol{y}$, mesurant l'écart à la gaussianité, et définie par :

$$J(p_{\boldsymbol{y}}) = K(p_{\boldsymbol{y}}, g_{\boldsymbol{y}}) = \int_{\boldsymbol{u}} p_{\boldsymbol{y}}(\boldsymbol{u}) \log \left[\frac{p_{\boldsymbol{y}}(\boldsymbol{u})}{g_{\boldsymbol{y}}(\boldsymbol{u})} \right] d\boldsymbol{u}$$

où $g_{\boldsymbol{y}}$ désigne la densité gaussienne ayant même moyenne et même variance que $p_{\boldsymbol{y}}$. Le jeu d'écriture (8) permet d'y voir alors plus clair. En effet, **(a)** $I(g_{\boldsymbol{y}})$ ne contient que des statistiques d'ordre 2, et sa maximisation assure la décorrélation entre les composantes de $\boldsymbol{y}$, **(b)** on peut montrer que $J(p_{\boldsymbol{y}})$ est invariant par transformations inversibles, c'est-à-dire constant lors de notre maximisation, et **(c)** la maximisation de $-J(p_{y_n})$ rend les composantes y_n les moins gaussiennes possible. Autrement dit, séparer les sources revient à "dégaussianiser" les sorties du séparateur ; ceci fait sens au vu du théorème de la limite centrale qui nous enseigne que mélanger des variables aléatoires indépendantes rend gaussien !

Il nous reste donc à minimiser $\sum_n J(p_{y_n})$. Or une intégrale sur u_n subsiste dans chacun des termes, et ceci reste coûteux même si nous n'avons plus que des intégrales monodimensionnelles.

Fort heureusement, on peut développer $J(p_{y_n})$ en série autour de la loi gaussienne, et ce développement fait apparaitre les cumulants de y_n ; en outre, si on adopte le développement de Edgeworth [3], les termes sont classés par importance au sens du théorème de la limite centrale, c'est-à-dire, par écart à la gaussianité. Plus précisément, en ne gardant que les termes les plus significatifs, on montre qu'après blanchiment spatial, on a [37, 40] :

$$J(p_{y_n}) \approx 4\gamma_{nnn}^2 + \gamma_{nnnn}^2 + 7\gamma_{nnn}^4 - 6\gamma_{nnn}^2\gamma_{nnnn} \tag{9}$$

où γ_{nnn} et γ_{nnnn} désignent les cumulants standardisés [4] d'ordre 3 et 4 de la variable y_n, respectivement.

Or, les cumulants d'ordre k de y_n dépendent linéairement des cumulants d'ordre k (inconnus) des sources, et sont des polynômes de degré k en les coefficients du mélange (inconnus aussi). Par exemple, si les sources sont symétriquement distribuées, alors les cumulants γ_{nnn} sont tous nuls, et le premier terme significatif, hormis $J(g_{\boldsymbol{y}})$, est :

$$\Psi(\boldsymbol{B}) = \sum_p \gamma_{nnnn}^2. \tag{10}$$

Maximiser une telle fonction est beaucoup moins difficile. Un aperçu de l'algorithme décrit dans [40] [43] est donné dans la section 5.6.

5.3 Vraisemblance

Dans [22], J.-F. Cardoso proposait d'estimer $\boldsymbol{A}$ en maximisant une vraisemblance empirique construite sur les échantillons $\boldsymbol{x}(t)$. Cette approche est très éclairante car elle permet d'inclure la connaissance éventuelle de la distribution des sources, et fait le lien avec *Infomax* et l'information mutuelle (7), toutefois en l'absence de bruit seulement comme nous allons le voir maintenant.

Si nous disposons de T réalisations indépendantes $\{x(1), \ldots, x(T)\} \overset{\text{def}}{=} \boldsymbol{X}_T$, alors la vraisemblance $p_{\boldsymbol{X}_T|\boldsymbol{A}}(\boldsymbol{u}_T|\boldsymbol{A})$ peut s'écrire, en faisant le changement de variable $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$:

$$p_{\boldsymbol{X}_T|\boldsymbol{A}}(\boldsymbol{u}_T|\boldsymbol{A}) = \frac{1}{|\det \boldsymbol{A}|} p_{\boldsymbol{s}}(\boldsymbol{A}^{-1}\boldsymbol{u}_T) \tag{11}$$

où $p_{\boldsymbol{s}}(\boldsymbol{u}) = \prod_n p_{s_n}(u_n)$ désigne la distribution conjointe des sources statistiquement indépendantes. En l'absence de bruit, le séparateur optimal est $\boldsymbol{B} = \boldsymbol{A}^{-1}$, de sorte qu'on peut écrire $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x}$, et $p_{\boldsymbol{X}_T|\boldsymbol{B}}(\boldsymbol{u}_T|\boldsymbol{B}) = |\det \boldsymbol{B}|\, p_{\boldsymbol{s}}(\boldsymbol{B}\boldsymbol{u}_T)$. Si maintenant T tend vers l'infini, la log-vraisemblance normalisée converge vers une limite familière :

$$\mathcal{L}_T(\boldsymbol{A}) \overset{\text{def}}{=} \frac{1}{T} \log p_{\boldsymbol{X}_T|\boldsymbol{A}} \to \int_{\boldsymbol{u}} p_{\boldsymbol{x}}(\boldsymbol{u}) \, \log p_{\boldsymbol{X}_T|\boldsymbol{A}}(\boldsymbol{u}|\boldsymbol{A}) d\boldsymbol{u}.$$

Il suffit maintenant d'ajouter et de soustraire $\log p_{\boldsymbol{x}}(\boldsymbol{u})$ dans l'intégrale pour faire apparaître une divergence de Kullback [22] [47, section 3.4] :

$$\mathcal{L}_\infty(\boldsymbol{A}) = -K(p_{\boldsymbol{x}}, p_{\boldsymbol{x}|\boldsymbol{A}}) + \int_{\boldsymbol{u}} p_{\boldsymbol{x}}(\boldsymbol{u}) \log p_{\boldsymbol{x}}(\boldsymbol{u}) du. \tag{12}$$

2. Les fonctions objectif vérifiant les propriétés d'invariance et de discrimination sont baptisées *contrastes* dans [40].

3. Le développement de Gram-Charlier [73], contrairement à celui de Edgeworth, ne classe pas les termes du développement par importance décroissante.

4. Si une variable $\boldsymbol{z}$ est standardisée, alors $\boldsymbol{y} = \boldsymbol{Q}\boldsymbol{z}$ l'est aussi, dès lors que $\boldsymbol{Q}$ est unitaire.

Le dernier terme est l'entropie de $\boldsymbol{x}$ et ne dépend pas de $\boldsymbol{B}$. Maximiser la vraisemblance revient donc à minimiser la divergence de Kullback $K(p_{\boldsymbol{y}}, p_{\boldsymbol{s}})$ [47, section 4.2]. En d'autres termes, on cherche $\boldsymbol{B}$ de manière à ce que la distribution des sorties $\boldsymbol{y}$ ressemble à celle des sources $\boldsymbol{s}$. Si on connaît la distribution des sources, ou même seulement les signes des kurtosis des sources [167], c'est un atout. Si on ne la connaît pas, on risque d'augmenter le biais d'estimation.

Enfin, en présence de bruit, le séparateur optimal n'est plus $\boldsymbol{A}^{-1}$ [47, sections 1.4 et 4.7], l'étude de la vraisemblance se complique et nous renvoyons à [47, section 4.7].

5.4 Lien entre Information Mutuelle et Vraisemblance

Pour y voir plus clair, on peut relier $K(p_{\boldsymbol{y}}, p_{\boldsymbol{s}})$ à l'information mutuelle. En effet, il est facile de montrer [22] [47, section 3.4] que :

$$K(p_{\boldsymbol{y}}, p_{\boldsymbol{s}}) = I(p_{\boldsymbol{y}}) + \sum_{n=1}^{N} K(y_n, s_n). \tag{13}$$

D'où on peut conclure qu'en l'absence de bruit, la vraisemblance est la somme de l'information mutuelle et de la divergence entre la distribution des sorties et des sources. L'équation (13) met clairement en évidence l'apport de la connaissance de la distribution des sources. Si cette dernière est trop mal connue, mieux vaut utiliser l'information mutuelle (section 5.2).

On peut aussi montrer que le critère *Infomax* introduit dans [11] est équivalent au maximum de vraisemblance en l'absence de bruit [47, section 4.2] [21].

5.5 Equivariance et performances

Le concept d'équivariance est bien décrit dans [22]. Il stipule qu'en l'absence de bruit, les performances d'un séparateur de sources ne devrait pas dépendre de la matrice de mélange $\boldsymbol{A}$ dès lors que celle-ci est inversible. Il y a deux limites à ce concept très séduisant de *performances uniformes* :

(i) il y a toujours un peu de bruit dans les mesures, et ce dernier peut être amplifié de manière importante si le séparateur est de la forme $\boldsymbol{B} = \boldsymbol{A}^{-1}$, ceci étant d'autant plus vrai que la matrice de mélange $\boldsymbol{A}$ est mal conditionnée ;

(ii) l'ensemble des "matrices inversibles" est lui-même mal défini, car il n'est pas fermé. Autrement dit, une suite de matrices inversible peut converger vers une matrice singulière, de sorte qu'un algorithme d'optimisation itératif ne peut pas fonctionner correctement sous une contrainte d'inversibilité ; il est nécessaire de réduire la recherche à un sous-ensemble plus petit [46, section VI].

Toutefois, si le bruit est suffisamment faible et n'est pas amplifié par le séparateur, l'hypothèse "sans bruit" peut être raisonnablement admise.

5.6 Quelques algorithmes de maximisation de contraste après blanchiment

Après blanchiment spatial, la matrice de covariance des observations est théoriquement l'identité, et on cherche une matrice séparante sous la forme d'une matrice orthogonale, de manière à ne pas modifier la covariance. Il existe maintenant de nombreuses approches au problème de séparation de sources par matrice orthogonale. Nous allons mentionner deux familles d'approches, qui exécutent des raffinements multiplicatifs. La première écrit la matrice orthogonale comme produit de rotations planes, et ramène le problème général à une succession de problèmes en dimension 2×2 pouvant être résolus algébriquement. Nous donnons cinq exemples d'algorithmes, chacun maximisant un critère de contratse différent. La deuxième reste en dimension supérieure à deux et effectue sa mise à jour à l'aide d'un gradient relatif.

5.6.1 Résolution en dimension $2 \rtimes 2$

Dans cette section, dans un souci de généralité, on admet que les données peuvent être complexes. Supposons que nous ayons 2 capteurs, et que la matrice de séparation $\boldsymbol{Q}$ recherchée soit unitaire. On note

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \boldsymbol{Q} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \; \boldsymbol{Q} = \begin{pmatrix} c & s \\ -s^* & c \end{pmatrix} \tag{14}$$

avec $c = \cos\alpha$ et $s = \sin\alpha . e^{\jmath\varphi}$, et $c^2 + |s|^2 = 1$. Ici, $\boldsymbol{z}$ est de covariance identité et est a priori un mélange bruité de $N > 2$ sources.

On note les cumulants d'ordre 3 et 4 de $\boldsymbol{z}$:

$$\mathcal{C}_{abc} = Cum\{z_a, z_b^*, z_c^*\} \text{ et } \mathcal{C}_{abcd} = Cum\{z_a, z_b^*, z_c^*, z_d\}$$

respectivement, et γ_{abc} et γ_{abcd} ceux de $\boldsymbol{y}$, sur le même format. Nous avons alors les relations multi-linéaires suivantes entre les cumulants d'entrée et de sortie :

$$\gamma_{pqr} = \sum_{abc} Q_{pa} Q_{qb}^* Q_{rc}^* \mathcal{C}_{abc} \tag{15}$$

$$\gamma_{pqrs} = \sum_{abcd} Q_{pa} Q_{qb}^* Q_{rc}^* Q_{sd} \mathcal{C}_{abcd} \tag{16}$$

Les cumulants $\mathcal{C}$ pouvant être estimés à partir des observations blanchies, les cumulants de sortie γ_{pqr} et γ_{pqrs} sont donc des polynômes trigonométriques en α, d'après (14).

Il existe plusieurs algorithmes, se différenciant les uns des autres selon la manière dont le problème en dimension 2×2 est traité et selon le choix de la fonction de contraste [45]. Nous décrivons ci-après 5 approches algébriques du problème $2 \rtimes 2$.

a) Ordre 3 réel. On peut adopter comme critère d'optimisation le contraste $\psi_3^{CoM2} = \sum_p |\gamma_{ppp}|^2$, si les cumulants γ_{ppp} des sources sont non nuls [5]. L'acronyme CoM2 a été introduit en 1994 et signifie "Contrast Maximization", le chiffre 2 précisant que les cumulants sont élevés au carré. Par ailleurs, l'indice 3 indique que ce sont les cumulants d'ordre 3 qui sont utilisés. En utilisant (15) et (14), on peut alors montrer que l'angle α optimal maximisant ψ_3^{CoM2} est donné par la résolution de deux trinômes du second degré en cascade [41], ou bien ce qui revient au même, par le calcul du vecteur propre dominant $\boldsymbol{v}$ d'une matrice 2×2 symétrique [47, p.167] de la forme :

$$\boldsymbol{v} = [\cos 2\alpha \,;\, \sin 2\alpha]^T. \tag{17}$$

5. En particulier, si une variable aléatoire est symétriquement distribuée par rapport à sa moyenne, son cumulant d'ordre 3 est nul.

b) Ordre 4 réel. Si les cumulants d'ordre 3 de certaines sources sont nuls, il convient de recourrier à ceux d'ordre 4, ce qui est plus coûteux mais bien plus fiable. On peut alors choisir comme critère de contraste : $\psi_4^{CoM2} = \sum_p |\gamma_{pppp}|^2$. Ici l'indice 4 précise que ce sont les cumulants d'ordre 4 qui sont utilisés. On montre alors que l'angle α optimal peut être obtenu en deux étapes. D'abord en résolvant un polynôme de degré 4 en la variable $\xi = \theta - 1/\theta$, où $\theta = \tan\alpha$. Parmi les 4 racines potentielles du polynôme de degré 4, on choisira celle qui maximise ψ_4^{CoM2}. Rappelons que la méthode de Ferrari permet de calculer explicitement les racines d'un polynôme de degré 4. Puis on résoud un polynôme de degré 2, $\theta^2 - \xi\theta - 1 = 0$, pour remonter à l'angle optimal $\alpha = \tan^{-1}\theta$ [37] [40] [47, p.169]. Les deux solutions obtenues sont équivalentes (elles se déduisent l'une de l'autre par permutation-signe).

c) Ordre 3 complexe. Dans le cas complexe et si les cumulants d'ordre 3 des sources sont non nuls, on peut suivre une procédure similaire à celle du (a), basée sur ψ_3^{CoM2} ; la solution est alors donnée par le vecteur propre dominant $\boldsymbol{v}$ d'une matrice 3×3 réelle symétrique, puis par les angles à travers la relation :

$$\boldsymbol{v} = [\cos 2\alpha\,;\ \sin 2\alpha \cos\varphi\,;\ \sin 2\alpha \sin\varphi]^T. \qquad (18)$$

comme indiqué dans [43] [47, p.168],

d) Ordre 4 complexe. Dans le cas complexe et si certains cumulants d'ordre 3 des sources sont nuls, la procédure décrite en (b) peut encore fonctionner, mais le polynôme de degré 4 a désormais deux variables (module et phase par exemple), ce qui rend le calcul des solutions plus coûteux [40, section 7.A.19].

Une seconde approche consiste à adopter le critère de contraste suivant : $\psi_4^{JAD} = \sum_{p,i,j} |\gamma_{ppij}|^2$, revenant à diagonaliser conjointement et approximativement des tranches matricielles du tenseur cumulant de sortie. L'acronyme JAD signifie "Joint Approximate Diagonalization". La maximisation de ce contraste permet d'accéder à une solution moins coûteuse [25] [26] [47, p.173], bien qu'aussi itérative *stricto sensu*. En effet, les angles α et φ sont obtenus par le calcul du vecteur propre dominant $\boldsymbol{v}$ d'une matrice 3×3 hermitienne, à travers une relation de type (18).

Une troisième approche consiste à diagonaliser conjointement et approximativement un ensemble de tranches tensorielles d'ordre 3 du cumulant d'ordre 4 des sorties. Pour ce faire, on maximise $\psi_4^{STD} = \sum_{p,i} |\gamma_{pppi}|^2$. L'acronyme STD signifie ici "Simultaneous Third order tensor Diagonalization". La solution peut être obtenue là encore par le calcul du vecteur propre dominant d'une matrice 3×3 [98] et une relation de type (18).

Ces trois solutions n'ont pas le même pouvoir discriminant car on montre que [6] :

$$\psi_4^{CoM2} \leq \psi_4^{STD} \leq \psi_4^{JAD}. \qquad (19)$$

Autrement dit, ψ_4^{CoM2} présente de meilleures performances que ψ_4^{STD} mais est plus coûteux. Et ψ_4^{STD} est à son tour plus coûteux que ψ_4^{JAD} mais ses performances sont meilleures. Cependant, la différence de performances est peu visible dans la plupart des problèmes courants.

6. La démonstration est immédiate, puisque tous les termes figurant dans ces sommes sont réels positifs, et que CoM2 a moins de termes que STD, qui, à son tour, a moins de termes que JAD.

e) Maximisation de la trace. Si les cumulants d'ordre 4 des sources sont de même signe ε, alors il est possible de calculer les angles α et φ à moindre coût. En effet le critère suivant est alors un contraste [45] [47, p.85] : $\psi_4^{CoM1} = \varepsilon \sum_p \gamma_{pppp}$. L'absence de carrés diminue le degré de la fraction rationnelle à maximiser, de sorte que la solution peut être obtenue en calculant le vecteur propre dominant d'une matrice 3×3 réelle symétrique, et une relation de type (18), comme précisé dans [43] [47, p.174],

Remarques Les codes Matlab de ces algorithmes sont en ligne depuis leur parution [35]. En outre, il existe des versions adaptatives [32].

5.6.2 Résolution en dimension $P > 2$ par balayage des paires

Rappelons que nous nous plaçons après blanchiment spatial, et qu'on cherche une matrice de séparation unitaire $\boldsymbol{Q}$ de manière à ce que le vecteur $\boldsymbol{y} = \boldsymbol{Q}\,\tilde{\boldsymbol{x}}$ soit une estimation du vecteur source à une permutation et à un facteur d'échelle près.

Or toute matrice unitaire $\boldsymbol{Q}$ peut se décomposer en produits de rotations planes [97] ; par conséquent, à l'instar de l'*algorithme de Jacobi* de diagonalisation des matrices hermitiennes, on peut exécuter une mise à jour itérative multiplicative de la matrice de séparation, de manière à maximiser un contraste [45]. Cette matrice orthogonale (unitaire dans le cas complexe) est paramétrée comme le produit d'un nombre arbitraire de rotations planes de la forme :

$$\boldsymbol{Q}[k,\ell] = \begin{pmatrix} \boldsymbol{I} & 0 & 0 & 0 & 0 \\ 0 & c & 0 & s & 0 \\ 0 & 0 & \boldsymbol{I} & 0 & 0 \\ 0 & -s^* & 0 & c & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{I} \end{pmatrix} \begin{matrix} \\ \leftarrow k \\ \\ \leftarrow \ell \\ \\ \end{matrix}$$

où les seuls éléments différents de la matrice identité se trouvent aux positions (k,k), $(k.\ell)$, (ℓ,k) et (ℓ,ℓ). En maximisant tour à tour la fonction de contraste par rapport à une rotation plane, on se ramène à une succession de problèmes d'optimisation résolus dans la section précédente. Il suffit alors de balayer toutes les paires d'indices, plusieurs fois s'il le faut [47, p.167-175]. L'important est que la solution de chacun des sous-problèmes puisse être obtenue rapidement (en fait, algébriquement).

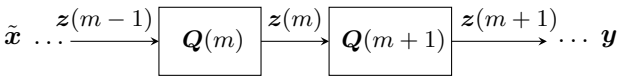


FIGURE 5 – Traitement par balayage de paires successives.

On suppose que les données ont été pré-blanchies spatialement, et on note $\tilde{\boldsymbol{x}}$ le vecteur d'observation blanchi. On applique une suite de rotations planes $\boldsymbol{Q}(m)$ à $\tilde{\boldsymbol{x}}$, et on note $\boldsymbol{z}(m)$ la sortie du filtre $\boldsymbol{Q}(m)$; autrement dit, $\boldsymbol{z}(0) = \tilde{\boldsymbol{x}}$, et $\boldsymbol{Q}(m)$ est une certaine matrice de rotation plane $\boldsymbol{Q}[k,\ell]$ qui agit dans le plan (k,ℓ). La relation entre m et (k,ℓ) dépend l'ordre de balayage des paires, et on pourra adopter un balayage cyclique par ligne, par exemple ; il n'est pas utile de rentrer dans le détail de cette relation pour comprendre le principe. A l'étape m, on calcule la

meilleure rotation plane maximisant le critère choisi (cf. section 5.6.1), on met à jour la sortie, $\boldsymbol{z}(m) = \boldsymbol{Q}(m)\boldsymbol{z}(m-1)$, et on accumule la matrice de séparation $\boldsymbol{Q} \leftarrow \boldsymbol{Q}(m)\boldsymbol{Q}$. Cette procédure, assez courante en algèbre linéaire [97], est illustrée par la figure 5.

5.6.3 Gradient relatif

Le concept de gradient relatif a été popularisé pour la séparation de sources principalement par Cardoso [24] [47, p.130, p.199]. Le point fort mis en avant était que les performances sont équivariantes (cf. section 5.5), mais c'est en réalité le cas de toutes les mises à jour multiplicatives (comme celles de la section précédente). Dans la suite de cette section, nous nous limiterons la présentation au corps des réels.

Dans le contexte de la séparation de sources, conformément à (3), on cherche la matrice séparante $\boldsymbol{B}$ qui maximise un critère $\Psi(\boldsymbol{B})$. Ici, on ne suppose plus nécessairement que les données ont été pré-blanchies spatialement, de sorte que $\boldsymbol{B}$ n'est pas forcément orthogonale. Pour simplifier, on va ici la supposer carrée inversible. Si on perturbe multiplicativement $\boldsymbol{B}$ par une matrice $\boldsymbol{I} + \boldsymbol{E}$ (où les valeurs de $\boldsymbol{E}$ sont petites), en faisant un simple développement [24] [47, §6.7.1], le critère s'écrit :

$$\Psi(\boldsymbol{B} + \boldsymbol{E}\boldsymbol{B}) = \Psi(\boldsymbol{B}) + \mathrm{trace}\{\nabla\Psi(\boldsymbol{B})^T \boldsymbol{E}\boldsymbol{B}\} + o(\boldsymbol{E}).$$

L'argument de la trace peut être remplacé, en appliquant la permutation circulaire $\mathrm{trace}\{abc\} = \mathrm{trace}\{cab\}$, par $\boldsymbol{B}\,\nabla\Psi(\boldsymbol{B})^T\boldsymbol{E}$, ce qui montre que si on définit le *gradient relatif* $\tilde{\nabla}\Psi$ en fonction du gradient par $\tilde{\nabla}\Psi(\boldsymbol{B}) = \nabla\Psi(\boldsymbol{B})\boldsymbol{B}$, alors

$$\Psi(\boldsymbol{B} + \boldsymbol{E}\boldsymbol{B}) = \Psi(\boldsymbol{B}) + \mathrm{trace}\{\tilde{\nabla}\Psi(\boldsymbol{B})^T\boldsymbol{E}\} + o(\boldsymbol{E})$$

et l'ascente du gradient relatif peut s'écrire multiplicativement

$$\boldsymbol{B} \leftarrow (\boldsymbol{I} + \mu\tilde{\nabla}\Psi(\boldsymbol{B}))\,\boldsymbol{B}. \tag{20}$$

Version stochastique. Admettons que le critère est l'espérance mathématique d'une fonction de la distribution de $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x}$, ce que l'on note $\Psi(\boldsymbol{B}) = \mathbb{E}\{f(\boldsymbol{y})\}$. Le critère ne dépend donc de $\boldsymbol{B}$ qu'à travers $\boldsymbol{y}$, ce qui est légitime en séparation de sources. Alors on peut décrire la version stochastique de ce gradient relatif comme suit [47, p.201] [24].

Tout d'abord $\Psi(\boldsymbol{B} + \boldsymbol{E}\boldsymbol{B}) = \mathbb{E}\{f(\boldsymbol{y} + \boldsymbol{E}\boldsymbol{y})\}$, et au premier ordre $\Psi(\boldsymbol{B} + \boldsymbol{E}\boldsymbol{B}) = \mathbb{E}\{f(\boldsymbol{y})\} + \mathbb{E}\{\nabla f(y)^T\boldsymbol{E}\boldsymbol{y}\} + o(\boldsymbol{E})$. En d'autres termes, ceci veut dire par définition de $\tilde{\nabla}$ que

$$\tilde{\nabla}\Psi(\boldsymbol{B}) = \mathbb{E}\{\nabla f(\boldsymbol{y})\,\boldsymbol{y}^T\}$$

et la mise à jour s'écrit $\boldsymbol{B} \leftarrow (\boldsymbol{I} + \mu\,\mathbb{E}\{\nabla f(\boldsymbol{y})\,\boldsymbol{y}^T\})\boldsymbol{B}$. La version stochastique s'obtient par simple suppression de l'espérance mathématique :

$$\boldsymbol{B}_{k+1} = [\boldsymbol{I} + \mu_k\nabla f(\boldsymbol{y}_k)\,\boldsymbol{y}_k^T]\boldsymbol{B}_k \tag{21}$$

Version pour matrice orthogonale. Si la matrice $\boldsymbol{B}$ est orthogonale, il convient d'en imposer la contrainte, en remarquant que la différentielle d'une matrice orthogonale est antisymétrique [39]. Ceci conduit à l'expression suivante [24] :

$$\boldsymbol{B}_{k+1} = [\boldsymbol{I} + \mu_k(\nabla f(\boldsymbol{y})\,\boldsymbol{y}^T - \boldsymbol{y}\,\nabla f(\boldsymbol{y})^T)]\boldsymbol{B}_k \tag{22}$$

Une version normalisée, plus stable, est décrite dans [24], exécutant conjointement le blanchiment spatial.

Choix de la fonction $f(\boldsymbol{y})$. Le choix du critère d'optimisation est important, comme nous l'avons déjà vu dans la section 5.6. Ce critère peut faire intervenir les cumulants de $\boldsymbol{y}$, ou bien, si la distribution des sources attendues est connue, la fonction score des sources appliquée à $\boldsymbol{y}$ [47, p.203] [21] .

5.7 Autres algorithmes

Dans [52], l'algorithme de *déflation* proposé est différent des précédents bien que s'appliquant aussi après pré-blanchiment spatial. La déflation s'opère alors naturellement en réduisant la dimension de l'espace, par suppression d'une colonne de la matrice orthonormée [47, section 6.9].

L'algorithme FastICA s'applique pour des mélanges instantanés, soit par déflation après blanchiment spatial, ou sans blanchiment et sans déflation [88]. La forme particulière de la fonction objectif essaie de prendre en compte un *a priori* sur la distribution des sources (comme dans la section 5.6.3). En dépit de son nom, FastICA n'est pas si rapide que ça [166] [165].

Enfin, si les sources sont à valeurs discrètes, on peut signaler la technique de *déflation parallèle* proposée dans [162], qui exploite explicitement la connaissance de l'alphabet de chaque source, et peut être utilisée dans les mélanges sous-déterminés. Chaque alphabet fait l'objet d'une déflation séparée, ce qui limite l'accumulation d'erreurs (toujours présente malgré la projection sur l'alphabet).

6 Mélanges convolutifs

Le modèle linéaire (1) demande à être amélioré pour certaines applications. Par exemple, dans le cas de signaux audio (musique, parole) enregistrés sur des microphones, ou en télécommunications non coopératives, il faut tenir compte de la propagation des ondes entre les sources et les récepteurs. La matrice de mélange $\boldsymbol{A}$ doit être remplacée par une matrice dont les coefficients sont des filtres : on a alors des mélanges linéaires convolutifs [45] [47, ch.8].

6.1 Mélange SISO

Historiquement, c'est le problème de l'égalisation aveugle SISO (une entrée et une sortie) qui a été étudié en premier, principalement dans le contexte des télécommunications. Nous n'allons pas en faire la revue, mais seulement mentionner ce qui nous sera utile pour la séparation de sources. Si on cherche à compenser l'effet d'un canal qui a mélangé une suite de variables i.i.d. (égalisation aveugle), on cherche en réalité à séparer un mélange particulier. Rendre les variables aléatoires décorrélées ne suffit pas, si le canal n'est pas à *minimum de phase*; il existe en effet une infinité de manières de rendre un signal blanc à l'ordre 2, toutes équivalentes à un filtre passe-tout près. La solution doit donc imposer une blancheur au sens fort (i.i.d.).

L'article précurseur en la matière est sans aucun doute celui de Donoho [58]. C'est lui qui a établi le fait que "démélanger revient à rendre moins gaussien". Le kurtosis peut être une mesure

d'écart à la gaussianité, de même que l'entropie. Les concepts développés ultérieurement dans [140] dans le cas SISO convolutif et dans le cas MIMO instantané [37, 40] peuvent rétrospectivement être vues comme des extensions – si on omet le fait que [58] n'a été découvert qu'après-coup.

6.2 Mélange SIMO

Quand le signal source n'est pas une séquence i.i.d. (c'est-à-dire blanche), on ne peut pas retrouver le canal ni la source sans hypothèse supplémentaire. En revanche, ceci devient possible si on reçoit le signal source sur *plusieurs* récepteurs distincts, après propagation à travers des canaux *différents*. C'est le modèle SIMO (une source et plusieurs capteurs). Dans le cas non bruité, le modèle est le suivant : $x_p(t) = h_p(t) \star s(t)$, où $h_p(t)$ désigne la réponse impulsionnelle du canal reliant la source $s(t)$ au récepteur p. Notons $\boldsymbol{x}(t)$ le vecteur de composantes $x_p(t)$ et $\boldsymbol{h}(t)$ le vecteur de composantes $h_p(t)$. Alors en prenant la transformée en z, on peut écrire la relation précédente sous forme compacte :

$$\boldsymbol{x}(z) = \boldsymbol{h}(z)\, s(z)$$

où $s(z)$ est scalaire. L'idée sous-jacente à l'identification aveugle SIMO est la suivante : si les canaux $h_i(z)$ sont des polynômes premiers entre eux, alors ils peuvent être identifiés, et on obtiendra théoriquement $s(z)$ en calculant le PGCD des $x_p(z)$. En pratique, on ne prend pas la transformée en z, mais on forme l'équation d'observation avec une matrice Töplitz par blocs, $\boldsymbol{T}(\boldsymbol{h})$:

$$\boldsymbol{x}_T = \boldsymbol{T}(\boldsymbol{h})\, \boldsymbol{s}_T \tag{23}$$

où $\boldsymbol{x}_T$ et $\boldsymbol{s}_T$ contiennent les échantillons des observations et des sources empilés de la manière adéquate [112].

On peut identifier $\boldsymbol{h}$ et estimer $s(t)$ si la matrice $\boldsymbol{T}(\boldsymbol{h})$ est de rang plein ; une condition nécessaire est précisément que ces polynômes $h_p(z)$ n'aient pas de zéros communs. Sous cette condition, diverses approches ont été proposées, soit par *prédiction linéaire* [2], soit par *méthode de sous-espace* [112] ; le premier type d'approche étant souvent considéré comme plus robuste [2], mais avec quelques réserves [101]. Les solutions proposées dans le cas SIMO ne recourrent qu'aux moments d'ordre 2.

6.3 Mélange MIMO

En séparation de sources, on s'intéresse évidemment au cas MIMO (le cas SIMO ne comporte qu'une seule source). Un point important à signaler est que supposer que le filtre égaliseur de canal est à réponse impulsionnelle finie (RIF), quand le canal (mélange) est RIF aussi, n'est pas restrictif dans le cas *multivarié*, contrairement à la dimension 1 ; la condition est que tous les mineurs de la matrice de mélange $\boldsymbol{A}(z)$ soient premiers entre eux [128], ce qui est vrai avec probabilité 1. Il faut aussi que le nombre de capteurs soit au moins égal au nombre de sources, bien évidemment. Ceci est une conséquence directe de l'identité de Bézout pour les matrices polynômiales [94, Lemme 6.3-9]. En d'autres termes, si $\boldsymbol{A}(z)$ est carrée et que son déterminant ne dépend pas de z, alors elle admet une inverse à gauche polynômiale [7]. Autrement dit, il existe généralement un inverse RIF à un canal RIF. Outre les premiers travaux de Nguyen Thi et al. [116] étendant la règle d'adaptation (5) à des mélanges convolutifs, on peut mentionner trois familles de méthodes.

1. La première idée est sans doute celle présentée en 1994 dans [41] où un modèle MA multivarié *non monique* est identifié en aveugle à l'aide des statistiques d'ordre supérieur à deux, par exemple en suivant l'algorithme de [38] ; le coefficient de retard nul est identifié ensuite par Analyse en Composantes Indépendantes. Les sources peuvent être extraites dans un second temps par régression entre les entrées et les sorties (innovations théoriquement blanches). On retrouve l'idée de *prédiction linéaire* MIMO plus tard dans [2] sous une autre forme, ainsi qu'une analyse de performances ; dans [79] ces performances sont optimisées grâce à une pondération optimale.

2. Dans [1] l'identification aveugle par *sous-espace* est étendue au cas MIMO. Outre une analyse asymptotique de performances, on y trouvera des commentaires sur le lien avec l'identification de modèles MA multivariés.

3. Une troisième approche est celle par *déflation*. L'idée consiste à extraire les sources une par une (i) en maximisant une fonction de contraste, puis (ii) en soustrayant leur contribution à l'observation par régression linéaire [152]. Notons qu'on peut éviter la deuxième étape de régression en effectuant un préblanchiment spatial et en utilisant un algorithme approprié [52].

La maximisation est faite localement avec un algorithme d'ascente itérative de type gradient dans [152], et par maximisation globale 1D par recherche de zéros d'un polynôme dans [161] ; cette technique est qualifiée de *RobustICA* dans [165]. Notons que cette dernière technique de recherche du maximum global selon des directions successives peut aussi être appliquée dans d'autres contextes, tels que la séparation de sources de module constant [164].

6.4 Approche dans le domaine fréquentiel ou temps-fréquence

Dans le cas de mélanges convolutifs, le signal observé satisfait le modèle :

$$\boldsymbol{x}(t) = \boldsymbol{A}(t) * \boldsymbol{s}(t), \tag{24}$$

où $*$ représente le produit de convolution, et $\boldsymbol{A}(t)$ est maintenant une matrice dont les coefficient sont des filtres inconnus. Le signal reçu sur le capteur p peut s'écrire :

$$x_p(t) = \sum_{n=1}^{N} \sum_{\tau=1}^{L} a_{pn}(\tau) s_n(t-\tau), \tag{25}$$

où les $a_{pn}(\tau)$ sont des filtres inconnus de longueur L.

Ce modèle temporel peut être transformé dans le domaine fréquentiel, ce qui peut conduire à différentes approches [139], qui ont été largement étudiées pour la séparation de signaux audio, notamment de parole et de musique.

En appliquant une transformée de Fourier à court terme (STFT) sur le modèle (25), on obtient l'équation :

$$x_{ij,p} = \sum_{n=1}^{N} a_{i,pn} s_{ij,n}, \tag{26}$$

7. De telles matrices sont dites *unimodulaires* [94, section 6.3].

où $x_{ij,p}$ représente le coefficient associé au signal reçu sur le capteur p, dans la bande de fréquence i et dans la fenêtre temporelle j, $a_{i,pn}$ est le coefficient du filtre entre le capteur p et la source n dans la bande de fréquence i, et $s_{ij,n}$ est la coefficient associé à la source n dans la bande de fréquence i et dans la fenêtre temporelle j. Notons que ces coefficients sont à valeurs complexes.

Le problème revient alors à résoudre pour chaque paire d'indices ij un problème de séparation instantané... Mais, les indéterminations de permutation et d'échelle rendent compliquée l'estimation des sources.

Une solution simple et efficace a été apportée par l'Independent Vector Analysis (IVA) [95] qui consiste à considérer conjointement toutes les fréquences.

Une autre approche est de considérer le spectre de puissance $\|x_{ij,p}\|^2$ du signal observé pour chaque paire d'indices ij. Cette quantité est évidemment non négative et on peut chercher à l'exprimer comme une somme pondérée (avec des coefficients positifs) des spectres de puissance des sources. Ceci conduit à un problème de factorisation de matrices non négatives, dont on peut trouver des solutions dans [19, 70, 64, 117].

Enfin, il est possible de ne pas travailler dans le domaine fréquenciel en utilisant un modèle ARMA multivarié. Comme dans IVA, l'intérêt est de ne pas créer artificiellement de multiples indéterminations de permutation-échelle. La principale difficulté réside dans la partie MA mais se résoud bien [38]. La procédure consiste à estimer les processus d'innovation (égalisation), puis à récupérer les sources ensuite par régression linéaire entre les innovations et les observations [33] [36].

6.5 Autres approches plus spécifiques

Lorsque les sources sont complexes et de module constant [164], ou lorsqu'elles appartiennent à un alphabet fini connu, tel que PSK-n [160], courantes en télécommunications, on peut adapter la fonction objectif pour égaliser un canal SISO de manière plus efficace. Cette idée s'étend au cas MIMO et à des alphabets quelconques et des approches dédiées existent [133] [45].

Lorsque les sources sont simplement retardées et amorties dans le mélange observé, ce qui est courant en traitement d'antenne, on peut (et même on doit) aborder le problème différemment [67] [68]. Ces approches peuvent même fonctionner dans les mélanges sous-déterminés (cf. section 7).

La technique de *déflation parallèle* (mentionnée plus haut dans le cas instantané) permet de réduire l'accumulation des erreurs de la déflation : on extrait les sources de modulations différentes en parallèle. Elle exploite explicitement la connaissance de l'alphabet de chaque source, et peut être utilisée dans les mélanges convolutifs sous-déterminés [133].

7 Mélanges linéaires sous-déterminés

7.1 Identification aveugle

Lorsque le nombre (P) de capteurs (taille de $\boldsymbol{x}$) est plus petit que le nombre (N) de sources (taille de $\boldsymbol{s}$), la matrice de mélange $\boldsymbol{A}$ est rectangulaire et non inversible. Dans un tel cas, elle peut être *identifiée* malgré tout sous des hypothèses assez faibles [47, ch.9]. Par contre, même si $\boldsymbol{A}$ est connue, l'*estimation* de $\boldsymbol{s}$ reste un problème très difficile, qui requiert des hypothèses supplémentaires. L'exemple théorique des mélanges 2×3 en est une bonne illustration [44]. C'est pourquoi, selon l'article précurseur de Tong [151], on préfèrera faire la distinction entre les problèmes d'identification du mélange, et celui de la séparation (extraction) des sources, pour plus de clarté.

Notons que les mélanges particuliers où les sources apparaissent seulement retardées et amorties ne demandent que l'estimation d'un nombre réduit de paramètres [68]. En télécommunications, ce problème est bien connu lorsque les canaux de transmission sont non sélectifs, et résolu de manière ad-hoc [109].

7.2 Extraction/séparation aveugle

Dans le cas sous-déterminé, on peut résoudre le problème de l'extraction de sources principalement de deux manières, en considérant : (i) soit le caractère discret des sources [45], (ii) soit la propriété de parcimonie (présence intermittente des sources).

L'exploitation de la parcimonie a été abordée par de nombreux chercheurs, avec des résultats d'identifiabilité de Donoho et al. et de Gribonval et al. Nous renvoyons au chapitre [47, ch.10] pour plus de détails théoriques, ainsi qu'une description des algorithmes et des applications en séparation audio ou en imagerie hyperspectrale.

Concernant le caractère discret, on doit en premier lieu adopter un critère d'optimisation adéquat prenant en compte la connaissance de l'alphabet des sources, tel que le contraste proposé dans [45, Théorème 16]. Cette connaissance conduit à des algorithmes itératifs [162]. Les solutions algébriques sont rares, et une solution est de créer des capteurs virtuels pour se ramener au cas sur-déterminé [42] [44] [29]. Une autre solution est celle de la *déflation parallèle*, bien adaptée au cas sous-déterminé, qui fonctionne même en convolutif [133].

Exemple. Intuitivement, l'identifiabilité (toujours à une permutation de colonnes près) peut se comprendre en un coup d'œil avec la figure 6 représentant un mélange instantané de 3 sources binaires de $\{-1,\ 1\}$; à gauche la matrice de mélange vaut

$$\boldsymbol{A} = \begin{bmatrix} 07 & 0.2 & 0.4 \\ 1 & -0.5 & 0.1 \end{bmatrix}$$

et est identifiable, et les sources peuvent être extraites. Par contre à droite la matrice de mélange est

$$\boldsymbol{A} = \begin{bmatrix} 0,2 & 0,2 & 0,4 \\ 0,5 & -0,5 & 0,1 \end{bmatrix}$$

et les sources ne peuvent pas être extraites (bien que la matrice de mélange soit d'ailleurs encore identifiable) car deux clusters sont superposés. Le niveau du bruit aussi intervient : s'il avait été trop fort, la position des clusters n'aurait pas pu être identifiée.

7.3 Sources parcimonieuses

De nombreux signaux sont parcimonieux dans la mesure où de nombreuses valeurs sont nulles. C'est par exemple le cas de

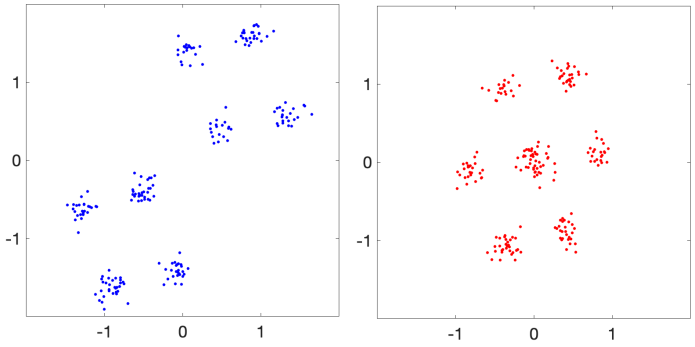


FIGURE 6 – Exemple d'un mélange de $N = 3$ sources binaires reçues sur $P = 2$ capteurs (le premier en abscisses et le second en ordonnées). À gauche : la matrice de mélange et le niveau de bruit sont tels que l'identification du mélange et l'extraction des sources par clustering sont possibles et aisées. À droite la matrice de mélange ne le permet pas car deux clusters sont superposés.

spikes dans les signaux cérébraux, de signaux électrocardiographiques réduits aux pics R, d'images hyperspectrales dans lesquelles un pixel peut ne contenir qu'un seul matériau. La parcimonie peut aussi être vue par blocs, par exemple, dans des signaux de parole qui sont une alternance de zones actives et de silences.

Par définition, la pseudo-norme ℓ_0 d'un vecteur $\boldsymbol{v} \in \mathbb{R}^N$, notée $\|\boldsymbol{v}\|_0$, est égale au nombre de composantes non nulles du vecteur. Si $\|\boldsymbol{v}\|_0 = k$, on définit alors le niveau de parcimonie par le rapport :

$$\alpha = \frac{k}{N}.$$

Le niveau de parcimonie ainsi défini est en réalité un niveau d'activité : la parcimonie est grande quand α est proche de zéro,

Un processus aléatoire $s(t)$ sera considéré comme parcimonieux si la plupart de ses réalisations sont nulles, c'est-à-dire si la probabilité $Pr(|s(t)| > 0)$ est proche de 0. On pourra ainsi définir le niveau de parcimonie (sparsity level) par le facteur :

$$\alpha = Pr(|s(t)| > 0).$$

Considérons maintenant un vecteur $\boldsymbol{s}_t = [s_1(t), \ldots, s_N(t)]^T$, dont les compopsantes sont N sources statistiquement indépendantes, supposées parcimonieuses et de même niveau de parcimonie α. Soit $\boldsymbol{s}_t \in \mathbb{R}^N$ une mesure de ce vecteur aléatoire à un indice t, le nombre moyen de sources actives est égal à αN. De façon plus précise, on peut calculer la probabilité que ce vecteur $\boldsymbol{s}_t$ ait k composantes non nulles, c'est-à-dire $Pr(\|\boldsymbol{s}_t\|_0 = k)$. Si les observations aux différents indices t sont indépendantes et identiquement distribuées, on a :

$$Pr(\|\boldsymbol{s}_t\|_0 = k) = \binom{N}{k} \alpha^k (1-\alpha)^{N-k}. \qquad (27)$$

On peut alors en déduire la probabilité que $\boldsymbol{s}_t$ ait moins de K composantes non nulles :

$$Pr(\|\boldsymbol{s}(t)\|_0 \leq K) = \sum_{k=1}^{K} \binom{N}{k} \alpha^k (1-\alpha)^{N-k}. \qquad (28)$$

Même pour des sources très parcimonieuses, le nombre de sources actives pour un indice t augmente avec N. Par exemple, pour un vecteur $\boldsymbol{s}_t$ de $N = 20$ sources très parcimonieuses avec $\alpha = 0,1$ (10% de valeurs non nulles), on a :

— $Pr(\|\boldsymbol{s}_t\|_0 = 1) = 0,27$,
— $Pr(\|\boldsymbol{s}_t\|_0 = 2) = 0,29$,
— $Pr(\|\boldsymbol{s}_t\|_0 = 3) = 0,19$.

7.3.1 Mélanges de sources parcimonieuses

Considérons le modèle de mélange sous-déterminé $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$, où $\boldsymbol{A}$ est une matrice de mélange rectangulaire, de taille $P \times N$, avec $P < N$. Si les N sources (les composantes de $\boldsymbol{s}$) sont très parcimonieuses, le nombre de composantes actives à chaque indice t est petit :

— si toutes les sources sont nulles, $\boldsymbol{s}_t = \boldsymbol{0}$ et $\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{s}_t = \boldsymbol{A}\,\boldsymbol{0} = \boldsymbol{0}$;
— si une seule source (composante) de $\boldsymbol{s}_t$ est non nulle, par exemple la i-ème : s_{it}, l'observation $\boldsymbol{x}_t$ est égale à :

$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{s}_t = \sum_{j=1}^{N} \boldsymbol{a}_j s_{jt} = \boldsymbol{a}_i s_{it}. \qquad (29)$$

Ces observations sont portées par le vecteur $\boldsymbol{a}_i$ dans l'espace à P dimensions des observations.
— si deux sources (ou plus) de $\boldsymbol{s}_t$ sont non nulles, l'observation $\boldsymbol{x}_t$ est engendrée par plusieurs vecteurs colonnes de la matrice $\boldsymbol{A}$.

La figure 7 illustre cette situation dans le cas d'une matrice $\boldsymbol{A}$ de dimension 2×3 (2 mélanges de 3 sources), avec des sources de niveau de parcimonie $\alpha = 0,1$ pour des mélanges non bruités (à gauche) et bruités (à droite).

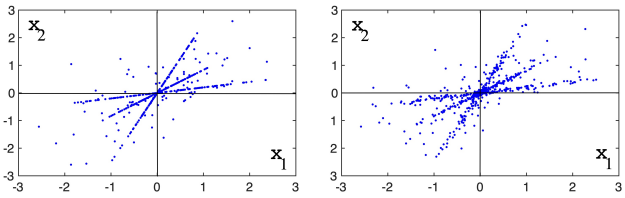


FIGURE 7 – Dans le plan cartésien des observations ($P = 2$), en raison de la parcimonie des sources, la plupart des points sont localisés sur les 3 sous-espaces de dimension 1, engendrés par les 3 vecteurs colonnes de la matrice de mélange. Les autres points, pour des indices t tels que 2 ou 3 sources sont actives simultanément, se trouvent dispersés dans le plan.

La figure 7 montre comment on peut facilement estimer, par de simples algorithmes de clustering, les colonnes de la matrice de mélange $\boldsymbol{A}$, lorsque les sources sont suffisamment parcimonieuses. Cette parcimonie des sources a conduit à de nombreux algorithmes dans lesquels on recherche les observations où une seule source est active : on parle de source dominante en audio [3], ou de pixel pur en imagerie hyperspectrale [115].

Si les sources ne sont pas très parcimonieuses, à chaque indice t, deux composantes (ou plus) de $\boldsymbol{s}_t$ peuvent être non nulles. Dans ce cas, les observations $\boldsymbol{x}_t$ sont situées dans un sous-espace, à deux dimensions ou plus, engendré par plusieurs colonnes de

la matrice de mélange $\boldsymbol{A}$. Si n sources sur N (le nombre total de sources) sont actives, le nombre de sous-espaces possibles est égal à $\binom{N}{n}$. Par exemple, dans [114] les auteurs procèdent en deux étapes :

— Une première étape consiste à estimer les sous-espaces, notés $\boldsymbol{B}_i, i = 1, \ldots, N_{se}$, dans lesquels les observations sont concentrées.

— La seconde étape consiste, à partir des sous-espaces $\boldsymbol{B}_i$, à identifier les vecteurs colonnes de la matrice de mélange en calculant pour tout vecteur $\boldsymbol{v}$ la fonction :

$$g_\sigma(\boldsymbol{v}) = \sum_{i=1}^{N_{se}} \exp\Big(-\frac{d^2(\boldsymbol{v}, \boldsymbol{B}_i)}{2\sigma^2}\Big), \tag{30}$$

où $d(\boldsymbol{v}, \boldsymbol{B}_i)$ est la distance entre le vecteur $\boldsymbol{v}$ et le sous-espace $\boldsymbol{B}_i$.

Dans la seconde étape, si le vecteur $\boldsymbol{v}$ est tel que $d(\boldsymbol{v}, \boldsymbol{B}_i)$ est petit par rapport à σ, le terme exponentiel est proche de 1. Ainsi, $g_\sigma(\boldsymbol{v})$ est approximativement égale au nombre de sous-espaces $\boldsymbol{B}_i$ qui contiennent $\boldsymbol{v}$. Dans le cas où n sources parmi N sont actives, il y a $\binom{N}{n}$ sous-espaces $\boldsymbol{B}_i$. Pour un vecteur colonne donné de la matrice $\boldsymbol{A}$, il y a donc $n-1$ vecteurs à choisir parmi $N-1$, c'est-à-dire $\binom{N-1}{n-1}$ sous-espaces : on a donc $g_\sigma(\boldsymbol{v}) \leq \binom{N-1}{n-1}$. Les vecteurs qui atteignent ce maximum sont les vecteurs colonnes de $\boldsymbol{A}$, à un facteur d'échelle près. Cette approche nécessite une recherche exhaustive des vecteurs $\boldsymbol{v}$ maximisant $g(\boldsymbol{v})$, donc en dimension P. Dans [114], le vecteur $\boldsymbol{v}$, normalisé, est représenté en coordonnées sphériques que l'on balaie avec un pas suffisamment fin. La complexité de recherche exhaustive explose avec P et c'est une limitation de cette méthode.

Par exemple, dans le cas où 2 sources parmi 4 sont actives, il y a $\binom{4}{2} = 6$ sous-espaces $\boldsymbol{B}_i$. Le nombre de sous-espaces $\boldsymbol{B}_i$ dans lesquels peut se trouver un vecteur colonne particulier de $\boldsymbol{A}$, est égal à $\binom{4-1}{2-1} = 3$ qui est le maximum atteint par $g_\sigma(\boldsymbol{v})$. Des détails sur ce type d'approches, sur la méthode présentée rapidement ci-dessus et sur les algorithmes sont donnés dans [114] ainsi que dans les références qui y sont citées.

7.3.2 Estimation des sources après estimation de $\boldsymbol{A}$

Supposons que l'on dispose d'une estimation $\widehat{\boldsymbol{A}}$ de $\boldsymbol{A}$. Dans le cas sous-déterminé, la matrice $\widehat{\boldsymbol{A}} \in \mathbb{R}^{P\times N}$ n'est pas inversible car $N > P$, et il reste encore à résoudre le problème mal posé et difficile :

$$\boldsymbol{x}_t = \widehat{\boldsymbol{A}}\boldsymbol{s}_t$$

pour tous les $t = 1, \ldots, T$.

En fait ce problème est mathématiquement similaire au problème de codage parcimonieux, qui consiste à calculer la représentation parcimonieuse (code) $\boldsymbol{c}_t$ d'un signal $\boldsymbol{x}_t \in \mathbb{R}^P$ à partir d'un dictionnaire redondant $\boldsymbol{D} \in \mathbb{R}^{P\times N}$ avec $N \gg P$:

$$\boldsymbol{x}_t = \boldsymbol{D}\,\boldsymbol{c}_t. \tag{31}$$

Malgré les différences dans l'interprétation physique, ces deux problèmes sont mathématiquement similaires et consistent à résoudre les équations précédentes avec une contrainte de parcimonie sur les sources $\boldsymbol{s}_t$:

$$\min \|\boldsymbol{s}_t\|_0 \;\; s.t. \;\; \|\boldsymbol{x}_t - \boldsymbol{A}\,\boldsymbol{s}_t\|_2^2 < \epsilon, \tag{32}$$

ou sur le code $\boldsymbol{c}_t$

$$\min \|\boldsymbol{c}_t\|_0 \;\; s.t. \;\; \|\boldsymbol{x}_t - \boldsymbol{D}\,\boldsymbol{c}_t\|_2^2 < \epsilon, \tag{33}$$

où ϵ est un seuil d'erreur acceptable.

Cette remarque permet d'utiliser les algorithmes de codage parcimonieux pour résoudre le problème de séparation de sources parcimonieuses dans le cas sous-déterminé. Mais, mieux, nous pouvons utiliser les résultats d'unicité formulés dans le cas de codage parcimonieux.

7.3.3 Unicité de la solution

La résolution du problème de codage parcimonieux (33), et les conditions d'unicité de la solution ont été abordées par de nombreux chercheurs depuis la fin des années 90, [78, 60, 59, 80]. Les conditions d'unicité sont formulées à partir de deux mesures sur la matrice de mélanges (ou son estimée) : spark($\boldsymbol{A}$) et la cohérence mutuelle $\mu(\boldsymbol{A})$, dont voici les définitions.

Définition 1 (spark($\boldsymbol{A}$)) [8] *Soit la matrice $\boldsymbol{A} \in \mathbb{R}^{P\times N}$, spark($\boldsymbol{A}$) est le plus petit nombre de colonnes de $\boldsymbol{A}$ qui sont linéairement dépendantes.*

Le calcul de spark($\boldsymbol{A}$) est NP-difficile, mais on peut facilement montrer les propriétés suivantes :

P1 spark($\boldsymbol{A}$) $= 1$ si et seulement si une colonne de $\boldsymbol{A}$ est nulle ;

P2 si toutes les sous-matrices de $\boldsymbol{A}$ de dimension $P \times P$ sont régulières, spark($\boldsymbol{A}$) $= P + 1$;

P3 dans tous les cas, spark($\boldsymbol{A}$) $\leq$ rank($\boldsymbol{A}$) $+ 1$.

Définition 2 (Cohérence mutuelle : $\mu(\boldsymbol{A})$) *Soit la matrice $\boldsymbol{A} \in \mathbb{R}^{P\times N}$, $\mu(\boldsymbol{A})$ est le maximum des valeurs absolues des coefficients de corrélation entre les colonnes de $\boldsymbol{A}$:*

$$\mu(\boldsymbol{A}) \triangleq \max_{i\neq j} \frac{|\boldsymbol{a}_i^\mathsf{T}\boldsymbol{a}_j|}{\|\boldsymbol{a}_i\|\,\|\boldsymbol{a}_j\|}. \tag{34}$$

où $\|\boldsymbol{u}\|$ est la norme euclidienne (ℓ_2) du vecteur $\boldsymbol{u}$.

La cohérence mutuelle caractérise le conditionnement de la matrice de mélange. Si deux colonnes, $\boldsymbol{a}_i$ et $\boldsymbol{a}_j$, sont colinéaires, la cohérence mutuelle $\mu(\boldsymbol{A}) = 1$: il sera impossible de séparer les sources i et j qui vivent dans le même espace. Si $\boldsymbol{a}_i$ et $\boldsymbol{a}_j$, sans être colinéaires, sont très proches, les sources i et j seront très difficiles à séparer.

Nous pouvons maintenant énoncer deux résultats d'unicité applicables en séparation de sources dans le cas de mélanges sous-déterminés. Considérons d'abord le problème avec la contrainte ℓ_0 :

$$\mathcal{P}_0 : \text{ minimiser } \|\boldsymbol{s}\|_0 \text{ sous la contrainte } \boldsymbol{A}\boldsymbol{s} = \boldsymbol{x}. \tag{35}$$

Théorème 4 (voir [60, 80]) *Si le système indéterminé d'équation linéaire $\boldsymbol{A}\boldsymbol{s} = \boldsymbol{x}$ possède une solution $\boldsymbol{s}^o$ telle que $\|\boldsymbol{s}^o\|_0 < \frac{1}{2}$spark($\boldsymbol{A}$), alors $\boldsymbol{s}^o$ est l'unique solution la plus parcimonieuse.*

8. spark($\boldsymbol{A}$) est lié au *rang de Kruskal* de $\boldsymbol{A}$, k-rank($\boldsymbol{A}$), par la relation [103] : spark($\boldsymbol{A}$) $=$ k-rank($\boldsymbol{A}$) $+ 1$. Ceci correspond à la propriété P2.

Ce théorème montre que s'il existe une solution $\boldsymbol{s}^o$ qui est suffisamment parcimonieuse, c'est l'unique solution.

Cependant, rechercher des solutions avec la contrainte de pseudo-norme ℓ_0 minimum relève d'un problème d'optimisation non convexe, qui est NP-difficile. De nombreux algorithmes proposent de relacher la contrainte ℓ_0 avec une contrainte ℓ_1, ce qui rend le problème d'optimisation convexe. Le problème à résoudre s'écrit alors :

$$\mathcal{P}_1 : \text{ minimiser } \|\boldsymbol{s}\|_1 \text{ sous la contrainte } \boldsymbol{A}\boldsymbol{s} = \boldsymbol{x}. \tag{36}$$

Mais, dans ce cas, à quelle condition la solution est-elle unique et identique pour les deux problèmes $\mathcal{P}_0$ et $\mathcal{P}_1$?

Donoho et al. [60], Gribonval et Nielsen [80] ont montré que la solution du problème avec contrainte ℓ_1 est équivalente à celle avec la contrainte ℓ_0 sous la condition suivante.

Théorème 5 *Si $\boldsymbol{A}\boldsymbol{s} = \boldsymbol{x}$ a une solution $\boldsymbol{s}^o$ qui satisfait $\|\boldsymbol{s}^o\|_0 < \frac{1+\mu(\boldsymbol{A})^{-1}}{2}$, cette solution est l'unique solution à la fois des problèmes (35) et (36).*

Remarquons cependant que la condition de parcimonie qui assure l'équivalence entre les problèmes $\mathcal{P}_0$ et $\mathcal{P}_1$, exige une plus grande parcimonie que la condition obtenue pour le problème (P_0). En effet, on a l'inégalité

$$1 + \mu(\boldsymbol{A})^{-1} \leq \text{spark}(\boldsymbol{A}),$$

et généralement $1 + \mu(\boldsymbol{A})^{-1} \ll \text{spark}(\boldsymbol{A})$.

7.3.4 Algorithmes pour l'estimation des sources

Dans le cas de mélanges sous-déterminés de sources parcimonieuses, après avoir estimé la matrice du mélange, on peut estimer les sources en utilisant plusieurs familles d'algorithmes :

- des algorithmes visant à résoudre le problème $\mathcal{P}_0$: ce problème est NP-difficile, mais il existe des algorithmes gloutons comme Matching Pursuit (MP) proposé par Mallat et Zhang en 93 [105] ou Orthogonal Matching Pursuit (OMP) proposé par Pati et al. [119] pour pallier l'accumulation d'erreurs de MP. On peut montrer [20] qu'OMP converge vers la bonne solution sous des conditions légères ;
- des algorithmes visant à résoudre le problème $\mathcal{P}_1$, qui lui est convexe : cette famille inclut les algorithmes de type LASSO [150] comme Basis Pursuit de Chen et Donoho [28]. Une synthèse de ces algorithmes est présentée dans [168].
- des algorithmes proposant une approximation différentiable de la contrainte ℓ_0, comme l'algorithme $SL0$ [110]. Pour un scalaire s, on peut utiliser l'approximation suivante (mais d'autres choix sont possibles) qui dépend d'un seul paramètre σ :

$$f_\sigma(s) = 1 - \exp\Big(-\frac{s^2}{2\sigma^2}\Big). \tag{37}$$

Pour un vecteur $\boldsymbol{s} = [s_1, \ldots, s_N]^\mathsf{T} \in \mathbb{R}^N$, l'approximation est alors :

$$\sum_{i=1}^{N} f_\sigma(s_i) = N - \sum_{i=1}^{N} \exp\Big(-\frac{s_i^2}{2\sigma^2}\Big) \approx \|\boldsymbol{s}\|_0, \tag{38}$$

et tend vers $\|\boldsymbol{s}\|_0$ si $\sigma \to 0$. Cette fonction, indéfiniment différentiable, dépend du seul paramètre σ qui contrôle la régularité et la précision de l'approximation. Par ailleurs, dans le cas bruité, le paramètre σ permet aussi de contrôler les valeurs du signal s_i qui seront considérées comme du bruit : pour $|s_i| < 3\sigma$, $f_\sigma(s_i) < 0,01$.

8 Mélanges non linéaires

Dans certaines applications, la relation entre les signaux mesurés sur les capteurs et les sources n'est pas linéaire. Citons trois exemples.

1. Les capteurs chimiques, utilisés pour mesurer les concentrations ioniques dans une solution, sont peu sélectifs et présentent une réponse non linéaire [61]. Selon le modèle de Nicolsky-Eisenman, la tension E fournie par le capteur (spécifique de l'ion i) s'écrit :

$$E = E_0 + d\log\Big(c_i + \sum_{j,j\neq i} K_{ij} a_j^{z_i/z_j}\Big), \tag{39}$$

où E_0 est une constante, a_i représente l'activité de l'ion principal i et les a_j les activités des ions interférents, K_{ij} est le coefficient de sélectivité du capteur, et z_i est la valence de l'ion i.

2. Considérons maintenant une feuille de papier mince avec des inscriptions sur le recto et le verso. En raison de la transparence du papier, l'image scannée d'une face de la feuille est un mélange non linéaire des images recto et verso, que l'on peut modéliser [108] :

$$\left\{\begin{array}{rcl} x_r^s(m,n) & = & f_r^i(m,n) + b_1 f_v^i(m,n)\exp(c_1 f_r^i(m,n)) \\ f_v^s(m,n) & = & f_v^i(m,n) + b_2 f_r^i(m,n)\exp(c_2 f_v^i(m,n)) \end{array}\right.$$

où les indices r et v correspondent à recto et verso, et les indices i et s à idéal et scannée. Ce mélange complexe peut être approximé par un mélange bilinéaire après développement de Taylor des termes exponentiels.

3. En imagerie satellitaire, les capteurs produisent des images hyperspectrales (plusieurs centaines de bandes de fréquences). La résolution spatiale est telle que chaque pixel représente une surface de quelques dizaines de mètres de côtés, parfois plus. Aussi, le spectre reçu sur le satellite est un mélange des spectres des différents éléments (végétation, eau, asphalte, constructions) présents dans la zone associée au pixel. Lorsque la surface est plane, un mélange linéaire est un bon modèle du spectre mesuré. Mais lorsque la surface présente un relief important, ou lorsqu'il y a de multiples réflexions (comme en milieu urbain), un modèle non linéaire est indispensable pour modéliser le mélange avec suffisamment de précision [56].

8.1 Les mélanges non linéaires sont-ils identifiables ?

Dans le cas général, le mélange non linéaire (NL) de sources $\boldsymbol{s}$ peut s'écrire :

$$\boldsymbol{x} = \mathcal{F}(\boldsymbol{s}). \tag{40}$$

Pour compenser le mélange, on cherche donc une transformation NL $\mathcal{G}$ telle que :

$$\hat{\boldsymbol{s}} = (\mathcal{G} \circ \mathcal{F})(\boldsymbol{s}) = \mathcal{H}(\boldsymbol{s}). \tag{41}$$

Pour que les sorties estimées, $\hat{\boldsymbol{s}}$, soient égales aux sources $\boldsymbol{s}$ à un facteur d'échelle et une permutation près, une condition nécessaire (non suffisante) est que la matrice Jacobienne de la transformation $\mathcal{H}$ soit diagonale, à une permutation près. Taleb [146] qualifie cette transformation de triviale, et la définit comme suit.

Définition. Une bijection $\mathcal{H}$ de $\mathbb{R}^N$ est dite triviale si et seulement si il existe des fonctions scalaires h_i, $1 \leq i \leq N$, et une permutation σ de l'ensemble $\{1, 2, \ldots, N\}$ telles que ses composantes $\mathcal{H}_i$ sont fonctions d'une seule source :

$$\mathcal{H}_i(s_1, \ldots, s_N) = h_i(s_{\sigma(i)}),\ i = 1, \ldots, N, \qquad (42)$$

On voit qu'une transformation triviale ne mélange pas les sources, mais applique une transformation non linéaire diagonale et une permutation arbitraires. La transformation non linéaire peut modifier considérablement le signal source : une telle indétermination n'est évidemment pas acceptable.

Par ailleurs, la préservation de l'indépendance statistique n'est pas suffisante pour assurer la séparation dans le cas non linéaire. En effet, à partir d'une simple méthode constructive, Darmois avait prouvé dans les années 50 que l'on peut trouver une infinité de mélanges non linéaires avec une matrice Jacobienne non diagonale qui préserve l'indépendance (donc correspondant à un mélange non trivial, qui ne fournit pas des sources séparées). Cette méthode constructive est présentée simplement dans [47, ch. 14 pp. 553-554].

Ainsi, dans le cas non linéaire général, l'indépendance statistique n'est pas suffisante pour séparer les sources. Par ailleurs, dans le cas où la matrice Jacobienne est diagonale, les sources estimées sont des fonctions NL inconnues des sources. En effet, si les sources $s_i(t)$ et $s_j(t)$ sont statistiquement indépendantes, $f_i(s_i(t))$ et $f_j(s_j(t))$, où $f_1(.)$ et $f_2(.)$ sont deux fonctions non linéaires quelconques, sont encore indépendantes. Ainsi, dans le cas général de mélanges NL, l'indépendance statistique ne permet pas d'estimer les sources ! Pour que les modèles non linéaires soient identifiables et que l'on puisse séparer les sources, il faut imposer d'autres contraintes.

Certains auteurs ont montré que la séparation n'était pas unique à moins de prendre en compte la structure temporelle des échantillons [89]. Almeida [5] postulait que la séparation était possible pour des fonctions non linéaires lisses, mais Babaie-Zadeh a montré que c'était inexact avec un contre-exemple [6].

Avec plusieurs doctorants (A. Taleb, M. Babaie-Zadeh, S. Hosseini et L. Duarte), Ch. Jutten s'est concentré sur des mélanges particuliers, notamment les mélanges post-nonlinéaires (PNL) et bilinéaires, pour lesquels des résultats d'identifiabilité [4] et des algorithmes ont été développés [47, ch.14]. Sous des conditions faibles (par exemple les fonctions non linéaires dans le mélange PNL doivent être inversibles), les sources sont identifiables avec les mêmes indéterminations d'échelle et de permutation que dans les mélanges linéaires. Une synthèse de ces méthodes se trouve dans l'article [53] et dans le livre [54].

Une approche récente originale [66] a été explorée par B. Ehsandoust durant sa thèse. Si la non-linéarité est dérivable en tous points, le mélange non linéaire peut être approché par un mélange localement linéaire. Bien sûr, l'approximation linéaire du mélange change en chaque point. Ainsi, on remplace un problème de séparation de sources dans un *mélange non linéaire invariant*, par un problème de séparation de sources dans *un mélange linéaire variant*. Cette approche permet de séparer les sources dans un cadre très général, mais les sources sont estimées à une fonction non linéaire près. En exploitant des informations a priori sur les sources, on peut concevoir des méthodes permettant de compenser cette distorsion [57, 62].

8.2 Mélanges post-nonlinéaires (PNL)

Ce type de mélange correspond à un mélange linéaire suivi, composante par composante, par une tranformation non linéaire (NL). Il peut être associé à un système dans lequel les capteurs sont non linéaires (par exemple, à cause de saturations, etc.). On a représenté schématiquement ce type de mélange et la structure de séparation à la figure 8.

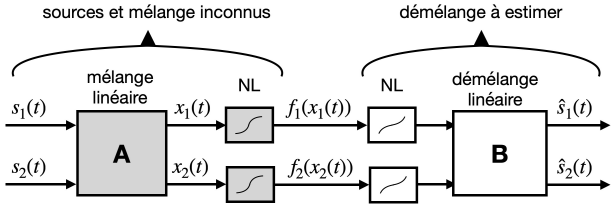


FIGURE 8 – Mélange post-nonlinéaire et la structure de séparation à estimer. Les sources et le mélange (en gris) sont inconnus. La structure de démélange comporte des non-linéarités pour compenser f_1 et f_2) suivies par une partie linéaire, notée $\boldsymbol{B}$.

Ce mélange a été étudié initialement par A. Taleb et Ch. Jutten [147, 148]. Avec ce mélange et la structure proposée, pourvu que la matrice de mélange $\boldsymbol{A}$ soit non diagonale, il existe des solutions avec les mêmes indéterminations (facteur d'échelle et permutation) que les mélanges linéaires. En effect, si la matrice $\boldsymbol{A}$ est diagonale (à une permutation près), chaque sortie du mélange PNL est une fonction non linéaire d'une seule source : les sorties sont donc déjà statistiquement indépendantes. On ne peut donc rien faire de mieux, à moins de connaitre les fonctions f_1 et f_2. Dans les premiers travaux, la partie NL de la structure de démélange est un perceptron multi-couches (MLP), ce qui permet d'approximer toute fonction continue et bornée, la partie linéaire est une simple matrice $\boldsymbol{B}$. Les paramètres du MLP et de la matrice de séparation $\boldsymbol{B}$ sont calculés de sorte que les sorties estimées, $\hat{s}_1(t)$ et $\hat{s}_2(t)$, soient statistiquement indépendantes. Dans [148], la mise en oeuvre algorithmique s'effectue en utilisant le critère d'information mutuelle qui avait été proposé par Comon quelques années auparavant [40].

Ces mélanges PNL ont donné lieu à de nombreux travaux, et ont été étendus à des mélanges convolutifs post-non linéaires [6, 7].

8.3 Mélanges bilinéaires et bilinéaires-quadratiques multivariés

Les mélanges bilinéaires-quadratiques sont caractérisés par l'équation de mélange suivante :

$$x_i(t) = \sum_{j=1}^{N} a_i j s_j(t) + \sum_{k=1}^{N} \sum_{l=k}^{N} b_{ikl} s_k(t) s_l(t). \qquad (43)$$

Dans les mélanges bilinéaires, les termes quadratiques (s_k^2, $k = 1, \ldots, N$) sont nuls, ce qui est obtenu en modifiant les bornes dans le second terme de droite : $\sum_{k=1}^{N-1} \sum_{l=k+1}^{N} b_{ikl} s_k(t) s_l(t)$.

Dans le cas de ces mélanges, on peut proposer une architecture de séparation qui est directement fondée sur cette équation, à savoir que les sources estimées, $\hat{s}_i(t)$, $i = 1, \ldots, N$; doivent s'écrire :

$$\hat{s}_i(t) = x_i(t) - \sum_{j=1, j\neq i}^{N} l_i j \hat{s}_j(t) - \sum_{k=1}^{N} \sum_{l=k}^{N} q_{ikl} \hat{s}_k(t) \hat{s}_l(t). \quad (44)$$

Dans le cas simple d'un mélange bilinéaire de deux sources et deux capteurs ($N = P = 2$), on a l'architecture présentée à la figure 9.

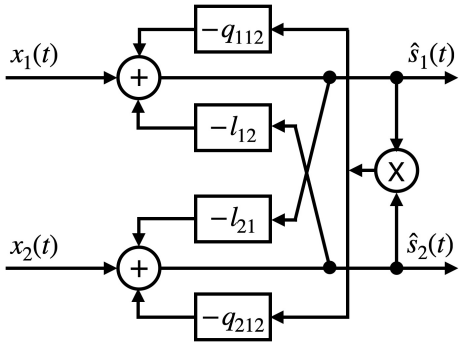


FIGURE 9 – Architecture de séparation pour un mélange bilinéaire de 2 sources par 2 capteurs. Les paramètres sont ajustés de façon à ce que les sources estimées deviennent statistiquement indépendantes.

Pour séparer un mélange linéaire quadratique, il faudrait ajouter dans l'architecture de la figure 9 les deux termes quadratiques, $\hat{s}_1^2(t)$ et $\hat{s}_2^2(t)$: il y aurait donc deux paramètres supplémentaires à estimer.

Pour calculer les paramètres de séparation, on utilise un critère d'indépendance statistiques des sources estimées. Par exemple, on peut annuler des statistiques d'ordres supérieurs [86], maximiser la vraisemblance [87] ou minimiser l'information mutuelle $I(\hat{s}(t)$ [111].

On pourrait utiliser un filtre polynomial (e.g. multilinéaire) de degré supérieur à deux. Le formalisme des filtres de Volterra le permettrait [30, 134, 121]. Mais à notre connaissance cela n'a pas été mis en œuvre pour la séparation de sources.

9 De la séparation de sources à la factorisation en matrices non négatives

9.1 Cas de sources et mélanges non négatifs

Soit un ensemble d'observations de la forme $\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{s}_t$, $t = 1, \ldots, T$. En posant $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]^\mathsf{T}$ et $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T]^\mathsf{T}$, la séparation de sources revient à résoudre le problème de factorisation matricielle :

$$\boldsymbol{X} = \boldsymbol{A}\,\boldsymbol{S}, \quad (45)$$

où $\boldsymbol{A}$ et $\boldsymbol{S}$ sont inconnues. Sous l'hypothèse de non-négativité des sources et des coefficients du mélange, c'est-à-dire lorsque tous les éléments de $\boldsymbol{X}$, $\boldsymbol{A}$ et $\boldsymbol{S}$) sont non négatifs, (45) revient à la factorisation de $\boldsymbol{X}$ en matrices non négatives (NMF pour Nonnegative Matrix Factorization) [47, ch.13].

On observe que le problème de factorisation, $\boldsymbol{X} = \boldsymbol{A}\,\boldsymbol{S}$, est mal posé. En effet, il existe une infinité de factorisations possibles,

$$\boldsymbol{X} = \boldsymbol{A}\,\boldsymbol{S} = \boldsymbol{A}\boldsymbol{Q}\,\boldsymbol{Q}^{-1}\boldsymbol{S}, \quad (46)$$

où $\boldsymbol{Q}$ est une matrice inversible de taille $N \times N$ (N est le nombre de sources). Avec la contrainte de non-négativité, on restreint les indéterminations, puisque cette contrainte impose aussi

$$\boldsymbol{A}\boldsymbol{Q} \geq \boldsymbol{0} \text{ et } \boldsymbol{Q}^{-1}\boldsymbol{s} \geq \boldsymbol{0}, \quad (47)$$

où la notation $\boldsymbol{M} \geq \boldsymbol{0}$ signifie que tous les éléments de la matrice $\boldsymbol{M}$ sont non négatifs.

Cependant, cette restriction n'est en général pas suffisante pour garantir l'unicité (aux indéterminations d'échelle et de permutation) de la solution en séparation de sources.

Le problème de factorisation en matrices non négatives (NMF, pour Nonnegative Matrix Factorization) a été étudié dans un cadre algébrique [149, 14, 27]. Il a ensuite été popularisé par des applications en traitement du signal et sciences des données, notamment pour la chimiométrie [118], la reconnaissance de visages [99] et la fouille de données [71]. Il a été considéré en séparation de sources [122, 113], en particulier dans le cas de sources audio [122, 70] et d'images hyperspectrales [115]. En effet, lorsque les sources audio (parole ou musique) sont représentées par les spectres de puissance, les sources (spectres) sont non négatives ainsi que les coefficients du mélange. Dans le cas d'images hyperspectrales, les spectres (sources) sont non négatifs, et les coefficients de mélanges sont associés aux proportions (abondances) des éléments (eau, végétation, asphalte, etc.) sur chaque pixel de l'image, qui sont aussi des quantités non négatives.

Pour exploiter et préserver le caractère non négatif des sources et des mélanges, il faut bien sûr éviter de centrer les observations, comme on le fait généralement dans les méthodes statistiques de séparation de sources.

9.2 Géométrie des sources et mélanges non négatifs

Si l'on suppose que les observations sont des mélanges non négatifs de sources non négatives, le modèle $\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{s}_t$, $t = 1, \ldots, T$ est tel que les éléments de la matrice de mélange (inconnue) $\boldsymbol{A}$ et des sources (inconnues) $\boldsymbol{s}_t$ sont non négatifs. On observe alors des propriétés géométriques intéressantes.

L'équation $\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{s}_t$ peut aussi s'écrire :

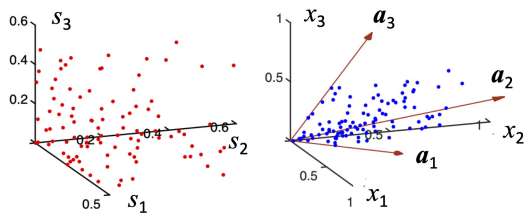
$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{s}_t = \sum_{n=1}^{N} \boldsymbol{a}_n s_{nt}, \quad (48)$$

où $\boldsymbol{a}_n$ est la n-ième colonne de $\boldsymbol{A}$ et $s_{nt} \geq 0$ est l'échantillon de la source n à l'indice t. Cette équation montre que les observations $\boldsymbol{x}_t \in \mathbb{R}^P$, $\forall t = 1, \ldots, T$, se trouvent à l'intérieur du cône engendré par les vecteurs colonnes de la matrice de mélange $\boldsymbol{A}$.

En tenant compte de la non-négativité des quantités $\boldsymbol{A}$ et $\boldsymbol{s}_t$, on déduit :

- $s_{nt} \geq 0, \ \forall n, t$, donc les vecteurs $\boldsymbol{s}_t$ sont associés à un point dans l'orthant non négatif de dimension N (N est le nombre de sources) : $\mathbb{R}_+^N \cup \{0\}$.
- les composantes des colonnes $\boldsymbol{a}_n$ de la matrice $\boldsymbol{A}$ ayant des valeurs non négatives, les vecteurs $\boldsymbol{a}_n$ sont situés dans l'orthant non négatif de dimension P : $\mathbb{R}_+^P \cup \{0\}$,
- les observations $\boldsymbol{x}_t$ sont donc situées à l'intérieur du cône cone($\boldsymbol{A}$) dont les arêtes sont les $\boldsymbol{a}_n$ et qui est inclus dans l'orthant non négatif de dimension P : $\mathbb{R}_+^P \cup \{0\}$.

Ceci est illustré à la figure 10, dans le cas simple de 3 mélanges de 3 sources ($N = P = 3$).



(a) (b)

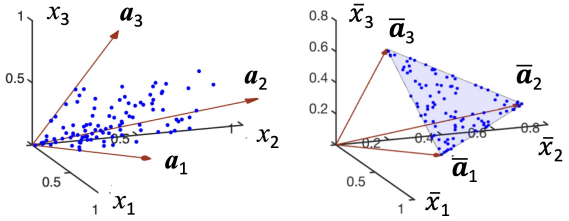
FIGURE 10 – Représentation de sources $\boldsymbol{s}_t$ non négatives et de leurs mélanges $\boldsymbol{A}$ non négatifs, pour $N = P = 3$. Les points associés aux vecteurs sources $\boldsymbol{s}_t$ (points rouges à gauche) sont situés dans l'orthant non négatif. Ceux associés aux observations $\boldsymbol{x}_t$ (points bleus à droite) sont engendrés par les 3 vecteurs colonnes $\boldsymbol{a}_n$, et situés à l'intérieur du cône dont les arêtes (en rouge) sont les $\boldsymbol{a}_n$, cône qui est inclus dans l'orthant non négatif.

On peut également représenter les sources, $\boldsymbol{x}_t$, et les observations, $\boldsymbol{x}_t$, en les normalisant :

$$\boldsymbol{s}_t^{nor} = \frac{\boldsymbol{s}_t}{\sum_{n=1}^N s_{nt}} \quad \text{et} \quad \boldsymbol{x}_t^{nor} = \frac{\boldsymbol{x}_t}{\sum_{p=1}^P x_{pt}}. \tag{49}$$

Cette normalisation correspond à une norme ℓ_1 car la non-négativité des signaux implique $\sum_{n=1}^N s_{nt} = \|\boldsymbol{s}_t\|_1$ et $\sum_{p=1}^P x_{pt} = \|\boldsymbol{x}_t\|_1$. Elle est aussi associée à une hypothèse de "somme à un" sur les sources, aussi appelée "column stochastic" : ceci s'applique lorsque, à chaque indice t, les valeurs s_{nt} sont par exemple des probabilités (data mining) ou bien des abondances (en imagerie hyperspectrale) dont la somme vaux 1. On peut trouver les détails du calcul de normalisation de l'équation $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$ dans [71, encadré page 66].

Les observations normalisées appartiennent donc au sous-espace affine de dimension $P-1$ et d'équation $\sum_{p=1}^P x_p = 1$. La projection du cone associé à $\boldsymbol{A}$ est alors un simplexe (si $\boldsymbol{A}$ est de rang plein) dont les sommets correspondent aux vecteurs colonnes de $\boldsymbol{A}$ (la Fig. 11 illustre cette situation dans le cas $N = P = 3$). Ce simplexe (c'est-à-dire l'enveloppe convexe) contenant toutes les observations $\boldsymbol{x}_t$ est noté convh($\boldsymbol{X}$) (pour convex hull en anglais) et vérifie : convh($\boldsymbol{X}$) $\subseteq$ convh($\boldsymbol{A}$).



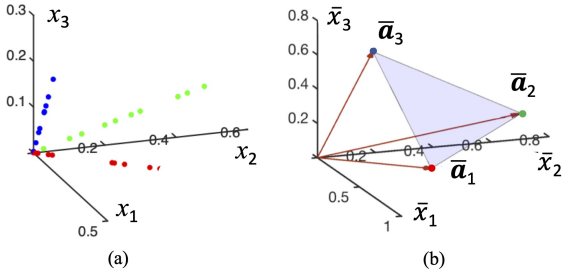
(a) (b)

FIGURE 11 – Dans le cas $N = P = 3$, les observations $\boldsymbol{x}_t$ (points bleus à droite) sont à l'intérieur du cone généré par les trois colonnes $\boldsymbol{a}_n$ de la matrice de mélange. Après normalisation ℓ_1, ces points sont projetés dans le simplexe (à droite) dont les 3 sommets correspondent aux 3 colonnnes normalisées de $\boldsymbol{A}$.

9.3 Condition d'unicité de la factorisation

Dans ces représentations géométriques, certains points particuliers suggèrent des conditions d'unicité de la factorisation et deux familles d'algorithmes.

Points sur les rayons extrêmes du cone. Si une observation $\boldsymbol{x}_t$ est telle que : $\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{s}_t = \boldsymbol{a}_n s_{nt}$, le point appartient à l'arête $\boldsymbol{a}_n$ du cone. Dans le simplexe associé, le point normalisé se projette sur le sommet correspondant à $\boldsymbol{a}_n / \|\boldsymbol{a}_n\|_1$. Ceci est illustré pour $N = P = 3$ à la figure 12. S'il existe des observations $\boldsymbol{x}_t$ situées sur chacune des arêtes du cone, on peut identifier les colonnes de la matrice de mélange : la factorisation est donc unique à un facteur d'échelle et une permutation près.



(a) (b)

FIGURE 12 – Les observations $\boldsymbol{x}_t$ sont de la forme $\boldsymbol{x}_t = \alpha \boldsymbol{a}_n$ ($\alpha > 0$). Elles sont localisées sur les arêtes du cone. Après normalisation ℓ_1, ces points sont projetés sur les sommets du simplexe correspondant aux colonnnes normalisées de $\boldsymbol{A}$.

Points situés sur les facettes du cone. Ces points correspondent à des observations $\boldsymbol{x}_t$ associées à des sources $\boldsymbol{s}_t$ avec au moins une valeur nulle. De tels points sont situés sur les facette du cone, ou après normalisation, sur les facettes du simplexe. Dans le cas où $N = P = 3$, les points sur les facettes correspondent à : $\boldsymbol{x}_t = \boldsymbol{a}_i s_{it} + \boldsymbol{a}_j s_{jt}$, avec $i, j \in \{1, 2, 3\}$. Ceci est illustré à la figure 13.

Dans les deux situations que nous venons de voir, les observations $\boldsymbol{x}_t$ permettent d'identifier $\boldsymbol{A}$ car cone($\boldsymbol{X}$) = cone($\boldsymbol{A}$), ou après normalisation convh($\boldsymbol{X}$) = convh($\boldsymbol{A}$).

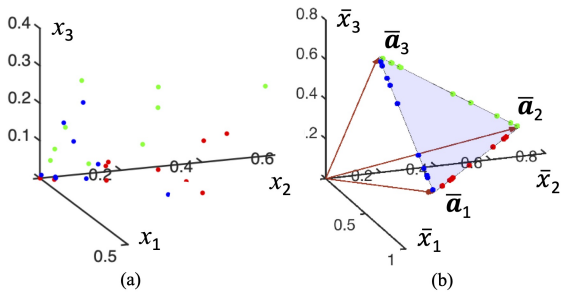


FIGURE 13 – Les observations $\boldsymbol{x}_t$ sont de la forme $\boldsymbol{x}_t = \boldsymbol{a}_i s_{it} + \boldsymbol{a}_j s_{jt}$. Elles sont localisées sur les facettes du cone. Après normalisation ℓ_1, ces points sont projetés sur les facettes du simplexe.

Dispersion des points. Si pour un index t, les valeurs des sources sont assez similaires, le point $\boldsymbol{x}_t$ se trouvera au milieu du cone ou du simplex. Si tous les $\boldsymbol{x}_t$ sont ainsi, l'ensemble des observations est concentré au mileu du cone ou du simplex, et il n'y a pas de point proche des arêtes ou des facettes : l'estimation des colonnes de $\boldsymbol{A}$ n'est alors pas possible. Dans de telles situations, la factorisation ne sera pas unique et la séparation de sources est impossible. Cette situation est illustrée à la figure 14 dans le cas $N = P = 3$.

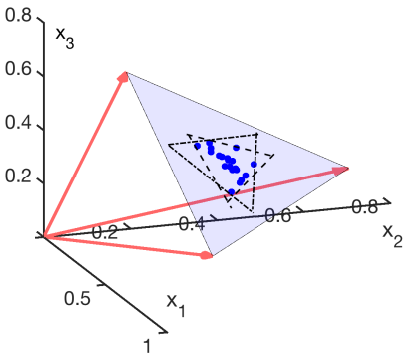


FIGURE 14 – Si les sources sont non nulles avec des valeurs similaires, les observations $\boldsymbol{x}_t$ sont localisées à l'intérieur du cone ou du simplex, loin des arêtes et des sommets. Dans de tels cas de faible dispersion, on ne peut pas identifier de façon unique les colonnes de $\boldsymbol{A}$: la factorisation et la séparation sont impossibles.

Ces illustrations suggèrent que la factorisation (et donc la séparation) sera possible si les sources sont suffisamment dispersées à chaque indice t. Avec des sources parcimonieuses, la situation est encore plus favorable, car pour chaque indice t, l'existence d'une unique source active, ou d'une partie seulement des sources actives, est probable, ce qui entraine une large dispersion des points $\boldsymbol{x}_t$ avec certains points localisés près des arêtes et des facettes du cone ou du simplexe. Ceci a conduit à l'énoncé de conditions d'unicité (aux indéterminations d'échelle et de permutation) exploitant l'existence d'observations sur les arêtes du cone$(\boldsymbol{A})$ (ou les sommets du simplexe) ou la dispersion des $\boldsymbol{x}_t$ à l'intérieur du cone ou du simplexe. Le lecteur curieux pourra trouver ces résultats dans l'ouvrage de N. Gillis [77] et dans ces tutoriels [17, 76, 71].

9.4 Principes de quelques algorithmes

Les approches classiques de NMF consistent à résoudre :

$$\min_{\boldsymbol{A},\boldsymbol{S}} \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{S}\|_F^2 \text{ avec } \boldsymbol{A} \geq \boldsymbol{0} \text{ et } \boldsymbol{S} \geq \boldsymbol{0}. \tag{50}$$

Ce problème de minimisation (conjointe sur $\boldsymbol{A}$ et $\boldsymbol{S}$) est non convexe et on procède souvent par une minimisation alternée. A l'itération k, on fait :

$$\begin{array}{lll} (1)\ \min_{\boldsymbol{A}} \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{S}^{k-1}\|_F^2 & \text{avec} & \boldsymbol{A} \geq \boldsymbol{0}, \\ (2)\ \min_{\boldsymbol{S}} \|\boldsymbol{X} - \boldsymbol{A}^k\boldsymbol{S}\|_F^2 & \text{avec} & \boldsymbol{S} \geq \boldsymbol{0}. \end{array} \tag{51}$$

Cependant, ce type d'algorithme additif ne garantit pas la non-négativité qui doit être vérifiée à chaque itération. Pour l'éviter, Lee et Seung en 1999 [99] ont proposé une mise à jour multiplicative qui préserve intrinsèquement la non-négativité :

$$\begin{array}{lll} (1)\ a_{ij} & \leftarrow & a_{ij} \frac{(\boldsymbol{X}\boldsymbol{S}^\top)_{ij}}{(\boldsymbol{A}\boldsymbol{S}\boldsymbol{S}^\top)_{ij}} \\ (2)\ s_{jk} & \leftarrow & s_{jk} \frac{(\boldsymbol{A}^\top\boldsymbol{X})_{jk}}{(\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{S})_{jk}}. \end{array} \tag{52}$$

Cet algorithme est simple, mais converge lentement. De plus, si un paramètre à estimer devient égal à 0, il ne variera plus : pour pallier ce problème, on peut remplacer cette valeur nulle par un petit nombre positif aléatoire.

Au-delà de ces algorithmes classiques, il existe deux familles d'algorithmes exploitant les particularités géométriques présentées ci-dessus. Nous insistons sur le fait qu'outre la non-négativité des sources, ces deux familles exploitent également leur parcimonie.

— La première famille d'algorithmes suppose l'existence de mesures $\boldsymbol{x}_t$ qui sont dues à une source dominante, c'est-à-dire que $\boldsymbol{s}_t$ ne comporte qu'une composante non nulle : ces points sont donc portés sur les arêtes de cone$(\boldsymbol{A})$ ou les sommets du simplexe convh$(\boldsymbol{A})$. Si un tel point existe pour chacune des colonnes de $\boldsymbol{A}$, l'estimation de $\boldsymbol{A}$ devient très simple. Pour y parvenir, il suffit donc d'identifier les observations $\boldsymbol{x}_t$ correspondant à ces arêtes. De nombreux algorithmes ont été proposés pour cela, en particulier pour le démélange spectral. Le principe des algorithmes recherchant le Pixel Purity Index remonte à Boardman en 1994 [18], mais de nombreuses variantes ont été proposées, comme Successive Projection Algorithm ou Vertex Component Analysis[115].

— La seconde famille consiste à rechercher le N-simplexe de volume minimal contenant toutes les observations $\boldsymbol{x}_t$ normalisées. Si les observations sont suffisamment dispersées, on peut prouver [104, 71] qu'il existe une solution unique (aux indéterminations d'échelle et de permutation près) vers laquelle ces algorithmes convergent. De nombreux algorithmes ont été conçus, en particulier pour le démélange spectral [100] : MVSA, SISAL, MVES, MVC-NMF, ICE, pour n'en citer que quelques-uns.

Les codes de certains de ces algorithmes sont accessibles sur les sites de collègues scientifiques, comme ceux de N. Gillis [75] ou celles de notre regretté collègue J. Bioucas [16].

10 Vers plus de diversité

La séparation aveugle de sources est connue principalement sous hypothèse de sources indépendantes. Mais nous avons vu que d'autres hypothèses, telles que la non-stationnarité ou la couleur des sources, permettent de se passer de leur indépendance statistique au sens fort (*i.e.* aux ordres supérieurs à 2). Il existe encore bien d'autres approches fonctionant sous d'autres hypothèses. C'est le cas des approches tensorielles déterministes, dont nous allons parler brièvement maintenant.

Lorsque les signaux enregistrés dépendent de plus de 2 variables physiques, disons 3 pour fixer les idées, alors les données peuvent être rangées dans un tableau de la forme $T(i,j,k) = f(x_i, y_j, z_k)$, puisque les variables sont nécessairement discrétisées dans des mesures numérisées.

Dans certaines situations, les données mesurées sont la superposition de fonctions à variables séparées, et le modèle d'observation suivant est réaliste [46] :

$$T(i,j,k) = \sum_{r=1}^{R} a_r(x_i) b_r(y_j) c_r(z_k) + bruit \qquad (53)$$

Ce modèle est connu sous le nom de décomposition canonique tensorielle [84]. Le principal intérêt de recourir à ce modèle et d'identifier les matrices représentant les fonctions $a(\cdot)$, $b(\cdot)$ et $c(\cdot)$ est que, sous des hypothèses assez faibles (notamment R pas trop grand), ces matrices peuvent être identifiées de manière unique à une permutation et un facteur d'échelle près [145].

On peut suivre cette idée dans de nombreux domaines, notamment en chimiométrie [144] ou en traitement d'antenne [141] [102] [126] ; par exemple on peut exploiter la diversité de réponse spatiale de capteurs [127], la diversité de vitesses de propagation [125], ou celle de polarisation des ondes [124]. Notons qu'il existe de nombreuses autres manières d'exprimer sous forme tensorielle la structure des données, par exemple Hankel, Löwner [153] [51], Prony multivariée [81] [138], nonstationnarité en temps ou inhomogéneité en espace [8]...

La prise en compte des informations sur la nature des données confère généralement à ce type d'approche de meilleures performances. La contrepartie est souvent un coût calcul plus élevé, étant donné que décomposer un tenseur en somme de termes de rang 1, ou approximer un tenseur par un autre de rang plus faible, sont des problèmes difficiles [143] [50].

Il est important de signaler que la contrainte de non-négativité permet à l'approximation tensorielle de rang faible d'exister et d'être unique [123], sans empêcher un calcul relativement rapide pour autant [31].

En revanche, le problème de l'approximation tensorielle de rang faible est généralement mal posé et des contraintes appropriées sont nécessaires ; la non-négativité en est une parmi d'autres. En traitement d'antenne par exemple, on préfèrera une contrainte sur la cohérence mutuelle (ou le *spark*) [103] [137].

Dans certaines applications, la parcimonie est attendue dans une ou plusieurs des matrices-facteurs de la décomposition (53) [9]. On peut l'imposer avec les normes ℓ_1 [159] ou $\ell_{1,2}$ [10], éventuellement conjointement avec la non-négativité. On impose cette contrainte matricielle en calculant la décomposition canonique tensorielle, par exemple avec l'algorithme ADMM, sans oublier d'ajouter une pénalité assurant l'existence de l'approximation de rang faible.

11 Discussion

La séparation de sources est un problème qui trouve des applications dans de nombreux domaines [47], dont certains impliquent le recours explicite aux décompositions tensorielles : biomédical [51], imagerie hyperspectrale et télédétection [154], télécommunications [142] [133], sonar/radar [126] [137], chimie [144]. Ce problème a aussi été considérablement étudié pour la séparation de signaux audio, en particulier parole et musique [116, 55, 13, 155, 74, 65], avec l'organisation des compétitions internationales SISEC (voir par exemple {https ://sisec.inria.fr}) conduites notamment par Emmanuel Vincent et plus tard par Antoine Liutkus, et soutenues initialement par un projet du GdR ISIS [157].

Les solutions exploitent toujours plusieurs formes de diversité : diversité spatiale en jouant sur le nombre de capteurs, diversité temporelle en jouant sur la coloration ou la non-stationarité des sources, ou d'autres a priori comme la non négativité, la parcimonie, ou le caractère discret.

Les travaux dans ce domaine ont été primés avec de nombreux Best Paper Awards, plusieurs grand prix, deux médailles d'argent du CNRS, et deux projets ERC. N'oublions pas la une du Washington Post en 2013 avec une image du rayonnement cosmique de l'univers obtenue par les travaux de J.-F. Cardoso à partir des données de la mission Planck.

Références

[1] K. Abed-Meraim, P. Loubaton, and E. Moulines. A subspace algorithm for certain blind identification problems. *IEEE Trans. Inf. Theory*, pages 499–511, March 1997.

[2] K. Abed-Meraim, E. Moulines, and P. Loubaton. Prediction error method for second-order blind identification. *IEEE Trans. Sig. Proc.*, 45(3) :694–705, March 1997.

[3] F. Abrard and Y. Deville. A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources. *Signal Processing*, 85(7) :1389 – 1403, 2005.

[4] S. Achard and C. Jutten. Identifiability of post-nonlinear mixtures. *IEEE Signal Processing Letters*, 12(5) :423—-426, 2005.

[5] L. Almeida. Linear and nonlinear ICA based on mutual information. In *Proceedings IEEE 2000 Adaptive systems for Signal Processing*, pages 117–122, Lake Louise, Canada, 2000.

[6] M. Babaie-Zadeh. *On Blind Source Separation in Convolutive and Non-Linear mixtures*. PhD thesis, Grenoble INP et Sharif Univ. of Technology, 1981.

[7] M. Babaie-Zadeh, C. Jutten, and K. Nayebi. A minimization-projection (mp) approach for blind separating convolutive mixtures. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages V–533, 2004.

[8] H. Becker, L. Albera, P. Comon, R. Gribonval, F. Wendling, and I. Merlet. Brain source imaging : from sparse to tensor models. *IEEE Sig. Proc. Magazine*, 32(6) :100–112, November 2015. hal-01190559.

[9] H. Becker, L. Albera, P. Comon, M. Haardt, G. Birot, et al. EEG extended source localization : Tensor-based vs conventional methods. *NeuroImage*, 96 :143–157, August 2014. hal-01011856.

[10] H. Becker, L. Albera, P. Comon, J.-C. Nunes, R. Gribonval, et al. SISSY : an efficient and automatic algorithm for the analysis of EEG sources based on structured sparsity. *Neuroimage*, 157 :157–172, August 2017. hal inserm-01617155.

[11] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6) :1129–1159, November 1995.

[12] A. Belouchrani, K. Abed-Meraim, J. F. Cardoso, and E. Moulines. A blind source separation technique using second order statistics. *IEEE Trans. SP*, 45(2) :434–444, February 1997.

[13] L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1) :191–199, 2006.

[14] A. Berman and R. J. Plemmons. *Non-Negative Matrices in Mathematical Sciences*. Academic Press (New-York), 1979.

[15] G. Bienvenu and L. Kopp. Optimality of high-resolution array processing using the eigensystem approach. *IEEE Trans. ASSP*, 31(5) :1235–1248, October 1983.

[16] J. M. Bioucas-Dias. Codes matlab. http://www.lx.it.pt/~bioucas/code.htm, 2009.

[17] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot. Hyperspectral unmixing overview : Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2) :354–379, 2012.

[18] J. W. Boardman. Geometric mixture analysis of imaging spectrometery data. In *Proc. Intl. Geoscience and Remote Sensing Symp.*, volume 4, pages 2369–2371, 1994.

[19] P. Bofill. Underdetermined blind separation of delayed sound sources in the frequency domain. *Neurocomputing*, 55(3) :627–641, 2003. Evolving Solution with Neural Networks.

[20] T. Tony Cai and Lie Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory*, 57(7) :4680–4688, 2011.

[21] J. F. Cardoso. Infomax and Maximum Likelihood for source separation. *IEEE Sig. Proc. Letters*, 4(4) :112–114, April 1997.

[22] J. F. Cardoso. Blind signal separation : statistical principles. *Proc. of the IEEE*, 90 :2009–2025, October 1998.

[23] J. F. Cardoso and P. Comon. Tensor-based independent component analysis. In *Proc. EUSIPCO*, pages 673–676, Barcelona, Spain, September 18-21 1990.

[24] J. F. Cardoso and B. Laheld. Equivariant adaptive source separation. *IEEE Trans. SP*, 44(12) :3017–3030, December 1996.

[25] J. F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. *IEE Proceedings - Part F*, 140(6) :362–370, December 1993.

[26] J. F. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM Jour. matrix Analysis*, 17(1) :161–164, January 1996.

[27] J.-C. Chen. The nonnegative rank factorizations of nonnegative matrices. *Linear Algebra and its Applications*, 62 :207–217, 1984.

[28] Shaobing Chen and D. Donoho. Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44 vol.1, 1994.

[29] P. Chevalier, L. Albera, A. Ferreol, and P. Comon. On the virtual array concept for higher order array processing. *IEEE Trans. Sig. Proc.*, 53(4) :1254–1271, April 2005.

[30] P. Chevalier, B. Picinbono, and P. Duvaut. Le filtrage de volterra transverse réel et complexe. *Traitement du Signal*, 7(5) :451–476, 1990.

[31] J. E. Cohen, R. Cabral-Farias, and P. Comon. Fast decomposition of large nonnegative tensors. *IEEE Sig. Proc. Letters*, 22(7) :862–866, July 2015. hal-01069069.

[32] P. Comon. Separation de melanges de signaux. In *XIIème Colloque Gretsi*, pages 137–140, Juan les Pins, 12 -16 juin 1989.

[33] P. Comon. Separation of sources using high-order cumulants. In *SPIE Conf. Adv. Alg. Archi. Sig. Proc.*, pages 170–181, San Diego, Ca, August 8-10 1989. vol. Real-time signal processing XII.

[34] P. Comon. Separation of stochastic processes. In *Proc. Workshop on Higher-Order Spectral Analysis*, pages 174–179, Vail, Colorado, June 28-30 1989. IEEE-ONR-NSF.

[35] P. Comon. Codes matlab. https://www.gipsa-lab.grenoble-inp.fr/user/pierre.comon/miscellaneous, 1989-2014.

[36] P. Comon. Analyse en Composantes Indépendantes et identification aveugle. *Traitement du Signal*, 7(3) :435–450, December 1990. Numero special non lineaire et non gaussien.

[37] P. Comon. Independent Component Analysis. In *Proc. 2nd Int. Sig. Proc. Workshop on Higher-Order Statistics*, pages 111–120, Chamrousse, France, July 10-12 1991. Keynote address. Republished in *Higher-Order Statistics*, J.L.Lacoume ed., Elsevier, 1992, pp 29–38.

[38] P. Comon. MA identification using fourth order cumulants. *Signal Processing*, 26(3) :381–388, 1992. hal-00347140.

[39] P. Comon. Remarques sur la diagonalisation tensorielle par la methode de Jacobi. In *XIVème Colloque Gretsi*, pages 125–128, 13-16 Septembre 1993.

[40] P. Comon. Independent Component Analysis, a new concept ? *Signal Processing, Elsevier*, 36(3) :287–314, April 1994.

[41] P. Comon. Tensor diagonalization, a useful tool in signal processing. In M. Blanke and T. Soderstrom, editors, *10th IFAC Symp. System Ident.*, pages 77–82, Copenhagen, Denmark, July 4-6 1994. hal-00561523.

[42] P. Comon. Blind channel identification and extraction of more sources than sensors. In *SPIE Conf. Adv. Sig. Proc. VIII*, volume 3461, pages 2–13, San Diego, July 19-24 1998. keynote. hal-00499421.

[43] P. Comon. From source separation to blind equalization, contrast-based approaches. In *1st Int. Conf. on Image and Signal Processing (ICISP'01)*, Agadir, Morocco, May 3-5, 2001. keynote. hal-01825729.

[44] P. Comon. Blind identification and source separation in 2x3 under-determined mixtures. *IEEE Trans. Signal Processing*, 52(1) :11–22, January 2004.

[45] P. Comon. Contrasts, independent component analysis, and blind deconvolution. *Int. J. Adapt. Control Sig. Proc.*, 18(3) :225–243, April 2004.

[46] P. Comon. Tensors : a brief introduction. *IEEE Sig. Proc. Magazine*, 31(3) :44–53, May 2014. hal-00923279.

[47] P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press, Oxford UK, Burlington USA, 2010. ISBN : 978-0-12-374726-6.

[48] P. Comon, C. Jutten, and J. Herault. Separation of sources, part II : Problems statement. *Signal Processing, Elsevier*, 24(1) :11–20, July 1991.

[49] G. Darmois. Analyse générale des liaisons stochastiques. *Rev. Inst. Internat. Stoch.*, 21 :2–8, 1953.

[50] J. H. de Morais Goulart, R. Couillet, and P. Comon. A random matrix perspective on random tensors. *Journal of Machine Learning Research*, 23(264) :1–36, 2022. open access.

[51] P. M. R. de Oliveira, J. H. de M. Goulart, C. A. R. Fernandes, and V. Zarzoso. Blind source separation in persistent atrial fibrillation electrocardiograms using block-term tensor decomposition with löwner constraints. *IEEE J. Biomed. Health Informatics*, 26(4) :1538–1548, April 2022.

[52] N. Delfosse and P. Loubaton. Adaptive blind separation of independent sources : a deflation approach. *Signal Processing*, 45 :59–83, 1995.

[53] Y. Deville and L. T. Duarte. An overview of blind source separation methods for linear-quadratic and post-nonlinear mixtures. In E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, editors, *Proceedings Int. Conf. Latent Variable Analysis and Signal Separation*, pages 155–167. Springer International Publishing, 2015.

[54] Y. Deville, L. T. Duarte, and S. Hosseini. *Nonlinear Blind Source Separation and Blind Mixture Identification - Methods for Bilinear, Linear-quadratic and Polynomial Mixtures*. Springer Cham, 2021.

[55] Y. Deville, M. Puigt, and B. Albouy. Time-frequency blind signal separation : extended methods, performance evaluation for speech sources. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, volume 1, pages 255–260, 2004.

[56] N. Dobigeon, J.-Y. Tourneret, C. Richard, J. C. M. Bermudez, S. McLaughlin, and A. O. Hero. Nonlinear unmixing of hyperspectral images : Models and algorithms. *IEEE Signal Processing Magazine*, 31(1) :82–94, 2014.

[57] K. Dogancay. Blind compensation of nonlinear distortion for bandlimited signals. *IEEE Transactions on Circuits and Systems I : Regular Papers*, 52(9) :1872–1882, 2005.

[58] D. Donoho. On minimum entropy deconvolution. In *Applied time-series analysis II*, pages 565–609. Academic Press, 1981.

[59] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Information Theory*, 52(1) :6–18, 2006.

[60] D.L. Donoho and M. Elad. Optimally sparse representation in general (non-orthogonal) dictionaries via L1 minimization. *Proc. Nat. Aca. Sci*, 100(5) :2197–2202, 2003.

[61] L. T. Duarte, C. Jutten, and S. Moussaoui. A bayesian nonlinear source separation method for smart ion-selective electrode arrays. *IEEE Sensors Journal*, 9(12) :1763–1771, October 2009.

[62] L. T. Duarte, R. Suyama, R. Attux, J. M. T. Romano, and C. Jutten. A sparsity-based method for blind compensation of a memoryless nonlinear distortion : Application to ion-selective electrodes. *IEEE Sensors Journal*, 15(4) :2054–2061, 2015.

[63] D. Dugué. Analycité et convexité des fonctions caractéristiques. *Annales de l'Institut Henri poincarré*, XII :45–56, 1951.

[64] Ngoc Q. K. Duong, Emmanuel Vincent, and Rémi Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7) :1830–1840, 2010.

[65] Jean-Louis Durrieu, Bertrand David, and Gaël Richard. A musically motivated mid-level representation for pitch estimation and musical audio source separation. *IEEE Journal of Selected Topics in Signal Processing*, 5(6) :1180–1191, 2011.

[66] B. Ehsandoust, M. Babaie-Zadeh, B. Rivet, and C. Jutten. Blind source separation in nonlinear mixtures : Separability and a basic algorithm. *IEEE Transactions on Signal Processing*, 65(16) :4339–4352, 2017.

[67] B. Emile and P. Comon. Estimation of time delays between unknown colored signals. *Signal Processing*, 69(1) :93–100, August 1998.

[68] B. Emile, P. Comon, and J. Leroux. Estimation of time delays with fewer sensors than sources. *IEEE Trans. Sig. Proc.*, 46(7) :2012–2015, July 1998.

[69] W. Feller. *An Introduction to Probability Theory and its Applications*, volume II. Wiley, 1971.

[70] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence : With application to music analysis. *Neurocomputing*, 21(3) :783–830, 2009.

[71] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma. Nonnegative matrix factorization for signal and data analytics : Identifiability, algorithms, and applications. *IEEE Signal Processing Magazine*, 36(2) :59–80, 2019.

[72] K. Fukushima. Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4) :193–202, April 1980.

[73] M. Gaeta and J. L. Lacoume. Source separation without a priori knowledge : the maximum likelihood solution. In *Proc. EUSIPCO*, pages 621–624, Barcelona, Spain, September 18-21 1990.

[74] Olivier Gillet and Gaël Richard. Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(3) :529–540, 2008.

[75] N. Gillis. Code. https://sites.google.com/site/nicolasgillis/code.

[76] N. Gillis. The why and how of nonnegative matrix factorization. In J. A. K. Suykens, M. Signoretto, and A. Argyriou, editors, *Regularization, Optimization, Kernels, and Support Vector Machines*, pages 257–291. International Society for Optics and Photonics, Chapman & Hall/CRC, 2014.

[77] N. Gillis. *Nonnegative Matrix Factorization*. SIAM, 2020.

[78] I. F. Gorodnitsky and B. D. Rao. Sparse signal reconstruction from limited data using FOCUSS : a re-weighted minimum norm algorithm. *IEEE Trans. Signal Processing*, 45(3) :600–616, 1997.

[79] A. Gorokhov and P. Loubaton. Blind identification of MIMO-FIR systems : a generalized linear prediction approach. *Signal Processing, Elsevier*, 73 :105–124, February 1999.

[80] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, pages 3320–3325, 2003.

[81] J. Harmouch, H. Khalil, and B. Mourrain. Structured low rank decomposition of multivariate hankel matrices. *Linear Algebra Appl.*, 542 :162–185, April 2018.

[82] J. Hérault and C. Jutten. Space or time adaptive signal processing by neural networks models. In *Int. Conf. Neural Networks for Computing*, pages 206–211, Snowbird (Utah, USA), 1986.

[83] J. Hérault, C. Jutten, and B. Ans. Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. In *GRETSI*, pages 1017–1022, Nice, May 1985.

[84] F. L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *J. Math. and Phys.*, 6(1) :165–189, 1927.

[85] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. Sciences (PNAS)*, 79(8) :2554–2558, April 1982.

[86] S. Hosseini and Y. Deville. Blind separation of linear-quadratic mixtures of real sources using a recurrent structure. In *Proceedings of the 7th International Work-Conference on Artificial and Natural Neural Networks (IWANN 2003)*, IWANN '03, page 241–248, Berlin, Heidelberg, 2003. Springer-Verlag.

[87] S. Hosseini and Y. Deville. Blind maximum likelihood separation of a linear-quadratic mixture. In C. G. Puntonet and A. Prieto, editors, *Proceedings of the 5th International Conference on Independent Component Analysis and Blind Signal Separation, ICA 2004, Granada, Spain, September 22-24, 2004*, volume 3195 of *Lecture Notes in Computer Science*, pages 694–701. Springer, 2004.

[88] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks*, 10(3) :626–634, 1999.

[89] A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis : Existence and uniqueness results. *Neural Networks*, 12(3) :429–439, 1999.

[90] C. Jutten. *Calcul neuromimétique et traitement du signal : analyse en composantes indépendantes*. PhD thesis, Univ. Joseph Fourier et INP Grenoble, 1981. Thèse d'état.

[91] C. Jutten and J. Hérault. Analog implementation of a permanent unsupervised learning algorithm. In *NATO Workshop on Neurocomputing*, Les Arcs, France, 1989.

[92] C. Jutten and J. Hérault. Blind separation of sources, part I : An adaptive algorithm based on neuromimetic architecture. *Signal Processing, Elsevier*, 24(1) :1–10, July 1991.

[93] A. M. Kagan, Y. V. Linnik, and C. R. Rao. *Characterization Problems in Mathematical Statistics*. Probability and Mathematical Statistics. Wiley, New York, 1973.

[94] T. Kailath. *Linear Systems*. Prentice-Hall, 1980.

[95] T. Kim and T.-W. Eltoft, T.and Lee. Independent vector analysis : An extension of ICA to multivariate components. In *Proceedings ICA 2006 (LNCS 3889)*, page 165–172. Springer-Verlag, 2006.

[96] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1) :59–69, January 1982.

[97] P. Lascaux and R. Theodor. *Analyse numérique matricielle appliquée à l'art de l'ingénieur*, volume 2. Masson, Dunod, 1987, 2004.

[98] L. De Lathauwer, B. De Moor, and J. Vandewalle. Independent Component Analysis and (simultaneous) third-order tensor diagonalization. *IEEE Trans. Sig. Proc.*, pages 2262–2271, October 2001.

[99] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401 :788–791, 1999.

[100] J. Li and J. M. Bioucas-Dias. Minimum volume simplex analysis : A fast algorithm to unmix hyperspectral data. In *IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium*, volume 3, pages III – 250–III – 253, 2008.

[101] P. Liavas, P. A. Regalia, and J-P. Delmas. On the robustness of the linear prediction method for blind channel identification. *IEEE Trans. Sig. Proc.*, 48 :1477–1481, May 2000.

[102] L-H. Lim and P. Comon. Multiarray signal processing : Tensor decomposition meets compressed sensing. *Compte-Rendus Mécanique de l'Academie des Sciences*, 338(6) :311–320, June 2010. hal-00512271.

[103] L.-H. Lim and P. Comon. Blind multilinear identification. *IEEE Trans. Inf. Theory*, 60(2) :1260–1280, February 2014.

[104] C.-H. Lin, W.-K. Ma, W.-C. Li, C.-Y. Chi, and A. Ambikapathi. Identifiability of the simplex volume minimization criterion for blind hyperspectral unmixing : The no-pure-pixel case. *IEEE Transactions on Geoscience and Remote Sensing*, 53(10) :5530–5546, 2015.

[105] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Processing*, 41(12) :3397–3415, 1993.

[106] P. Mccullagh. *Tensor Methods in Statistics*. Monographs on Statistics and Applied Probability. Chapman and Hall, 1987.

[107] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5(4) :115–133, December 1943.

[108] F. Merrikh-Bayat, M. Babaie-Zadeh, and C. Jutten. A nonlinear blind source separation solution for removing the show-through effect in the scanned documents. In *EUSIPCO 2008 - 16th European Signal Processing Conference*, volume CD-Rom, page 5 pages, Lausanne, Switzerland, August 2008. EURASIP. 5 pages.

[109] H. Meyr, M. Moeneclaey, and S. A. Fechtel. *Digital communication receivers : synchronization, channel estimation, and signal processing*. Wiley, 1998.

[110] G. H. Mohimani, M. Babaie-Zadeh, and C. Jutten. A fast approach for overcomplete sparse decomposition based on smoothed ℓ_0 norm. *IEEE Trans. Signal Processing*, 57(1) :289–301, 2009.

[111] F. Mokhtari, M. Babaie-Zadeh, and C. Jutten. Blind separation of bilinear mixtures using mutual information minimization. In T. Adali, J. Chanussot, C. Jutten, and J. Larsen, editors, *MLSP 2009 - IEEE 19th International Workshop on Machine Learning for Signal Processing*, page 6 pages, Grenoble, France, September 2009.

[112] E. Moulines, P. Duhamel, J. F. Cardoso, and S. Mayrargue. Subspace methods for the blind identification of multichannel FIR filters. *IEEE Trans. Sig. Proc.*, 43(2) :516–525, February 1995.

[113] S. Moussaoui, D. Brie, and J. Idier. Non-negative source separation : range of admissible solutions and conditions for the uniqueness of the solution. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 5, pages v/289–v/292 Vol. 5, 2005.

[114] F. Movahedi Naini, G. Hosein Mohimani, M. Babaie-Zadeh, and Christian Jutten. Estimating the mixing matrix in sparse component analysis (sca) based on partial k-dimensional subspace clustering. *Neurocomputing*, 71(10) :2330–2343, 2008. Neurocomputing for Vision Research Advances in Blind Signal Processing.

[115] J. M. P. Nascimento and J. M. Bioucas-Dias. Vertex component analysis : a fast algorithm to unmix hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(4) :898–910, 2005.

[116] H. Nguyen Thi, J. Caelen, and Christian Jutten. Réhaussement de la parole par la séparation de sources dans un mélange convolutif. *Journal de Physique IV Proceedings*, 04(C5) :C5–541–C5–544, 1994.

[117] A. Ozerov and C. Fevotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3) :550–563, 2010.

[118] P. Paatero and U. Tapper. Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5 :111–126, 1994.

[119] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit : recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, pages 40–44 vol.1, 1993.

[120] D. T. Pham and J-F. Cardoso. Blind separation of instantaneous mixtures of nonstationary sources. *IEEE Trans. SP*, 49(9) :1837–1848, September 2001.

[121] B. Picinbono. Higher-order statistical signal processing with Volterra filters. In *Proc. Workshop on Higher-Order Spectral Analysis,* Vail, Colorado, pages 62–67, June 1989.

[122] M. Plumbley. Conditions for nonnegative independent component analysis. *IEEE Signal Processing Letters*, 9(6) :177–180, 2002.

[123] Y. Qi, P. Comon, and L. H. Lim. Uniqueness of nonnegative tensor approximations. *IEEE Trans. Inf. Theory*, 62(4) :2170–2183, April 2016. hal-01015519, arXiv :1410.8129.

[124] F. Raimondi and P. Comon. Tensor decomposition of polarized seismic waves. In *GRETSI'2015*, Lyon, France, September 8-11 2015. hal-01164363.

[125] F. Raimondi, P. Comon, O. Michel, S. Sahnoun, and A. Helmstetter. Tensor decomposition exploiting diversity of propagation velocities ; application to localization of icequake events. *Signal Processing*, 118 :75–88, January 2016.

[126] F. Raimondi, R. Cabral Farias, O. Michel, and P. Comon. Wideband multiple diversity tensor array processing. *IEEE Trans. Sig. Proc.*, 65 :5334–5346, October 2017. hal-01350549.

[127] F. E. D. Raimondi and P. Comon. Tensor DoA estimation with directional elements. *IEEE Signal Processing Letters*, 24(5) :648–652, May 2017. hal-01369713.

[128] R. Rajagopal and L. C. Potter. Multivariate MIMO FIR inverses. *IEEE Trans. Image Proc.*, 12(4) :458–465, April 2003.

[129] B. Ramachandran and K-S. Lau. *Functional Equations in Probability Theory*. Wiley, 1991.

[130] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Math. Statist.*, pages 400–407, 1951.

[131] J.-P. Roll. *Contribution à la proprioception musculaire, à la perception et au contrôle du mouvement chez l'homme*. PhD thesis, Université d'Aix-Marseille 1, 1981.

[132] F. Rosenblatt. The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6) :386–408, 1958.

[133] L. Rota, V. Zarzoso, and P. Comon. Parallel deflation with alphabet-based criteria for blind source extraction. In *IEEE SSP'05*, pages 751–756, Bordeaux, France, July, 17-20 2005.

[134] W. J. Rugh. *Non Linear Systems Theory, Volterra and Wiener Approaches*. Hopkins, 1989.

[135] P. Ruiz and J. l. Lacoume. Extraction of independent sources from correlated inputs. In *Proc. Workshop on Higher-Order Spectral Analysis*, pages 146–151, Vail, Colorado, June 28-30 1989. ONR-NSF, IEEE.

[136] D. Rumelhart, G. Hinton, and R. Williams. Learning representations by back-propagating errors. *Nature*, 323 :533–536, October 1986.

[137] S. Sahnoun and P. Comon. Joint source estimation and localization. *IEEE Trans. Sig. Proc.*, 63(10) :2485–2495, May 2015. hal-01005352.

[138] S. Sahnoun, K. Usevich, and P. Comon. Multidimensional ESPRIT for damped and undamped signals : Algorithm, computations and perturbation analysis. *IEEE Trans. Sig. Proc.*, 65(22) :5897–5910, November 2017. hal-01360438.

[139] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari. A review of blind source separation methods : two converging routes to ilrma originating from ica and nmf. *APSIPA Transactions on Signal and Information Processing*, 8(1) :–, 2019.

[140] O. Shalvi and E. Weinstein. New criteria for blind deconvolution of nonminimum phase systems. *IEEE Trans. Inf. Theory*, 36(2) :312–321, March 1990.

[141] N. D. Sidiropoulos, R. Bro, and G. B. Giannakis. Parallel factor analysis in sensor array processing. *IEEE Trans. Sig. Proc.*, 48(8) :2377–2388, August 2000.

[142] N. D. Sidiropoulos, G. B. Giannakis, and R. Bro. Blind PARAFAC receivers for DS-CDMA systems. *IEEE Trans. on Sig. Proc.*, 48(3) :810–823, March 2000.

[143] V. De Silva and L-H. Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis Appl.*, 30(3) :1084–1127, 2008.

[144] A. Smilde, R. Bro, and P. Geladi. *Multi-Way Analysis*. Wiley, Chichester UK, 2004.

[145] A. Stegeman and N. Sidiropoulos. On Kruskal's uniqueness condition for the CP decomposition. *Lin. Alg. Appl.*, 420(2-3) :540–552, January 2007.

[146] A. Taleb. A generic framework for blind source separation in structured nonlinear models. *IEEE Transactions on Signal Processing*, 50(8) :1819–1830, 2002.

[147] A. Taleb and C. Jutten. Nonlinear source separation : the post-nonlinear mixtures. In *The European Symposium on Artificial Neural Networks (ESANN)*, 1997.

[148] A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Transactions on Signal Processing*, 47(10) :2807–2820, 1999.

[149] L. B. Thomas. Rank factorization of nonnegative matrices (A. Berman). *SIAM Review*, 16(3) :393–394, 1974.

[150] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1) :267–288, 1996.

[151] L. Tong, R. Liu, and V. C. Soon. Indeterminacy and identifiability of blind identification. *IEEE Trans. on Circuits and Syst.*, 38(5) :499–509, May 1991.

[152] J. K. Tugnait. Identification and deconvolution of multichannel non-Gaussian processes using higher order statistics and inverse filter criteria. *IEEE Trans. Sig. Proc.*, 45 :658–672, March 1997.

[153] Z. Vavrin. A unified approach to Loewner and Hankel matrices. *Linear Algebra Appl.*, 143 :171–222, January 1991.

[154] M. Veganzones, J. E. Cohen, R. Cabral Farias, J. Chanussot, and P. Comon. Nonnegative tensor CP decomposition of hyperspectral data. *IEEE Trans. Geoscience and Remote Sensing*, 54(5) :2577–2588, May 2016.

[155] E. Vincent. Musical source separation using time-frequency source priors. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1) :91–98, 2006.

[156] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Trans. ASLP*, 14(4) :1462–1469, July 2006.

[157] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4) :1462–1469, 2006.

[158] E. A. Vittoz and X. Arreguit. Cmos integration of herault-jutten cells for separation of sources. In *Analog VLSI Implementation of Neural Systems*, 1989.

[159] D. Wang, Z. Chang, and F. Cong. Sparse nonnegative tensor decomposition using proximal algorithm and inexact block coordinate descent scheme. *Neural Computing and Applications*, 33 :17369–17387, 2021.

[160] V. Zarzoso and P. Comon. Blind and semi-blind equalization based on the constant power criterion. *IEEE Trans. Sig. Proc.*, 53(11) :4363–4375, November 2005.

[161] V. Zarzoso and P. Comon. Blind channel equalization with algebraic optimal step size. In *Eusipco'05*, Antalya, Turkey, Sept. 4-8 2005.

[162] V. Zarzoso and P. Comon. Alphabet-based deflation for blind source extraction in underdetermined mixtures. In *Proc. ICA Research Network International Workshop*, pages 21–24, Liverpool, UK, Sept. 18–19, 2006.

[163] V. Zarzoso and P. Comon. Comparative speed analysis of FastICA. In *7th Int. Conf. ICA*, LNCS 4666, pages 293–300, London, UK, September 9-12 2007. Springer.

[164] V. Zarzoso and P. Comon. Optimal step-size constant modulus algorithm. *IEEE Trans. Com.*, 56(1) :10–13, January 2008.

[165] V. Zarzoso and P. Comon. Robust independent component analysis by iterative maximization of the kurtosis contrast with algebraic optimal step size. *IEEE Trans. NN*, 21(2) :248–261, February 2010.

[166] V. Zarzoso, P. Comon, and M. Kallel. How fast is FastICA ? In *XIV European Signal Processing Conference, Eusipco'06*, Florence, Italy, Sept. 4-8 2006.

[167] V. Zarzoso, R. Phlypo, and P. Comon. A contrast for independent component analysis with priors on source kurtosis signs. *IEEE Signal Processing Letters*, 15 :501–504, 2008.

[168] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang. A survey of sparse representation : Algorithms and applications. *IEEE Access*, 3 :490–530, 2015.