

Séparation de sources non-négatives : Une méthode géométrique robuste au bruit

Wendyam Serge Boris OUEDRAOGO¹, Antoine SOULOUMIAC², Meriem JAIDANE³, Christian JUTTEN¹

¹GIPSA-lab, UMR 5216 CNRS

11 rue des mathématiques, BP 46, F-38402 Grenoble Cedex, France

²CEA, LIST, Laboratoire d’Outils pour l’Analyse de Données
Gif-sur-Yvette, F-91191, France ;

³Unité de recherche Signaux et Systèmes

École Nationale d’Ingénieurs de Tunis, BP 37, 1002 Tunis, Tunisie

wendyam.ouedraogo@gipsa-lab.grenoble-inp.fr, antoine.souloumiac@cea.fr

Résumé – Nous avons proposé récemment une méthode géométrique de séparation des mélanges linéaires instantanés de sources non-négatives, dénommée SCSA-UNS pour Simplicial Cone Shrinking Algorithm for Unmixing Non-negative Sources. Cette méthode qui fonctionne en absence de bruit, estime la matrice de mélange et les sources en cherchant le cône simplicial d’ouverture minimale, contenant le nuage des signaux observés. Nous proposons ici une extension de la méthode SCSAUNS pour les mélanges bruités, dans le cas où les taux de parcimonie des sources sont connues a priori. L’idée est d’éliminer progressivement les observations qui, avec l’effet du bruit, se dispersent en dehors du cône simplicial engendré par la matrice de mélange. Des simulations sur des données synthétiques montrent l’efficacité de la méthode proposée.

Abstract – Recently, we proposed a geometrical method for unmixing linear instantaneous mixtures of non-negative sources, denoted SCSA-UNS for Simplicial Cone Shrinking Algorithm for Unmixing Non-negative Sources. The latter method operates in noiseless case, and estimates the mixing matrix and the sources by finding the minimum aperture simplicial cone, containing the scatter plot of mixed data. In this paper, we propose an extension of SCSA-UNS to tackle noisy mixtures, when the sparsity degrees of the sources are known a priori. The proposed method iteratively eliminates the noisy mixed data points, which lie outside of the simplicial cone generated by the mixing matrix. Simulations on synthetic data show the effectiveness of the proposed method.

1 Introduction

Le problème de séparation de mélanges linéaires instantanés de sources non-négatives (SAS non-négatives) se rencontre dans de nombreuses situations, par exemple lorsqu’il s’agit d’estimer les spectres et les profils de concentration de différents composants d’une substance chimique [1], lors de l’estimation des abondances et des signatures spectrales des différents éléments d’une scène observée par un spectro-imageur [6], ou encore pour l’estimation des pharmacocinétiques d’un traceur radioactif dans un organe [2]. Le modèle de mélange est donné par l’équation (1) :

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{B} \quad (1)$$

où $\mathbf{X} \in \mathbb{R}_+^{m \times K}$ est la matrice des données mesurées, $\mathbf{A} \in \mathbb{R}_+^{m \times n}$ est la matrice de mélange, $\mathbf{S} \in \mathbb{R}_+^{n \times K}$ est la matrice des sources et $\mathbf{B} \in \mathbb{R}^{m \times K}$ est la matrice de bruit. n , m et K sont respectivement le nombre de sources, le nombre d’observations et le nombre d’échantillons. Notons par $x_i(k)$ le k -ième échantillon de l’observation \mathbf{x}_i , $s_j(k)$ le k -ième échantillon de la source \mathbf{s}_j , et a_{ij} la contribution de la source j à la mesure i . Nous supposons que $m \geq n$ et que la matrice de mélange est de rang plein. L’objectif est d’estimer \mathbf{A} et \mathbf{S} à partir de \mathbf{X} .

Parmi les méthodes de séparation de sources non-négatives, les méthodes géométriques présentent un grand intérêt du fait de leur efficacité et de leur faible coût de calcul. Ces méthodes ne nécessitent pas l’indépendance mutuelle des sources, ni de connaissances (même a priori) sur leur lois de probabilités. Dans ces approches, chaque colonne de la matrice des données mesurées \mathbf{X} est représentée par un point dans l’espace des observations; l’ensemble des observations forment ainsi un nuage de points dans l’espace. Ces méthodes utilisent les propriétés géométriques de ce nuage de points pour concevoir les algorithmes de séparation. Le premier type de méthodes géométriques suppose la dominance locale des sources (c’est-à-dire que $\forall 1 \leq i \leq n, \exists k_i$ tel que $s_i(k_i) \neq 0$ et $s_j(k_i) = 0$ pour $j \neq i$), ces méthodes estiment la matrice de mélange et les sources en cherchant les points extrêmes du nuage des observations [3, 6]. L’autre type de méthodes géométriques est fondé sur l’estimation de la matrice de mélange par recherche du simplexe de volume minimal contenant le nuage des observations [9, 10, 8]. Cependant les performances des méthodes géométriques se dégradent très fortement en présence de bruit (même à “fort” rapport signal sur bruit).

Dans ce travail, nous présentons une technique d’amélioration

de la robustesse au bruit, de la méthode géométrique décrite dans [4]. Outre le fait qu'elle ne nécessite pas la dominance locale des sources, la nouvelle méthode proposée présente une meilleure robustesse au bruit, par rapport aux autres méthodes géométriques.

2 Rappel de la méthode SCSA-UNS en absence de bruit

Nous présentons la méthode SCSA-UNS décrite dans [4], dans le cas où le mélange déterminé (c'est-à-dire quand $m = n$). Si le mélange est sur-déterminé ($m > n$), une première étape de réduction de dimensions doit être effectuée, par exemple en utilisant la décomposition en valeurs singulières, cela est aussi décrit dans [4].

Pour une matrice $\mathbf{U} \in \mathbb{R}_+^{n \times n}$, nous définissons le cône simplicial engendré par la matrice \mathbf{U} par :

$$\mathbf{Span}^+(\mathbf{U}) = \{ \mathbf{z} \mid \mathbf{z} = \mathbf{U}\mathbf{y} \text{ avec } \mathbf{y} \in \mathbb{R}_+^n \} \quad (2)$$

Nous définissons aussi l'ouverture de $\mathbf{Span}^+(\mathbf{U})$ par l'équation (3), où \mathbf{u}_i est la i -ième colonne de \mathbf{U} .

$$\mathcal{A}(\mathbf{Span}^+(\mathbf{U})) = \frac{|\det(\mathbf{U})|}{\|\mathbf{u}_1\| \times \|\mathbf{u}_2\| \times \dots \times \|\mathbf{u}_n\|} \quad (3)$$

En absence de bruit, le nuage des observations, est contenu dans le cône simplicial engendré par la matrice de mélange:

$$\{ \mathbf{x}(k), 1 \leq k \leq K \} \subseteq \mathbf{Span}^+(\mathbf{A}) \quad (4)$$

La figure 1.a illustre le cône engendré par la matrice de mélange, ainsi que le nuage des observations dans le cas $m = n = 3$. La figure 1.b illustre la projection de $\mathbf{Span}^+(\mathbf{A})$ et du nuage des observations sur le plan défini par:

$$\mathcal{P} = \left\{ \mathbf{y} = [y_1, y_2, y_3]^T \in \mathbb{R}^3 \mid \sum_{i=1}^3 y_i = 1 \right\} \quad (5)$$

Les observations et le cône projetés sont obtenus par :

$$\tilde{x}_i(k) = x_i(k) / \sum_{l=1}^3 x_l(k) \text{ and } \tilde{a}_{ij} = a_{ij} / \sum_{l=1}^3 a_{il} \quad (6)$$

En supposant que $\mathbf{Span}^+(\mathbf{A})$ est l'unique cône simplicial d'ouverture minimale contenant le nuage des observations, l'idée de la méthode SCSA-UNS est d'estimer la matrice de mélange et les sources en cherchant ce cône. Cela revient à résoudre l'équation (7) :

$$\hat{\mathbf{A}} = \arg \min_{\mathbf{A} \geq 0, \mathbf{A}^{-1}\mathbf{X} \geq 0} \mathcal{A}(\mathbf{Span}^+(\mathbf{A})) \quad (7)$$

Dans [4], nous proposons un algorithme qui permet de résoudre l'équation (7). En partant d'un cône simplicial initial, contenant le nuage des observations, la méthode proposée permet de réduire itérativement l'ouverture de ce cône, jusqu'à ce qu'il s'ajuste au nuage des observations.

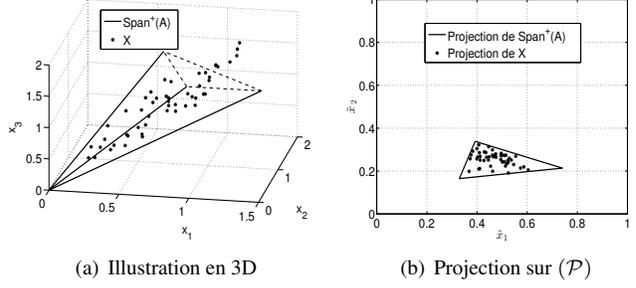


FIG. 1: Illustration de $\mathbf{Span}^+(\mathbf{A})$ et du nuage des observations pour $m = n = 3$

3 Amélioration de la robustesse au bruit de SCSA-UNS: Robust SCSA-UNS

En présence de bruit additif, certains points du nuage des observations peuvent se retrouver en dehors du cône simplicial engendré par la matrice de mélange, ceci est illustré sur la figure 2.

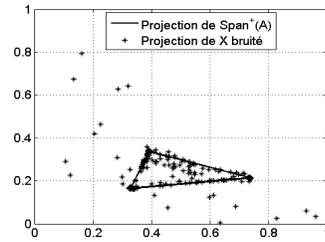


FIG. 2: Illustration de l'effet du bruit sur le nuage des observations, pour $m = n = 3$

Lors de l'estimation de la matrice de mélange, il est nécessaire de ne pas considérer les points qui se situent à l'extérieur du cône qu'elle engendre. Une analyse nous a permis de voir que les points les plus susceptibles d'être entraînés hors de $\mathbf{Span}^+(\mathbf{A})$ par le bruit sont ceux, qui en absence de bruit, sont obtenus à partir de sources de coordonnées nulles¹ ou sont à dominance locale. Nous proposons ici une méthode de réduction de l'effet du bruit sur les performances de la séparation, dans le cas où l'on connaît a priori les pourcentages de valeurs nulles (aussi appelé taux parcimonie) de chaque source.

Soit τ_i le taux de parcimonie de la source i ($0 \leq \tau_i \leq 1$), $\tau_i = 0$ si aucune valeur de la source i n'est nulle, et $\tau_i = 1$ si toutes les valeurs de la source i sont nulles. Ainsi, la probabilité que la k -ième valeur de la source i soit nulle vaut:

$$Pr[s_i(k) = 0] = \tau_i = Pr_i \quad (8)$$

Ainsi la probabilité qu'un point du nuage des sources soit de coordonnées nulles vaut:

$$P_1 = Pr[s_1(k) = 0, s_2(k) = 0, \dots, s_n(k) = 0] \quad (9)$$

¹Un point $\mathbf{s}(k) = [s_1(k), s_2(k), \dots, s_n(k)]^T$ est de coordonnées nulles si $s_i(k) = 0, \forall 1 \leq i \leq n$

Un point $\mathbf{s}(k) = [s_1(k), s_2(k), \dots, s_n(k)]^T$ est à dominance locale, s'il existe $1 \leq i \leq n$ tel que $s_i(k) \neq 0$ et $s_j(k) = 0$ pour $j \neq i$. La probabilité qu'un point du nuage des sources soit à dominance locale vaut donc:

$$P_2 = \sum_{i=1}^n Pr[s_1(k) = 0, \dots, s_i(k) \neq 0, \dots, s_n(k) = 0] \quad (10)$$

Si de plus nous supposons que les sources sont "indépendantes" alors P_1 et P_2 valent respectivement

$$P_1 = \prod_{i=1}^n Pr_i = \prod_{i=1}^n \tau_i \quad (11)$$

$$P_2 = \sum_{i=1}^n \left[(1 - Pr_i) \prod_{j=1, j \neq i}^n Pr_j \right] = \sum_{i=1}^n \left[(1 - \tau_i) \prod_{j=1, j \neq i}^n \tau_j \right] \quad (12)$$

Il faut préciser que l'hypothèse d'indépendance des sources faite ici est une "hypothèse technique", qui permet d'exprimer la probabilité jointe en fonction du produit des probabilités marginales. Elle ne constitue cependant pas une condition nécessaire au fonctionnement de la méthode SCSCA-UNS.

Les nombres de points de coordonnées nulles et à dominance locale peuvent ainsi être estimés respectivement par :

$$K_1 = P_1 K \text{ et } K_2 = P_2 K \quad (13)$$

La méthode Robust SCSCA-UNS consiste à exécuter itérativement l'algorithme SCSCA-UNS [4] et à éliminer à chaque fois les points situés sur l'enveloppe convexe du cône engendré par l'estimée de la matrice de mélange, obtenue à l'itération courante. L'algorithme s'arrête lorsque le nombre total de point éliminés vaut $K_1 + K_2$. Le pseudo-code de la méthode Robust SCSCA-UNS est donné par l'Algorithme 1 ci-dessous:

Algorithm 1 : Robust SCSCA-UNS

- 1: Initialisation: mettre $K_b = 0$, et rouver un cône simplicial initial contenant le nuage des observations (voir [4])
- 2: **repeat**
- 3: A partant du cône simplicial courant, exécuter l'algorithme SCSCA-UNS pour estimer le cône simplicial d'ouverture minimal contenant le nuage des points restants.
- 4: Sélectionner et éliminer tous les points situés sur l'enveloppe convexe du cône simplicial estimé à l'étape précédente.

$$K_b = K_b + \text{le nombre de point sélectionné}$$

- 5: **until** $K_b \geq K_1 + K_2$
-

4 Simulations

Nous présentons ici des résultats de simulations sur données synthétiques. Pour cela, nous fixons : $m = 10$, $n = 3$ et $K =$

5000. Pour simplifier l'interprétation des résultats, chaque source est générée avec le même taux de parcimonie τ , les valeurs nulles étant tirées aléatoirement suivant une distribution uniforme entre 0 et 1. Les coefficients de la matrice de mélange sont générés aussi aléatoirement suivant une distribution uniforme entre 0 et 1. Le bruit sur chaque observation est généré suivant une distribution gaussienne de moyenne nulle et de variance unité, le rapport signal sur bruit défini par l'équation (14) est fixé à la même valeur pour toutes les observation. Pour maintenir la non-négativité des observations, nous mettons la valeurs négatives à zéro.

$$RSB_i = 10 \log_{10} \left(\sum_k \left[\sum_{j=1}^n a_{ij} s_j(k) \right]^2 / \sum_{k=1}^K [b_i(k)]^2 \right) \quad (14)$$

La méthode Robust SCSCA-UNS proposée ici, est comparée aux méthodes VCA [6], SISAL [7] et SCSCA-UNS [4], suivant l'erreur de séparation, E_{sep} , définie par l'équation (15), où $\Delta = \hat{\mathbf{A}}^\dagger \mathbf{A}$ et $\hat{\mathbf{A}}^\dagger$ est la pseudo-inverse de l'estimée de la matrice de mélange.

$$E_{sep} = \frac{1}{2n(n-1)} \sum_{i=1}^n \left[\sum_{j=1}^n \left(\frac{|\Delta_{ij}|}{\max_l |\Delta_{il}|} + \frac{|\Delta_{ji}|}{\max_l |\Delta_{li}|} \right) - 2 \right] \quad (15)$$

La figure 3 montre les évolutions des valeurs moyennes de l'erreur de séparation (obtenues pour 50 réalisations indépendantes Monte Carlo), en fonction du rapport signal sur bruit (RSB) pour différentes valeurs de τ . Cette figure montre que la méthode Robust SCSCA-UNS présente les meilleurs résultats de séparation par rapport aux autres méthodes, et les performances s'améliorent d'autant plus que le taux de parcimonie est élevé. Des résultats similaires (non présentés ici) ont été observés lorsqu'on augmente le nombre de sources.

5 Conclusions et Perspectives

Nous avons présenté dans cet article une amélioration de la méthode SCSCA-UNS pour la séparation de mélanges bruités de sources non-négatives. Une analyse de l'effet du bruit sur le nuage des observations nous a montré que les observations les plus susceptibles d'être entraînées en dehors du cône simplicial engendré par la matrice de mélange sont celle obtenues à partir de sources de coordonnées nulles et des sources à dominance locale. La méthode proposée opère dans le cas où les taux de parcimonie des sources sont connus a priori, elle consiste à éliminer progressivement les observations situés en dehors du cône simplicial engendré par la matrice de mélange, au cours de l'estimation de cette matrice. Des résultats de séparation sur données synthétiques ont montré une meilleur robustesse de la méthode proposée par rapport aux autres méthodes géométriques.

Les futures travaux consisteront à estimer les taux de parcimonie des sources, dans le cas où il ne seront pas connus a pri-

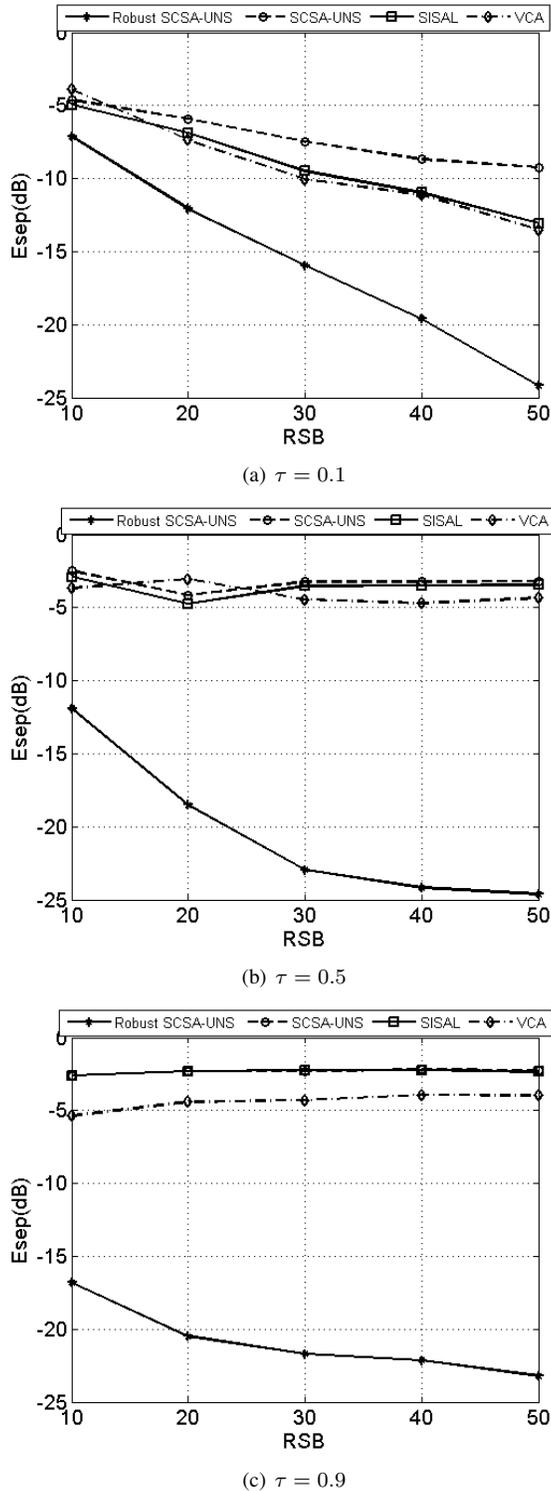


FIG. 3: Illustration de $\text{Span}^+(\mathbf{A})$ et du nuage des observations pour $m=10$ et $n=3$

ori, et à étudier l'effet d'une mauvaise estimation de ces taux sur les performances de séparation. L'application de la méthode Robust SCSA-UNS à des données réelles, notamment de spectrométrie de masse, est également envisagée.

References

- [1] S. Moussaoui. *Séparation de sources non-négatives. Application au traitement des signaux de spectroscopie*. PhD thesis, Université Henri Poincaré, Nancy 1, 2005.
- [2] J.S. Lee, D.D. Lee, S. Choi et D.S. Lee. *Application of Non-negative Matrix Factorization to Dynamic Positron Emission Tomography*. In 3rd Int. Conf. on Independent Component Analysis and Blind Signal Separation. pp. 556-562, 2001
- [3] C. Lazar, D. Nuzillard et P. Billaudel. *Une Approche Géométrique de la Séparation Aveugle de Sources* GRETSI, ID 502, GRETSI, Dijon, septembre 2009.
- [4] W.S.B. Ouedraogo, A. Souloumiac, M. Jaidane et C. Jutten. *Geometrical method using simplicial cones for overdetermined nonnegative blind source separation: application to real PET images*. LVA-ICA'12. Lecture Notes in Computer Science. vol. 7191, pp 494-501, 2012.
- [5] J. Li et J.M. Bioucas-Dias. *Minimum Volume Simplex Analysis : A Fast Algorithm to Unmix Hyperspectral Data*. IEEE IGARSS'08. vol. 3, pp. III-250 -III-253, 2008.
- [6] J.M.P. Nascimento et J.M. Bioucas-Dias. *Vertex Component Analysis: A Fast Algorithm to Unmix Hyperspectral Data*. IEEE IGARSS'05. Vol. 43, pp. 898-910, 2005.
- [7] J.M. Bioucas-Dias. *A Variable Splitting Augmented Lagrangian Approach to Linear Spectral Unmixing*. IEEE WHISPERS'09, pp.1-4, 2009.
- [8] T.H. Chan, C.Y. Chi, Y.M. Huang et W.K. Ma. *A Convex Analysis-Based Minimum-Volume Enclosing Simplex Algorithm for Hyperspectral Unmixing*. IEEE Transactions on Signal Processing. vol 57. pp. 4418-4432, 2009.
- [9] M.D. Craig. *Minimum-Volume Transforms for Remotely Sensed Data*. IEEE Transactions on Geoscience and Remote Sensing, vol 32. pp. 542-552, 1994.
- [10] J. Li, J.M. Bioucas-Dias *Minimum Volume Simplex Analysis : A Fast Algorithm to Unmix Hyperspectral Data*. IEEE International Symposium on Geoscience and Remote Sensing Symposium, 2008.
- [11] Y. Zhu, T.H. Chan, E.P. Hoffman, Y. Wang. *Gene Expression Dissection by Non-negative Well-Grouped Source Separation*. IEEE Workshop on Machine Learning for Signal Processing, pp. 255 - 260, 2008