Séparation aveugle de sources sonores par factorisation en matrices positives avec pénalité sur le volume du dictionnaire

Valentin LEPLAT, Nicolas GILLIS, Xavier SIEBERT, Andersen M.S. ANG

Département de Mathématique et Recherche Opérationnelle de l'Université de Mons
9 rue de Houdain, 7000 Mons, Belgique
valentin.leplat@umons.ac.be, nicolas.gillis@umons.ac.be
xavier.siebert@umons.ac.be, manshun.ang@umons.ac.be

Résumé – La séparation de sources désigne les techniques visant à retrouver des signaux inconnus appelés sources à partir d'une observation de leur mélange. Dans ce papier, nous considérons la situation où le signal mélangé a été enregistré avec un seul capteur. La séparation aveugle consiste à isoler et extraire chacun des signaux sonores sources sur base d'un nombre limité d'informations; habituellement la seule information plus ou moins maitrisée concerne le nombre de sources à priori présentes dans le signal mélangé. Sur base d'une représentation temps-fréquence du signal, une des méthodes les plus répandues se base sur l'utilisation de techniques de séparation telle que la NMF (Factorisation en matrices positives). Les méthodes NMF consistent classiquement en la minimisation d'une fonction de coût telles que les divergences de Kullback-Leibler et d'Itakura-Saito appartenant à la famille des divergences β. Dans ce papier, nous présentons un nouveau modèle de séparation basé sur la minimisation d'une divergence de Kullback-Leibler incluant une pénalité favorisant des solutions pour la matrice « dictionnaire » de volume minimum. Afin de résoudre ce problème, la fonction de coût est remplacée par une fonction auxiliaire séparable et convexe à minimiser. On montre alors que la minimisation de cette fonction objectif conduit à des résultats plus interprétables, notamment dans le cas où le rang de la factorisation est surestimé en regard du nombre de sources réellement présentes dans le signal.

Abstract – Audio source separation concerns techniques used to extract unknown signals called sources from a mixed signal. In this paper, we assume that the audio signal is recorded with a single microphone. Considering a mixed signal composed of various audio sources, the blind audio source separation consists in isolating and extracting each of the sources on the basis of a single recording. Usually, the only known information is the number of estimated sources present in the mixed signal. Based on a time-frequency representation of the signal, classical source separation techniques integrate algorithms such as nonnegative matrix factorization (NMF). Optimization problems in blind audio source separation are based on the minimization of criteria such as the Kullback-Leibler and Itakura-Saito divergences, both divergences belonging to the family of β-divergences. In this paper, we present a new model of separation based on the minimization of the Kullback-Leibler including a penalty term promoting the columns of the dictionary matrix to have small volume. In order to solve this problem, the global cost function is replaced by a convex and separable auxiliary function that will be minimized. We will show that we obtain more interpretable results in the case where the factorization rank (that is, the number of sources present into the mixed signal) is overestimated.

1 Introduction

La factorisation en matrices en positives est une technique d'approximation de rang faible utilisée pour la décomposition de données positives. Etant donnée une matrice $V \in \mathbb{R}_+^{F\times N}$ et un entier positif K, la NMF consiste à trouver une matrice positive W avec K colonnes et une matrice positive H avec K lignes telles que $V \approx WH$. Cette relation signifie que chaque colonne de V est approximée par une combinaison linéaire des colonnes de W pondérée par les éléments des colonnes correspondantes de H. Dans le cas où V correspond au spectrogramme d'amplitude (ou de puissance) d'un signal audio, la matrice W est appelée dictionnaire dont chaque colonne contient la signature spectrale d'un composant, les lignes de la matrice H représentent les coefficients d'activation de chaque composant le long de la dimension N (au cours du temps dans notre cas). Notons que la méthode de séparation de source présentée dans ce papier s'applique aux réprésentations temps-fréquence quadratiques qui satisfont la propriété de positivité.

La factorisation est habituellement recherchée en considérant le problème de minimisation suivant :

$$\begin{aligned} \min_{W\in\mathbb{R}^{F\times K}, H\in\mathbb{R}^{K\times N}} \quad & D(V|WH) = \sum_{fn} d(V_{fn}|[\,WH\,]_{fn}) \\ \text{tel que} \quad & H \geq 0, W \geq 0\,, \end{aligned} \tag{1}$$

où la notation $A \geq 0$ exprime donc la contrainte de positivité sur les entrées de A et où $d(x|y)$ est une mesure d'écart entre les scalaires x et y. Pour la séparation aveugle de sources sonores, une fonction de coût communément utilisée est la divergence-β discrète notée $d_\beta(x,y)$ définie par :

$$d_\beta(x,y) = \begin{cases} \frac{1}{\beta(\beta-1)}\left(x^\beta + (\beta-1)\,y^\beta - \beta x y^{\beta-1}\right) \\ \qquad \text{pour} \quad \beta \in \mathbb{R}\setminus(0,1)\,, \\ x\log\frac{x}{y} - x + y \quad \text{pour} \quad \beta = 1, \\ \frac{x}{y} - \log\frac{x}{y} - 1 \quad \text{pour} \quad \beta = 0. \end{cases}$$

La divergence β est ainsi définie par la valeur particulière donnée à β, et correspond à la norme de Frobenius, la divergence de Kullback-Leibler et la divergence d'Itakura-Saito dans les cas particuliers où β=2, 1 et 0, respectivement. Dans ce cas, la fonction objectif de (1) s'écrit comme suit $D_\beta(V|WH) = \sum_{fn} d_\beta(V_{fn}|[WH]_{fn})$.

La factorisation en matrices positives est dans la plupart des cas mal posée car la solution optimale n'est pas unique. Afin de faire en sorte que la solution du problème (1) soit unique (aux permutations et mises à l'échelle près sur les lignes de H et les colonnes de W) rendant ainsi le problème bien posé et facteurs (W, H) identifiables, une technique est de rechercher une solution pour W de volume (engendré par l'espace colonne) minimum ; voir par exemple [1].

2 Modèle β-NMF de volume minimum

Dans ce papier, nous présentons la formulation suivante pour la β-NMF de volume minimum :

$$\min_{W(:,j)\in\Delta^F \forall j, H\geq 0} \quad F(W,H) = D_\beta(V|WH) + \lambda \text{vol}(W)\,, \tag{2}$$

où $\Delta^F = \left\{x \in \mathbb{R}_+^F | \sum_i x_i = 1\right\}$, λ est le poids du terme de pénalité et vol(W) est une fonction de mesure du volume engendré par les colonnes de W. Notez qu'une normalisation est considérée pour les colonnes de W afin d'éviter que W ne tende vers zéro sachant que $WH = (W/a)(Ha)$ pour n'importe quel $a > 0$. Dans ce papier, nous utiliserons

$$\text{vol}(W) = \text{logdet}(W^TW + \delta I),$$

dans le problème (2), où I est la matrice identité d'ordre K et δ est un scalaire positif qui empêche le terme logdet(W^TW) de tendre vers $-\infty$ lorsque W tend vers une matrice de rang incomplet ($r = \text{rank}(W) < K$). La raison d'utiliser une telle mesure est que $\sqrt{\det(W^TW)}/K!$ est le volume de l'enveloppe convexe des colonnes de W et de l'origine. Une seconde motivation importante pour l'utilisation de logdet$(W^TW + \delta I)$ en tant que régularisation sur le volume plutôt que det(W^TW) est sa plus grande simplicité de calcul : bien que les deux fonctions soient non-convexes et conceptuellement aussi complexes à gérer, la première permet de trouver des mises à jours plus simples car elle possède une borne supérieure dite serrée alors que la seconde non, voir [6] pour plus de détails. Dans le cas sans bruit et sous certaines conditions sur $V = WH$, ce modèle permettra d'identifier les facteurs latents $\left(W^\#, H^\#\right)$ qui ont généré V. Ces conditions particulières nécessitent que les colonnes de V soient suffisamment bien réparties dans l'enveloppe convexe générée par les colonnes de W, voir [2], [3] et [4] ; il s'agit de la condition de dispersion suffisante (« Sufficiently scattered condition » dans la littérature anglaise). En particulier, les données (colonnes de V) doivent être localisées sur les facettes de l'enveloppe convexe, ce qui revient à dire que H doit être « suffisamment » creuse. A notre connaissance, ces résultats théoriques ne s'appliquent que dans le cas exact (pas de bruit dans les données), par conséquent la robustesse au bruit du modèle (2) doit encore être rigoureusement étudiée [8]. La condition de dispersion suffisante est une généralisation de la condition de séparabilité qui nécéssite que $W = V(:, \kappa)$ pour un ensemble d'indices κ de taille K. La séparabilité rend la résolution du problème NMF plus aisée. Remarquons néanmoins que bien que la NMF de volume minimum garantisse l'identifiabilité des facteurs latents, le problème (2) est toujours difficile à résoudre dans la plupart des cas ; comme l'est la NMF originale [5].

3 Algorithme pour min-vol KL-NMF (2)

Une stratégie d'optimisation populaire pour la NMF est basée sur une série d'itérations au cours desquelles les matrices W et H sont mises à jour et optimisées de manière alternative, nous avons adopté cette stratégie dans ce papier. Afin de résoudre le problème, la fonction de coût du modèle (2) est remplacée par une fonction auxiliaire séparable (c'est-à-dire que les variables sont découplées et peuvent ainsi être optimisées indépendemment) qui consitue une borne supérieure convexe que l'on va minimiser. Il s'agit donc d'un algorithme de majorisation-minimisation.

Dû à la limitation en taille de ce papier, nous ne présentons ici que la méthode générale pour la construction de cette fonction auxiliaire séparable et convexe. L'intégralité des développements sera disponible dans un article à paraître prochainement. Le principe repose sur la construction d'une fonction auxiliaire pour chacun des deux termes de (2).

Pour le terme $D_\beta(V|WH)$ de (2), nous avons utilisé la fonction auxiliaire présentée dans [7].

En ce qui concerne le terme logdet, la construction repose sur les éléments suivants :

— tout d'abord la construction d'une borne supérieure strictement convexe à partir de l'approximation de Taylor limité au premier ordre de la fonction logdet comme utilisé par exemple dans [6],

— ensuite la construction d'une borne supérieure séparable à l'approximation du point précédent.

Pour $\beta = 1$ (la mesure d'écart correspond donc à la divergence de Kullback-Leibler), les mises à jour multiplicatives suivantes pour W et H garantissent la décroissance de la fonction objectif $F(W, H)$:

$$W \leftarrow W \odot \frac{\left[\left[\left[\Phi\right]^{\cdot 2} + 2\Theta \odot \left(\frac{[V]}{[WH]}H^T\right)\right]^{\cdot \frac{1}{2}} - \Phi\right]}{[\Theta]} \tag{3}$$

où $\Phi = J_{F,N}H^T - 4\lambda(WY^-)$, $\Theta = 4\lambda W(Y^+ + Y^-)$, $\odot$ représente le produit matriciel de Hadamard, $\frac{[.]}{[.]}$ est l'opérateur de division élément par élément, $(.)^{(.)}$ est l'opérateur de puissance élément par élément, $J_{F,N}$ est une matrice de uns de dimensions $F \times N$, $Y = (W^TW + \delta I)^{-1}$ avec $\delta > 0$, et on définit $Y^+ = \max(Y, 0)$ et $Y^- = \max(-Y, 0)$ de telle manière que $Y = Y^+ - Y^-$. La mise à jour pour les entrées de H est

$$H \leftarrow H \odot \frac{\left[W^T\left(\frac{[V]}{[WH]}\right)\right]}{[W^TJ_{F,N}]}, \tag{4}$$

comme dans l'article original de Lee et Seung [9]. Les mises à jour définies dans (3) et (4) garantissent la décroissance de la fonction objectif de (2). Cependant, lors de la normalisation des colonnes de W, le terme divergence-β de F ne varie pas (si les lignes de H sont mises à l'échelle de manière appropriée) mais le terme logdet changera et par conséquent la fonction objectif F pourrait augmenter. Pour parer à ce problème, nous intégrons une procédure de recherche en ligne. L'algorithme 1 implémente cette stratégie et sera désigné par le nom « min-vol KL-NMF » dans la suite de ce papier.

Algorithm 1 min-vol KL-NMF

Require: matrice $V \in \mathbb{R}^{M\times T}$, une initialisation pour $H \in \mathbb{R}_+^{K\times T}$, une initialisation pour $W \in \mathbb{R}^{M\times K}$, rang de factorisation K, un nombre maximum d'itérations maxiter, le poids de la pénalité $\lambda > 0$ et $\delta > 0$

Ensure: une factorisation NMF (W, H) de rang K de $V \approx WH$ avec $W \geq 0$ et $H \geq 0$.

$\gamma = 1, Y = \left(W^T W + \delta I\right)^{-1}$,
for k = 1 : maxiter **do**
  % Mise à jour de H
  $H \leftarrow H \odot \frac{\left[W^T\left(\frac{[V]}{[WH]}\right)\right]}{[W^T J_{F,N}]}$
  $\Phi \leftarrow J_{F,N} H^T - 4\lambda\left(WY^-\right)$
  $\Theta \leftarrow 4\lambda W\left(Y^+ + Y^-\right)$
  % Mise à jour de W
  $W^+ \leftarrow W \odot \frac{\left[\left[[\Phi]^{\cdot 2} + 2\Theta\odot\left(\frac{[V]}{[WH]} H^T\right)\right]^{\cdot\frac{1}{2}} - \Phi\right]}{[\Theta]}$
  $W_\gamma^+ = \text{normaliser}\left(W^+\right)$
  % Recherche en ligne
  while $F\left(W_\gamma^+, H\right) > F\left(W, H\right)$ **do**
    $\gamma \leftarrow \gamma \times 0.8$
    $W_\gamma^+ \leftarrow \text{normaliser}\left(\left(1-\gamma\right)W + \gamma W^+\right)$
  end while
  $W \leftarrow W_\gamma^+$
  % Mise à jour de Y
  $Y \leftarrow \left(W^T W + \delta I\right)^{-1}$
  % Mise à jour de γ
  $\gamma \leftarrow \min\left(1, \gamma \times 1.2\right)$
end for

4 Résultats numériques

Dans cette section nous présentons les résultats obtenus avec l'algorithme min-vol KL-NMF appliqué à un morceau de piano comprenant les 30 premières secondes de "Prelude et Fugue no.1 en do majeur" de Jean-Sebastien Bach interprété par Glenn Gould [1]. Ce morceau de piano est composé des treize notes suivantes : si_3, do_4, $ré_4$, mi_4, $fa_4^\#$, sol_4, la_4, do_5, $ré_5$, mi_5, fa_5,sol_5, la_5. Le morceau de piano a été enregistré avec une fréquence d'échantillonnage $f_s = 11025$Hz (fréquence maximum exploitable = 5513 Hz) produisant un nombre d'échantillons temporels $T = 330750$. La TFCT (Transformée de Fourier à court terme) du signal audio est tout d'abord déterminée en utilisant des fenêtres de Hamming d'une longueur $F = 1024$; la résolution temporelle est donc de 46 ms et la résolution fréquentielle est de 10.76 Hz. Un recouvrement typique de 50% entre deux fenêtres successives a été considéré conduisant à la génération de 647 fenêtres (=N).

La Figure 1 présente la partition du morceau, le signal audio dans le domaine temporel et sa représentation temps-fréquence sous la forme du spectrogramme d'amplitude.

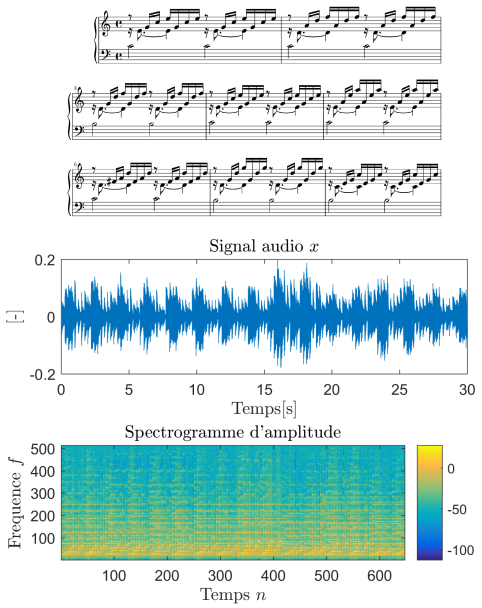

FIGURE 1 – Trois représentations des données : (Au-dessus) la partition. (Au milieu) Le signal enregistré échantillonné dans le domaine temporel. (En bas) Le spectrogramme d'amplitude V exprimé en dB.

La Figure 2 présente les résultats obtenus pour W et H avec un rang de factorisation $K = 16$, donc surestimé par rapport au nombre de notes (13). Les résultats présentés ont été obtenus avec une initialisation aléatoire pour les matrices W et H et un nombre maximum d'itérations fixé à 300. Les meilleurs résultats sur 5 cinq analyses ont été retenus. On observe que trois composantes sont mises à zéro (voir symbole *) tandis que le modèle est capable d'identifier 13 notes. Après analyses des fréquences fondamentales des 13 sources estimées, celles-ci correspondent aux fondamentales des 13 notes citées précédemment. Notez que peu d'harmoniques sont visibles dans la Figure 2, ceci est dû au mode d'affichage condensé des résultats (sur une même figure), en générant des graphiques individuels pour chaque colonne de W avec une échelle logarithmique, on observe un nombre plus important d'harmoniques comme habituellement observé pour des signatures spectrales de sources sonores. Notez également qu'en utilisant la β-NMF standard ou la β-NMF avec une contrainte de parcimonie pour analyser ce même extrait musical dans la même configuration de test,

1. https ://www.youtube.com/watch ?v=ZlbK5r5mBH4

ces deux modèles génèrent autant de composants que la valeur du rang de factorisation, subdivisant ainsi une ou plusieurs sources alors que l'algorithme min-vol KL-NMF préserve l'intégrité des 13 sources présentes dans le signal audio. Des simulations supplémentaires incluant des comparaisons de résultats obtenus entre le modèle présenté dans ce papier et des modèles sans pénalité ou avec pénalités classiques telles que la parcimonie seront intégrées dans un article à paraître prochainement.

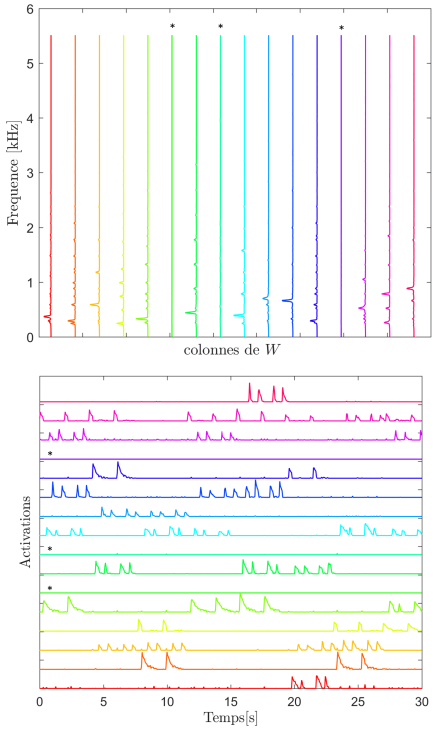

FIGURE 2 – Résultats pour W et H.

En ce qui concerne la séquence des sources estimées, la Figure 3 montre (sur un intervalle de temps limité à la première mesure) qu'elle suit la séquence théorique de la partition, notez que pour plus de clarté un seuillage a été appliqué aux lignes de H (activations) de même qu'une permutation.

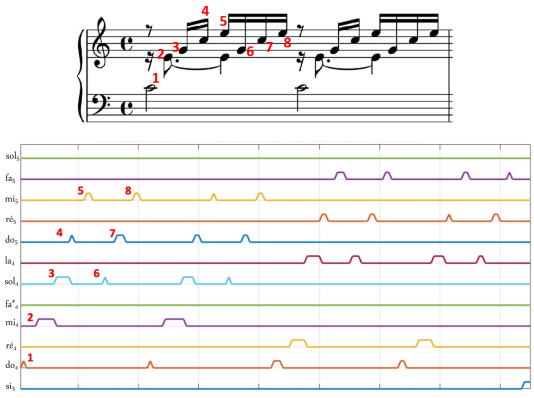

FIGURE 3 – Validation de la séquence des sources estimées.

5 Conclusion et perspectives

Dans ce papier nous avons présenté un nouveau modèle de séparation aveugle de sources sonores monophoniques basé sur la minimisation d'une fonction objectif intégrant une mesure d'écart de la famille des divergences beta et un terme de pénalité favorisant des solutions pour W de volume minimum. On a proposé un algorithme simple pour résoudre ce problème et nous avons illustré le comportement de cette méthode sur des données réelles. On a spécialement mis l'emphase sur la capacité qu'a ce modèle à faire tendre vers zéro certains composants de la factorisation lorsque le rang est mal choisi et surestimé en regard du nombre de sources présentes dans le signal. Ce travail est préliminaire et d'importantes questions restent encore ouvertes : peut-on prouver la robustesse de ce modèle au bruit ? Peut-on concevoir des algorithmes plus rapides ?

Références

[1] X. Fu, K. Huang, N.D. Sidiropoulos et W-K. Ma. *Nonnegative matrix factorization for signal and data analytics : Identifiability, algorithms and applications*. IEEE Signal Processing Magazine, 2018.

[2] C-H. Lin, W-K. Ma, W-C. Li, C-Y. Chi et 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,vol. 53,no. 10,pp. 5530-5546 2015.

[3] X. Fu, W-K. Ma, K. Huang et N.D. Sidiropoulos. *Blind separation of quasi-stationary sources : Exploiting convex geometry in covariance domain*. IEEE Transactions Signal Processing, vol. 63, no. 9,pp. 2306-2320, 2015.

[4] X. Fu, K. Huang et N.D. Sidiropoulos. *On identifiability of nonnegative matrix factorization*. IEEE Signal Processing Letters, vol. 25, no. 3,pp. 328-332, 2018.

[5] S. Vavasis. *On the complexity of nonnegative matrix factorization*. SIAM Journal on Optimization, vol. 20, no. 3,pp. 1364-1377, 2010.

[6] X. Fu, K. Huang, B. Yang, W-K. Ma et N.D. Sidiropoulos. *Robust Volume Minimization-Based Matrix Factorization for Remote Sensing and Document Clustering*. IEEE Transactions Signal Processing, vol. 64, pp. 6254 - 6268, 2016.

[7] C. Févotte et J. Idier. *Algorithms for nonnegative factorization with the beta-divergence*. Neural Computation, 2011.

[8] V. Leplat, N. Gillis et A.M.S Ang. *Minimum-volume rank-deficient nonnegative matrix factorizations*. IEEE-ICASSP, 2019.

[9] D.D. Lee et H.S. Seung, *Algorithms for non-negative matrix factorization*. In Advances in neural information processing systems, pp. 556-562, 2001.