Factorisation de matrices structurée en groupes avec la divergence d'Itakura-Saito.

Augustin LEFÈVRE[†‡], Francis BACH[†]Cédric FÉVOTTE[‡]

[†]INRIA / ENS - projet Sierra
23, avenue d'Italie, 23, avenue d'Italie, CS 81321, 75214 Paris Cedex 13, France

[‡]CNRS LTCI / Telecom ParisTech
37-39, rue Dareau, 75014 Paris
augustin.lefevre@ens.fr

Résumé – Les techniques de factorisation en matrices positives sont très populaires en séparation de sources aveugle : les différentes composantes de la factorisation sont assimilées à des sons récurrents dans un enregistrement audio. Reste ensuite à regrouper ces composantes élémentaires pour former une estimation des différentes sources sonores présentes dans l'enregistrement. Ce regroupement est d'ordinaire laissé à la charge de l'utilisateur, mais cette tâche, aisée pour des enregistrements d'une dizaine de secondes où l'on estime cinq à dix composantes, devient difficile et longue lorsqu'on passe à des enregistrements de plusieurs minutes et plus de vingt composantes. Nous proposons ici d'apprendre une factorisation de matrices structurée par groupes, qui permet ainsi d'effectuer le regroupement en même temps que l'estimation des composantes élémentaires. Nous appliquons notre algorithme à la séparation de sources non-supervisée d'enregistrements mono-canal et la segmentation d'une minute d'entretien radiophonique.

Abstract – Nonnegative matrix factorization has gained popularity in the blind signal separation community : for simple signals, individual components of NMF were found to retrieve meaningful signals such as notes or events. However, when applied to more complex signals, such as music instruments, it is more reasonable to suppose that each sound source corresponds to a subset of components. Grouping is usually done either by the user, or based on heuristics, but as the number of components grows large, this task becomes even more time-consuming than the parameter inference task. In this paper, we argue that grouping may be incorporated in the inference of the the matrix factorization as part of a structured statistical model. We apply our algorithm to source separation of single-channel recordings and segmentation of a 1-minute long excerpt from a radio interview.

1 Introduction

Nous considérons dans cet article le problème de la séparation non-supervisée de sources sonores mono-canal. Pour cette tâche, la factorisation en matrices positives est utile pour « fragmenter » un mélange en signaux élémentaires (les composantes de la NMF). Reste ensuite à assigner à chaque source un groupe de composantes. Cette tâche est complexe, puisqu'elle implique de considérer l'ensemble des permutations de K éléments.

Pour remédier à ce problème, nous proposons d'incorporer un critère de regroupement à l'étape de la factorisation, en imposant un modèle statistique structuré au mélange observé, sous la forme d'un problème de maximum de vraisemblance pénalisée.

La parcimonie structurée est actuellement un sujet de recherche actif. Dans le cas de la norme Euclidienne, il a été montré que l'emploi de normes mixtes permet d'identifier des groupes [1] et même des structures hiérarchiques plus complexes [4]. Avec la divergence d'Itakura-Saito, nous avons proposé[5] un terme de pénalité dont la structure est très similaire à une norme mixte. Nous montrons expérimentalement qu'il favorise la parcimonie par groupes.

Nous présentons ici notre modèle et le problème d'inférence par maximum de vraisemblance pénalisée qui lui est associé. Nous reprenons un algorithme implémenté précédemment [5], et proposons une extension de notre pénalité pour des structures temporelles plus fines. Notre principale contribution dans le présent article est une analyse détaillée de l'effet de notre algorithme pour la séparation de sources dans des mélanges sonores musicaux, notamment des pistes sonores issues de la campagne de séparation de sources SiSEC 2010, volet « Professional Music Recordings ». Enfin nous illustrons, sur un extrait d'une minute d'entretien radiophonique, l'intérêt de l'extension de notre pénalité à une structure par blocs temporels.

Notations. Les matrices sont en majuscules et en gras (par exemple $\mathbf{X} \in \mathbb{R}^{F \times N}$), les vecteurs en minuscules et en gras (par exemple, $\mathbf{x} \in \mathbb{R}^{F}$), les scalaires en minuscules (par exemple, $x \in \mathbb{R}$). Pour tout vecteur et toute matrice $\mathbf{X}$, $\mathbf{X} \geq 0$ signifie que toute ses entrées sont positives. Enfin, nous écrirons pour alléger les notations $\hat{\mathbf{V}} = \mathbf{WH}$. Sauf indication contraire les sommes sont pour $k = 1, \ldots, K$, $f = 1, \ldots, F$, $n = 1, \ldots, N$.

2 Modèle statistique structuré et estimation

2.1 Modèle génératif

Étant donnée la transformée de Fourier à fenêtre glissante $\mathbf{X} \in \mathbb{C}^{F \times N}$ d'un signal sonore, nous supposons que $\mathbf{X}$ est un mélange linéaire additif instantané de variables gaussiennes complexes i.i.d. :

$$x_{fn} = \sum_k x_{fn}^{(k)} \quad \text{où} \quad x_{fn}^{(k)} \sim \mathcal{N}_c(0, w_{fk} h_{kn}) \,. \tag{1}$$

Cela revient à supposer que le signal observé est stationnaire par morceaux, et que sa densité spectrale est générée par une somme de Gaussiennes avec un structure bien spécifique sur les paramètres de variance. Ainsi, nous avons $\mathbb{E}(\mathbf{V}) = \mathbf{WH}$ où $\mathbf{V} = |\mathbf{X}|^2$ est la densité spectrale observée. L'estimation par maximum de vraisemblance de $(\mathbf{W}, \mathbf{H})$ équivaut à minimiser la divergence d'Itakura-Saito entre $\mathbf{V}$ et $\hat{\mathbf{V}}$. La divergence d'Itakura-Saito est définie pour les réels strictement positifs par : $d_{IS}(x, y) = \frac{x}{y} - \log \frac{x}{y} - 1$, et pour les vecteurs et les matrices en sommant sur les composantes : $D_{IS}(\mathbf{X}, \mathbf{Y}) = \sum_{f,n} d_{IS}(x_{fn}, y_{fn})$.

Si nous assignons à une source le sous-ensemble de composantes g, alors le filtre de Wiener est un estimateur sans biais du signal émis par la source g, c'est-à-dire $\mathbf{X}^{(g)} = \sum_{k \in g} \mathbf{X}^{(k)}$ (on confondra dans la suite une source et son groupe de composantes) :

$$\mathbb{E}\left(x_{fn}^{(g)} | x_{fn}, \mathbf{W}, \mathbf{H}\right) = \frac{\sum_{k \in g} w_{fk} h_{kn}}{\sum_{g, k \in g} w_{fk} h_{kn}} x_{fn} \,. \tag{2}$$

2.2 Maximum de Vraisemblance et pénalité induisant la parcimonie

Nous cherchons une partition $\mathcal{G}$ des K composantes en un nombre G de groupes disjoints. On confondra dans la suite une source et son groupe de composantes g. Nous partons du principe que les composantes d'un même groupe sont inactives en même temps i.e., nous cherchons à imposer de la parcimonie par groupe dans les colonnes de $\mathbf{H}$. Partant d'un modèle probabiliste simple sur $\mathbf{H}$ expliqué plus en détail dans [5], nous obtenons un terme de pénalité $\Psi(\mathbf{H})$. L'inférence de $(\mathbf{W}, \mathbf{H})$ par maximum de vraisemblance pénalisée nous conduit à résoudre :

$$\begin{array}{ll} \min & D_{IS}(\mathbf{V}, \mathbf{WH}) + \lambda \Psi(\mathbf{H}), \\ \mathbf{W} \geq 0, \mathbf{H} \geq 0 & \\ \forall k,\ \|\mathbf{w}_{\cdot k}\|_1 = 1 & \end{array} \tag{3}$$

où $\Psi(\mathbf{H}) = \sum_{g,n} \psi(\|\mathbf{h}_{gn}\|_1)$, et $\psi(x) = \log(a + x)$. Nous appelons le problème (3) GIS-NMF (group Itakura-Saito NMF), et désignons par $\mathcal{L}(\mathbf{W}, \mathbf{H})$ sa fonction objectif. L'équation (3) généralise IS-NMF car si $\lambda = 0$ on retrouve le problème IS-NMF standard. Un compromis est recherché entre l'attache aux données et un critère de regroupement déterminé par Ψ. Bien que nous imposions un choix particulier pour Ψ, notons que du point de vue de l'optimisation nous supposons seulement que ψ soit différentiable, concave, et croissante.

Généralisation à des pénalités par blocs temporels On peut encore renforcer notre a priori en favorisant des zéros consécutifs dans les colonnes de $\mathbf{H}$, en considérant un terme de pénalité de la forme :

$$\Psi(\mathbf{H}) = \sum_{n=0}^{N-1} \sum_{g \in \mathcal{G}} \psi(\|\mathbf{h}_{g\,n} + \mathbf{h}_{g\,n+1}\|_1) \,. \tag{4}$$

Les ingrédients pour implémenter cette pénalité sont les mêmes qu'avec la précédente. Ainsi, au prix d'une modification mineure de notre algorithme, il est possible d'imposer une forme de parcimonie par blocs aux coefficients d'activation de $\mathbf{H}$. C'est une forme de lissage légèrement différente de celle proposé, par exemple, dans [2, 7]. En séparation de sources sonores, imposer la parcimonie par blocs correspond à l'intuition qu'une source est en général muette pendant plusieurs fenêtres de temps d'affilée (au moins une demi-seconde en musique, et plusieurs secondes pour la parole, sachant qu'une fenêtre de temps équivaut à une vingtaine de milli-secondes).

2.3 Inférence et sélection de paramètres

Notre algorithme [5] généralise les mises à jour multiplicatives pour la NMF classique avec la divergence d'Itakura-Saito. Sa complexité en mémoire et en temps est $O(FKN)$. Il respecte une propriété de descente i.e., toute suite d'estimateurs $(\mathbf{W}^{(t+1)}, \mathbf{H}^{(t+1)})_{t \geq 0}$ obtenue par notre algorithme vérifie $\mathcal{L}(\mathbf{W}^{(t+1)}, \mathbf{H}^{(t+1)}) \leq \mathcal{L}(\mathbf{W}^{(t)}, \mathbf{H}^{(t)})$. Par conséquent, la fonction objectif converge, mais pas nécessairement $(\mathbf{W}, \mathbf{H})$. Initialiser l'algorithme plusieurs fois et garder le résultat ayant la plus faible fonction objectif permet déviter les problèmes de minima locaux. On observe dans nos simulatins que $(\mathbf{W}, \mathbf{H})$ convergent [3]. Dans nos expériences nous arrêtons l'algorithme au bout d'un nombre fixe d'itérations.

En pratique, le choix de l'hyperparamètre λ est crucial (celui de a l'est moins, en revanche). Nous avons aussi proposé dans [5] une statistique de test de type Kolmogrov-Smirnov pour sélectionner automatiquement λ parmi plusieurs valeurs à partir des estimateurs de $\hat{\mathbf{V}}$ obtenus pour chaque valeur (cf. Figure 1 dans [5]).

3 Application à la séparation de sources sonores

Nous avons montré, sur des données sythétiques, que la pénalité proposée produit bien l'effet souhaité, à savoir que si les niveaux d'activation des sources (mesurés par $\|h_{gn}\|_1$) se chevauchent peu, alors notre pénalité permet de les identifier correctement[5]. Nous appliquons dans cette partie notre algorithme à des signaux réels, d'une part pour vérifier si cette propriété se vérifie encore, d'autre part pour examiner en quoi cela contribue à améliorer la qualité de la séparation de sources.

	o=0		o=0.5		o=1	
	sNMF	heur.	sNMF	heur.	sNMF	heur.
piste 1						
guitare	**-16.74**	-22.61	**-4.33**	-18.65	-15.03	**-13.71**
basse	**3.81**	-48.51	**6.90**	-2.56	**4.17**	-7.44
piano	**-7.05**	-25.80	-15.13	**-13.59**	**-5.11**	-8.49
piste 2						
guitare	**-0.80**	-45.37	**1.29**	-16.50	-3.39	**-0.22**
voix	**2.39**	-3.03	**0.50**	0.14	**0.59**	-5.36
piste 3						
guitare	**-4.12**	-49.50	**-3.70**	-14.10	**0.34**	-0.56
voix	**-2.24**	-9.36	**0.13**	-5.01	**-2.24**	-8.03

TAB. 1: Ratios source à distortion (SDR) pour des pistes d'une dizaine de secondes (source : SiSEC 2010).

	o=0		o=0.66		o=1	
	sNMF	heur.	sNMF	heur.	sNMF	heur.
piste 4						
guitare	**8.88**	-67.53	**1.47**	- 5.29	- 5.13	**- 4.16**
basse	**13.60**	3.77	**7.72**	- 8.11	**- 0.21**	- 2.68
piste 5						
voix	**3.74**	3.33	2.37	**2.88**	**2.30**	- 7.53
guitare	**0.31**	- 0.88	- 4.03	**-3.41**	- 5.37	**- 2.43**

TAB. 2: Ratios source à distortion (SDR) pour des pistes d'une trentaine de secondes (source Internet Archive).

Protocole. Nous expérimentons notre algorithme sur deux types de pistes sonores : d'une part, les pistes de la campagne de séparation de sources SiSEC 2010 (volet « Professional Music Recordings »), d'une durée d'environ dix secondes chacune. D'autre part, des pistes sonores trouvées sur Internet Archive[1], dont nous avons pris des extraits de 20 à 30 secondes.

Pour la plupart des pistes nous considérons des mélanges de deux sources. Pour la piste 1 de la campagne SiSEC 2010 nous en avons mélangé trois. Pour chaque piste, nous disposons des sources et produisons un mélange linéaire instantané tel que les niveaux sonores de chaque source soient similaires.

Par ailleurs, afin d'évaluer la pertinence de notre pénalité, nous considérons plusieurs degrés de chevauchement entre les sources. Dans la plupart des pistes présentées, les sources se chevauchent totalement, nous avons donc artificiellement mis à zéros des trames de façon à obtenir le degré de chevauchement désiré. Le but est d'examiner jusqu'à quel degré de chevauchement p le critère de parcimonie par groupes permet d'identifier les sources correctement.

Nous prenons le ratio source à distortion (SDR) comme critère de performance (cf. [6]). Les résultats complets pour tous les mélanges, fichiers audio inclus, sont disponibles en ligne [2]. Nous comparons dans les Tables 1 et 2, notre algorithme (sNMF) avec une méthode heuristique simple : estimer $(\mathbf{W}, \mathbf{H})$ par NMF, puis chercher la permutation qui minimise $\Psi(\mathbf{H})$. Le lecteur intéressé pourra trouver en ligne une comparaison avec un masque binaire aléatoire, et une performance oracle (meilleure permutation possible).

Paramètres. Les pistes sont toutes échantillonneées à 22 kHz. Pour les transformées de Fourier, nous utilisons des fenêtres de 512 échantillons (soit environ 20 millisecondes et 256 échantillons d'espacement. Pour la NMF, nous fixons $a = 10^{-2}$ et choisissons, le même nombre de composantes par groupes K. Les paramètres K et λ sont ensuite choisis à l'aide de notre statistique de test. Pour le méthode heuristique, nous prenons 5 composantes par source lorsqu'il y a deux sources, et 2 lorsqu'il y en a 3, au-delà parcourir toutes les permutations possibles prend trop de temps.

Analyse des résultats. Pour les pistes sonores de la campagne SiSEC 2010, on constate que dans tous les cas, notre méthode produit de meilleures estimations que la méthode heuristique de référence.

Première observation, dans le cas de la piste 5, où l'on sépare voix et basse, lorsque les supports sont disjoints ($o = 0$), nous obtenons une bonne séparation (ce qui se confirme à l'écoute). Constatons également que lorsque les supports fréquentiels des sources sont moins clairement distincts, les résultats se détériorent (cf. piste 5).

Par ailleurs, notons que les résultats de séparation s'améliorent nettement quand les enregistrements fournis sont plus longs (Table 2), pour la méthode heuristique comme pour la nôtre.

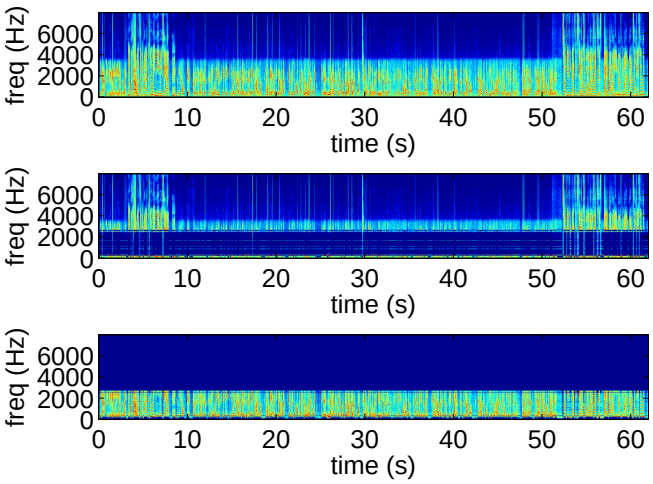
Notons également que la qualité de la séparation s'améliore losrque les sources se chevauchent partiellement : en effet dans ce cas les données sont plus diversifiées ce qui permet d'apprendre $\mathbf{W}$ plus finement.

Au vu de l'ensemble des résultats, nous concluons que la pénalité que nous proposons ne suffit pas toujours à identifier correctement les supports de chaque source, même dans le cas où ils sont complètement disjoints. Pour obtenir une séparation satisfaisante, il faut soit faire des hypothèses supplémentaires sur $\mathbf{W}$, soit réfléchir à des structures temporelles plus fines. C'est l'objet d'une seconde expérience que nous présentons maintenant.

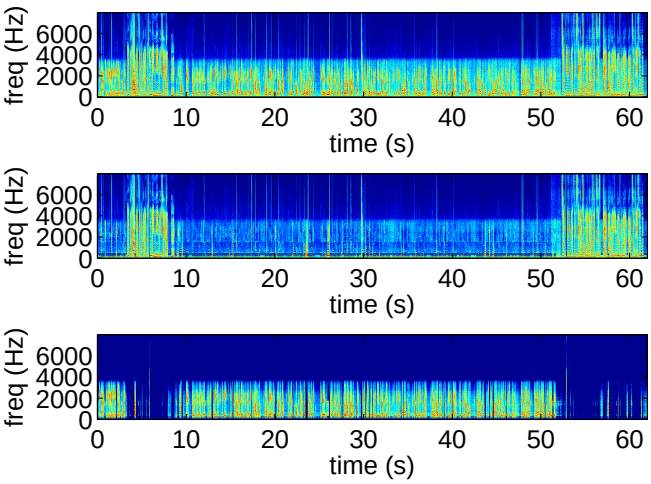
Pénalité par blocs temporels. Nous présentons Figure 1 le résultat de notre algorithme avec la pénalité par blocs présentée en Section 2.2 : sur un extrait d'une minute d'interview radiophonique, nous parvenons ainsi à segmenter la voix d'un homme et celle d'une femme. Nous avons choisi cet extrait pour deux raisons : d'une part, il n'y a aucun chevauchement entre les deux sources, ce qui illustre le bien-fondé de notre pénalité pour des mélanges sonores réels. D'autre part les supports fréquentiels des deux sources se chevauchent nettement, mais il y a aussi toute une bande de fréquence pour laquelle seule la voix de femme est présente, ce qui laisse à penser qu'il est possible d'obtenir une séparation sans faire d'hypothèses sur $\mathbf{W}$.

[1] www.archive.org

[2] www.di.ens.fr/~lefevrea/demos.html



(a) sans pénalité



(b) pénalité par blocs temporels

FIG. 1: Application de la pénalité par blocs à une tâche de segmentation. (Haut) mélange (Milieu) estimation de la voix de femme (Bas) estimation de la voix d'homme.

Pour cette expérience, vue la durée de l'enregistrement nous n'avons pas pu valider le choix des paramètres sur une grille suffisamment grande. À partir de plusieurs valeurs essayées nous avons retenu $K = 15$ et $\lambda = 1$. Imposer des zéros par blocs contigüs permet d'obtenir des supports quasi-disjoints pour chaque source, ce qui n'est pas le cas pour une NMF standard (cf. Figure 1).

Ne disposant pas des sources originales, nous ne pouvons pas mesurer la qualité de la séparation obtenue, les sources estimées seront disponibles en ligne.

4 Conclusion

Nous avons présenté une procédure d'inférence pour regrouper les composantes en NMF avec la divergence d'Itakura-Saito. Au lieu de regrouper les composantes après l'estimation de la NMF, nous incorporons le critère de regroupement dans l'optimisation. Appliqué à des mélanges sonores réels, notre algorithme pemet d'identifier les supports de chaque source à condition que les enregistrements fournis soient suffisamment longs. De plus, une simple extension de notre pénalité par blocs temporels permet de trouver des sources qui s'interrompent pendant des intervalles de temps entier, ce qui peut être intéressant dans des enregistrements sonores longs (un album entier, une émission radio, la bande son d'un film, etc.). Ceci fera l'objet de travaux à venir.

Références

[1] S. Bengio, F. Pereira, Y. Singer, and D. Strelow. Group sparse coding. In *Adv. NIPS*, 2010.

[2] C. Févotte. Majorization-minimization algorithm for smooth itakura-saito nonnegative matrix factorizations. In *In Proc. Int. Conf. on Acous., Speech, and Sig. Proc. (ICASSP)*, Prague, Czech Republic, May 2011.

[3] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence : With application to music analysis. *Neural Comput.*, 21(3) :793–830, 2009.

[4] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Int. Conf. on Mach. Learn. (ICML)*, 2010.

[5] A. Lefèvre, F. Bach, and C. Févotte. Itakura-Saito nonnegative matrix factorization with group sparsity. *In Proc. Int. Conf. on Acous., Speech, and Sig. Proc. (ICASSP)*, 2011.

[6] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Trans. on Audio, Speech and Lang. Proc.*, 14(4), 2006.

[7] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans. on Audio, Speech, and Lang. Proc.*, 15(3), March 2007.