Une Approche Alternée pour la Factorisation avec Dictionnaire de Matrices Non Négatives.

Jérémy E. COHEN, Nicolas GILLIS

Département de Mathématiques et Recherche opérationelle
Rue de Houdain, 9, 7000, Mons, Belgique
jeremy.cohen@umons.ac.be, nicolas.gillis@umons.ac.be

Résumé – Nous proposons un nouveau modèle pour la factorisation de matrices non négatives mettant en jeu un dictionnaire. Ce modèle induit un algorithme glouton alterné, comparé à l'état de l'art pour le démélange spectral d'images hyperspectrales satisfaisant l'hypothèse de séparabilité.

Abstract – In this paper, we propose a new model along with an algorithm for dictionary-based nonnegative matrix factorization. We show its effectiveness on spectral unmixing of hyperspectral images using self dictionary compared to state-of-the-art methods.

1 Introduction

Les problèmes d'approximation de rang faible de matrices ont récemment gagné en importance, et on les retrouve dans une grande variété d'applications ; voir par exemple [1] pour quelques références. De façon générale, on peut formuler ce problème de la façon suivante : étant donné une matrice de données $X \in \mathbb{R}^{p\times n}$, dont chaque colonne $X(:,j)$ est un vecteur de données dans un espace de dimension p, et étant donné un rang r, on souhaite obtenir une base représentée par la matrice $U \in \mathbb{R}^{p\times r}$ et des coefficients $V \in \mathbb{R}^{r\times n}$ tels que

$$X \approx UV \iff X(:,j) \approx \sum_{k=1}^{R} U(:,k)V(k,j)\ \forall j.$$

Il existe de nombreuses variantes à ce problème, en fonction de la mesure de l'erreur et des contraintes imposées aux facteurs (U,V) et/ou à l'approximation de rang faible UV.

Cette communication s'intéresse en particulier au cas de la factorisation de matrices non négatives, pour laquelle les facteurs U et V sont contraints, élément par élément, à être non négatifs [2]. De plus, seul le cas particulier où les colonnes du facteur U appartiennent à un dictionnaire $D \in \mathbb{R}^{p\times d}$ est étudié, où $d \gg r$ est le nombre d'atomes. Formellement, cela signifie que $U = D(:,\mathcal{K})$ pour un ensemble d'indices $\mathcal{K} \subset \{1,2,\ldots,d\}$ avec $|\mathcal{K}| = r$.

Ainsi en quantifiant la norme de l'erreur d'approximation par la norme de Frobenius, on peut formuler la NMF avec dictionnaire de la façon suivante :

$$\begin{array}{ll} \min_{\mathcal{K},V\geq 0} & ||X - D(:,\mathcal{K})V||_F^2 \\ \text{tel que} & \mathcal{K} \subset \{1,2,\ldots,d\} \text{ et } |\mathcal{K}| = r. \end{array} \quad (1)$$

Le modèle (1) est particulièrement utile pour le démélange spectral d'images hyperspectrales (HSI). Ces images hyperspectrales sont des images pour lesquelles de la réflectance (fraction de la lumière réfléchie) des matériaux présents sur la scène est mesurée pour de nombreux canaux fréquentiels (environ 100 à 200 pour chaque pixel). En d'autres termes, chaque pixel est en fait un vecteur contenant des valeurs de réflectances à différentes longueur d'ondes, le spectre obtenu étant souvent nommé signature spectrale. Le modèle de mélange linéaire fait l'hypothèse que chaque pixel est une combinaison linéaire des signatures spectrales des matériaux constituant la scène, appelés endmembers, pour expliquer les données par un modèle de rang faible. C'est un modèle valide pour des mélanges macroscopiques des matériaux sur la scène et avec peu de reliefs.

Pour des images hyperspectrales bien résolues spatialement, il est possible qu'au moins un pixel ne contiennent qu'un seul matériaux, et ce pour chaque matériaux présent sur la scène. Sous cette hypothèse de séparabilité ou bien de pixels purs, le démélange spectral par un modèle de rang faible se ramène à résoudre (1) en utilisant un dictionnaire propre $D = X$. Un dictionnaire est également disponible lorsque les endmembers sont contenus dans une librairie spectrale. Le lecteur intéressé est invité à consulter [3, 4] pour davantage de détails.

1.1 NMF avec dictionnaire propre

Dans cet article court, seul le cas d'un dictionnaire propre $D = X$ est étudié. Il y a eu un regain récent des recherches sur les algorithmes d'identification de pures pixels, développé principalement dans les années 90' et 2000, en raison de nouveaux résultats théoriques du comportement de ces algorithmes en présences de bruit (preuve de robustesse). Il existe, à notre connaissance, principalement deux types d'approches pour résoudre (1).

Les approches géométriques Elles consistent à sélectionner les atomes par un critère géométrique, par exemple le volume de l'enveloppe convexe de $X(:,\mathcal{K})$. Cette famille d'algorithmes regroupe notamment l'analyse des sommets composants (VCA) [5] et l'algorithme de projections successives (SPA) [6, 7, 8, 9]. Ces algorithmes sont généralement rapides avec une complexité en $\mathcal{O}(pnr)$. Les atomes sélectionnés ne conduisent parfois pas cependant à une faible valeur de l'erreur de recons-

truction $||X - X(:,\mathcal{K})V||_F$ car ce coût n'est pas directement minimisé. De façon notable, ces méthodes sont sensibles aux données aberrantes.

Les approches de régressions parcimonieuses Ces méthodes sont basées sur la reformulation suivante de (1) imposant la parcimonie par ligne aux scores Y :

$$\begin{array}{ll}\min_{Y\in\mathbb{R}^{d\times n}} & ||X - XY||_F^2 \\ \text{tel que} & Y \text{ a } r \text{ lignes non nulles.}\end{array}$$

La parcimonie par ligne de Y peut être imposée de différentes manières, en particulier en convexifiant la norme $\ell_{0,\infty}$ avec, par exemple, la norme $\ell_{1,2}$ [10], la norme $\ell_{1,\infty}$ [11], ou en utilisant la programmation linéaire [12, 13] ou des algorithmes proximaux [14].

Ces méthodes offrent l'avantage de mieux modéliser le modèle (1) puisqu'elles prennent en compte de façon explicite le terme d'attache aux données. Elles offrent en général de bonnes solutions mais au prix d'un temps de calcul important, un problème d'optimisation à dn variables devant être résolu. En particulier lorsque $D = X$, le nombre de variable est n^2. En démélange spectral, n peut atteindre plusieurs millions de pixels et ces approches deviennent inextricables. Une solution est de choisir en premier lieu les pixels contenant l'information recherchée avec une approche géométrique de clustering [15], ou de sélectionner aléatoirement un sous-ensemble de pixels. De plus, le problème résolu n'est qu'un problème approché, et les résultats pourraient ne pas être aussi proches que voulu de la solution du problème non convexe.

1.2 Contribution et structure du papier

Nous proposons un nouvel algorithme pour résoudre (1). Il tire avantage des deux approches décrites plus haut : il est rapide, nécessitant $\mathcal{O}(pnr)$ opérations, mais prend en compte le terme d'attache aux données $||X - X(:,\mathcal{K})V||_F^2$ de façon explicite. La complexité est donc du même ordre de grandeur que celle de méthodes géométriques comme VCA ou SPA, mais avec une constante plus grande du fait de la nature itérative de l'algorithme. La section 2 contient une description du nouvel algorithme, tandis que la section 3 montre qu'il est favorablement compétitif avec l'état de l'art pour le démélange spectral avec dictionnaire propre.

2 Algorithme alterné proposé

Il est difficile en pratique de résoudre le problème combinatoire (1) directement. Nous proposons une stratégie alternée :

— Estimation de V. Pour un $\mathcal{K}$ fixé, V est la solution d'un problème convexe, obtenue de façon efficace par l'estimateur au sens des moindres carrées non négatifs. Notez l'utilisation de $D(:,\mathcal{K})$, et non de U, pour estimer V. En utilisant U, nous avons observé que l'algorithme n'est en général pas capable de modifier significativement $D(:,\mathcal{K})$ au fil des itérations, qui reste proche de sa valeur initiale. En d'autres termes, notre choix permet à l'algorithme d'explorer une plus grande partie de l'ensemble admissible et ainsi générer de meilleures solutions.

— Estimation de $\mathcal{K}$. C'est un problème difficile car combinatoire et potentiellement de grande taille. Afin de répondre à ce problème, une variable auxiliaire $U = D(:,\mathcal{K})$ peut être introduite. Le problème d'optimisation peut alors être réécrit de la façon suivante :

$$\begin{array}{ll}\min_{\mathcal{K},U\geq 0,V\geq 0} & ||X - UV||_F^2 + \delta||U - D(:,\mathcal{K})||_F^2 \\ \text{tel que} & \mathcal{K} \subset \{1,2,\ldots,d\} \text{ et } |\mathcal{K}| = r,\end{array} \quad (2)$$

pour un paramètre de régularisation $\delta > 0$. En alternant également entre U et $\mathcal{K}$, on obtient pour U un problème de moindre carrés non négatifs tandis qu'il est facile de choisir les atomes les plus proches des colonnes de U pour estimer $\mathcal{K}$. Le paramètre δ est augmenté au fur et à mesure pour imposer la convergence du modèle (2) vers le modèle (1).

Un autre bénéfice de l'introduction d'un variable auxiliaire U est de permettre une correction des atomes sélectionnés, puisque U minimise réellement le terme d'attache aux données tout en étant proche des atomes sélectionnés $D(:,\mathcal{K})$. Cette correction introduit une flexibilité dans les signatures spectrales pouvant prendre en compte la variabilité spectrale [16] de façon partielle.

Il est également important de noter que les moindres carrés non négatifs ne doivent pas nécessairement être résolus avec une grande précision du fait de la nature alternée de l'algorithme proposé. Nous avons choisi d'utiliser 10 itérations d'un algorithme de descente par blocs [17]. Enfin, aucune preuve de convergence n'est disponible puisque la fonction de coût peut augmenter lors du choix glouton des atomes du dictionnaire. L'algorithme proposé est résumé ci-dessous.

Initialisation L'algorithme 1 tente de résoudre un problème non-linéaire et combinatoire. Par conséquent, et nous le confirmons plus loin, il est très sensible au choix des paramètres initiaux. Dans ce papier, les valeurs initiales du problème sont déterminées de la façon suivante :

1. Estimer $\mathcal{K}$ avec un algorithme au choix. Dans la section dédié aux expériences numériques seront utilisés les algorithmes de l'état de l'art ainsi qu'une initialisation aléatoire.
2. Poser $U = D(:,\mathcal{K})$, résoudre pour V le problème de moindre carrés sans contraintes puis projeter sur l'orthant non négatif (en Matlab, `V=max(0,U\M)`).
3. Améliorer (U, V) avec 10 itérations d'un algorithme de NMF classique, ici A-HALS [17].
4. Initialiser

$$\delta = 0.01\,\frac{||X - UV||_F^2}{||U - D(:,\mathcal{K})||_F^2},$$

de telle sorte que le terme d'attache aux données ait initialement le plus d'importance dans la fonction de coût. De fait, lorsque δ est grand, l'algorithme permet moins facilement de changer $\mathcal{K}$.

Algorithm 1 Algorithme Alterné pour (2)

Input: $X \in \mathbb{R}^{p\times n}$, entier r, `maxiter`.
Output: $U \in \mathbb{R}_+^{p\times r}, V \in \mathbb{R}_+^{r\times n}$ et un ensemble d'indices $\mathcal{K}$ tels que $||X - D(:,\mathcal{K})V||_F$ est petit et $U \approx D(:,\mathcal{K})$.

Choisir des facteurs initiaux $U \geq 0, V \geq 0$ et un ensemble d'indices $\mathcal{K}$, et une valeur initiale pour δ.
for $k = 1 : \texttt{maxiter}$ **do**
  Estimation de $V : \min_{V\geq 0} ||X - D(:,\mathcal{K})V||_F^2$.
  Estimation de U :

$$\min_{U\geq 0} ||X - UV||_F^2 + \delta||U - D(:,\mathcal{K})||_F^2.$$

  $\mathcal{K} = \emptyset$.
  for $k = 1 : r$ **do**
    $\mathcal{K} = \mathcal{K} \cup \text{argmax}_k \frac{D(:,k)^T U(:,k)}{||D(:,k)||_2}$.
  end for
  if $||U - D(:,\mathcal{K})||_F > 0.01||U||_F$ **then**
    Augmenter δ.
  end if
  if $||U - D(:,\mathcal{K})||_F < 0.05||U||_F$ et K n'a pas changé sur 5 itérations **then**
    L'algorithme a convergé.
  end if
end for

3 Expériences numériques sur images hyperspectrales

Dans cette section, l'algorithme 1 est comparé avec trois algorithmes géométriques (VCA [5], SPA [6], SNPA [18]), un algorithme hiérarchique (H2NMF [19]), et un algorithme résolvant le problème relaxé avec pré-sélection de 100 à 500 atomes (FGNSR [15]).

L'algorithme 1 est initialisé avec chacun des algorithmes mentionnés, et les méthodes ainsi obtenues sont notées d-X, où X désigne l'algorithme ayant servi à l'initialisation. Par exemple d-VCA désigne l'algorithme 1 initialisé par l'algorithme VCA. A titre de comparaison, 10 initialisations aléatoires sont également utilisées (r pixels tirés aléatoirement). Sont présentés plus bas le pire cas, la moyenne et le meilleur cas (en terme d'erreur de reconstruction), respectivement notés RAND-wo, RAND-av et RAND-be. Pour toutes ces expériences numériques, δ est multiplié aux itérations idoines par un facteur 1.5. En effet afin de borner δ, il n'est augmenté que lorsque $||U - D(:,\mathcal{K})||_F$ est trop grand. Le temps de calcul reporté plus loin pour l'algorithme 1 ne prend pas en compte l'initialisation ni la pré-sélection pour FGNSR.

Ces différentes approches sont comparées sur un sous-ensemble des données utilisées dans [15] [1] :

— Urban HSI avec 162 bandes spectrales et 309 × 309 pixels, $r = 6$.
— San Diego airport HSI avec 158 bandes spectrales et 400 × 400 pixels, $r = 8$.

Les résultats sont rapportés dans la table 1. Pour un ensemble d'indices $\mathcal{K}$ extrait par un algorithme donné, l'erreur de reconstruction relative (Err. rel.) fournie **en pourcentage** dans la table est donnée par $\min_{V\geq 0} \frac{||X - X(:,\mathcal{K})V||_F}{||X||_F}$. L'erreur de reconstruction minimale est mise en gras. Le nombre d'itérations nécessaire à la convergence est donné pour l'algorithme 1 entre parenthèses. [2]

	Urban		San Diego	
	Temps (s.)	Err. rel.	Temps (s.)	Err. rel.
RAND-wo	0.00	7.87	0.00	10.63
d-RAND-wo	22.46 (13)	5.09	38.55 (9)	4.86
RAND-av	0.02	11.51	0.02	9.41
d-RAND-av	23.91 (13)	4.65	49.50 (14)	4.21
RAND-be	0.00	13.77	0.02	8.49
d-RAND-be	22.01 (11)	4.36	59.17 (18)	**3.57**
VCA	2.01	18.38	3.51	7.47
d-VCA	26.89 (15)	5.83	68.45 (22)	5.15
SPA	0.30	9.58	0.45	12.62
d-SPA	24.37 (13)	4.67	68.45 (22)	4.08
SNPA	24.34	9.63	64.96	12.84
d-SNPA	23.04 (13)	4.94	64.04 (18)	3.75
H2NMF	19.02	5.81	36.77	4.75
d-H2NMF	26.66 (15)	**4.05**	44.18 (10)	4.13
FGNSR-100	2.73	5.58	2.55	3.73
d-FGNSR-100	26.72 (14)	4.36	43.85 (11)	3.63
FGNSR-500	40.11	5.07	38.70	4.05
d-FGNSR-500	25.07 (13)	4.40	43.88 (11)	3.67

TABLE 1 – Résultats numériques pour les données URBAN (gauche) et San Diego Airport (droite).

La Figure 1 montre les signatures spectrales désignées comme endmembers pour l'image hyperspectrale Urban avec la méthode d-H2NMF. En utilisant les cartes d'abondances associées, il est possible d'identifier grossièrement les matériaux sur la scène.

On peut observer que

— Dans tous les cas, l'algorithme 1 permet d'améliorer la solution initiale fournie par VCA, SPA, H2NMF et FGNSR, ainsi que l'initialisation aléatoire.
— Dans tous les cas, l'algorithme proposé converge en moins de 20 itérations, grâce à l'augmentation agressive de δ.
— Même l'initialisation aléatoire de l'algorithme 1 permet d'obtenir une erreur de reconstruction très faible. De façon surprenante, le pire cas donne des résultats supérieurs à VCA, SPA et H2NMF. De plus, l'initialisation aléatoire fournit la meilleure estimation pour l'image hyperspectrale San Diego Airport. Cette observation vient nuancer la dépendance *a priori* des performances de l'algorithme proposé par rapport à l'initialisation.
— Dans tous les cas, l'algorithme 1 donne la meilleure erreur de reconstruction.

Une raison possible des bonnes performances de l'algorithme

1. Pour une comparaison complète : `https://hal.archives-ouvertes.fr/hal-01493420`

2. Cette expérience a été réalisée en utilisant Matlab sur un ordinateur portable Intel CORE i5-3210M CPU @2.5GHz 6GB RAM.

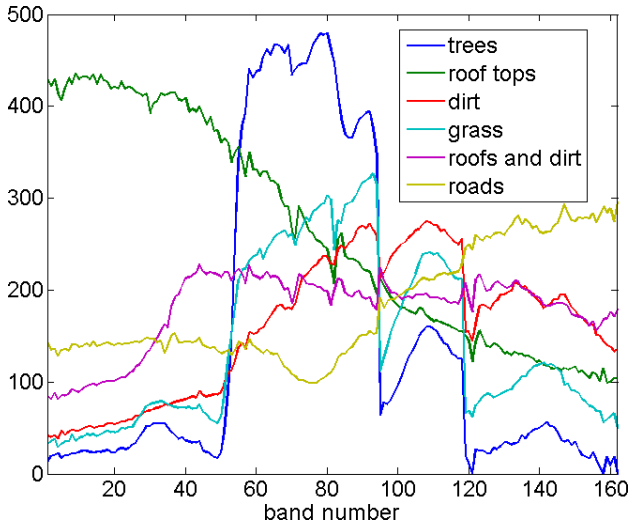

FIGURE 1 – Signatures spectrales des endmembers extraits par d-H2NMF pour Urban HSI.

proposé pour ces jeux de données est la densité du nuage des signatures spectrales, permettant une mise à jour progressive de $\mathcal{K}$ à partir de ses voisins. De futurs travaux s'attacheront à l'analyse du comportement de l'algorithme 1 dans d'autres scénarios.

4 Conclusion

Un nouvel algorithme pour calculer la factorisation non négative de matrices avec dictionnaire est proposé dans cette communication. Comme les méthodes géométriques, il est rapide, nécessitant $\mathcal{O}(mnr)$ opérations, et peut donc être utilisé sur des problèmes de grandes dimensions. Comme les méthodes de régression parcimonieuses, il prend en compte de façon explicite l'attache aux données, et permet donc de sélectionner des atomes générant une faible erreur de reconstruction. L'efficacité de cet algorithme est illustrée sur le démélange spectral de plusieurs images hyperspectrales. Dans tous les cas étudiés, il identifie les signatures spectrales avec la plus faible erreur de reconstruction. Ce papier s'est intéressé uniquement au cas de dictionnaires propres, mais la formulation proposée permet de traiter le cas général, qui sera étudié dans une version étendue de cette communication.

Acknowledgments

Les auteurs remercient le F.R.S.-FNRS (incentive grant for scientific research no F.4501.16). NG remercie également l'ERC (starting grant no 679515).

Références

[1] M. Udell, C. Horn, R. Zadeh, and S. Boyd, "Generalized low rank models," *Foundations and Trends in Machine Learning*, vol. 9, no. 1, pp. 1–118, 2016.

[2] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[3] J. 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*, vol. 5, no. 2, pp. 354–379, 2012.

[4] W. K. Ma, J. M. Bioucas-Dias, T. H. Chan, N. Gillis, P. Gader, A. J. Plaza, A. Ambikapathi, and C. Y. Chi, "A signal processing perspective on hyperspectral unmixing : Insights from remote sensing," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 67–81, 2014.

[5] J. Nascimento and J. Dias, "Vertex component analysis : a fast algorithm to unmix hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 898–910, 2005.

[6] M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvão, T. Yoneyama, H. C. Chame, and V. Visani, "The successive projections algorithm for variable selection in spectroscopic multicomponent analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 57, no. 2, pp. 65–73, 2001.

[7] H. Ren and C.-I. Chang, "Automatic spectral target recognition in hyperspectral imagery," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1232–1249, 2003.

[8] T.-H. Chan, W.-K. Ma, A. Ambikapathi, and C.-Y. Chi, "A simplex volume maximization framework for hyperspectral endmember extraction," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4177–4193, 2011.

[9] N. Gillis and S. A. Vavasis, "Fast and robust recursive algorithmsfor separable nonnegative matrix factorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 698–714, 2014.

[10] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few : Sparse modeling for finding representative objects," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. Institute of Electrical & Electronics Engineers (IEEE), 2012. [Online]. Available : http://dx.doi.org/10.1109/cvpr.2012.6247852

[11] E. Esser, M. Moller, S. Osher, G. Sapiro, and J. Xin, "A convex model for nonnegative matrix factorization and dimensionality reduction on physical space," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3239–3252, 2012.

[12] V. Bittorf, B. Recht, E. Ré, and J. Tropp, "Factoring nonnegative matrices with linear programs," in *Advances in Neural Information Processing Systems (NIPS '12)*, 2012, pp. 1223–1231.

[13] N. Gillis and R. Luce, "Robust near-separable nonnegative matrix factorization using linear optimization," *J. Mach. Learn. Res.*, vol. 15, pp. 1249–1280, 2014.

[14] R. Ammanouil, A. Ferrari, C. Richard, and D. Mary, "Glup : Yet another algorithm for blind unmixing of hyperspectral data," *Proc. IEEE WHISPERS*, pp. 1–4, 2014.

[15] N. Gillis and R. Luce, "A fast gradient method for nonnegative sparse regression with self dictionary," *arXiv :1610.01349*, 2016.

[16] A. Zare and K. Ho, "Endmember variability in hyperspectral analysis : Addressing spectral variability during spectral unmixing," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 95–104, 2014.

[17] N. Gillis and F. Glineur, "Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization," *Neural Comput.*, vol. 24, no. 4, pp. 1085–1105, 2012.

[18] N. Gillis, "Successive nonnegative projection algorithm for robust nonnegative blind source separation," *SIAM J. Imaging Sci.*, vol. 7, no. 2, pp. 1420–1450, 2014.

[19] N. Gillis, D. Kuang, and H. Park, "Hierarchical clustering of hyperspectral images using rank-two nonnegative matrix factorization," *IEEE Trans. Geosci. Remote Sensing*, vol. 53, no. 4, pp. 2066–2078, 2015.