Factorisation en Matrices Non négatives basée Dictionnaire avec l'aide du transport optimal

Rémi CORNILLET[1], Jérémy E. COHEN[1], Nicolas COURTY[2]

[1]Université Lyon, INSA-Lyon, UCBL, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, Villeurbanne, France

[2]Université Bretagne Sud, CNRS UMR 6074 IRISA, 56000 Vannes, France

remi.cornillet@creatis.insa-lyon.fr, jeremy.cohen@cnrs.fr, nicolas.courty@irisa.fr

Résumé – La Factorisation en Matrices Non négatives (NMF) est une méthode bien connue de réduction de dimensions et d'approximation de rang faible, permettant de représenter des données sous la forme d'une combinaison linéaire positive de templates interprétables. Dans un cadre semi-supervisé où ces templates sont des distributions et sont à sélectionner parmi les colonnes d'un dictionnaire connu, le transport optimal est tout indiqué pour définir une distance entre ces templates et le dictionnaire. Dans cet article, nous introduisons le modèle NMF basée Dictionnaire avec régularisation Wasserstein (NMFDW) et discutons son utilité pour le démélange d'images hyperspectrales. Un algorithme de descente par bloc de coordonnées est proposé pour estimer les paramètres du modèle, et ses performances sont évaluées sur des données semi-synthétiques de télédétection. Sur ces données, le modèle NMFDW apparaît plus robuste aux perturbations, mais ses performances sont inférieures à l'état de l'art.

Abstract – Nonnegative Matrix Factorization (NMF) is a well known method for dimensionality reduction and low-rank approximation that allows to represent data as linear nonnegative combination of interpretable templates. In a semi-supervised setting where templates are distributions and should be picked from the columns of a known dictionary matrix, a tool of choice to define adequate metrics between templates and the dictionary is optimal transport. In this article, the Wasserstein Dictionary-based NMF (WDNMF) model is introduced and its use in spectral unmixing of hyperspectral images is discussed. A bloc coordinate descent algorithm is proposed, which performance is studied on a semi-synthetic remote sensing dataset. In these experiments, WDNMF is more robust to perturbations but performs relatively poorly compared to its competitors.

1 Introduction

Pour une matrice non négative $\mathrm{Y} \in \mathbb{R}_+^{n\times m}$, la factorisation en matrices non négatives (*Nonnegative Matrix Factorization, NMF*) consiste en l'approximation $\mathrm{Y} \approx \mathrm{WH}$ telle que $\mathrm{W} \in \mathbb{R}_+^{n\times r}$ et $\mathrm{H} \in \mathbb{R}_+^{r\times m}$ avec $r < \min(n, m)$ appelé rang non négatif. En pratique, on cherche un produit WH qui soit une bonne approximation de Y, au sens où l'on résout un certain problème d'optimisation non convexe, par exemple :

$$\mathrm{W}, \mathrm{H} \in \underset{\mathrm{W}\in\mathbb{R}_+^{n\times r}, \mathrm{H}\in\mathbb{R}_+^{r\times m}}{\operatorname{argmin}} \|\mathrm{Y} - \mathrm{WH}\|_F^2 \tag{1}$$

avec ici la norme de Frobenius utilisée comme fonction de perte. Intuitivement, la NMF suppose que les données sont des combinaisons linéaires positives de templates. On va ici s'intéresser à l'un de ses usages en imagerie hyperspectrale : le démélange spectral, qui consiste en l'obtention des différents spectres caractéristiques des matériaux composant les pixels d'une image hyperspectrale (les colonnes de W) ainsi que les abondances de ces matériaux (les lignes de H) sous l'hypothèse d'un mélange linéaire.

Supposons maintenant que l'on dispose d'une base de données de spectres étiquetés représentée sous la forme d'une matrice $\mathrm{D} \in \Sigma_{n\times d}$ qu'on appellera dictionnaire, où $\Sigma_{n\times d}$ est l'ensemble des matrices dont les colonnes sont dans le simplexe de dimension n. Il est naturel de chercher à comparer les spectres estimés dans la matrice W avec *D a posteriori*. Récemment, il a été proposé d'utiliser cette information directement durant l'estimation des matrices W et H de la NMF [3]. On cherche alors une approximation de Y sous la forme d'une factorisation en matrices non négatives telle que W soit une collection de spectres issus de D. On introduit le problème de factorisation en matrices non négatives basée dictionnaire :

$$\underset{\sigma\in I(r,d), \mathrm{H}\in\mathbb{R}_+^{r\times m}}{\operatorname{argmin}} \| \mathrm{Y} - \mathrm{D}_\sigma \mathrm{H} \|_F^2 \tag{2}$$

Avec $\sigma \in I(r, d)$ l'ensemble des injections de $[\![1,\ r]\!]$ dans $[\![1,\ d]\!]$ et en notant D_σ la matrice formée par les colonnes de D d'indice $(\sigma(i))_{i\in[\![1,\ r]\!]}$. Ce problème ainsi posé suppose de prendre comme templates directement les spectres du dictionnaire D, mais cette approche peut être trop restrictive. Par exemple le dictionnaire pourrait contenir des spectres échantillonnés différemment des données. On se propose donc de relâcher l'égalité stricte entre W et certaines colonnes de D, ce qui conduit au modèle de NMF basée Dictionnaire régularisé :

$$\underset{\substack{\mathrm{W}\in\mathbb{R}_+^{n\times r},\ \mathrm{H}\in\mathbb{R}_+^{r\times m}\\ \sigma\in I(r,d)}}{\operatorname{argmin}} \| \mathrm{Y}-\mathrm{WH}\|_F^2+\epsilon\sum_{i=1}^{r}\mathcal{L}(\mathrm{W}_i,\mathrm{D}_{\sigma(i)}) \quad (3)$$

où $\mathcal{L}$ est une fonction de coût, appelée régularisation. Un travail antérieur a considéré une norme de Frobenius pour cette régularisation [3] avec des résultats mitigés. Cependant, après normalisation, un spectre est une distribution de probabilité, d'où l'utilisation de la distance de Wasserstein comme fonction de régularisation, distance plus respectueuse de la géométrie des espaces de probabilité que la norme de Frobenius.

La suite de cet article s'intéressera donc au modèle de NMF basée Dictionnaire régularisé par Wasserstein (NMFDW) et propose notamment un algorithme de descente par bloc de coordonnées. On l'étudiera sur un exemple issu de l'imagerie hyperspectrale grâce à des données synthétiques. Après avoir évalué l'algorithme proposé et l'avoir comparé à d'autres algorithmes de la littérature [3], plusieurs pistes d'amélioration sont discutées.

2 Rappels sur le transport optimal

Le transport optimal permet de définir des distances entre deux distributions de probabilité en élevant des distances point à point. Formellement, soit α et β deux histogrammes de Σ_n, un couplage entre α et β est la donnée d'une mesure P sur $P(\Sigma_n,\Sigma_n)$ dont les marginales sont α et β. Soit $\mathcal{U}(\alpha,\beta)$ l'ensemble des couplages. En notant α et β comme étant des vecteurs, les couplages peuvent être représentés par des matrices P vérifiant $\mathrm{P}\mathbb{1}_n=\alpha$ et $\mathrm{P}^T\mathbb{1}_n=\beta$. La distance du transport optimal, aussi appelée distance de Wasserstein, est la solution au problème de Kantorovitch [5] :

$$\mathcal{W}_C(\alpha,\beta)=\min_{P\in\mathcal{U}(\alpha,\beta)}\sum_{i,j}C_{i,j}P_{i,j} \quad (4)$$

avec $C\in\mathbb{R}^{n_\alpha\times n_\beta}$ une matrice de coût entre chaque élément du support de α vers celui de β. On utilise dans cet article la distance L_2 entre les éléments du support de α et de β. Dans le cas unidimensionnel, on a une forme close définie comme suit :

$$\mathcal{W}_{L_2}(\alpha,\beta)^2=\left\|\mathcal{C}_\alpha^{-1}-\mathcal{C}_\beta^{-1}\right\|_{L^2}^2=\int_0^1|\mathcal{C}_\alpha^{-1}(r)-\mathcal{C}_\beta^{-1}(r)|^2dr \quad (5)$$

avec $\mathcal{C}_\alpha^{-1}$ la fonction quantile inverse associée à α [5]. Dans le cas des histogrammes, $\mathcal{C}_\alpha^{-1}$ est une fonction en escalier et l'ensemble n'est pas différentiable, seulement sous-différentiable.

3 Le modèle NMFDW

3.1 Présentation du modèle

Estimer les paramètres du modèle NMFDW revient à résoudre le problème d'optimisation suivant :

$$\underset{\substack{\mathrm{W}\in\Sigma_{n\times r},\ \mathrm{H}\in\mathbb{R}_+^{r\times m}\\ \sigma\in I(r,d)}}{\operatorname{argmin}} \| \mathrm{Y}-\mathrm{WH}\|_F^2+\epsilon\sum_{i=1}^{r}\mathcal{W}_{L_2}(\mathrm{W}_i,\mathrm{D}_{\sigma(i)}) \quad (6)$$

où ϵ est un hyperparamètre du modèle à régler. Notons qu'afin d'assurer que les colonnes de W sont bien des distributions de probabilité, nous imposons que ces colonnes appartiennent au simplex Σ_n de dimension n grâce à une contrainte convexe.

3.2 Algorithme d'optimisation

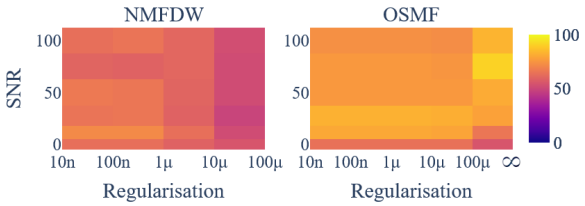
Ce problème est non convexe et NP-dur car plus général que la NMF [6]. Cependant, la fonction de perte dans l'équation (6) est convexe par rapport à chaque bloc W, H, il est donc naturel d'opter pour une stratégie alternée. On procède donc par descente par bloc de coordonnées. L'algorithme proposé prend en entrée les données Y, les matrices initiales W et H, le dictionnaire D et la valeur initiale de σ. On renvoie alors le coût total, les matrices estimées $\hat{\mathrm{W}}$ et $\hat{\mathrm{H}}$ et le support estimé $\hat{\sigma}$. Les trois blocs de variables sont W, H et σ, et on procède en effectuant un pas d'optimisation successivement sur chaque bloc. La mise à jour de H consiste en 1 itération de descente de gradient avec un pas $\frac{1}{\|W\|_F^2}$. La mise à jour de W est formellement plus compliquée, on cherche à résoudre le problème

$$\underset{W\in\Sigma_W}{\operatorname{argmin}} \| \mathrm{Y}-\mathrm{WH}\|_F^2+\epsilon\sum_{i=1}^{r}\mathcal{W}_{L_2}(\mathrm{W}_i,\mathrm{D}_{\sigma(i)}) \quad (7)$$

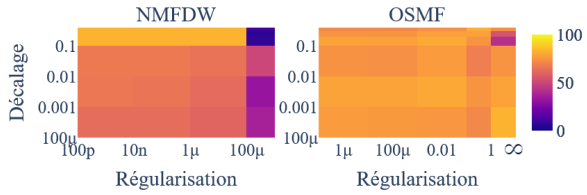
dont le terme de régularisation de Wasserstein n'est pas différentiable car continu et affine par morceau. L'algorithme retenu pour cette communication pour ce sous problème est une descente de sous-gradient projeté dont on effectue 1 itération avec un pas $\frac{1}{\|H\|_F^2}$. Pour σ, on suppose W et H fixé et on calcule la distance entre chaque spectre de W et chaque spectre de D dans une matrice $\mathrm{M}\in\mathbb{R}^{r\times d}$, telle que $\mathrm{M}_{i,j}=\mathcal{W}_C(\mathrm{W}_i,\mathrm{D}_j)$. On cherche alors les indices des colonnes de D minimisant leur distance à celles de W, ce qui peut s'obtenir en résolvant le problème d'assignement linéaire de coût la matrice M. Ce problème de transport optimal est un programme linéaire très étudié pouvant être résolu par l'algorithme de Jonker-Volgenant[4] L'algorithme de descente par bloc de coordonnées s'arrête après 100 itérations ou si la fonction de perte varie de façon relative de moins de 10^{-8}.

Pour la suite, on compare cet algorithme à deux algorithmes similaires proposés dans [3] OSMF et OSMF-pen [1], respectivement basé sur (2) et sur (3) avec la norme de Frobenius comme régularisation plutôt que la distance de Wasserstein.

1. La publication originale [3] nomme ces algorithmes MPALS et Flex-MPALS, mais nous choisissons ici de les renommer en One-Sparse Matrix Factorization (-penalized) en lien avec la publication plus récente [2] et le dépôt `https://github.com/cohenjer/dlraos-essentials`



(a) Taux d'identification moyen selon ϵ et η, à gauche NMFDW et à droite OSMF-pen



(b) Taux d'identification moyen selon ϵ et λ, à gauche NMFDW et à droite OSMF-pen

FIGURE 1 – Taux d'identification moyen en fonction de ϵ et de respectivement η à gauche et λ à droite.

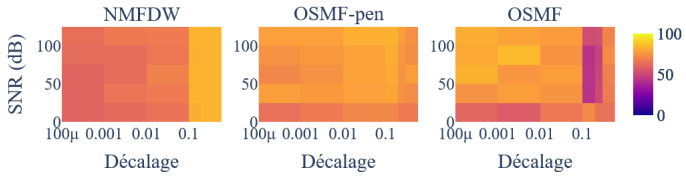


FIGURE 2 – Taux d'identification moyen en fonction de λ et de η à $\epsilon = 10^{-6}$ pour NMFDW et $\epsilon = 10^{-3}$.

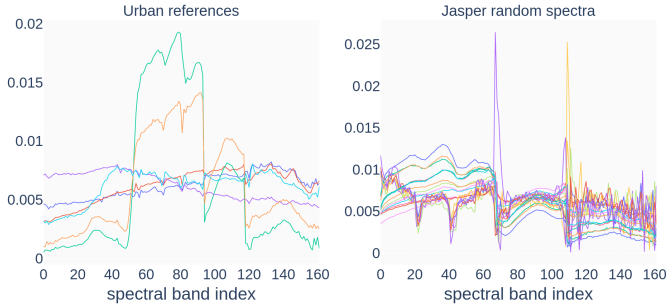


FIGURE 3 – Un exemple de dictionnaire utilisé dans les expériences. A gauche, les six premiers atomes issus de la vérité terrain d'*Urban*. A droite, vingt spectres aléatoirement choisis dans *Jasper*.

4 Évaluation

4.1 Méthodologie d'évaluation

Présentation des bases de données : On utilise deux bases de données classiques d'imagerie hyperspectrale pour évaluer la performance des différents algorithmes : *Urban* et *Jasper Ridge* [7]. *Urban* est une image de 307×307 pixels sur 210 canaux correspondant à des longueurs d'onde de $400nm$ à $2500nm$, tous les $10nm$. Les canaux 1-4, 76, 87, 101-111, 136-153 et 198-210 sont supprimés à cause de problèmes d'acquisition, il reste 162 canaux disponibles. On n'utilise pas directement l'image *Urban* dans cette expérience, mais plutôt six spectres de référence issus de cette image, correspondant à l'asphalte, l'herbe, les arbres, les toits, la terre et le métal [7].

Jasper Ridge est une image de 512×614 pixels de 224 canaux chacun correspondant à des longueurs d'onde de $380nm$ à $2500nm$, tous les $9.46nm$. Les canaux 1-3, 108-112, 154-166, 220-224 sont supprimés, il ne reste que 198 canaux. Afin d'aligner les spectres des deux jeux de données, on se contentera de tronquer les 36 premiers canaux de *Jasper Ridge*. On construit alors un dictionnaire D en ajoutant aux six spectres de référence de *Urban* vingt spectres choisis aléatoirement dans l'image *Jasper Ridge*. Le dictionnaire ainsi construit contient donc des spectres réalistes, que nous allons utiliser pour générer des données synthétiques, une réalisation d'un tel dictionnaire est présentée à la Figure 3.

Données synthétiques : Les données synthétiques sont générées selon le procédé suivant : $\mathrm{Y} = \tilde{\mathrm{W}}\mathrm{H} + \eta\mathrm{N}$ avec N une matrice de bruit dont les entrées sont des gaussiennes centrées réduites i.i.d. et H une matrice dont les entrées sont échantillonnées selon une loi uniforme sur $[0, 1]$ puis normalisées pour ℓ_1 selon les colonnes. Chaque colonne $\tilde{\mathrm{W}}_i$ de la matrice $\tilde{\mathrm{W}}$ est le barycentre au sens du transport optimal entre le vecteur $e_n = [0, 0, \ldots, 0, 1]$ et la colonne D_i de D, de poids respectif λ et 1-λ. Cette manipulation décale la masse d'un atome tout en conservant sa géométrie, créant une inadéquation entre templates de la NMF et templates de référence.

Trois hyperparamètres sont à étudier : le paramètre de régularisation ϵ, le niveau de bruit η et le paramètre de décalage λ ; le paramètre η est ici présenté sous la forme d'un rapport signal sur bruit (SNR) en dB. Le paramètre ϵ correspond à l'impact de la pénalisation dans le calcul du coût. Un paramètre ϵ faible permet de plus facilement changer les atomes sélectionnés au cours de l'algorithme, a contrario une valeur élevée force W à vivre près des templates sélectionnés au sens de l'OT, au risque de tomber dans un minimum local. On prendra comme critère d'évaluation la capacité qu'aura chaque algorithme à retrouver les spectres de références d'Urban dans un dictionnaire, c'est à dire le pourcentage d'éléments de la liste d'indices $[\![1,\ 6]\!]$ que l'on retrouve dans $\{\sigma(i)\}_{i=1..6}$. On parlera par la suite de taux d'identification.

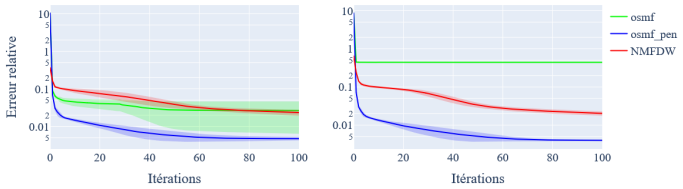
4.2 Résultats

Conditions expérimentales : On présente ici le taux d'identification moyen de chaque algorithme sur 25 données synthé-

tiques différentes, en faisant varier les différents paramètres. Afin de présenter une comparaison honnête des différentes méthodes, tous les algorithmes effectuent 100 itérations extérieures sans critère d'arrêt autre. On affichera les résultats de l'algorithme OSMF sur la heatmap associée à OSMF-pen, pour $\epsilon = \infty$. La première expérience étudie l'impact du niveau de bruit η pour différentes valeur de ϵ. La seconde expérience est similaire mais fait varier λ et ϵ en fixant η à 0. La troisième expérience fixe $\epsilon = 10^{-3}$ pour OSMF-pen et $\epsilon = 10^{-6}$ pour NMFDW. On étudie ici le cas où le signal subit à la fois un bruit gaussien et un décalage.

Chaque algorithme est initialisé avec un même jeu de données, généré aléatoirement pour chaque donnée synthétique. σ_{init} est une injection aléatoirement choisie de $I(r, d)$. On pose $\mathrm{W}_{init} = \mathrm{D}_{\sigma_{init}}$ et H_{init} est généré en tirant uniformément ses composants sur $[0, 1]$, puis chacune de ses colonnes est normalisée selon L_1.

Résultats : L'algorithme NMFDW est en l'état moins performant qu'OSMF et qu'OSMF-Pen dans le cas de données avec bruit gaussien mais y semble plus robuste, selon la Figure 1(a). A contrario NMFDW est plus sensible au choix de ϵ là où OSMF-pen y est plus robuste pour la grille choisie. Augmenter le paramètre λ a, cependant, un effet inattendu : le taux d'identification augmente pour l'algorithme NMFDW pour atteindre jusqu'à 80%, effet visible Figure 2 et Figure 1(b). En l'état, il n'est pas encore possible d'expliquer cette amélioration, il peut s'agir d'une conséquence des jeux de données choisies. En prenant comme critère l'évolution de l'erreur relative durant l'exécution de l'algorithme, tant OSMF-pen que NMFDW semblent robustes au paramètre λ. OSMF converge quasi-instantanément vers une erreur relative plus importante.



(a) Cas SNR = 50, $\lambda = 0$ (b) Cas SNR = 50, $\lambda = 0.3$

FIGURE 4 – Moyenne et écart type de l'évolution de l'erreur relative en fonction du nombre d'itérations sur 25 jeux de données synthétiques

5 Perspectives

Il y a plusieurs pistes d'améliorations à explorer vis à vis du modèle proposé. Le choix de la distance de Frobenius pour estimer la qualité de l'approximation (1) n'est pas forcément la plus adaptée compte tenu des données étudiées. Expérimenter avec la divergence de Kullback-Leiber à la place pourrait apporter des résultats différents et pertinents. De même, le choix de n'utiliser que la distance L_2 comme coût sous-jacent au transport optimal dans (4), s'il est pratique en terme d'implémentation car permettant l'utilisation de la forme close (5), n'est pas forcément le plus pertinent. Également, l'algorithme utilisé pour estimer les paramètres de NMFDW pourrait être amélioré ou bénéficier de preuves de convergence.

La version actuelle du code utilise le transport optimal balancé, à comprendre qu'on restreint les entrées à être des distributions de probabilités, donc positives et de somme à 1. Dans le cas de l'imagerie hyperspectrale, on perd actuellement l'information de la luminosité des pixels lorsqu'on normalise les données pour l_1. Le transport non balancé éviterait cette perte d'information en contre partie de temps de calculs plus élevés.

La dernière piste qu'on évoquera sera celle de l'option barycentrique : on sélectionne actuellement chaque template comme étant à proximité d'un atome du dictionnaire. En supposant des banques de données grandes et étiquetées, il est envisageable d'utiliser les barycentres au sens du transport optimal [1] des atomes du dictionnaire. Pour reprendre l'exemple de l'imagerie hyperspectrale, cela revient à supposer qu'on possède plusieurs spectres typiques d'un élément et de choisir comme template pour celui-ci un élément du barycentre formé par un certain nombre de ces spectres. Chaque template pourra alors être choisi dans des espaces plus grands, tout en restant totalement identifiable.

Références

[1] Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein Barycentric Coordinates : Histogram Regression Using Optimal Transport. *ACM Transactions on Graphics*, 35(4) :71 :1–71 :10, April 2016.

[2] Jeremy E. Cohen. Dictionary-based low-rank approximations and the mixed sparse coding problem. *Frontiers in Applied Mathematics and Statistics*, 8, 2022.

[3] Jérémy E. Cohen and Nicolas Gillis. Dictionary-based tensor canonical polyadic decomposition. *IEEE Transactions on Signal Processing*, 66(7) :1876–1889, 2018.

[4] David F. Crouse. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4) :1679–1696, 2016.

[5] Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020.

[6] Stephen A Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3) :1364–1377, 2010.

[7] Feiyun Zhu, Ying Wang, Bin Fan, Gaofeng Meng, and Chunhong Pan. Effective spectral unmixing via robust representation and learning-based sparsity. *ArXiv*, abs/1409.0685, 2014.