Minimisation d'une forme générale de divergence sous contrainte de non-négativité Application à la factorisation en matrices non-négatives

Henri LANTÉRI, Céline THEYS, Cédric RICHARD[+]

Université de Nice Sophia-Antipolis, UMR Fizeau, CNRS, Observatoire de la Côte d'Azur
[+] Institut Universitaire de France
henri.lanteri@unice.fr, celine.theys@unice.fr, cedric.richard@unice.fr,

Résumé – La factorisation en matrices non-négatives (Nonnegative Matrix Factorization, NMF) est une méthode dédiée à la réduction de dimension de données non-négatives. Initialement formulée dans [1], elle a récemment connu un regain d'intérêt et été appliquée avec succès dans nombre d'applications [2]. Etant donnée une matrice $\boldsymbol{V}$ de dimension $(F \times N)$, à composantes non-négatives, elle consiste à rechercher une factorisation de la forme $\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H}$ où $\boldsymbol{W}$ et $\boldsymbol{H}$ sont également à composantes non-négatives, de dimension $(F \times K)$ et $(K \times N)$ respectivement. La dimension K est généralement choisie de sorte que $FK + KN \ll FN$. Diverses fonctions d'écart $\mathcal{D}(\boldsymbol{V}, \boldsymbol{W}\boldsymbol{H})$ ont été considérées dans la littérature, individuellement, pour formuler le problème d'optimisation

$$\min_{\boldsymbol{W},\boldsymbol{H}} \mathcal{D}(\boldsymbol{V}, \boldsymbol{W}\boldsymbol{H}) \quad \text{s.c.} \quad [\boldsymbol{W}]_{ij} \geq 0, [\boldsymbol{H}]_{ij} \geq 0 \tag{1}$$

et définir une méthode de résolution associée [1, 2, 3]. Dans le cadre de cette communication, nous proposons une forme générale de divergence incluant un grand nombre de fonctions d'écarts classiques, et nous présentons un algorithme général pour traiter le problème de minimisation considéré.

Abstract – Factoring in non-negative matrices (Nonnegative Matrix Factorization, NMF) is a method dedicated to the reduction of non-negative data. Originally formulated in [1], it has recently experienced a renewed interest and has been successfully applied in many applications [2]. Given a matrix $\boldsymbol{V}$ of dimension $(F \times N)$ with non-negative components , it is to find a factorization of the form $\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H}$, where $\boldsymbol{W}$ and $\boldsymbol{H}$ are also non-negative components, of dimension $(F \times K)$ and $(K \times N)$ respectively. Size K is usually chosen so that $FK + KN \ll$ FN. Various functions of deviation $\mathcal{D}(\boldsymbol{V}, \boldsymbol{W}\boldsymbol{H})$ were considered in the literature, individually, to formulate the optimization problem

$$\min_{\boldsymbol{W},\boldsymbol{H}} \mathcal{D}(\boldsymbol{V}, \boldsymbol{W}\boldsymbol{H}) \quad \text{s.c.} \quad [\boldsymbol{W}]_{ij} \geq 0, [\boldsymbol{H}]_{ij} \geq 0 \tag{2}$$

and define a resolution method associated [1, 2, 3]. As part of this paper, we propose a general form of divergence, including many classic divergences, and we present a general algorithm to treat the minimization problem in question.

1 Forme générale de divergence

La divergence proposée ci-dessous permet de retrouver essentiellement des divergences appartenant à la classe de Csiszär [4], et constitue une généralisation au sens des divergences d'Arimoto, [5] et de Sharma-Mittal, [6]. Les divergences de Bregman [7] et de Jensen Burbea-Rao [8, 9] sont exclues de cette étude, sauf cas particuliers qui seront mentionnés. Ainsi par exemple, la norme de Frobenius et la divergence d'Itakura-Saito [2] ne seront ici pas considérées. Afin d'introduire les notations utilisées, la forme générale de divergence considérée s'écrit :

$$\mathcal{D}(p,q) = g\left(\sum_{ij} h(p_{ij}, q_{ij})\right) - g\left(\sum_{ij} f(p_{ij}, q_{ij})\right) \tag{3}$$

avec

$$p_{ij} = [\boldsymbol{V}]_{ij}, \qquad q_{ij} = [\boldsymbol{W}\boldsymbol{H}]_{ij} \tag{4}$$

et

$$g(x) = \frac{1}{\alpha\,(\alpha-1)\,(s-1)}\, x^{\frac{s-1}{\delta-1}} \qquad 0 \leq \alpha \leq 1 \tag{5}$$

$$h(p_{ij}, q_{ij}) = \sum_{ij} (\alpha\, p_{ij}^{\gamma} + (1-\alpha)\, q_{ij}^{\gamma})^{\frac{1}{\gamma}}, \tag{6}$$

δ \ γ	-1	0	1	2
-1	0	GM-HM	AM-HM	SM-HM
0	HM-GM	0	AM-GM	SM-GM
1			Jensen-Shannon	
2	SM-HM	SM-GM	SM-AM	0

TABLE 1 – Divergences obtenues par (3) pour $s = \delta$

$$f(p_{ij}, q_{ij}) = \sum_{ij} (\alpha\, p_{ij}^{\delta} + (1-\alpha)\, q_{ij}^{\delta})^{\frac{1}{\delta}}. \tag{7}$$

Plus simplement, on a recours à des moyennes généralisées pondérées par α, les coefficients δ et γ permettant de passer en revue un certain nombre de moyennes classiques. La fonction $g(x)$ permet la généralisation au sens de Sharma-Mittal. Le paramètre s permet, par passage à la limite, d'atteindre deux sous-classes de divergences.

1.0.1 Sous-classe $\mathcal{D}_{s=\delta}$

Le cas correspondant est trivial et les divergences correspondantes s'écrivent :

$$\mathcal{D}(p,q) = \frac{1}{\alpha\,(\alpha-1)\,(\delta-1)} \left(\sum_{ij} (\alpha\, p_{ij}^{\gamma} + (1-\alpha)\, q_{ij}^{\gamma})^{\frac{1}{\gamma}} - \sum_{ij} (\alpha\, p_{ij}^{\delta} + (1-\alpha)\, q_{ij}^{\delta})^{\frac{1}{\delta}} \right). \tag{8}$$

Les divergences obtenues pour différentes valeurs des paramètres γ et δ sont résumées dans le Tableau 1. Les notations AM, GM, SM et HM désignent respectivement les moyennes arithmétiques, géométriques, racines carrées et harmoniques. Les deux combinaisons (GM-HM) et (HM-GM) ne sont pas considérées dans la suite car elles sont, soit négatives, soit non convexes. La divergence de Jensen-Shannon implique de faire tendre δ et γ vers 1, pour aboutir ainsi à la forme explicite :

$$\mathcal{D}(p,q) \simeq \frac{1}{\alpha\,(\alpha-1)} \sum_{ij} \alpha\, p_{ij} \log(p_{ij}) + (1-\alpha)\, q_{ij} \log(q_{ij}) - \sum_{ij} (\alpha\, p_{ij} + (1-\alpha)\, q_{ij}) \log(\alpha\, p_{ij} + (1-\alpha)\, q_{ij}). \tag{9}$$

On note par ailleurs que la divergence de Kullback-Leibler est obtenue à partir de la divergence (AM-GM) en effectuant le passage à la limite $\alpha \to 1$, et que sa duale est obtenue pour $\alpha \to 0$.

1.2 Sous-classe $\mathcal{D}_{s\to 1}$

Il est nécessaire dans ce cas de faire un développement limité à l'ordre 1 par rapport à s. On obtient un ensemble de divergences qui font apparaître des différences entre les logarithmes des différentes moyennes. Dans la mesure où les propriétés de convexité de ces divergences ne sont pas connues, l'analyse ci-après est limitée au seul cas classique de la divergence de Rényi [10]. Elle est obtenue en considérant $\gamma = 1$ et $\delta \to 0$, et s'écrit :

$$\mathcal{D}(p,q) = \frac{1}{\alpha\,(\alpha-1)} \log(\sum_{ij} p_{ij}^{\alpha}\, q_{ij}^{1-\alpha}) - \log(\sum_{ij} \alpha\, p_{ij} + (1-\alpha)\, q_{ij}). \tag{10}$$

On note que cette forme est utilisable avec des variables qui ne sont pas des densités de probabilité, et que l'expression classique est retrouvée lorsque les variables p_{ij} et q_{ij} sont normalisées. Par ailleurs, la divergence de Kullback-Leibler est obtenue à partir de la divergence de Rényi par passage à la limite $\alpha \to 1$, et que sa duale est obtenue pour $\alpha \to 0$.

2 Application à la factorisation en matrices non-négatives

Afin de résoudre le problème (2), la littérature ne préconise en l'état actuel que des méthodes d'optimisation alternée sur les éléments des matrices $\boldsymbol{W}$ et $\boldsymbol{H}$. On suggère d'adopter l'algorithme SGM, largement décrit dans [11], qui s'appuie sur une décomposition du gradient de la fonction coût de la forme

$$-\frac{\partial \mathcal{D}}{\partial [\boldsymbol{H}]_{ij}} = [\boldsymbol{R}]_{ij} - [\boldsymbol{S}]_{ij} \qquad -\frac{\partial \mathcal{D}}{\partial [\boldsymbol{W}]_{ij}} = [\boldsymbol{T}]_{ij} - [\boldsymbol{U}]_{ij}, \tag{11}$$

où les matrices $\boldsymbol{R}$, $\boldsymbol{S}$, $\boldsymbol{T}$ et $\boldsymbol{U}$ sont à termes positifs. La minimisation alternée prend alors la forme suivante

$$[\boldsymbol{W}]_{ij}^{k+1} = [\boldsymbol{W}]_{ij}^{k} + \mu^k [\boldsymbol{W}]_{ij}^{k} \left(\frac{[\boldsymbol{T}(\boldsymbol{W}^k, \boldsymbol{H}^k)]_{ij}}{[\boldsymbol{U}(\boldsymbol{W}^k, \boldsymbol{H}^k)]_{ij}} - 1 \right), \tag{12}$$

$$[\boldsymbol{H}]_{ij}^{k+1} = [\boldsymbol{H}]_{ij}^{k} + \rho^k [\boldsymbol{H}]_{ij}^{k} \left(\frac{[\boldsymbol{R}(\boldsymbol{W}^{k+1}, \boldsymbol{H}^k)]_{ij}}{[\boldsymbol{S}(W^{k+1}, \boldsymbol{H}^k)]_{ij}} - 1 \right). \tag{13}$$

Le principal intérêt de ce type d'algorithme est que, pour des pas μ^k et ρ^k égaux à 1, on aboutit à une formulation multiplicative particulièrement simple à mettre en œuvre. Cette forme est communément utilisée pour le problème NMF et, plus généralement, pour la minimisation de fonctions coût sous contraintes de non-négativité. Il est à noter que la littérature ignore largement la forme relaxée sous-jacente. La suite de l'article est consacrée au calcul des gradients (11) pour la forme générale de divergence (3)-(6).
Ainsi on a

$$-\frac{\partial \mathcal{D}}{\partial [\boldsymbol{H}]_{ij}} = \frac{1}{\alpha\,(\delta-1)} \left(\sum_{ij} (\alpha\, p_{ij}^{\gamma} + (1-\alpha)\, q_{ij}^{\gamma})^{\frac{1}{\gamma}} \right)^{\frac{s-\delta}{\delta-1}} [\boldsymbol{W}^\top \boldsymbol{A}]_{ij} - \frac{1}{\alpha\,(\delta-1)} \left(\sum_{ij} (\alpha\, p_{ij}^{\delta} + (1-\alpha)\, q_{ij}^{\delta})^{\frac{1}{\delta}} \right)^{\frac{s-\delta}{\delta-1}} [\boldsymbol{W}^\top \boldsymbol{B}]_{ij}, \tag{14}$$

$$-\frac{\partial \mathcal{D}}{\partial [\boldsymbol{W}]_{ij}} = \frac{1}{\alpha\,(\delta-1)} \left(\sum_{ij} (\alpha\, p_{ij}^{\gamma} + (1-\alpha)\, q_{ij}^{\gamma})^{\frac{1}{\gamma}} \right)^{\frac{s-\delta}{\delta-1}} [\boldsymbol{A}\, \boldsymbol{H}^\top]_{ij} - \frac{1}{\alpha\,(\delta-1)} \left(\sum_{ij} (\alpha\, p_{ij}^{\delta} + (1-\alpha)\, q_{ij}^{\delta})^{\frac{1}{\delta}} \right)^{\frac{s-\delta}{\delta-1}} [\boldsymbol{B}\boldsymbol{H}^\top]_{ij}, \tag{15}$$

avec

$$p_{ij} = [\boldsymbol{V}]_{ij}, \qquad q_{ij} = [\boldsymbol{W}\boldsymbol{H}]_{ij},$$

et

$$[\boldsymbol{A}]_{ij} = \left(\frac{\alpha\, p_{ij}^{\gamma} + (1-\alpha)\, q_{ij}^{\gamma}}{q_{ij}^{\gamma}} \right)^{\frac{1-\gamma}{\gamma}},$$

$$[\boldsymbol{B}]_{ij} = \left(\frac{\alpha\, p_{ij}^{\delta} + (1-\alpha)\, q_{ij}^{\delta}}{q_{ij}^{\delta}} \right)^{\frac{1-\delta}{\delta}}.$$

2.1 Sous-classe $\mathcal{D}_{s=\delta}$

On obtient dans ce cas :

$$-\frac{\partial \mathcal{D}}{\partial [\boldsymbol{H}]_{ij}} = \frac{1}{\alpha\,(\delta-1)}[\boldsymbol{W}^\top \boldsymbol{A}]_{ij} - \frac{1}{\alpha\,(\delta-1)}[\boldsymbol{W}^\top \boldsymbol{B}]_{ij}, \quad (16)$$

$$-\frac{\partial \mathcal{D}}{\partial [\boldsymbol{W}]_{ij}} = \frac{1}{\delta-1}[\boldsymbol{A}\boldsymbol{H}^\top]_{ij} - \frac{1}{\alpha\,(\delta-1)}[\boldsymbol{B}\boldsymbol{H}^\top]_{ij}. \quad (17)$$

Dans le cas particulier de la divergence de Jensen-Shannon, il est plus simple de calculer l'expression des gradients à partir de l'équation (9) directement. On obtient

$$-\frac{\partial \mathcal{D}}{\partial [\boldsymbol{H}]_{ij}} \simeq \Big[\boldsymbol{W}^\top \big(\log[\boldsymbol{W}\boldsymbol{H}] - \log[\alpha\,\boldsymbol{V} + (1-\alpha)\,\boldsymbol{W}\boldsymbol{H}]\big)\Big]_{ij}, \quad (18)$$

$$-\frac{\partial \mathcal{D}}{\partial [\boldsymbol{W}]_{ij}} \simeq \Big[\big(\log[\boldsymbol{W}\boldsymbol{H}] - \log[\alpha\,\boldsymbol{V} + (1-\alpha)\,\boldsymbol{W}\boldsymbol{H}]\big)\boldsymbol{H}^\top\Big]_{ij}. \quad (19)$$

Dans cette expression, la fonction logarithmique est appliquée à chacun des termes des matrices.

2.2 Sous-classe $\mathcal{D}_{s\to 1}$

Seul le cas de la divergence de Rényi est considéré, par manque de place. Dans ce cas, on a

$$-\frac{\partial \mathcal{D}}{\partial [\boldsymbol{H}]_{ij}} \simeq \Big[\boldsymbol{W}^\top \Big(\big(\sum_{ij} p_{ij}^{\alpha}\, q_{ij}^{1-\alpha}\big)^{-1} \frac{[\boldsymbol{V}]^{\alpha}}{[\boldsymbol{W}\boldsymbol{H}]^{\alpha}} - \big(\sum_{ij}(\alpha\, p_{ij} + (1-\alpha)\, q_{ij})\big)^{-1}\mathbf{1}\Big)\Big]_{ij}, \quad (20)$$

$$-\frac{\partial \mathcal{D}}{\partial [\boldsymbol{W}]_{ij}} \simeq \Big[\Big(\big(\sum_{ij} p_{ij}^{\alpha}\, q_{ij}^{1-\alpha}\big)^{-1} \frac{[\boldsymbol{V}]^{\alpha}}{[\boldsymbol{W}\boldsymbol{H}]^{\alpha}} - \big(\sum_{ij}(\alpha\, p_{ij} + (1-\alpha)\, q_{ij})\big)^{-1}\mathbf{1}\Big)\boldsymbol{H}^\top\Big]_{ij}. \quad (21)$$

Dans ces expressions, les opérations de rapport entre deux matrices et de puissance sont effectuées terme à terme.

3 Simulations

L'imagerie hyperspectrale fait l'objet d'une attention toute particulière ces dernières années, voir par exemple [12], [13] et les références associées. Celà consiste en l'acquisition de données à haute résolution dans des centaines de bandes spectrales, géo référencées. Avec sa capacité à produire des données très détaillées autant spatialement que spectralement, cette technologie offre de nouvelles possibiités dans l'acquisition et le traitement d'informations pour les applications civiles et militaires.

Chaque pixel d'une image hyperspectrale caractérise une signature spectrale. Il est classique de considérer que chaque pixel étendu sur un vecteur peut être modélisé comme un mélange linéaire de différentes composantes spectrales élémentaires, appelés "endmembers". Si l'on veut se référer à nos notations, chaque colonne de $\boldsymbol{V}$ peut être interprétée comme une signature spectrale obtenue par mélange linéaire des spectres élémentaires, c.a.d les colonnes de $\boldsymbol{W}$. Le problème est alors d'estimer les spectres élémentaires $\boldsymbol{W}$ et les coefficients contenus dans $\boldsymbol{H}$ à partir des signatures spectrales $\boldsymbol{V}$.

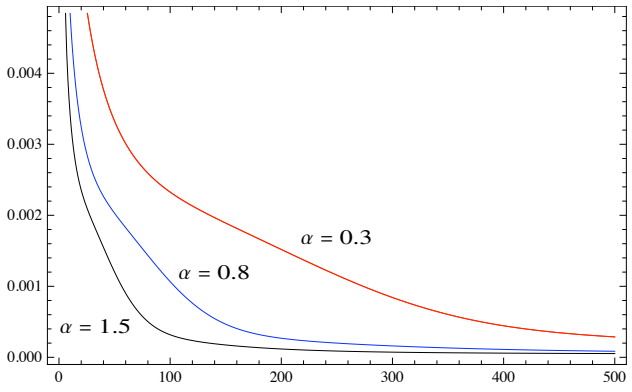

FIGURE 1 – Divergence de Renyi au cours des itérations pour différentes valeurs de α.

De nombreuses simulations ont été réalisées pour valider l'algorithme proposé, (12) et (13), on montrera les résultats obtenus dans le cas de la divergence de Rényi, (20) et (21) pour différentes valeurs de α. L'expérience présentée dans ce papier correspond à 8 mélanges linéaires de 6 "endmembers", la longueur de chaque spectre est de 826. Les spectres élémentaires utilisés dans cet exemple proviennent de la librairie ENVI, [14].

La figure (1) montre l'évolution de la divergence de Renyi au cours des itérations pour différentes valeurs du paramètre α, le paramètre α modifiant la vitesse de convergence. La figure (2) trace l'allure de deux des signatures spectrales reconstruites $\boldsymbol{V}$ pour une divergence "s'approchant" de Kullback-Leibler ($\alpha \to 1$). Les figures (3) et (4) tracent l'allure des mêmes signatures spectrales $\boldsymbol{V}$ pour une divergence de Renyi avec respectivement $\alpha = 0.8$ et $\alpha = 1.5$. On peut noter l'efficacité de l'algorithme dans tous les cas.

4 Conclusion

Dans ce travail, nous avons d'abord proposé une forme de divergence entre deux champs de données, qui est une généralisation de la distance de Csiszär. L'analyse de deux cas limites a permis de mettre en évidence un grand nombre de divergences classiques par action sur l'un des paramètres. La minimisation de ces divergences sous contrainte de non-negativité a été appliquée au problème de la factorisation en matrices non-négatives, et toutes les expressions nécessaires à la mise en œuvre de tels algorithmes ont été données. Une expérimentation relative au démélange des données hyperspectrales est

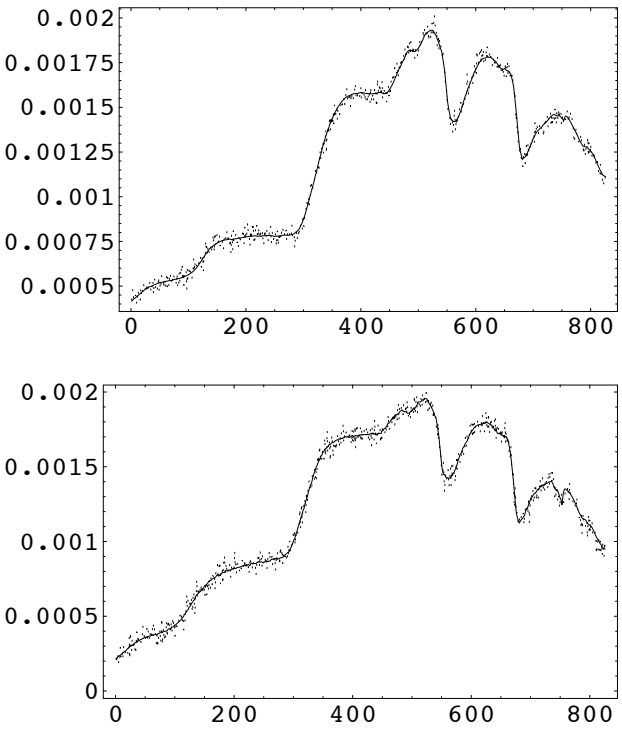

FIGURE 2 – Divergence de Kullback-Leibler ($\alpha \rightarrow 1$) - De haut en bas, colonnes 2 et 6 de $\boldsymbol{V}$, trait plein pour les vraies valeurs, trait pointillé pour les valeurs estimées à la fin des 500 itérations. $snr = 30dB$

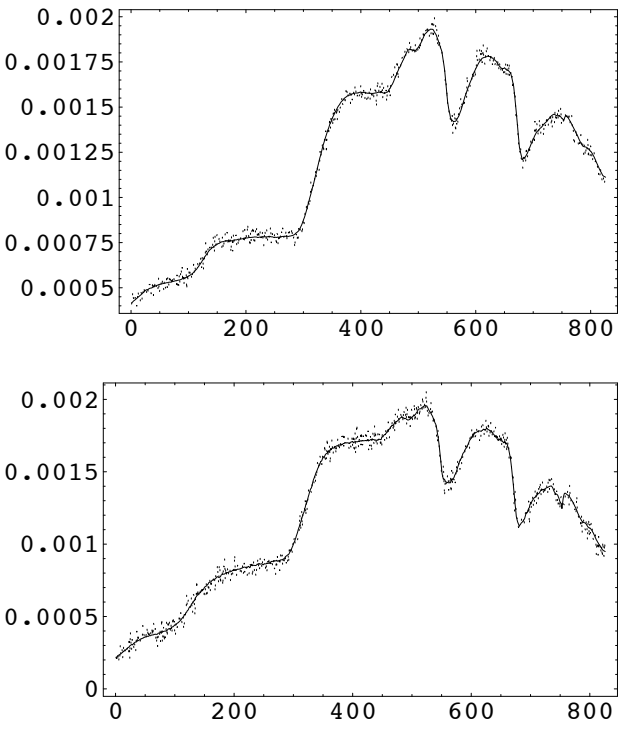

FIGURE 3 – Divergence de Renyi ($\alpha = 0.8$) - De haut en bas, colonnes 2 et 6 de $\boldsymbol{V}$, trait plein pour les vraies valeurs, trait pointillé pour les valeurs estimées à la fin des 500 itérations. $snr = 30dB$

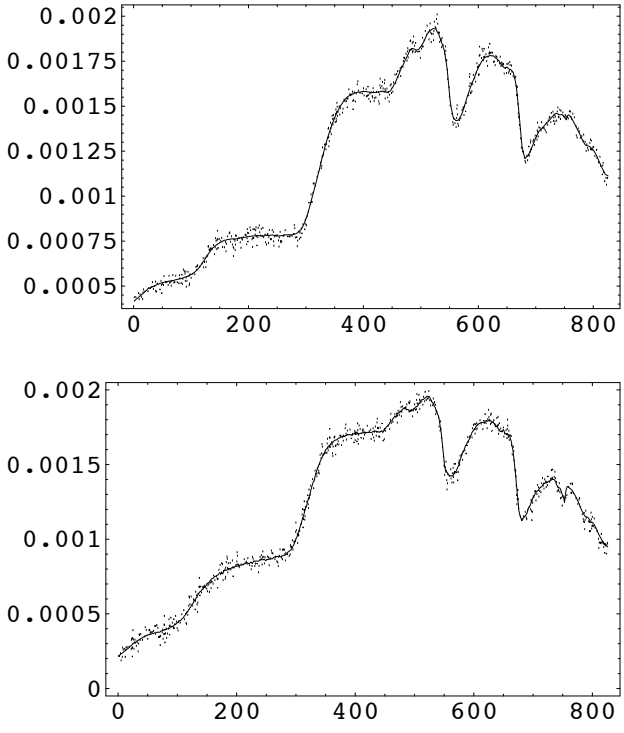

FIGURE 4 – Divergence de Renyi ($\alpha = 1.5$) - De haut en bas, colonnes 2 et 6 de $\boldsymbol{V}$, trait plein pour les vraies valeurs, trait pointillé pour les valeurs estimées à la fin des 500 itérations. $snr = 30dB$

proposée et les résultats obtenus dans le cas de la divergence de Renyi pour plusieurs valeurs du paramètre α sont donnés et montrent l'efficacité de l'algorithme.

Références

[1] D. D. Lee and S. Seung. Algorithms for Non-Negative Matrix Factorization. *Nature*, 6755 :788–791, 1999.

[2] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence with application to music analysis. *Neural Computation*, 2009.

[3] A. H. Phan A. Cichocki, R. Zdunek and S. Amari. *Nonnegative matrix and tensor factorizations*. Wiley, 2009.

[4] I. Csiszär. Information type measures of difference of probability distributions. *Studia Scientiarum Mathematicarum Hungarica*, 2 :299–318, 1967.

[5] F. Österreider. On a class of perimeter type distances of probability distributions. *Kybernetika*, 32(4) :389–393, 1996.

[6] C. Arndt. *Information measures*. Springer-Verlag, 2001.

[7] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7 :200–217, 1967.

[8] J. Burbea and C. R. Rao. On the convexity of some divergence measures based on entropy functions. *IEEE Transactions on Information Theory*, 28(3) :489–495, 1982.

[9] J. Burbea and C. R. Rao. On the convexity of higher order Jensen differences based on entropy function. *IEEE Transactions on Information Theory*, 28(6) :961–963, 1982.

[10] A. Renyi. On measures of entropy and information. In *Berkeley Symposium on Mathematics Statistica and Probabilities*, 1961.

[11] H. Lantéri, C. Theys, C. Richard, and C. Févotte. Split gradient method for nonnegative matrix factorization. In *Proc. 18th European Signal Processing Conference (EUSIPCO'10)*, Aalborg, Denmark, Aug. 2010.

[12] C. I. Chang. *Hyperspectral Imaging : Techniques for Spectral Detection and Classification*. Plenum Publishing Co., New York, 2003.

[13] D. A. Landgrebe. *Signal Theory Methods in Multispectral Remote Sensing*. Wiley, New York, 2003.

[14] RSI (Research Systems Inc.). *ENVI User's guide Version 4.0*. Boulder, CO 80301 USA, September 2003.