Co-factorisation douce en matrices non-négatives Application au regroupement multimodal de locuteurs

Nicolas Seichepine[1], Slim Essid[1], Cédric Févotte[2], Olivier Cappé[3]

[1]Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI
37-39 rue Dareau, 75014 Paris

[2]Laboratoire Lagrange (CNRS, OCA & Université de Nice)
Parc Valrose, 06000 Nice

[3]CNRS LTCI, Télécom ParisTech
37-39 rue Dareau, 75014 Paris

nicolas.seichepine@telecom-paristech.fr, slim.essid@telecom-paristech.fr
cfevotte@unice.fr, olivier.cappe@telecom-paristech.fr

Résumé – Nous présentons ici une nouvelle méthode pour une co-factorisation bi-modale en matrices non-négatives. Cette méthode est adaptée aux situations où deux modalités sont liées par une même information sous-jacente. Elle permet une co-factorisation dite douce, qui prend en compte la relation entre les modalités tout en évitant l'hypothèse forte d'un même facteur en commun. Cette méthode n'impose pas que les données associées à chaque modalité aient la même dimension ou soient exprimées dans un même espace. La co-factorisation est obtenue par résolution d'un problème d'optimisation, via une méthode de majoration-minimisation ; puis une application au regroupement multimodal de locuteurs est présentée, où la co-factorisation douce sert à exploiter la corrélation entre les pistes audio et vidéo dans des débats télévisés.

Abstract – This paper presents a new method for bi-modal nonnegative matrix factorization. This method is well-suited to situations where two streams of data (modalities) are related by a common underlying information. It allows for a *soft* co-factorization, which takes into account the relationship that exists between the modalities but avoids the strong hypothesis of a common *factor*. The data related with each modality can have different dimensionality and live in different features spaces. The co-factorization is obtained through an optimization algorithm, solved within the majorization-minimization framework. An application to multimodal speaker diarization is then presented, where the soft co-factorization is useful to exploit the correlation between audio and video modalities in edited talk-show videos.

1 Introduction

Il est fréquent que plusieurs modalités soient liées à une même information sous-jacente. Une utilisation optimale des données disponibles impose alors un traitement *conjoint*. Or ces données peuvent souvent être mises sous forme matricielle, leur exploitation passant par une factorisation. Plusieurs travaux existent donc déjà, qui cherchent à *co-factoriser* plusieurs modalités [1, 10]. Néanmoins, toutes ces méthodes partent d'une hypothèse qui est l'existence d'un facteur commun *identique*.

Pour autant, une même information sous-jacente n'est pas toujours idéalement modélisée par un facteur commun : dans le cas du regroupement multimodal de locuteurs (*multimodal speaker diarization* [8, 6]) par exemple, où l'on cherche à identifier quel locuteur parle à quel instant en s'appuyant à la fois sur l'audio et sur une piste vidéo associée, il est évident que la piste vidéo fournit des informations sur la personne qui parle puisque celle-ci est *généralement* à l'écran. Mais il n'existe pas un facteur commun à la vidéo et à l'audio puisque apparitions à l'écran et interventions à l'oral ne sont que partiellement corrélées.

On s'intéresse donc à une forme *douce* de co-factorisation en matrices non-négatives, qui *informe* la factorisation de chaque modalité (audio, vidéo) avec la factorisation de l'autre, mais qui n'impose pas que les factorisations résultantes incluent un facteur commun. Cette nouvelle factorisation offre de plus un contrôle sur la *force* du lien entre les deux modalités, et permet de choisir le cas échéant un couplage qui soit parcimonieux.

Cette *co-factorisation douce* en matrices non-négatives est ensuite appliquée à un problème de regroupement multimodal de locuteurs, abordé sous l'angle de la factorisation d'histogrammes de comptes comme introduit par [2].

Le modèle et l'algorithme sont décrits dans la section 2, puis les résultats de l'application pratique à la tâche de regroupement de locuteurs sont présentés section 3.

2 Construction d'un algorithme

La tâche de co-factorisation douce en matrices non-négatives est ici présentée dans le cadre de la résolution d'un problème d'optimisation de fonctionnelles de coût adaptées ; un algorithme est ensuite proposé pour la résolution de ce problème.

2.1 Factorisation en matrices non-négatives et fonctionnelles de coût

Étant donnée une matrice $V \in \mathbb{R}_+^{F\times N}$, la factorisation en matrices non-négatives (NMF) consiste à trouver deux matrices $W \in \mathbb{R}_+^{F\times K}$ et $H \in \mathbb{R}_+^{K\times N}$ telles que $V \simeq WH$; on souhaite généralement que $K \ll N$.

La NMF peut donc s'exprimer comme le problème d'optimisation $\min D(V \,|\, WH)$ par rapport W et H, où D une mesure de similarité, avec $W \geq 0$ et $H \geq 0$.

Dans le même ordre d'idée, nous proposons donc, étant données deux mesures de similarité D_1 et D_2, une fonction de pénalisation P, et des paramètres de pondération β_1, β_2 et β_j, une co-factorisation douce qui s'exprime comme suit :

$$\begin{cases} \min\limits_{W_1,H_1,W_2,H_2} \beta_1 D_1(V_1 \,|\, W_1H_1) \\ \qquad\qquad +\beta_2 D_2(V_2 \,|\, W_2H_2) \\ \qquad\qquad +\beta_j P\left(W_1,H_1,W_2,H_2\right) \\ \text{s.c.} \quad W_1 \geq 0,\, H_1 \geq 0,\, W_2 \geq 0,\, H_2 \geq 0. \end{cases} \tag{1}$$

Dans la suite, nous considérerons que les mesures de similarité D_1 et D_2 correspondent à la divergence de Kullback-Leibler, $D_{KL}\left(x,y\right) = x\log\left({}^{x}/_{y}\right) - x + y$. Par ailleurs, c'est l'écart entre les facteurs H_1 and H_2 uniquement qui sera pénalisé à l'aide d'une norme ℓ_1. On note que cette pénalité est linéairement affectée par les changements d'échelles, au même titre que la divergence de Kullback-Leibler utilisée comme mesure de similarité ; de la sorte, les différents termes de la fonctionnelle de coût restent « équilibrés » les un par rapport aux autres, indépendamment de l'échelle. Ces choix sont guidés par les caractéristiques de l'application pratique, mais d'autres sont parfaitement possibles et mèneraient à des algorithmes similaires.

2.2 Fonctionnelles de coût pour une NMF couplée et optimisation

Dès lors, une NMF avec un couplage doux peut s'exprimer comme :

$$\begin{cases} \min\limits_{W_1,H_1,W_2,H_2} C\left(W_1,H_1,W_2,H_2\right) = \beta_1 D_{KL}(V_1 \,|\, W_1H_1) \\ \qquad\qquad +\beta_2 D_{KL}(V_2 \,|\, W_2H_2) \\ \qquad\qquad +\beta_j \left\|H_1 - H_2\right\|_1 \\ \text{s.c.} \quad W_1 \geq 0,\, H_1 \geq 0,\, W_2 \geq 0,\, H_2 \geq 0. \end{cases} \tag{2}$$

Néanmoins, un tel problème d'optimisation ne permet pas d'obtenir des résultats pertinents.

1. Ce problème est affecté par une indétermination d'échelle ; il est en effet possible de diminuer le coût et de rendre la pénalité arbitrairement petite simplement en jouant sur l'échelle : avec $0 < \alpha < 1$, on a $C\left(\frac{1}{\alpha}W_1, \alpha H_1, \frac{1}{\alpha}W_2, \alpha H_2\right) < C\left(W_1,H_1,W_2,H_2\right)$ ce qui mène à des solutions dégénérées.
2. Rien n'impose que H_1 et H_2 aient des normes similaires ; on s'attend juste à ce que H_1 et H_2 aient des *formes* proches. Dès lors, il est indispensable de remettre à l'échelle ces matrices avant toute comparaison.
3. Il n'est pas possible de prendre en compte les situations où $H_1 \in \mathbb{R}^{K_1\times N}$ et $H_2 \in \mathbb{R}^{K_2\times N}$ avec $K_1 \neq K_2$.

Les situations où $K_1 \neq K_2$ peuvent être simplement traitées en ignorant le terme de pénalité pour les lignes de H_1 qui ne sont pas liées à H_2. En conséquence, K_1 et K_2 seront indifféremment notés K dans la suite. Il est en revanche nécessaire d'introduire la matrice diagonale $\Lambda \in \mathbb{R}^{K\times K}$, dont le n-ème coefficient diagonal est donné par $\lambda_n = \sum_f w_{fn}$, et la matrice diagonale $S \in \mathbb{R}_+^{K\times K}$ dont on note le k-ème coefficient diagonal s_k, pour résoudre les autres difficultés. Le problème est donc maintenant traduit par :

$$\begin{cases} \min\limits_{W_1,H_1,W_2,H_2,S} C\left(W_1,H_1,W_2,H_2\right) = \beta_1 D_{KL}(V_1 \,|\, W_1H_1) \\ \qquad\qquad +\beta_2 D_{KL}(V_2 \,|\, W_2H_2) \\ \qquad\qquad +\beta_j \left\|\Lambda_1 H_1 - S\Lambda_2 H_2\right\|_1 \\ \text{s.c.} \quad W_1 \geq 0,\, H_1 \geq 0,\, W_2 \geq 0,\, H_2 \geq 0. \end{cases} \tag{3}$$

Ce problème n'est de fait plus sujet aux mêmes problèmes numériques, puisque Λ_1 et Λ_2 empêchent que le coût soit diminué simplement à cause de l'échelle des matrices H_1 et H_2, et S permet une comparaison à l'échelle de H_1 et H_2. Au total, la seule hypothèse faite par ce modèle est que V_1 et V_2 ont le même nombre de colonnes, ce qui n'est que peu contraignant pour des données de même longueur.

Malheureusement, la fonctionnelle de coût (3) n'admet pas de solution analytique ; elle peut toutefois être optimisée en procédant à une descente par blocs et en optimisant itérativement par rapport à H_1, H_2, W_1, W_2 et S. Dans la pratique, il est possible de construire un algorithme de majoration-minimisation, qui assure que le coût décroît à chaque itération [4] et est aisément dérivable pour notre modèle [1].

2.3 Illustration sur données synthétiques

L'algorithme précédemment décrit est ici illustré sur un jeu de données synthétiques ; les tests sont menés comme

1. Les calculs et des codes sont disponibles en ligne : http ://perso.telecom-paristech.fr/~seichepi/gretsi2013

suit : on construit des matrices H_1 et $H_2 \in \mathbb{R}^{2\times 240}$, aux lignes faites de motifs de zéros et de uns. Puis l'on considère W_1 et $W_2 \in \mathbb{R}^{20\times 2}$, dont les coefficients sont tirés aléatoirement suivant une loi uniforme sur [1, 11]. On a alors $V_1 = W_1H_1$ et $V_2 = W_2H_2$. Pour illustration, W_1 et W_2 sont fixées dans l'algorithme à leur valeur réelle ; en revanche, l'initialisation se fait avec des valeurs aléatoires pour H_1 et H_2.

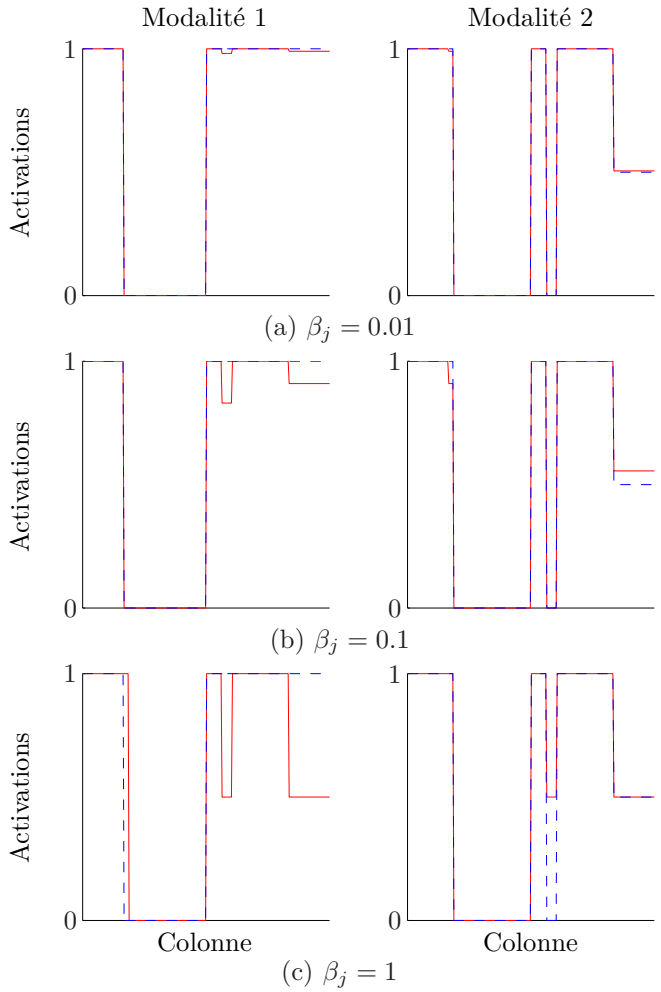

FIGURE 1 – Influence du couplage : les colonnes gauche et droite correspondent respectivement à la première et à la seconde modalité. Les lignes continues correspondant aux activations H retournées par l'algorithme, les lignes pointillées correspondant à la vérité terrain. Pour des raisons de lisibilité, on n'affiche qu'une composante par modalité.

On observe (Figure 1) que quand le paramètre de couplage β_j augmente, l'algorithme retourne des motifs d'activation pour la première modalité qui s'éloignent de la vérité terrain de la première modalité, pour se rapprocher de celle de la seconde, et réciproquement.

Il a également été possible d'observer que les motifs d'activations associés à la modalité avec la plus haute pondération (par exemple, β_1) étaient moins déformés par le couplage que ceux associés à la modalité avec la plus faible pondération (par exemple β_2).

Ces comportements permettent d'obtenir à volonté, et selon le réglages des hyperparamètres, des solutions où les motifs d'activations d'une modalité sont arbitraiement influencés par ceux de l'autre.

3 Application au regroupement de locuteurs

L'algorithme présenté ci-dessus est maintenant testé sur un problème réel, à savoir le regroupement de locuteurs sur un contenu multimédia, avec ou sans utilisation de la piste vidéo, pour comparaison.

3.1 Jeu de données et paramétrage

Les tests sont menés en utilisant les 33 premières vidéos de *Canal9 political debates database* [9]. Ce jeu de données est constitué des vidéos de différents débats politiques, tous organisés pour une même émission. Chaque vidéo comprend un modérateur et deux à quatre invités. Le regroupement de locuteurs est testé sur des segments de huit minutes par vidéo.

Tous les scores d'erreur sont calculés en utilisant la métrique NIST pour l'évaluation du regroupement de locuteurs [5]. Ce score est grossièrement une mesure de la fraction de temps de parole qui n'a pas été attribuée au bon locuteur ; meilleurs sont les résultats, plus ce score est faible.

Les factorisations de chaque flux (audio, vidéo) s'effectuent selon le processus décrit dans [2]. Des matrices de descripteurs V_{audio} et V_{video} sont construites pour représenter respectivement les flux audio et vidéo ; chaque colonne de V_{video} correspond à un histogramme d'occurences de « mots visuels », tandis que les colonnes de V_{audio} correspondent à des histogrammes d'états audio, estimés à partir de coefficients cepstraux. Les coefficients de ces matrices sont donc bien positifs, et l'emploi d'une divergence de Kullback-Leibler est justifié face à des histogrammes de compte [7, 3].

Pour privilégier l'audio – l'objectif restant le regroupement de locuteurs – tout en tenant compte de la vidéo, on applique une pondération $\beta_1 = 1$ et $\beta_2 = 5$, raisonnable au vu du comportement de l'algorithme sur des données synthétiques. Le poids du couplage β_j est lui fixé par réglage sur un ensemble d'apprentissage : parmi les 33 vidéos, 10 ont été aléatoirement choisies dans cet objectif. Ces vidéos ne sont pas utilisées dans les tests, mais ont permis de fixer $\beta_j = 0.1$ (table 1).

Les résultats peuvent varier au cours des initialisations puisque la fonctionnelle de coût n'est pas convexe. Tous les tests sont donc effectués avec 15 initialisations aléa-

TABLE 1 – Résultats moyens pour différentes valeurs de β_j (ensemble d'apprentissage).

β_j	0.01	0.1	1
Score d'erreur moyen	22.2	20.3	42.2

toires pour chaque vidéo, et seuls les résultats associés au meilleur coût de la fonctionnelle (optimisation ayant eu le plus de succès) sont retenus. Les scores d'erreur NIST associés aux solutions retenues sont ensuite moyennés sur les 23 vidéos de test, pour fournir un indice de performance global.

Enfin, si l'on note K le nombre de locuteurs : V_{audio} se factorise sur K canaux, mais V_{video} devrait être factorisée sur $K+1$ canaux [2]. Bien que cette différence ne soit pas problématique pour notre algorithme (section 2), nous proposons une comparaison avec une méthode naïve qui effectue une co-factorisation *dure* en factorisant $V_{stacked}$, une unique matrice qui regroupe $\beta_1 V_{video}$ et $\beta_2 V_{audio}$. Cette méthode naïve n'autorise pas des nombres de canaux différents, et l'on assure donc que la comparaison ne sera pas biaisée en utilisant cette méthode à la fois avec K *et* $K+1$ canaux.

3.2 Résultats

La table 2 résume les scores d'erreur (script NIST) obtenus en utilisant une NMF juste sur la piste audio (a), en utilisant la NMF pour factoriser une matrice compactant les descripteurs audio et vidéo sur K (b) ou $K+1$ canaux (c), et en utilisant l'algorithme que nous présentons pour une co-factorisation douce des modalités audio et vidéo (d).

Il ressort clairement qu'une co-factorisation douce est plus efficace, d'une part que les co-factorisations qui supposent qu'il existe un facteur exactement commun (b, c), d'autre part qu'une méthode qui ignore purement la modalité vidéo (a). L'information présente dans la vidéo est donc correctement exploitée pour informer la tâche de regroupement de locuteurs.

TABLE 2 – Résultats moyens des différentes méthodes sur les 23 vidéos de test.

Méthode	(a)	(b)	(c)	(d)
Score d'erreur moyen	21.4	25.1	18.9	16.8

4 Conclusion

Nous avons donc présenté un algorithme qui permet une co-factorisation douce en matrices non-négatives. Cet algorithme prend en compte l'existence d'un lien entre deux flux d'information sans pour autant supposer que ce lien est matérialisé par un facteur commun.

Sur un cas réel, la prise en compte d'une deuxième modalité (vidéo) apporte bien un surcroît d'information qui permet d'améliorer les résultats obtenus sur la première modalité (audio) ; ce surcroît d'information a pu être exploité plus utilement avec notre algorithme qu'avec la prise en compte d'un simple facteur commun, ce qui justifie le rejet de cette hypothèse pour l'application considérée.

2. Un par personne, et un pour les plans larges, voir [2].

Références

[1] Evrim Acar, Tamara Kolda, and Daniel Dunlavy. All-at-once Optimization for Coupled Matrix and Tensor Factorizations. *Computing Research Repository (CoRR)*, 2011.

[2] Slim Essid and Cédric Févotte. Nonnegative matrix factorization for unsupervised audiovisual document structuring. *IEEE Transactions on Multimedia*, 2012.

[3] Cédric Févotte and Ali Taylan Cemgil. Nonnegative matrix factorisations as probabilistic inference in composite models. In *Proc. 17th European Signal Processing Conference (EUSIPCO)*, pages 1913–1917, Glasgow, Scotland, 2009.

[4] David Hunter and Kenneth Lange. A Tutorial on MM Algorithms. *The American Statistician*, 58(1) :30–37, 2004.

[5] NIST. The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan, 2009.

[6] Athanasios Noulas, Gwenn Englebienne, and Ben Krose. Multimodal Speaker Diarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(1) :79–93, 2011.

[7] Jonathon Shlens. Notes on Kullback-Leibler Divergence and Likelihood Theory, 2007.

[8] Félicien Vallet, Slim Essid, and Jean Carrive. A multimodal approach to speaker diarization on TV talk-shows. *IEEE Transactions on Multimedia*, 2012.

[9] Alessandro Vinciarelli, Alfred Dielmann, Sarah Favre, and Hugues Salamin. Canal9 : A database of political debates for analysis of social interactions. In *IEEE International Workshop on Social Signal Processing*, Amsterdam, 2009. Ieee.

[10] Yusuf Kenan Yilmaz, Ali Taylan Cemgil, and Umut Simsekli. Generalized Coupled Tensor Factorization. In *Proc. 15th Advances in Neural Information Processing Systems (NIPS)*, 2011.