

Co-factorisation douce en matrices non-négatives

Application au regroupement multimodal de locuteurs

N. Seichepine, S. Essid, C. Févotte

Télécom ParisTech

10 juillet 2013

Sommaire

- 1 NMF - lien avec l'audio et la vidéo
 - Rappels sur la NMF
 - Application à des modalités audio/vidéo
- 2 Coupler des NMFs : pourquoi et comment ?
 - Le cas des vidéos éditées
 - Co-factorisation douce en pratique
- 3 Résultats en regroupement multimodal de locuteurs

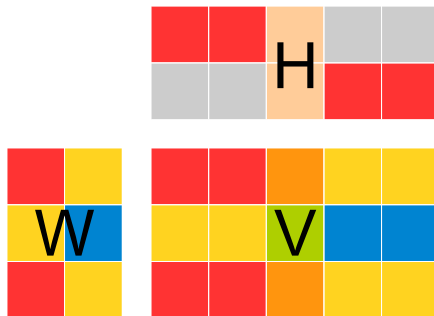
Principe et notations

- Approcher une matrice V à coefficients positifs par le produit de deux matrices W et H à coefficients positifs
- $V \in \mathbb{R}_+^{F \times N}$, $W \in \mathbb{R}_+^{F \times K}$, $H \in \mathbb{R}_+^{K \times N}$
- A priori, WH est de rang faible
- Peut se faire en résolvant le problème d'optimisation

$$\min_{W, H} D(V | WH)$$

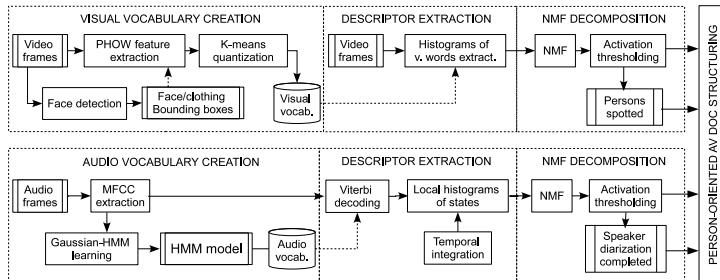
Principe et notations

- L'additivité « sans compensation » favorise des décompositions par parties
- Les colonnes de W constituent un dictionnaire, les lignes de H des patterns d'activation

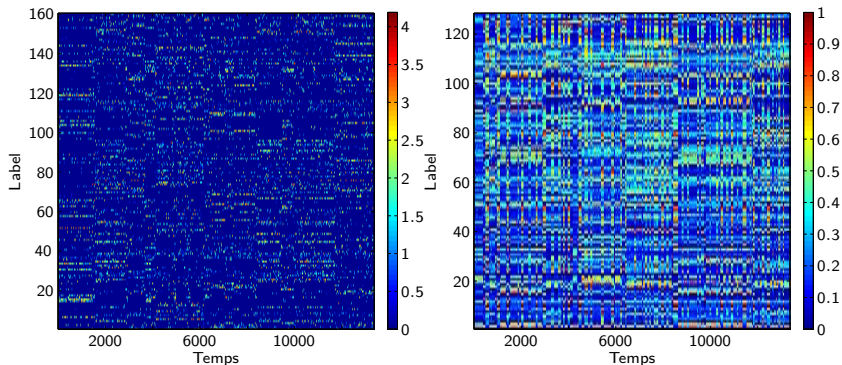


Construction de matrices non-négatives et représentatives

- L'idée est de construire des matrices qui *caractérisent* chaque locuteur
 - d'un point de vue audio, ou vidéo



Construction de matrices non-négatives et représentatives



$\log V_{\text{audio}}$ - gauche / $V_{\text{video}} / \max(V_{\text{video}}, \mathbb{I}, 2)$ - droite

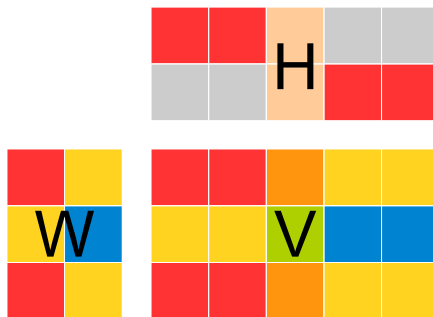
Regroupement de locuteurs et NMF

- Regroupement de locuteurs : qui parle quand ?



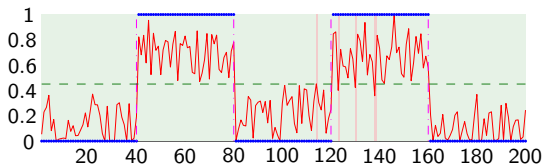
Regroupement de locuteurs et NMF

- Lien direct avec la NMF :



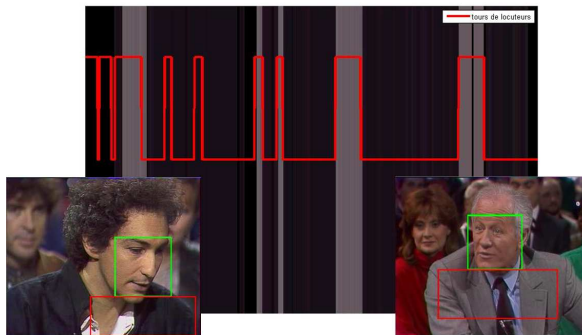
Regroupement de locuteurs et NMF

- Lien direct avec la NMF :



Regroupement de locuteurs et NMF

- Situation équivalente pour la vidéo (« qui est à l'écran quand ? »)
- Activations audio et vidéo corrélées, mais pas égales

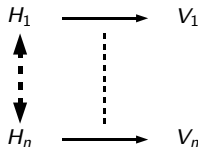


(Image : Slim Essid)

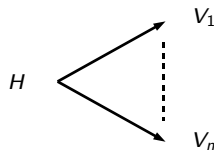
Co-factorisation douce et co-factorisation dure

- Co-factorisation dure : $V_1 = W_1 H$ et $V_2 = W_2 H$
 - ne correspond pas aux constatations

Co-factorisation douce

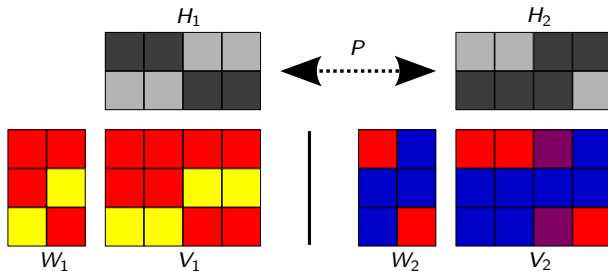


Co-factorisation dure



Co-factorisation douce et co-factorisation dure

- Co-factorisation douce : $V_1 = W_1 H_1$, $V_2 = W_2 H_2$, les différences entre H_1 et H_2 sont pénalisées



Construction d'une nouvelle fonctionnelle de coût

- Choix naturel :

$$\begin{aligned} \min_{W_1, H_1, W_2, H_2} \quad & C(W_1, H_1, W_2, H_2) = \beta_1 D_{KL}(V_1 | W_1 H_1) \\ & + \beta_2 D_{KL}(V_2 | W_2 H_2) + \beta_j \|H_1 - H_2\|_1, \\ \text{s.c.} \quad & W_1 \geq 0, H_1 \geq 0, W_2 \geq 0, H_2 \geq 0. \end{aligned}$$

- Distance de Kullback-Leibler adaptée aux histogrammes de comptes
- Pénalisation ℓ_1 adaptée à nos attentes sur le lien audio/vidéo
- Hyperparamètres $\beta_1, \beta_2, \beta_j$

Problèmes numériques

- Soit $\alpha \in]0, 1[$; on peut faire diminuer le coût en considérant

$$\beta_1 D_{KL}(V_1 | \frac{1}{\alpha} W_1 \alpha H_1) + \beta_2 D_{KL}(V_2 | \frac{1}{\alpha} W_2 \alpha H_2) + \beta_j \|\alpha H_1 - \alpha H_2\|_1$$

- On introduit les matrices diagonales $\Lambda_i \in \mathbb{R}^{K \times K}$ avec pour k -ème coefficient diagonal $\lambda_{i,k} = \sum_f w_{i,fk}$ et l'on considère

$$\begin{aligned} & \min_{W_1, H_1, W_2, H_2} C(W_1, H_1, W_2, H_2) \\ &= \beta_1 D_{KL}(V_1 | W_1 H_1) \\ &+ \beta_2 D_{KL}(V_2 | W_2 H_2) + \beta_j \|\Lambda_1 H_1 - \Lambda_2 H_2\|_1, \\ & \text{s.c. } W_1 \geq 0, H_1 \geq 0, W_2 \geq 0, H_2 \geq 0. \end{aligned}$$

Problèmes numériques

- Pour pouvoir comparer les matrices H_1 et H_2 « à l'échelle », il faut également introduire une matrice diagonale S ; le problème numériquement stable finalement étudié est

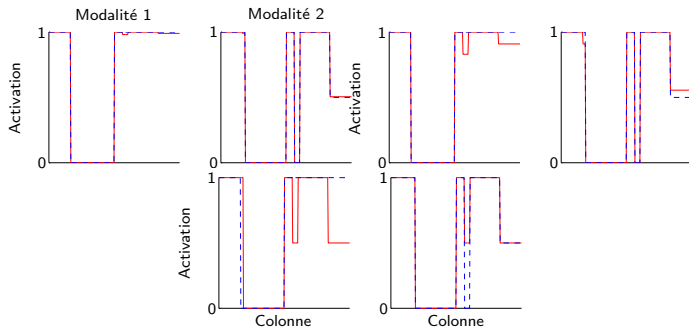
$$\begin{aligned}
 & \min_{W_1, H_1, W_2, H_2} C(W_1, H_1, W_2, H_2) \\
 &= \beta_1 D_{KL}(V_1 | W_1 H_1) \\
 &+ \beta_2 D_{KL}(V_2 | W_2 H_2) + \beta_j \|\Lambda_1 H_1 - S \Lambda_2 H_2\|_1, \\
 &\text{s.c. } W_1 \geq 0, H_1 \geq 0, W_2 \geq 0, H_2 \geq 0.
 \end{aligned}$$

- Soluble via une optimisation alternée en W et H par majoration-minimisation

Calibration

- β_1 est redondant
- β_2 est un choix direct de l'utilisateur
- β_j doit être optimisé à part
- Les réglages peuvent être adaptés directement pour des problèmes de dimensions différentes

Comportement effectif



De gauche à droite/haut en bas : $\beta_j = 0.01$, $\beta_j = 0.1$, $\beta_j = 1$

Données et scores

- 33 premières vidéos de *Canal9 political debates database*
 - débats à la télévision suisse, invités variés
 - trois à cinq personnes
- 10 vidéos aléatoirement sélectionnées pour le réglage de β_j , les autres sont utilisées pour les tests
- On mesure le « Diarization Error Rate »
 - script de scoring NIST standard
 - grossièrement, mesure de la fraction de temps-locuteur incorrectement attribuée (le plus petit est le meilleur)

Cadre de comparaison

- Du fait de la non-convexité, initialisations aléatoires, on conserve le « meilleur run »
- Comparaison entre :
 - une factorisation de la seule piste audio (a)
 - co-factorisation dure sur Q (b) ou $Q + 1$ (c) composantes où Q est le nombre de locuteurs
 - co-factorisation douce (d)

Méthode	(a)	(b)	(c)	(d)
DER moyen	21.4	25.1	18.9	16.8

Conclusion

- La co-factorisation douce est bien capable de prendre en compte l'information apportée par une modalité supplémentaire
- Elle peut le faire plus efficacement qu'une co-factorisation dure
- Et dans un cadre permettant aisément l'ajout de nouveaux a priori