

Factorisations en matrices non négatives

Non negative Matrix Factorization (NMF)

Paul Magron, Bertrand David

Telecom-ParisTech, LTCI CNRS UMR5141
département TSI, groupe AAO

27 novembre 2015



Introduction : problème standard

- NMF : Non Negative Matrix Factorization
- données $\rightarrow$ représentation temps-fréquence $X(f,t)$
- $|X(f,t)|^\alpha$: matrice $\mathbf{X}$ à coefficients positifs
- décomposition

$$\mathbf{X} \approx \mathbf{WH} \quad \text{ou} \quad \mathbf{X} = \mathbf{WH} + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E}$$

$$\mathbf{X} \approx \begin{bmatrix} \cdot \\ \cdot \\ \cdot \end{bmatrix} \begin{bmatrix} \cdot & \cdot & \cdot \end{bmatrix} + \begin{bmatrix} \cdot \\ \cdot \\ \cdot \end{bmatrix} \begin{bmatrix} \cdot & \cdot & \cdot \end{bmatrix} + \dots$$

Introduction : technique de résolution

Technique de résolution

- aveugle, en minimisant une *distance* ou une *divergence*

$$\mathcal{E}(\mathbf{W}, \mathbf{H}) = D(\mathbf{X} | \mathbf{WH})$$

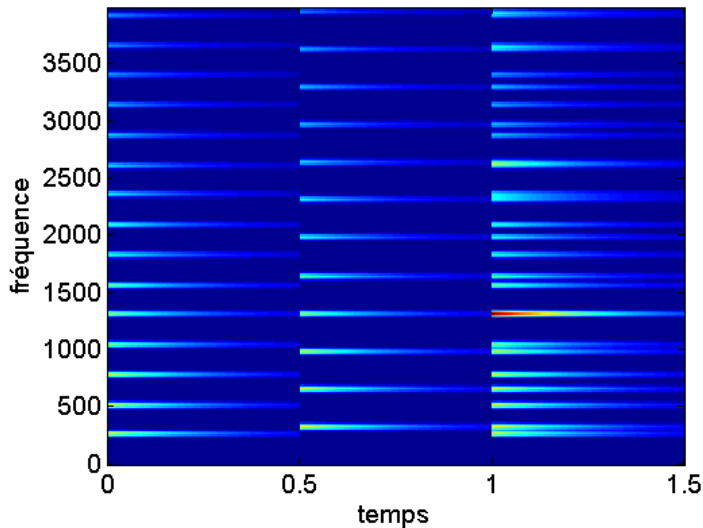
- la plupart du temps séparable :

$$D(\mathbf{X} | \mathbf{WH}) = \sum_{f=0}^{F-1} \sum_{t=0}^{T-1} D(X_{ft} | [\mathbf{WH}]_{ft})$$

- optimisation *itérative*

- $\mathbf{W}_n \leftarrow f_W(\mathbf{W}_{n-1}, \mathbf{H}_{n-1})$
- $\mathbf{H}_n \leftarrow f_H(\mathbf{W}_{n-1}, \mathbf{H}_{n-1})$

Exemple : DO-MI-accord DO/MI



Propriété importante : réduction de rang

la NMF s'appuie

- sur une **réduction de rang**
- c'est explicite : $\hat{\mathbf{X}}$ est de rang K
 - $\rightarrow$ réduction de la dimensionnalité $K(F + T) \ll FT$

$$\mathbf{X} \approx \underbrace{\begin{bmatrix} \cdot & \cdot & & & \\ \cdot & \cdot & & & \\ \cdot & \cdot & & & \\ \cdot & \cdot & & & \\ \cdot & \cdot & & & \\ \cdot & \cdot & & & \\ \cdot & \cdot & & & \end{bmatrix}_{F \times K} \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}_{K \times T}}_{\hat{\mathbf{X}}}$$

- sur la positivité des données $\rightarrow$ pas de création d'énergie noire

Remarque : méthodes HR

Pour les méthodes HR

- vecteur de données non bruitées

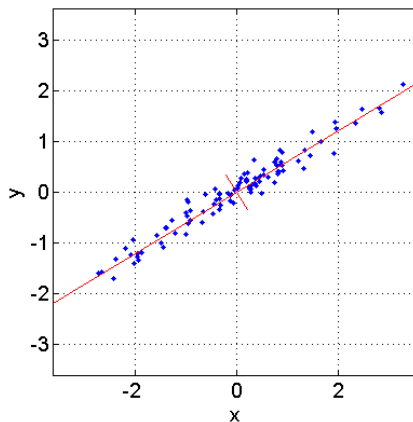
$$\mathbf{x} = \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(n-1) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 1 \\ z_0 & z_1 \\ \vdots & \vdots \\ z_0^{n-1} & z_1^{n-1} \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} b_0 \\ b_1 \end{bmatrix}}_{\mathbf{b}}$$

- $\mathbf{x} \in \text{span} \mathbf{V} = \mathcal{E}_{\text{sig}}$: espace **signal** (ici de dimension $K = 2$)
- Matrice de covariance : de rang $K = 2$!

$$\mathbf{C}_{xx} = \underbrace{\begin{bmatrix} \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \end{bmatrix}}_{\mathbf{V}} \begin{bmatrix} \cdot & \cdot \\ \cdot & \cdot \end{bmatrix}_{K \times K} \overbrace{\begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}}^{\mathbf{V}^T}_{K \times T}$$

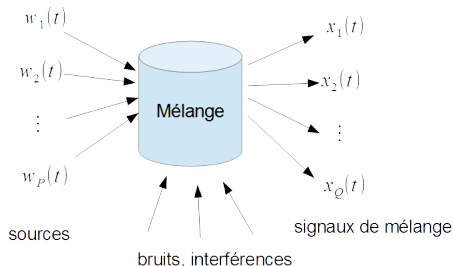
Remarque : ACP

- $\mathbf{u}$: vecteur aléatoire gaussien, IID, $\mathbf{v} = \mathbf{M}\mathbf{u}$
- ACP : *Analyse en Composantes Principales* : quelle(s) direction(s) explique(nt) le mieux la variance des données ?



Contexte plus général

- Application type : BSS (Blind Source Separation)
- exemple : MEG/EEG, Audio, astrophysique, reconnaissance automatique...
- On connaît la matrice des signaux de mélange : $\mathbf{X}$
- On veut la factoriser sous la forme : $\mathbf{X} = \mathbf{WH} + \mathbf{E}$, sans connaître $\mathbf{H}, \mathbf{W}, \mathbf{E}$!



Séparation de sources

- Il faut ajouter des *a priori* sur les sources, le système etc... pour pouvoir résoudre.

Les grands types d'hypothèses

- ICA : les sources sont *statistiquement indépendantes*
- NMF, NTF : les données et le système sont *positifs (non negative)*
- Méthodes parcimonieuses. ex : codage parcimonieux

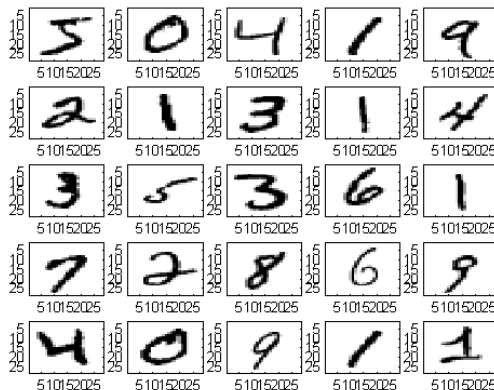
Codage parcimonieux

- $\mathbf{h}_k$ le + parcimonieux possible (avec bcp d'entrées = 0)
- la meilleure approximation possible
- parfois avec un grand dictionnaire $\mathbf{W}$
- On cherche

$$\underset{\mathbf{D}}{\operatorname{Argmin}} \sum_{k=1}^T \underset{\mathbf{h}_k}{\operatorname{Argmin}} \left\{ \|\mathbf{x}_k - \mathbf{W}\mathbf{h}_k\|_2^2 + \lambda \|\mathbf{h}_k\|_1 \right\}$$

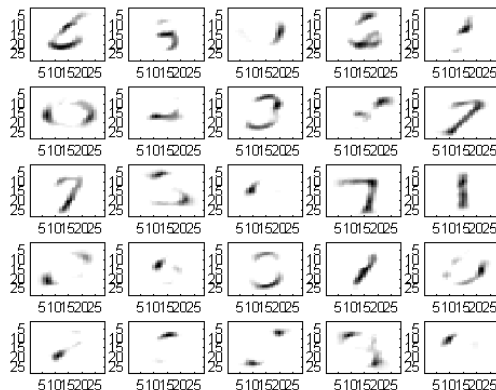
Intérêt de la NMF

- [Lee&Seung,Nature 99] : la NMF décompose en parties significatives
- Exemple chiffres tracés à la main (base NYU, Yann Lecun)



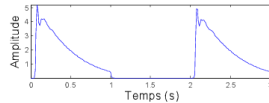
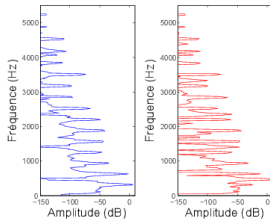
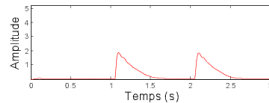
Intérêt de la NMF

- [Lee&Seung,Nature 99] : la NMF décompose en parties significatives
- Exemple chiffres tracés à la main (base NYU, Yann Lecun)

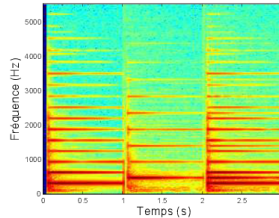


Intérêt de la NMF

- [Lee&Seung,Nature 99] : la NMF décompose en parties significatives
- Décomposition de signaux audio

 H

 W


Spectrogramme original



Différentes approches

- Approche heuristique : optimisation du critère
 - populaire : mises à jour multiplicatives
 - paramètre θ : $\frac{\partial \mathcal{L}}{\partial \theta} = \nabla_{\theta}^{+} - \nabla_{\theta}^{-}$
 - mise à jour $\theta \leftarrow \theta \frac{\nabla_{\theta}^{-}}{\nabla_{\theta}^{+}}$
- Modèles probabilistes :
 - on suppose que $X(f, t)$ est la réalisation d'une somme de v.a.
 - ou : on suppose $\mathbf{X}_{ft}$ est l'histogramme d'un tirage structuré de (f, t)

Distances et divergences

Rappel : Distance

- séparation : $d(x|y) = 0 \Leftrightarrow x = y$
- symétrie : $d(x|y) = d(y|x)$
- inégalité triangulaire : $d(x|y) \leq d(x|z) + d(z|y)$

Divergence :

- séparation : **oui**
- symétrie, inégalité triangulaire : **non**

Remarques :

- souvent : $D(\mathbf{X}|\mathbf{WH}) = \sum_f \sum_t d(\mathbf{X}_{ft} | [\mathbf{WH}]_{ft})$
- distance = cas particulier de divergence

β -divergences

- Le catalogue est très fourni : Amari, Csiszar, Bregman
- Cas particulier des divergences de Bregman : les β -divergences

Expressions et cas usuels

- $d_{\beta}(x|y) = \frac{1}{\beta(\beta-1)}(x^{\beta} + (\beta-1)y^{\beta} - \beta xy^{\beta-1})$
- cas $\beta = 2$: distance euclidienne

$$d_{\beta}(x|y) = \frac{1}{2}(x-y)^2$$

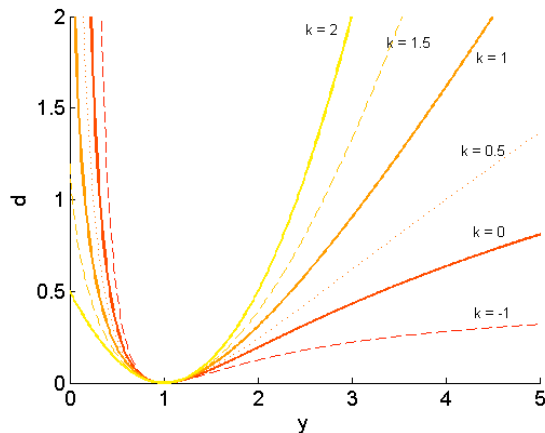
- cas $\beta = 1$: divergence de Kullback-Leibler (KL) [Kullback&Leibler, 1951]

$$d_{\beta}(x|y) = x \ln \frac{x}{y} + y - x$$

- cas $\beta = 0$: divergence d'Itakura-Saito (IS) [Itakura&Saito, 1968]

$$d_{\beta}(x|y) = \frac{x}{y} - \ln \frac{x}{y} - 1$$

β -divergences



- convexe pour $1 \leq \beta \leq 2$
- minimum en $x = y$ et croît avec $|x - y|$

β -divergences : différentiation

- $\frac{\partial d}{\partial y} = y^{\beta-2}(y-x)$
- mise à jour de $\mathbf{W}$:

$$\mathbf{W} \leftarrow \mathbf{W} \circledast \frac{\left[[\mathbf{W}\mathbf{H}]^{(\beta-2)} \circledast \mathbf{X} \right] \mathbf{H}^T}{[\mathbf{W}\mathbf{H}]^{(\beta-1)} \mathbf{H}^T}$$

- mise à jour de $\mathbf{H}$:

$$\mathbf{H} \leftarrow \mathbf{H} \circledast \frac{\mathbf{W}^T \left[[\mathbf{W}\mathbf{H}]^{(\beta-2)} \circledast \mathbf{X} \right]}{\mathbf{W}^T [\mathbf{W}\mathbf{H}]^{(\beta-2)}}$$

Algorithmes multiplicatifs

1. Initialisation (aléatoire ou mieux)
2. mise à jour multiplicative de $\mathbf{W}$ et $\mathbf{H}$
3. normalisation (colonnes de $\mathbf{W}$ ou lignes de $\mathbf{H}$)
4. retour au 2

Unicité de la décomposition

- Une solution existe

Invariances

- invariance par permutation
- invariance par changement d'échelle
- plus généralement : $\mathbf{WH} = (\mathbf{WT})(\mathbf{T}^{-1}\mathbf{H})$ avec $\mathbf{T}$ à coeff. ≥ 0
- On démontre que $\mathbf{T}$ est forcément un produit permutation x matrice diagonale
- $\rightarrow$ on normalise (svt) les colonnes de $\mathbf{W}$ (en norme L^2 par exemple)

Unicité : cône polyédrique

- **WH** : dans le cône polyédrique engendré par les colonnes de **W**
- $C_{\mathbf{W}} = \{ \sum_{k=1}^K h_k \mathbf{w}_k \text{ où } (h_1, h_2, \dots, h_K) \in \mathbb{R}^+ \}$
- svt : on peut agrandir/rétrécir $C_{\mathbf{W}}$ sans changer **WH**
- difficile de garantir l'unicité dans ces conditions !

Convergence : décroissance de $\mathcal{L}$

Décroissance du critère [Lee&Seung, 2001][Kompass, 2007]

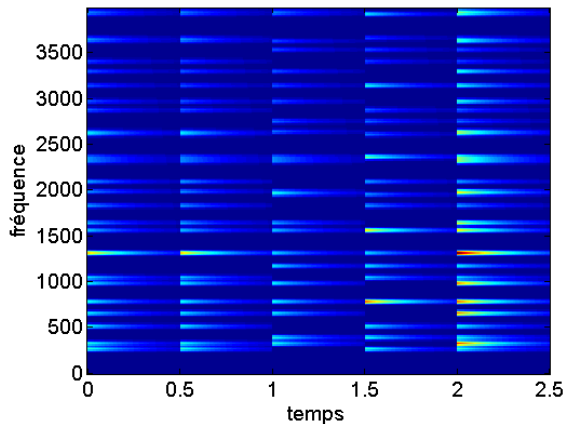
- démontré pour $1 \leq \beta \leq 2$ en s'appuyant sur la convexité de la divergence
- $\nRightarrow$ la convergence
- en pratique la vitesse de convergence est parfois plus importante

Minima locaux

- pas nécessairement des points stationnaires ($\nabla \mathcal{L} = 0$) car optimisation contrainte
- ex : lorsqu'on atteint un « bord » (une composante = 0)
- donc solution pas forcément un minimum
- certains auteurs ajoutent un ε aux dénominateurs

La solution peut n'être pas « musicale »

- Cas DO-MI/DO-MI/DO-SOL/SOL-MI/DO-MI-SOL



La valeur du critère n'est pas significative

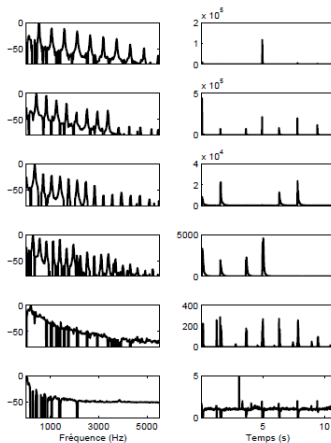


FIGURE IV.3 – Une factorisation réussie, $D_{IS}(\mathbf{V}, \mathbf{WH}) \approx 65000$.

La valeur du critère n'est pas significative

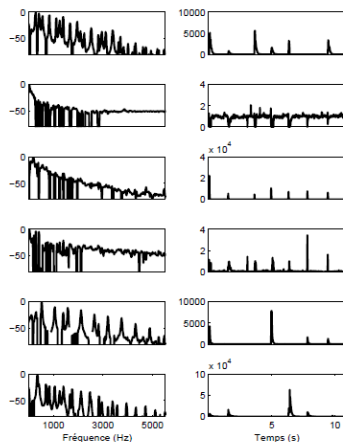
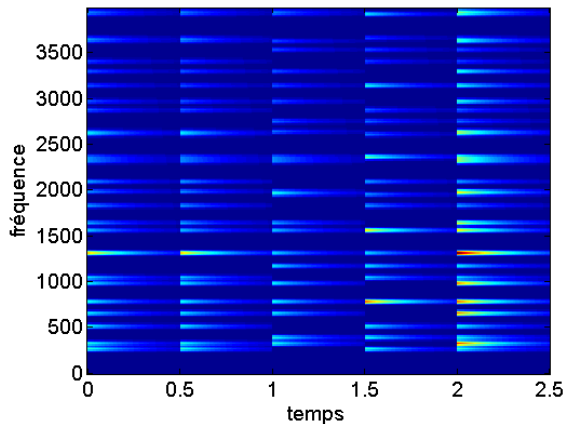


FIGURE IV.4 – Une factorisation ratée, $D_{IS}(\mathbf{V}, \mathbf{WH}) \approx 58000$

Amélioration : initialisation

- Cas DO-MI/DO-MI/DO-SOL/SOL-MI/DO-MI-SOL



Amélioration : contraintes, modèles

Aperçu des variantes contraintes

- parcimonie [Hoyer, 2004]
- régularité temporelle
 - en pénalisant les variations [Virtanen, 2007][Chen, 2006]
 - en introduisant des chaînes de Markov pour les activations [Fevotte, 2009]
- régularité spectrale
- harmonicité [Vincent, 2008][Rigaud, 2012]

Modèles de mélanges

Modèle gaussien à R composantes latentes [Fevotte, 2009]

- On s'intéresse à une colonne d'observation $\mathbf{x}_t$ complexe
- $\mathbf{x}_t = \sum_{r=1}^R \mathbf{c}_t^r$
- $\mathbf{c}_t^r$ identiquement distribués selon : $\mathbf{c}_t^r \sim \mathcal{N}_c(0, h_{rt} \text{diag} \mathbf{w}_r)$, soit encore $c_{ft}^r \sim \mathcal{N}(0, h_{rt} w_{fr})$

Modèle de Poisson [Virtanen, 2008]

- On s'intéresse à une colonne d'observation $\mathbf{x}_t$ *non negative*
- $\mathbf{x}_t = \sum_{r=1}^R \mathbf{c}_t^r$
- $\mathbf{c}_t^r$ identiquement distribués selon une loi de Poisson $c_{ft}^r \sim \mathcal{P}(h_{rt} w_{fr})$,
avec par exemple $c, P_c(c) = \mathcal{P}(\lambda) = \exp(-\lambda) \frac{\lambda^c}{c!}$

Maximum de vraisemblance

Fonction de coût « naturelle » = - log vraisemblance :

$$\mathcal{E}_{\text{MV}}(\mathbf{W}, \mathbf{H}) = -\log p(\mathbf{X}|\mathbf{W}, \mathbf{H})$$

Cas Gaussien :

- revient à minimiser $D_{\text{IS}} \rightarrow \text{IS-NMF}$

Cas de Poisson :

- revient à minimiser $D_{\text{KL}} \rightarrow \text{KL-NMF}$

Modèles à composantes latentes

PLCA [Shashanka, 2007] (Probabilistic Latent Component Analysis)

- C'est un modèle de « comptage » : $\mathbf{X}$ matrice non négative = histogramme d'un tirage du couple (f, t)
- Tirage selon une variable latente z
 - on tire z
 - on suppose f et t indépendantes conditionnellement à z :

$$P(f, t|z) = P(f|z)P(t|z)$$

- Toutes les variables latentes sont indépendantes :

$$P(f, t) = \sum_z P(z)P(f|z)P(t|z)$$

- Équivalence NMF : $P(f|z) \equiv \mathbf{W}$ (motifs spectraux normalisés) et $P(z)P(t|z) \equiv \mathbf{H}$ (activations temporelles)

Modèles à composantes latentes

- En supposant tout les tirages i indépendants, la log-Vraisemblance s'écrit

$$L = \sum_i \log P(f_i, t_i)$$

- $\mathbf{X}$ est l'histogramme $\rightarrow$ le tirage (f, t) apparaît X_{ft} fois dans la somme L
 $\rightarrow$

$$L = \sum_{f,t} X_{ft} \log P(f, t)$$

- Maximiser la vraisemblance $L \equiv$ minimiser la divergence KL

Intérêt de ces modèles

- Calcul non heuristique de la fonction de coût
- Possibilité de fixer des hypothèses directement dans le modèle
- Estimation basée sur l'algorithme EM

Estimation : algorithme EM

- On souhaite estimer les paramètres θ du modèle (ici les $\mathbf{W}$ et les $\mathbf{H}$ par maximisation de la log-vraisemblance des données x :

$$L(\theta) = \log(P(x; \theta))$$

.

→ Complicé analytiquement sur des cas pratiques.

- On utilise alors l'approche MM (Minoration Maximisation) : on introduit à l'itération (i) une fonction auxiliaire $l(\theta, \theta^{(i)})$ qui minore $L(\theta)$.
- La maximum de $l(\theta, \theta^{(i)})$ par rapport à θ conduit à l'estimation $\theta^{(i+1)}$, et on réitère la procédure.

→ Comment trouver la fonction minorante l ?

Estimation : algorithme EM

- Idée principale : utiliser les variables cachées z (par exemple les sources).
- On peut alors montrer que

$$l(\theta, \theta^{(i)}) = \mathbb{E}_{Z/X; \theta^{(i)}}(\log(p(x, z; \theta))) + \text{constante}$$

minore la log-vraisemblance et vérifie $l(\theta^{(i)}, \theta^{(i)}) = L(\theta^{(i)})$.

- L'algo EM est donc :
 - on initialise les paramètres θ au hasard : $\theta^{(0)}$
 - E : On calcul la vraisemblance jointe des données et des états étant donné les paramètres $\theta^{(i-1)}$

$$\mathbb{E}_{Z/X; \theta^{(i)}}(\log(x, z; \theta))$$

- M : on maximise la vraisemblance pour obtenir un nouveau jeu de paramètres

$$\theta^{(i+1)} = \underset{\theta}{\operatorname{Argmax}} \mathbb{E}_{Z/X; \theta^{(i)}}(\log(x, z; \theta))$$

Filtrage de Wiener

Rappels sur filtrage de Wiener (simple) :

- on observe $y(n) = x(n) + b(n)$: somme de processus mutuellement indépendants
- on cherche le filtre g tel que $y \star g \approx x$ (au sens des MC)
- résultat : $\hat{g} = \frac{S_{xx}}{S_{xx} + S_{bb}}$

Utilisation en NMF : masquage fréquentiel

Pour trouver la source k :

1. La TFCT d'un mélange est $\mathbf{Z}$, et $\mathbf{X} = |\mathbf{Z}|^2$
2. NMF : $\mathbf{X} \approx \mathbf{W}\mathbf{H}$
3. Masquage : $\mathbf{Y}_k = \frac{\mathbf{w}_k \mathbf{h}_k}{\mathbf{W}\mathbf{H}} \circledast \mathbf{Z}$
4. $y_k(n)$ obtenue par inversion de TFCT $\mathbf{Y}_k$

NMF : modèle de voix (J-L. Durrieu)

- *Représentation* temps-fréquences des données : Voix principale + accompagnement *Musical*

$$R(f, t) = V_{ft} + M_{ft}$$

- Modèle probabiliste pour V_{ft} :
 - Un état = une combinaison source/filtre (filtre k , fréquence fondamentale u) = une v.a. gaussienne circulaire

$$\sigma_{kut}^2(f) = b_{kut} \underbrace{\phi_k(f) e_u(f)}_{\text{un état}}$$

- $V_{ft} \sim$ mélange de gaussiennes
- l'excitation est calculée à l'aide d'un modèle d'onde glottique
- modèle = structuration de la variance
- accompagnement : $M_{ft} \sim \mathcal{N}_c(0, \sum_r h_{rt} w_r(f))$ (NMF « classique »)
- Optimisation : algorithme EM $\rightarrow$ mises à jour multiplicatives sur les paramètres

NMF : modèle de voix (J-L. Durrieu)

- Autre modèle possible

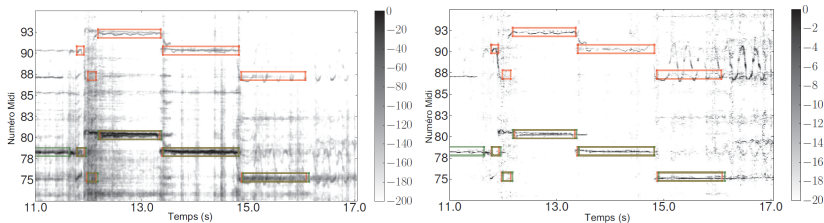
$$V_{ft} \sim \mathcal{N}_c \left(0, \left(\sum_k b_{kt}^{\phi} \phi_k(f) \right) \left(\sum_u b_{ut}^e e_u(f) \right) \right)$$

- Optimisation : algorithme EM $\rightarrow$ mises à jour multiplicatives

$$\theta \leftarrow \theta \frac{\nabla_{\theta}^{-}}{\nabla_{\theta}^{+}}$$

NMF : modèle de voix (J-L. Durrieu)

Résultats : Représentation mi-niveau (guitare et voix)



	Mix	Accompagnement	Voix
Original			
Estimé			

NMF : Structuration du spectre (R. Hennequin)

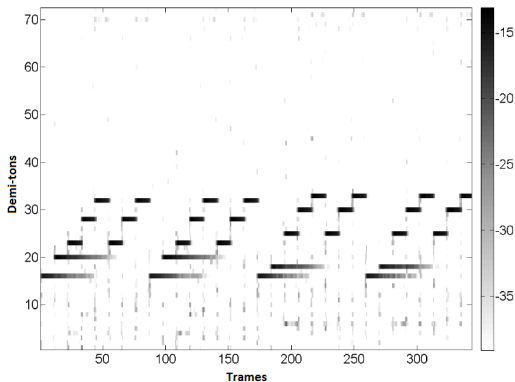
- On cherche souvent des notes = spectres harmoniques (ou quasi) $\approx$ somme de partiels
- $\rightarrow$ modèle additif d'atome spectral : pour une source

$$W_f^r(t) = \sum_{k=1}^{N_r} a_k g(f - k f_0^r(t))$$

- g est la TF de la fenêtre d'analyse
- permet de traduire **les modulations de** f_0
- calcul possible de la mise en jour en réduisant g au lobe principal

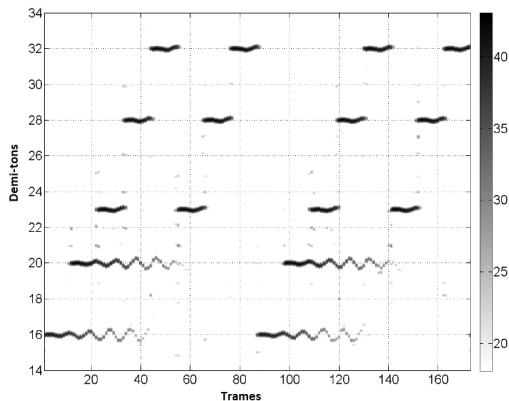
NMF : Structuration du spectre (R. Hennequin)

Résultats : exemple du premier prélude de Bach au synthétiseur



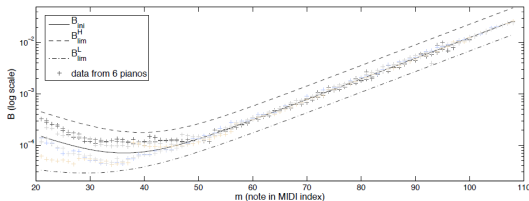
NMF : Structuration du spectre (R. Hennequin)

Résultats : exemple du premier prélude de Bach au synthétiseur



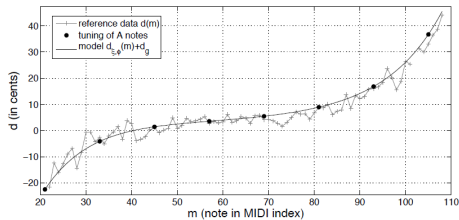
NMF : modèle pour le piano (F. Rigaud)

- Piano : cas particulier de spectres inharmoniques
 - partiels $f_n = nF_0(1 + bn^2)^{1/2}$
 - influence sur l'accord → courbes de « Railsback » (extension des octaves)



NMF : modèle pour le piano (F. Rigaud)

- Piano : cas particulier de spectres inharmoniques
 - partiels $f_n = nF_0(1 + bn^2)^{1/2}$
 - influence sur l'accord → courbes de « Railsback » (extension des octaves)



NMF : modèle pour le piano (F. Rigaud)

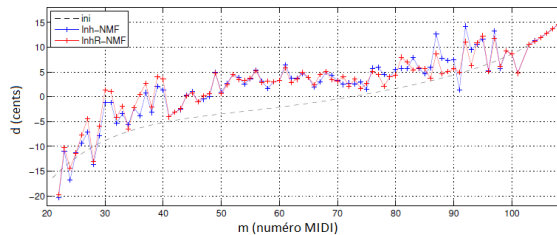
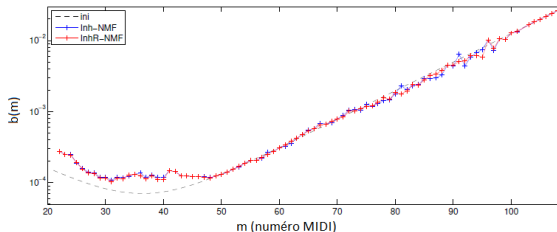
- Atome spectraux avec modèle additif

$$W_f^r = \sum_{k=1}^{N_r} a_k g(f - f_{kr}).$$

- pas de variation temporelle
 - structuration du spectre via les f_{kr}
- Deux méthodes
 - NMF inharmonique $f_{kr}^{Inh} = kF_0^r(1 + b_r k^2)^{1/2}$
 - NMF inharmonique relâchée : contrainte dans la fonction de coût : $D_\beta(\mathbf{R}|\mathbf{WH}) + \lambda \sum_{k,r} (f_{kr} - f_{kr}^{Inh})^2$.

NMF : modèle pour le piano (F. Rigaud)

Résultats :



« Does inharmonicity improve ... ? »

La réponse est « Yes » :

→ comparaison NMF-harmonique, strictement inharm. et inharm. relâchée.

	N_r	5	10	20	30
Ha	$\mathcal{P}$	39.3	51.9	58.9	62.1
	$\mathcal{R}$	41.0	49.3	56.6	60.6
	$\mathcal{F}$	38.7	49.0	55.9	59.7
Inh_1	$\mathcal{P}$	32.6	44.6	54.9	55.6
	$\mathcal{R}$	34.5	46.0	56.7	57.7
	$\mathcal{F}$	32.4	43.8	54.0	54.8
$InhR_1$	$\mathcal{P}$	34.0	43.0	53.4	55.2
	$\mathcal{R}$	36.0	44.9	55.7	57.1
	$\mathcal{F}$	33.7	42.6	52.7	54.4
Inh_2	$\mathcal{P}$	44.3	59.6	66.4	66.9
	$\mathcal{R}$	45.1	55.9	60.9	62.5
	$\mathcal{F}$	43.0	55.8	61.5	62.6
$InhR_2$	$\mathcal{P}$	43.3	57.5	64.1	64.6
	$\mathcal{R}$	45.2	54.9	60.3	61.1
	$\mathcal{F}$	42.5	54.1	60.2	60.7

- $\mathcal{P}$ = Précision
- $\mathcal{R}$ = Rappel
- $\mathcal{F}$ = F-mesure
- initialisation 1 = "à plat",
- initialisation 2 = moyenne sur plusieurs piano B, F_0

Conclusion

- NMF : une approche « bottom-up », guidée par les données
- cependant : nécessité d'informer la NMF, d'inclure des à priori, des contraintes...
- Un mot du TP