A TUTORIAL ON NONNEGATIVE MATRIX FACTORISATION WITH APPLICATIONS TO AUDIOVISUAL CONTENT ANALYSIS

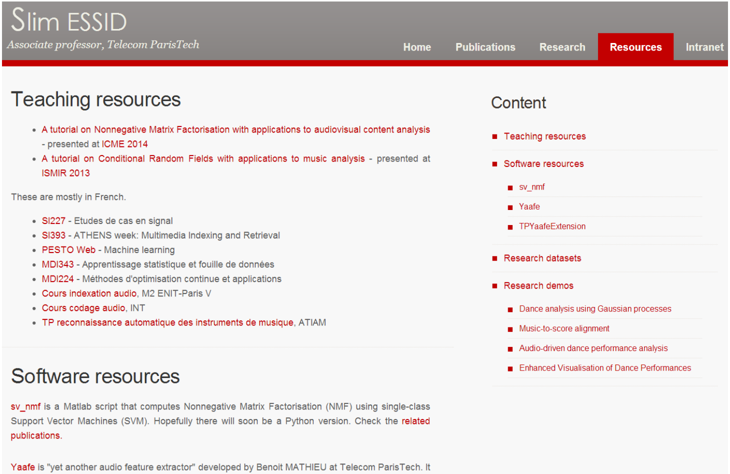
Slim ESSID & Alexey OZEROV

Telecom ParisTech / Technicolor

July 2014

Slides available online

http://www.telecom-paristech.fr/~essid/resources.htm



Support

The tutorial is partially supported by the European projects:

- FP7 AXES (Access to Audiovisual Archives) http://www.axes-project.eu
- FP7 REVERIE (REal and Virtual Engagement in Realistic Immersive Environments) http://www.reveriefp7.eu/

Credits

Some illustrations, slides and demos are reproduced courtesy of:

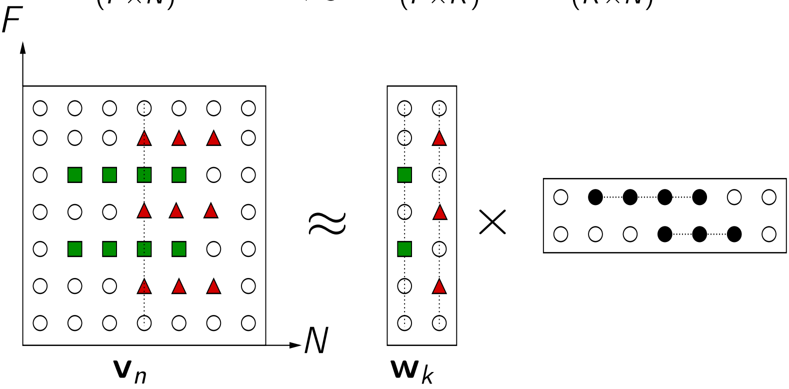
- C. Févotte,
- N. Seichepine,
- A. Masurelle,
- R. Hennequin,
- F. Vallet,
- A. Liutkus,
- G. Richard,
- E. Vincent,
- F. Bimbot,
- N. Q. K. Duong,
- D. El Badawy,
- L. Le Magoarou,
- L. Chevallier,
- J. Sirot,
- V. D. Blondel,
- L. de Vinci.

- Introduction
- NMF models
- Algorithms for solving NMF
- Constrained NMF schemes
- Multi-stream and cross-modal NMF schemes
- Applications
- Conclusion

Explaining data by factorisation

General formulation

$$\mathbf{V}_{(F\times N)} \approx \mathbf{W}_{(F\times K)} \times \mathbf{H}_{(K\times N)}$$



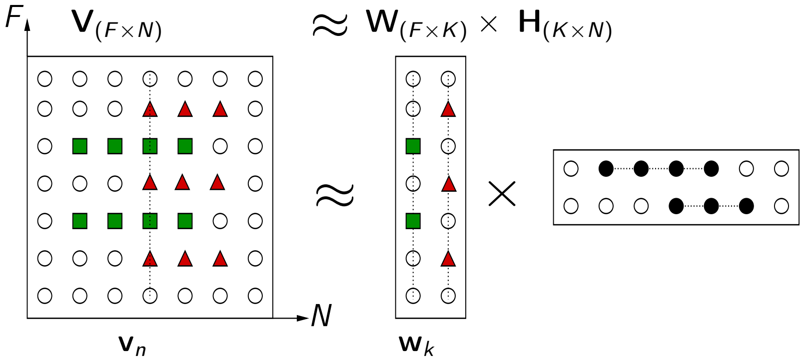
$$\mathbf{v}_n \approx \sum_{k=1}^{K} h_{kn}\mathbf{w}_k$$

Illustration by C. Févotte

Explaining data by factorisation

General formulation

$$\mathbf{V}_{(F\times N)} \approx \mathbf{W}_{(F\times K)} \times \mathbf{H}_{(K\times N)}$$



data matrix

"explanatory variables"
"basis", "dictionary",
"patterns", "topics"

"regressors",
"activation coefficients",
"expansion coefficients"

Illustration by C. Févotte

Principal Component Analysis (PCA)

Recalling the technique[1]

Assuming the data is real-valued ($\mathbf{v}_n \in \mathbb{R}^F$) and centered ($\mathbb{E}[\mathbf{v}] = 0$),

- PCA returns a dictionary $\mathbf{W}_{PCA} \in \mathbb{R}^{F \times K}$ such that the **least squares error** is minimized:

$$\mathbf{W}_{PCA} = \min_{\mathbf{W}} \frac{1}{N} \sum_n \|\mathbf{v}_n - \hat{\mathbf{v}}_n\|_2^2 = \frac{1}{N} \|\mathbf{V} - \mathbf{W}\mathbf{W}^T\mathbf{V}\|_F^2$$

- A solution is given by:

$$\mathbf{W}_{PCA} = \mathbf{E}_{1:K}$$

where $\mathbf{E}_{1:K}$ denotes the K dominant **eigenvectors** of $\mathbf{C}_{\mathbf{v}}$:

$$\mathbf{C}_{\mathbf{v}} = \mathbb{E}[\mathbf{v}\mathbf{v}^T] \approx \frac{1}{N} \sum_n \mathbf{v}_n \mathbf{v}_n.$$

[1] slide adapted from (Févotte, 2012).

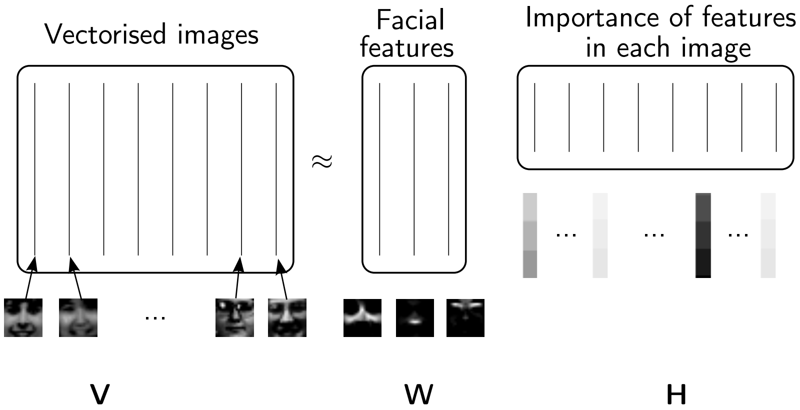
Explaining face images by PCA[2]

Image example: 49 images among 2429 from MIT's CBCL face dataset

[2]slide adapted from (Févotte, 2012).

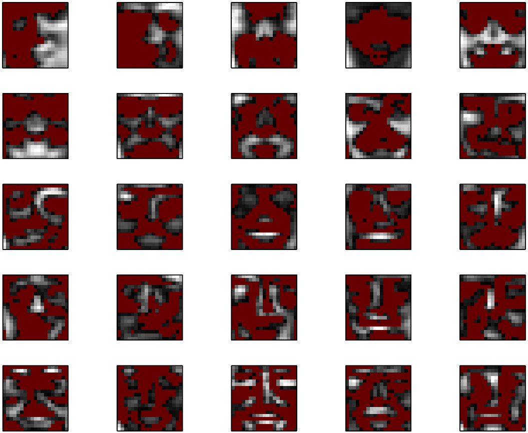
Explaining face images by PCA

Method



Explaining face images by PCA[3]

Eigenfaces



*Red pixels indicate **negative values**! How to interpret this?*

[3] slide adapted from (Févotte, 2012).

Data is often nonnegative by nature[4]

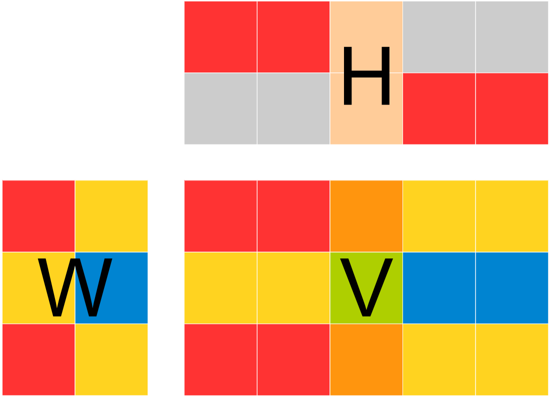
- pixel intensities;
- amplitude spectra;
- occurrence counts;
- food or energy consumption;
- user scores;
- stock market values;
- ...

For the sake of **interpretability** of the results, optimal processing of **nonnegative data** may call for processing under **nonnegativity constraints**.

[4] slide adapted from (Févotte, 2012).

The Nonnegative Matrix Factorisation model

NMF provides an unsupervised linear representation of the data:



$$\mathbf{V} \approx \mathbf{W}\mathbf{H};$$

- $\mathbf{W} = [w_{fk}]$ s.t. $w_{fk} \geq 0$ and
- $\mathbf{H} = [h_{kn}]$ s.t. $h_{kn} \geq 0$.

Illustration by N. Seichepine

Why nonnegative factors?

- Nonnegativity induces **sparsity**.
- Nonnegativity leads to **part-based decompositions**.

"Atoms energy cancellation" is not allowed: once an atom is selected with some energy, it cannot be further concealed by other atoms.

NMF outputs

Image example

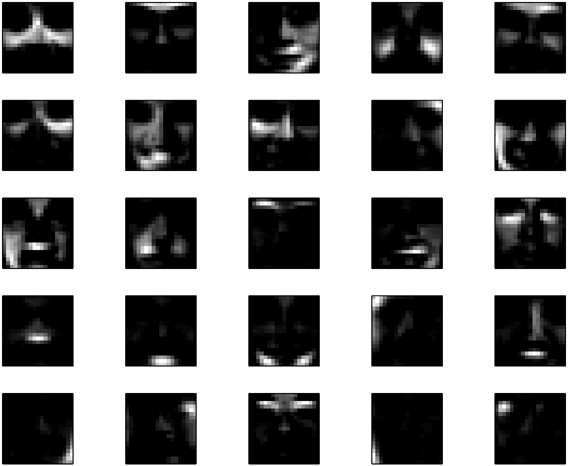


Illustration by C. Févotte

NMF outputs
Audio example

NMF produces **part-based** representations of the data:

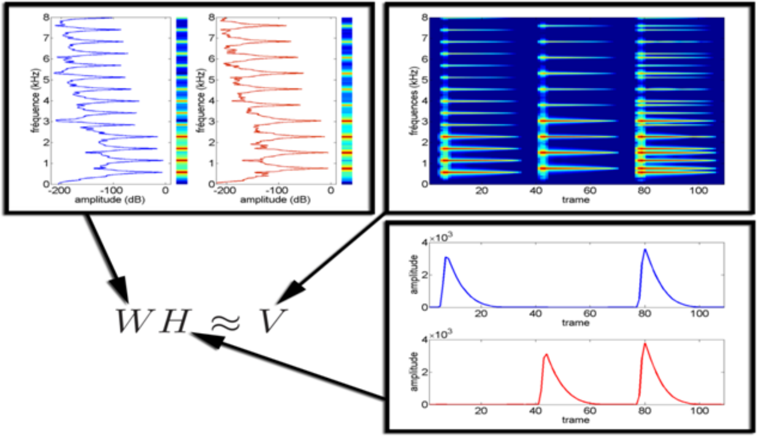


Illustration by R. Hennequin.

History

NMF is more than **30-year** old!

- previous variants referred to as:
  - **nonnegative rank fatorisation** (Jeter and Pye, 1981; Chen, 1984);
  - **positive matrix factorisation** (Paatero and Tapper, 1994);
- popularized by Lee and Seung (1999) for "**learning the parts of objects**".

Since then, widely used in various research areas for diverse applications.

Notations I

- $\mathbf{V}$: the $F \times N$ **data matrix**:
  - F features (rows),
  - N observations/examples/feature vectors (columns);
- $\mathbf{v}_n = (v_{1n}, \cdots, v_{Fn})^T$: the n-th **feature vector** observation among a collection of N observations $\mathbf{v}_1, \cdots, \mathbf{v}_N$;
- $\mathbf{v}_n$ is a column vector in $\mathbb{R}_+^F$; $\mathbf{v}_n$ is a row vector;

- $\mathbf{W}$: the $F \times K$ **dictionary matrix**:
  - w_{fk} is one of its coefficients,
  - $\mathbf{w}_k$ a dictionary/basis vector among K elements;

Notations II

- **H** : the $K \times N$ **activation/expansion matrix:**
  - $\mathbf{h}_n$: the **column vector** of activation coefficients for observation $\mathbf{v}_n$:

$$\mathbf{v}_n \approx \sum_{k=1}^{K} h_{kn} \mathbf{w}_k \ ;$$

  - $\mathbf{h}_{k:}$: the **row vector** of activation coefficients relating to basis vector $\mathbf{w}_k$.

General usages of NMF I

What for?

NMF is a non-supervised data decomposition technique, akin to **latent variable analysis**, that can be used for:

- **feature learning**: like Principal Component Analysis (PCA);
  - learn NMF on training dataset $\mathbf{V}_{train} \rightarrow$ dictionary $\mathbf{W}$
  - exploit $\mathbf{W}$ to decompose new test examples $\mathbf{v}_n$: $\mathbf{v}_n \approx \sum_{k=1}^{K} h_{kn}\mathbf{w}_k$; $h_{kn} \geq 0$
  - use $\mathbf{h}_n$ as **feature vector** for example n.

Evaluation for face recognition:

- **Dataset**: Olivetti faces, 40 classes
- **Classifiers**: LDA (Linear Discriminant Analysis)
- **Cross-validated results**:

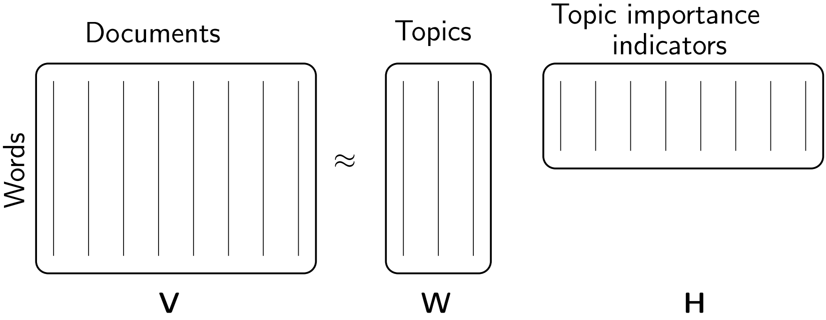
	Accuracy
PCA	93%
ICA	93%
NMF	**96%**

General usages of NMF II

What for?

- **topics recovery**:

  assume $\mathbf{V} = [v_{fn}]$ is a (scaled) **term-document** co-occurrence matrix: v_{fn} is the frequency of occurrences of word m_f in document d_n;



Topics recovery

NMF link to Probabilistic Latent Semantic Analysis (PLSA)

- **topics recovery**: like Probabilistic Latent Semantic Analysis (PLSA):

  assume $\mathbf{V} = [v_{fn}]$ is a (scaled) **term-document** co-occurrence matrix:
  v_{fn} is the frequency of occurrences of word m_f in document d_n;

PLSA model (Hofmann, 1999)

$$P(m_f, d_n) = \sum_{k=1}^{K} P(t_k) P(d_n|t_k) P(m_f|t_k)$$

$\rightarrow$ *the documents can be explained by some underlying topics t_k.*

Topics recovery

NMF link to Probabilistic Latent Semantic Analysis (PLSA)

- Let $w_{fk} = \hat{P}(t_k)\hat{P}(m_f|t_k)$ and $h_{kn} = \hat{P}(d_n|t_k)$;
- the model can be re-written as:

$$[\hat{P}(m_f, d_n)] = [\hat{v}_{fn}] = \mathbf{WH}$$

The $\mathbf{w}_k$ can be interpreted as **topics** explaining the data being analyzed to the extent given by related $\mathbf{h}_{k:}$.

Link between NMF and PLSA (Gaussier and Goutte, 2005)

- Any (local) maximum likelihood solution of PLSA is a solution of NMF with Kullback-Leibler (KL) divergence.
- Any solution of NMF with KL divergence yields a (local) maximum likelihood solution of PLSA.

Text document analysis example

After sklearn topics extraction demo (Pedregosa et al., 2011)

Analysing the 20 newsgroups dataset with NMF, the following topics are automatically determined:

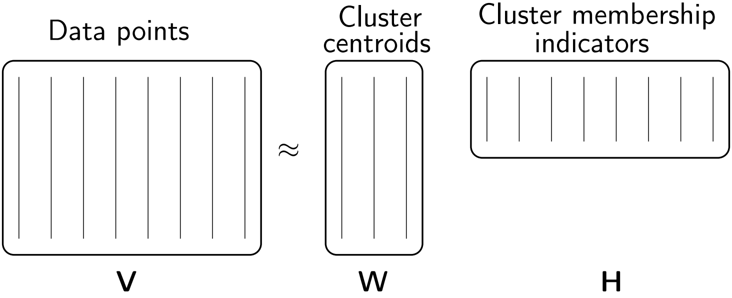
- **Topic #0:** god people bible israel jesus christian true moral think christians believe don say human israeli church life children jewish
- **Topic #1:** drive windows card drivers video scsi software pc thanks vga graphics help disk uni dos file ide controller work
- **Topic #2:** game team nhl games ca hockey players buffalo edu cc year play university teams baseball columbia league player toronto
- **Topic #3:** window manager application mit motif size display widget program xlib windows user color event information use events values
- **Topic #4:** pitt gordon banks cs science pittsburgh univ computer soon disease edu reply pain health david article medical medicine

Topics described by most frequent words in each dictionary element $\mathbf{W}_k$.

General usages of NMF III

What for?

- **clustering**: like K-means (Ding et al., 2005, 2010; Xu et al., 2003):



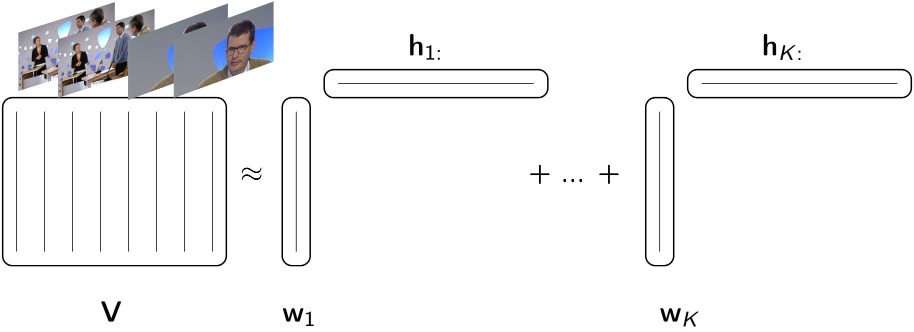
- NMF can handle overlapping clusters and provides *soft* cluster membership indications.

General usages of NMF IV

What for?

- **temporal segmentation**: like Hidden Markov Models (HMM);

  analysing temporal data sequences, *e.g.*, videos:

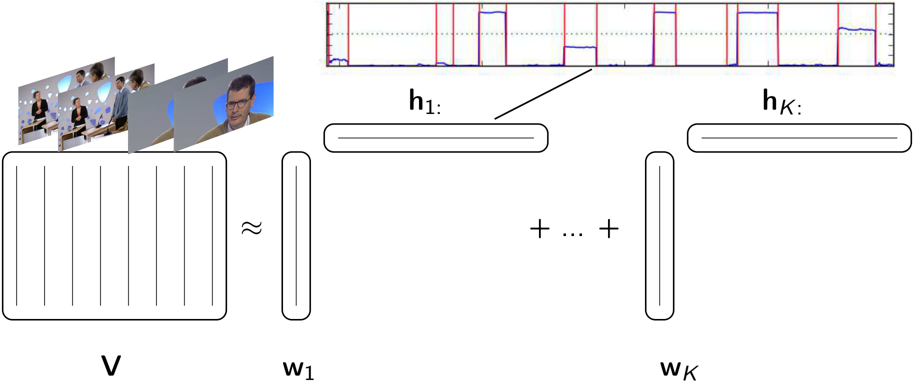


General usages of NMF IV

What for?

- **temporal segmentation**: like Hidden Markov Models (HMM);

  analysing temporal data sequences, *e.g.*, videos:

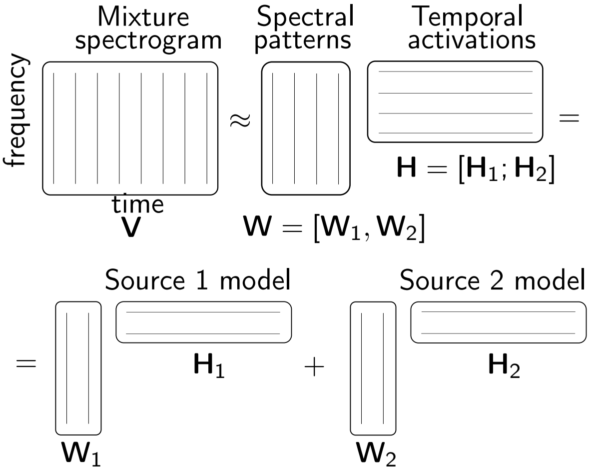


Temporal segmentation can be achieved by thresholding the temporal activations relating to components of interest.

General usages of NMF V

What for?

- **filtering and source separation**: as with Independent Component Analysis (ICA):



In summary...

What for?

NMF is a non-supervised data decomposition technique, akin to **latent variable analysis**, that can be used for:

- **topics recovery**: like Probabilistic Latent Semantic Analysis (PLSA);
- **feature learning**: like Principal Component Analysis (PCA);
- **clustering**: like K-means;
- **temporal segmentation**: like Hidden Markov Models (HMM);
- **filtering and source separation**: as with Independent Component Analysis (ICA);
- **coding** as with vector quantization.

Overview of NMF application domains I

A variety of successful applications:

- **Text mining**: (Xu et al., 2003; Berry and Browne, 2006; Kim and Park, 2008)
- **Images**:
  - unsupervised object discovery (Sivic et al., 2005)
  - object and face recognition (Soukup and Bajla, 2008)
  - tagging (Kalayeh et al., 2014)
  - denoising and inpainting (Mairal et al., 2010)
  - texture classification (Sandler and Lindenbaum, 2011)
  - spectral data (Berry et al.)
  - hashing (Monga and Mihcak, 2007)
  - watermarking (Lu et al., 2009)
- **Electroencephalography (EEG) data**:
  - feature extraction (Cichocki and Rutkowski, 2006; Lee et al., 2009)
  - artifact rejection (Damon et al., 2013a,b)

Overview of NMF application domains II

- **Bioinformatics**:
  - gene expression analysis (Brunet et al., 2004; Gao and Church, 2005)
  - protein interaction clustering (Greene et al., 2008)
- **Other**:
  - collaborative filtering (Melville and Sindhwai, 2010)
  - community discovery (Wang et al., 2010)
  - portfolio diversification (Drakakis et al., 2007)
  - food consumption analysis (Zetlaoui et al., 2010)
  - industrial source apportionment (Limem et al., 2013)
- **Audio and music**
- **Videos**

Audio and music processing

- **Source separation** (NMF is state-of-the art):
  - **speech**: separating voices in speech mixtures or voice from background (Virtanen, 2007; Virtanen and Cemgil, 2009; Mohammadiha et al., 2013)
  - **music**: separating singing voice/melody from accompaniment or musical instruments in polyphonic mixtures (Durrieu et al., 2009; Ozerov and Fevotte, 2010; Hennequin et al., 2011; Ozerov et al., 2013; Rafii et al., 2013)
- **Signal enhancement/denoising**:

  (Wilson et al., 2008; Schmidt et al., 2007; Sun and Mazumder, 2013)
- **Audio inpainting**

  (Roux et al., 2011; Yilmaz et al., 2011)

Audio and music processing

- **Compression**

  (Ozerov et al., 2011b; Nikunen et al., 2011)

- **Music transcription:** recognizing musical notes played by Piano, Drums or multiple instruments

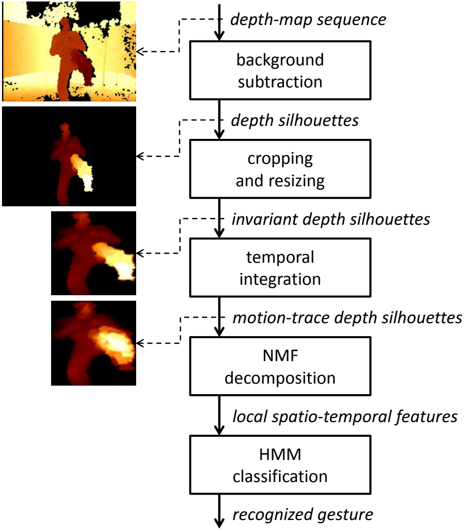
  (Smaragdis and Brown, 2003; Abdallah and Plumbley, 2004; Vincent et al., 2007; E. Vincent et al., 2008; Févotte et al., 2009; Bertin et al., 2010; Vincent et al., 2010)

Video processing

- NMF use for video processing remains quite limited, despite its potential.
- Known works:
  - Video summarization (Cooper and Foote, 2002)
  - Dynamic video content representation and scene change detection (Bucak and Gunsel, 2007)
  - Onscreen person spotting and shot-type classification (Essid and Fevotte, 2012, 2013)
  - Fingerprinting (Cirakman et al., 2010)
  - Action recognition (Krausz and Bauckhage, 2010; Masurelle et al., 2014)
  - Compression (Türkan and Guillemot, 2011)

Action recognition using depth silhouettes

Using NMF for feature learning (Masurelle et al., 2014)



Skeleton features	PCA	NMF
78%	89%	**91%**

Recognition accuracies

- considering Kinect recordings of 8 actions;
- using Huawei/3DLife grand challenge dataset for action recognition.

Video Structuring

Using NMF for temporal segmentation and soft-clustering (Essid and Fevotte, 2013)

Discovering the video editing structure (Essid and Fevotte, 2012)

Performing **speaker diarization** (Seichepine et al., 2013)

"Who spoke when?"

illustration by N. Seichepine

Using the **Canal9 political debates** database (Vinciarelli et al., 2009).

- Introduction
  - Motivation
  - First look at the model
  - General usages and applications
  - Difficulties in NMF
- NMF models
- Algorithms for solving NMF
- Constrained NMF schemes
- Multi-stream and cross-modal NMF schemes
- Applications
- Conclusions

Model order choice

A suitable choice of K is very important

Model order K corresponds to the number of rank-1 matrices within the approximation

The choice of K results in a compromise between

Data fitting

A greater K leads to a better data approximation

Model complexity

A smaller K leads to a less complex model (easier to estimate, less parameters to transmit, etc ...)

A right **model order choice is important** and it depends on the data $\mathbf{V}$ and on the application.

NMF is ill-posed

The solution is not unique

Given $\mathbf{V} = \mathbf{WH}$; $\mathbf{W} \geq 0$, $\mathbf{H} \geq 0$; any matrix $\mathbf{Q}$ such that:

- $\mathbf{WQ} \geq 0$
- $\mathbf{Q}^{-1}\mathbf{H} \geq 0$

provides an alternative factorisation $\mathbf{V} = \tilde{\mathbf{W}}\tilde{\mathbf{H}} = (\mathbf{WQ})(\mathbf{Q}^{-1}\mathbf{H})$.

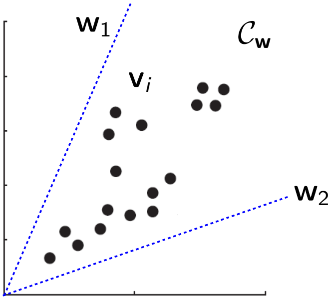
In particular, $\mathbf{Q}$ can be any **nonnegative generalised permutation matrix**; *e.g.*, in $\mathbb{R}^3$:

$$\mathbf{Q} = \begin{bmatrix} 0 & 0 & 2 \\ 0 & 3 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

This case is not so problematic: merely accounts for **scaling** and **permutation** of basis vectors $\mathbf{w}_k$.

Geometric interpretation and ill-posedness

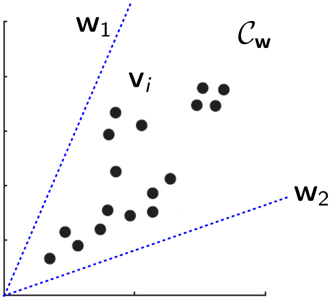
NMF assumes the data is well described by a **simplicial convex cone** $\mathcal{C}_{\mathbf{w}}$ generated by the columns of $\mathbf{W}$:



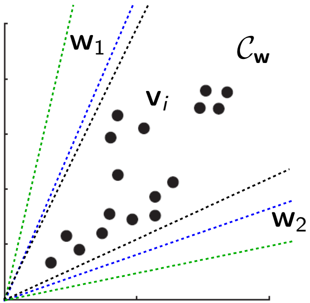
$$\mathcal{C}_{\mathbf{w}} = \left\{ \sum_{k=1}^{K} \lambda_k \mathbf{w}_k ;\ \lambda_k \geq 0 \right\}$$

Geometric interpretation and ill-posedness

NMF assumes the data is well described by a **simplicial convex cone** $\mathcal{C}_{\mathbf{w}}$ generated by the columns of $\mathbf{W}$:



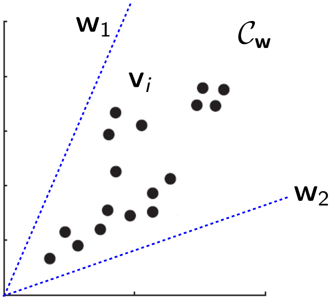
$$\mathcal{C}_{\mathbf{w}} = \left\{ \sum_{k=1}^{K} \lambda_k \mathbf{w}_k ; \; \lambda_k \geq 0 \right\}$$

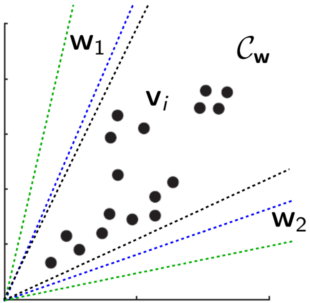


Problem: which $\mathcal{C}_{\mathbf{w}}$?

Geometric interpretation and ill-posedness

NMF assumes the data is well described by a **simplicial convex cone** $\mathcal{C}_{\mathbf{w}}$ generated by the columns of $\mathbf{W}$:





$$\mathcal{C}_{\mathbf{w}} = \left\{ \sum_{k=1}^{K} \lambda_k \mathbf{w}_k;\ \lambda_k \geq 0 \right\}$$

Problem: which $\mathcal{C}_{\mathbf{w}}$?

$\rightarrow$ Need to impose **constraints** on the set of possible solutions to select the most "useful" ones.

Constrained NMF methods

Different types of constraints have been considered in previous works:

- **Sparsity** constraints: either on $\mathbf{W}$ or $\mathbf{H}$ (*e.g.*, Hoyer, 2004; Eggert and Korner, 2004);
- **Shape** constraints on $\mathbf{w}_k$, *e.g.*:
  - **convex NMF**: $\mathbf{w}_k$ are convex combinations of inputs (Ding et al., 2010);
  - **harmonic NMF**: $\mathbf{w}_k$ are mixtures of harmonic spectra (Vincent et al., 2008).
- **Spatial coherence** or **temporal** constraints on $\mathbf{h}_k$: activations are **smooth** (Virtanen, 2007; Jia and Qian, 2009; Essid and Fevotte, 2013);
- **Cross-modal correspondence** constraints: factorisations of related modalities are related, *e.g.*, temporal activations are correlated (Seichepine et al., 2013; Liu et al., 2013; Yilmaz et al., 2011);
- **Geometric** constraints: *e.g.*, select particular cones $\mathcal{C}_{\mathbf{w}}$ (Klingenberg et al., 2009; Essid, 2012).

- Introduction
- NMF models
  - Cost functions
  - Weighted NMF schemes
- Algorithms for solving NMF
- Constrained NMF schemes
- Multi-stream and cross-modal NMF schemes
- Applications
- Conclusion

NMF optimization criteria

NMF approximation $\mathbf{V} \approx \mathbf{WH}$ is usually obtained through:

$$\min_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH}),$$

where $D(\mathbf{V}|\widehat{\mathbf{V}})$ is a *separable matrix divergence*:

$$D(\mathbf{V}|\widehat{\mathbf{V}}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d(v_{fn}|\hat{v}_{fn}),$$

and $d(x|y)$ defined for all $x, y \geq 0$ is a *scalar divergence* such that:

- $d(x|y)$ is continuous over x and y;
- $d(x|y) \geq 0$ for all $x, y \geq 0$;
- $d(x|y) = 0$ if and only if $x = y$.

Popular (scalar) divergences

Euclidean (EUC) distance (Lee and Seung, 1999)

$$d_{EUC}(x, y) = (x - y)^2$$

Kullback-Leibler (KL) divergence (Lee and Seung, 1999)

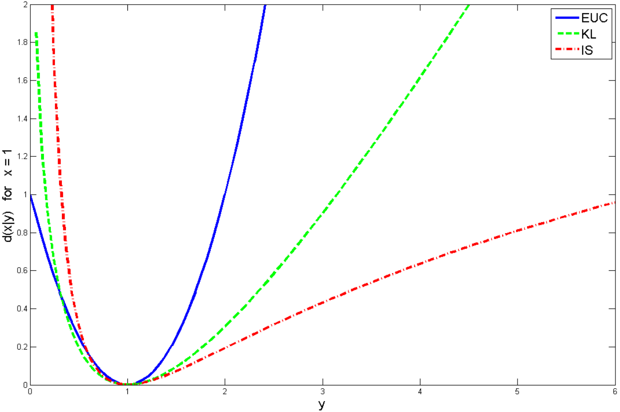
$$d_{KL}(x, y) = x \log \frac{x}{y} - x + y$$

Itakura-Saito (IS) divergence (Févotte et al., 2009)

$$d_{IS}(x, y) = \frac{x}{y} - \log \frac{x}{y} - 1$$

Convexity properties

Divergence $d(x\|y)$	EUC	KL	IS
Convex on x	yes	yes	yes
Convex on y	yes	yes	**no**



Scale invariance properties[5]

$$
\begin{aligned}
d_{EUC}(\lambda x|\lambda y) &= \lambda^2 d_{EUC}(x|y) \\
d_{KL}(\lambda x|\lambda y) &= \lambda d_{KL}(x|y) \\
d_{IS}(\lambda x|\lambda y) &= d_{IS}(x|y)
\end{aligned}
$$

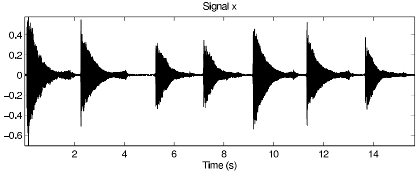
The IS divergence is **scale-invariant** $\rightarrow$ it provides higher accuracy in the representation of data with large dynamic range, such as audio spectra.

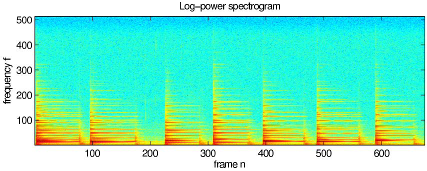
[5] slide adapted from (Févotte, 2012).

Music transcription demo

Demo slide courtesy of C. Févotte (Fevotte et al., 2009)

(MIDI numbers: 61, 65, 68, 72)



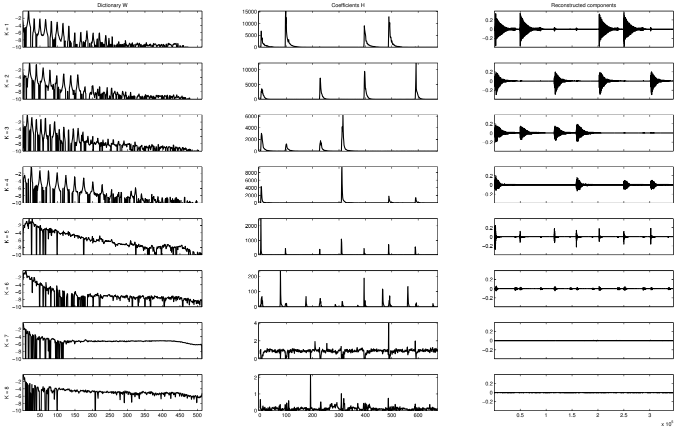


Three representations of the data.

Music transcription demo

Demo slide courtesy of C. Févotte (Fevotte et al., 2009)

NMF decomposition with $K = 8$



Pitch estimates: 65.0 68.0 61.0 72.0 0 0 0 0

(True values: 61, 65, 68, 72)

General parametric families of divergences

β-divergence (Eguchi and Kano., 2001)

$$d_\beta(x|y) = \begin{cases} \frac{1}{\beta(\beta-1)}\left(x^\beta + (\beta-1)y^\beta - \beta\, x\, y^{\beta-1}\right) & \beta \in \mathbb{R}\backslash\{0,1\} \\ x \log \frac{x}{y} - x + y & \beta = 1 \\ \frac{x}{y} - \log \frac{x}{y} - 1 & \beta = 0 \end{cases}$$

Generalizes IS ($\beta = 0$), KL ($\beta = 1$) divergences and EUC ($\beta = 2$) distance.

α-divergence (Cichocki et al., 2006, 2008)

$$d_\alpha(x|y) = \frac{1}{\alpha(\alpha-1)}\left(\alpha x + (1-\alpha)y - x^\alpha y^{1-\alpha}\right)$$

And many others ...

- Separable divergences:
  - Csiszar's divergence (generalizes α-divergence) (Cichocki et al., 2006)
  - Bregman divergence (generalizes β-divergence) (Bregman, 1967; I. S. Dhillon and S. Sra, 2005)
  - $\alpha\beta$-divergence (A. Cichocki et al., 2011)
  - etc ...
- Nonseparable divergences:
  - γ-divergence (Fujisawa and Eguchi, 2008)
  - ρ-(Rényi's) divergence (Devarajan and Ebrahimi, 2005)
  - etc ...

Which divergence to choose?

NMF divergence choice depends on the **data** and on the **application.**

One can choose the divergence as follows:

- by **intuition** or from some **prior knowledge of the application goal** (*e.g.*, NMF is used for predicting the unseen data while minimizing the mean squared error $\Longrightarrow$ EUC distance) or **invariances** (*e.g.*, scale invariance for music analysis with IS divergence) ;
- from some **probabilistic considerations** (presented in the upcoming section);
- **optimize the divergence** (*e.g.* from some parametric family) on some development data within a particular application.

Statistical viewpoint

For many divergences a probabilistic formulation is possible: the divergence minimization becomes equivalent to a maximum likelihood criterion (Févotte et al., 2009; Cemgil, 2009b):

$$D(\mathbf{V}|\widehat{\mathbf{V}}) = -\log p(\mathbf{V}|\widehat{\mathbf{V}}) + const$$

Examples:

	Divergence $D(\mathbf{V}\vert\widehat{\mathbf{V}})$	Probability distribution	p.d.f. $p(\mathbf{V}\vert\widehat{\mathbf{V}})$
EUC	$\sum_{f,n} (v_{fn} - \hat{v}_{fn})^2$	$v_{fn} \sim \text{Gaussian}\left(\hat{v}_{fn}, \sigma^2\right)$	$\prod_{f,n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(v_{fn}-\hat{v}_{fn})^2}{2\sigma^2}\right)$
KL	$\sum_{f,n} \left(v_{fn} \log \frac{v_{fn}}{\hat{v}_{fn}} - v_{fn} + \hat{v}_{fn}\right)$	$v_{fn} \sim \text{Poisson}\left(\hat{v}_{fn}\right)$	$\prod_{f,n} \frac{1}{\Gamma(v_{fn}+1)} \hat{v}_{fn}^{v_{fn}} \exp\left(-\hat{v}_{fn}\right)$
IS	$\sum_{f,n} \left(\frac{v_{fn}}{\hat{v}_{fn}} - \log \frac{v_{fn}}{\hat{v}_{fn}} - 1\right)$	$v_{fn} \sim \text{Exponential}\left(\frac{1}{\hat{v}_{fn}}\right)$	$\prod_{f,n} \frac{1}{\hat{v}_{fn}} \exp\left(-\frac{v_{fn}}{\hat{v}_{fn}}\right)$

Statistical viewpoint

Numerous advantages of a probabilistic NMF formulation:

- possibility of using efficient **probabilistic inference algorithms** such as the Expectation-Maximization (EM) algorithm (Févotte et al., 2009) and the Monte Carlo methods (Cemgil, 2009b; Schmidt et al., 2009);
- possibility of **introducing various constraints** into NMF modeling via prior distributions (Arngren et al., 2011);
- possibility of learning the NMF from **partially missing** (Roux et al., 2011) or **noisy** (Arberet et al., 2012) data;
- possibility of combining the NMF with **different probabilistic models** (Ozerov et al., 2012), *e.g.*, the hidden Markov models (HMMs) (Ozerov et al., 2009).

Weighted NMF

Conventional NMF optimization criterion (separable divergence case):

$$\min_{\mathbf{W},\mathbf{H}\geq 0} \sum_{f=1}^{F} \sum_{n=1}^{N} d(v_{fn}|\hat{v}_{fn}) \, .$$

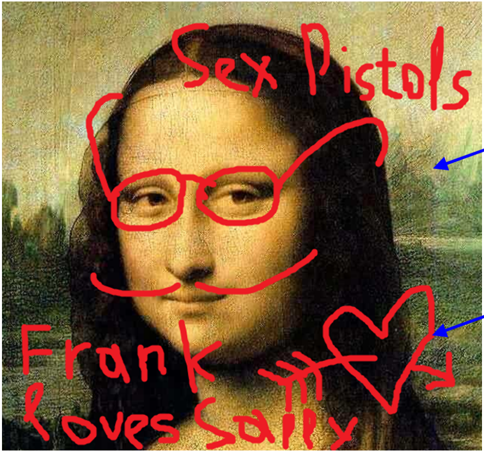
Weighted NMF optimization criterion:

$$\min_{\mathbf{W},\mathbf{H}\geq 0} \sum_{f=1}^{F} \sum_{n=1}^{N} b_{fn} d(v_{fn}|\hat{v}_{fn}) \, ,$$

where b_{fn} ($f = 1, \ldots, F$, $n = 1, \ldots, N$) are some nonnegative weights representing the contribution of data point v_{fn} into NMF learning.

Weighted NMF application example I

Learning from partial observations (e.g., for **image inpainting** as in (Mairal et al., 2010)):



Observed value

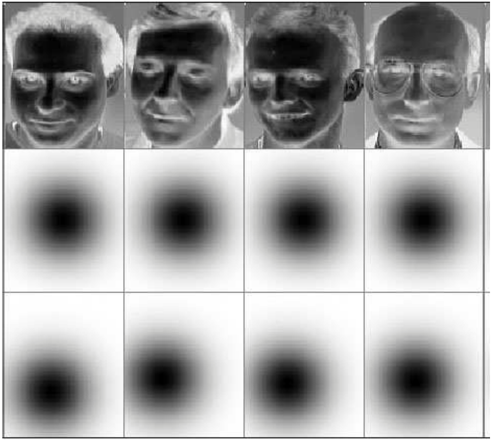
$b_{fn} = 1$

Missing value

$b_{fn} = 0$

Weighted NMF application example II

Face feature extraction (example and figure from (Blondel et al., 2008)):



Data $\mathbf{V}$

Weights $\mathbf{B} = \{b_{fn}\}_{f,n}$

Image-centered weights

Face-centered weights

- Introduction
- NMF models
- Algorithms for solving NMF
  - Preliminaries
  - Multiplicative update rules
  - Model order selection, initialization and stopping criteria
- Constrained NMF schemes
- Multi-stream and cross-modal NMF schemes
- Applications
- Conclusion

Optimization difficulties

An efficient solution of the NMF optimization problem

$$\min_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH}) \Leftrightarrow \min_{\boldsymbol{\theta}} C(\boldsymbol{\theta})\,;\ \ C(\boldsymbol{\theta}) \stackrel{\text{def}}{=} D(\mathbf{V}|\mathbf{WH})$$

(where $\boldsymbol{\theta} \stackrel{\text{def}}{=} \{\mathbf{W}, \mathbf{H}\}$ denotes the NMF parameters) must cope with the following difficulties:

- the **nonnegativity constraints** must be taken into account;
- **no uniqueness** of the solution is guaranteed in general;
- the optimization problem has usually a **multitude of local and global minima**.

Alternating optimization strategy

The problem is usually easier to optimize over one matrix (say $\mathbf{H}$) given the other matrix (say $\mathbf{W}$) is known and fixed.

Indeed, for several divergences $D(\mathbf{V}|\mathbf{WH})$ is even convex separately w.r.t. $\mathbf{H}$ and w.r.t. $\mathbf{W}$, but not w.r.t. $\{\mathbf{W}, \mathbf{H}\}$.

For this reason many state-of-the-art NMF optimization algorithms rely on the following iterative alternating optimization strategy.

Alternating optimization a.k.a block-coordinate descent (one iteration):

- update $\mathbf{W}$, given $\mathbf{H}$ fixed,
- update $\mathbf{H}$, given $\mathbf{W}$ fixed.

Multiplicative update rules

A heuristic approach introduced by (Lee and Seung, 2001) to solve $\min_{\boldsymbol{\theta}} C(\boldsymbol{\theta})$

Multiplicative update (MU) rule for H (similarly for W) is defined as:

$$h_{kn} \leftarrow h_{kn} \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_{-} / \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_{+},$$

where

$$\nabla_{h_{kn}} C(\boldsymbol{\theta}) = \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_{+} - \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_{-} \ ,$$

and the summands are both nonnegative.

NOTE: The nonnegativity of **W** and **H** is guaranteed by construction.

MU rules for the β-divergence

For example, in the case of the β-divergence (generalizing the three popular divergences) the following decomposition:

$$\nabla_y d_\beta(x|y) = \underbrace{y^{\beta-1}}_{[\nabla_y d_\beta(x|y)]_+} - \underbrace{x\, y^{\beta-2}}_{[\nabla_y d_\beta(x|y)]_-}$$

leads to the following MU rules (in matrix form) (Févotte et al., 2009):

MU rules for NMF with the β-divergence (one iteration):

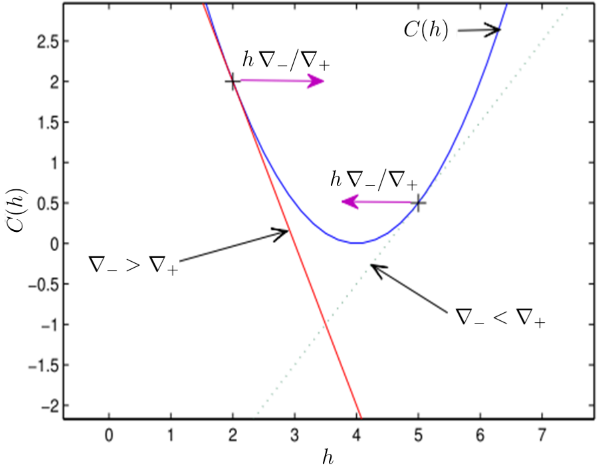
$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T \left((\mathbf{WH})^{\cdot[\beta-2]} \odot \mathbf{V} \right)}{\mathbf{W}^T (\mathbf{WH})^{\cdot[\beta-1]}},$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left((\mathbf{WH})^{\cdot[\beta-2]} \odot \mathbf{V} \right) \mathbf{H}^T}{(\mathbf{WH})^{\cdot[\beta-1]} \mathbf{H}^T},$$

Re-normalize **W** columns and **H** rows to address scale-invariance (see Févotte et al. 2009).

Intuitive explanation

We consider for simplicity $\nabla_h C(h) = \nabla_+ - \nabla_-$



Discussion

The only two things guaranteed by this approach:

- the newly updated value lies in the **direction of partial derivative decrease**;
- the newly updated value is **always nonnegative**.

Nothing more can be guaranteed in general, and all the other algorithm properties depend on the "**positive-negative**" **decomposition chosen:**

$$\nabla_{h_{kn}} C(\boldsymbol{\theta}) = \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_{+} - \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_{-} .$$

Gradient descent viewpoint

Each MU rule can be interpreted as a **diagonally rescaled gradient descent** (Lee and Seung, 2001):

$$h_{kn} \leftarrow h_{kn} - \mu_{kn} \nabla_{h_{kn}} C(\boldsymbol{\theta}),$$

where the step-size μ_{kn} is defined as $\mu_{kn} \triangleq h_{kn} / \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_+$.

Though this re-formulation does not bring any new properties for the algorithm (*e.g.*, the convergence).

Majorisation-minimisation viewpoint

For **many divergences** and **certain "positive-negative" decompositions** each MU rule can be interpreted as a **Majorisation-Minimisation (MM)** procedure (Hunter and Lange, 2004):

To minimise $C(s)$, *e.g.*, $s = w_{fk}$ or $s = h_{kn}$:

- build $G(s|\tilde{s})$ such that $G(s|\tilde{s}) \geq C(s)$ and $G(\tilde{s}|\tilde{s}) = C(\tilde{s})$;
- optimize iteratively $G(s|\tilde{s})$ instead of $C(s)$.

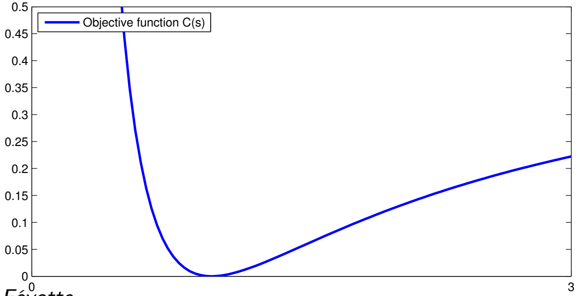


Illustration by C. Févotte

Majorisation-minimisation viewpoint

For **many divergences** and **certain "positive-negative" decompositions** each MU rule can be interpreted as a **Majorisation-Minimisation (MM)** procedure (Hunter and Lange, 2004):

To minimise $C(s)$, *e.g.*, $s = w_{fk}$ or $s = h_{kn}$:

- build $G(s|\tilde{s})$ such that $G(s|\tilde{s}) \geq C(s)$ and $G(\tilde{s}|\tilde{s}) = C(\tilde{s})$;
- optimize iteratively $G(s|\tilde{s})$ instead of $C(s)$.

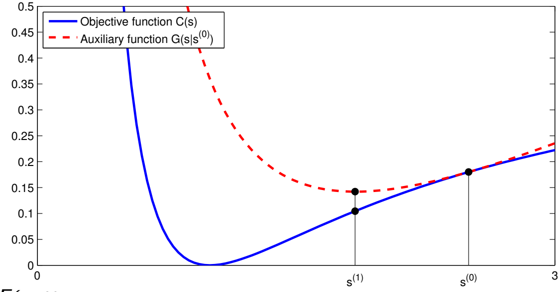


Illustration by C. Févotte

Majorisation-minimisation viewpoint

For **many divergences** and **certain "positive-negative" decompositions** each MU rule can be interpreted as a **Majorisation-Minimisation (MM)** procedure (Hunter and Lange, 2004):

To minimise $C(s)$, *e.g.*, $s = w_{fk}$ or $s = h_{kn}$:

- build $G(s|\tilde{s})$ such that $G(s|\tilde{s}) \geq C(s)$ and $G(\tilde{s}|\tilde{s}) = C(\tilde{s})$;
- optimize iteratively $G(s|\tilde{s})$ instead of $C(s)$.

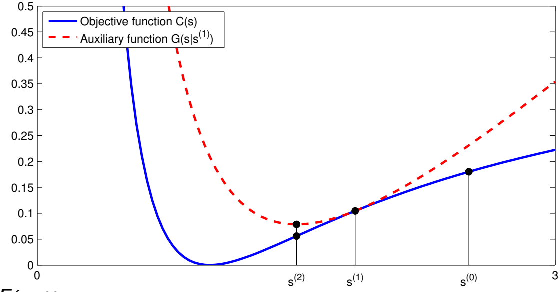


Illustration by C. Févotte

Majorisation-minimisation viewpoint

For **many divergences** and **certain "positive-negative" decompositions** each MU rule can be interpreted as a **Majorisation-Minimisation (MM)** procedure (Hunter and Lange, 2004):

To minimise $C(s)$, *e.g.*, $s = w_{fk}$ or $s = h_{kn}$:

- build $G(s|\tilde{s})$ such that $G(s|\tilde{s}) \geq C(s)$ and $G(\tilde{s}|\tilde{s}) = C(\tilde{s})$;
- optimize iteratively $G(s|\tilde{s})$ instead of $C(s)$.

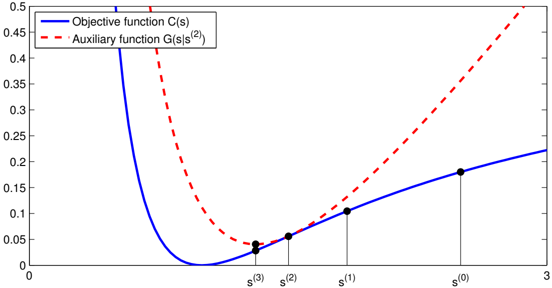


Illustration by C. Févotte

Majorisation-minimisation viewpoint

For **many divergences** and **certain "positive-negative" decompositions** each MU rule can be interpreted as a **Majorisation-Minimisation (MM)** procedure (Hunter and Lange, 2004):

To minimise $C(s)$, *e.g.*, $s = w_{fk}$ or $s = h_{kn}$:

- build $G(s|\tilde{s})$ such that $G(s|\tilde{s}) \geq C(s)$ and $G(\tilde{s}|\tilde{s}) = C(\tilde{s})$;
- optimize iteratively $G(s|\tilde{s})$ instead of $C(s)$.

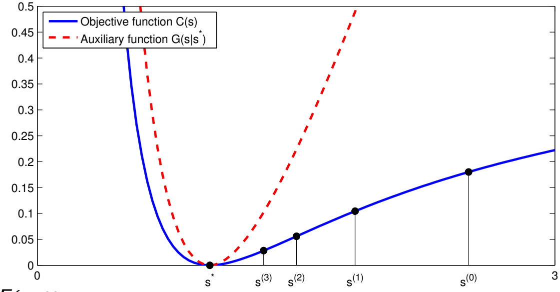


Illustration by C. Févotte

Majorisation-minimisation viewpoint

For **many divergences** and **certain "positive-negative" decompositions** each MU rule can be interpreted as a **Majorisation-Minimisation (MM)** procedure (Hunter and Lange, 2004):

To minimise $C(s)$, *e.g.*, $s = w_{fk}$ or $s = h_{kn}$:

- build $G(s|\tilde{s})$ such that $G(s|\tilde{s}) \geq C(s)$ and $G(\tilde{s}|\tilde{s}) = C(\tilde{s})$;
- optimize iteratively $G(s|\tilde{s})$ instead of $C(s)$.

- **NOTE:** The MM procedure guarantees the cost is non-increasing at each iteration:

$$C(s^{(t+1)}) \leq G(s^{(t+1)}|s^{(t)}) \leq G(s^{(t)}|s^{(t)}) = C(s^{(t)}).$$

Convergence analysis

Monotonicity (“convergence” in terms of **non-increase** of the cost):

- is not guaranteed in general for MU rules;
- is proven (via the majorisation-minimisation formulation) for some divergences (*e.g.*, α and β-divergences) with particular “positive-negative” decompositions (see, *e.g.*, Févotte and Idier 2010; Yang and Oja 2011).

Local convergence in parameters (whether the solution converges to a stationary point?)

- very few positive results for MU rules (see, e.g., Lin 2007a; Badeau et al. 2010);
- the main difficulty is due to non-uniqueness of the NMF.

Summary

Advantages:

- easy to implement;
- non-negativity of $\mathbf{W}$ and $\mathbf{H}$ is guaranteed.

Drawbacks:

- monotonicity is not always guaranteed;
- among other algorithms the convergence rate is not the highest one.

Other alternating optimization algorithms

Gradient-like algorithms (Lin, 2007b)

- Advantages: may **"converge" faster** than MU rules
- Drawbacks: nonnegativity **constraints must be explicitly handled.**

Newton-like algorithms (Zdunek and Cichocki, 2006)

- Advantages: **"converge" faster** than Gradient-like algos and MU rules
- Drawbacks: nonnegativity **constraints must be explicitly handled**; limited to **convex divergences**

Expectation-maximization (EM) algorithms (Févotte et al., 2009; Cemgil, 2009a)

- Advantages: nonnegativity **constraints are implicitly handled**; possibility of **introducing other constraints** via probabilistic priors
- Drawbacks: may **"converge" slower** than MU rules; limited to NMF with **probabilistic formulation**

Online algorithms

Online algorithms to handle **continuous data streams** (Bucak and Gunsel, 2009; Simon and Vincent, 2012)

Online algorithms to handle **big data** (stochastic gradient-like) (Mairal et al., 2010)

How to choose model order?

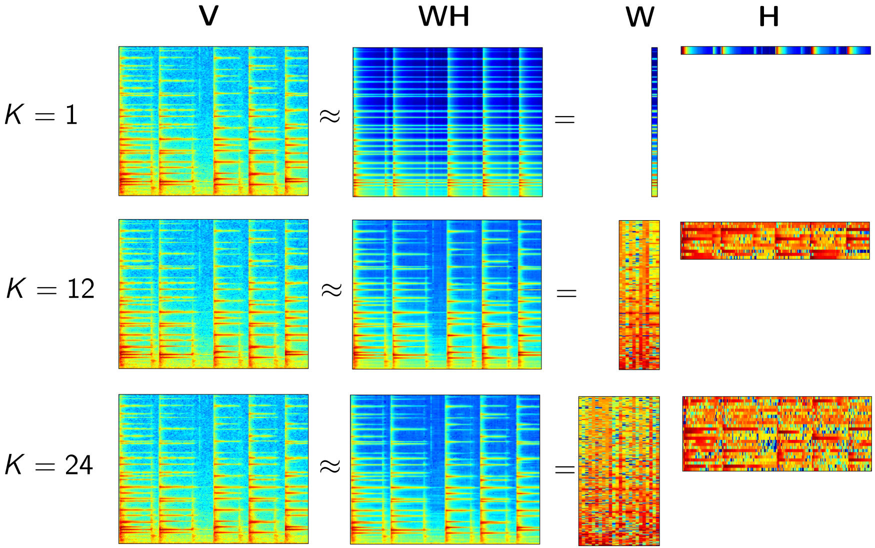
A right **model order choice is important** and it depends on the data $\mathbf{V}$ and on the application.

The following strategies are usually used to set up an appropriate model order:

- **Model order K is fixed** during the NMF decomposition, and it was
  - either chosen by intuition,
  - either chosen based on some prior knowledge (*e.g.*, known number of clusters for clustering),
  - or trained on some development data within a particular application.
- **Model order K is estimated automatically** within the NMF decomposition (Tan and Févotte, 2013; Schmidt and Morup, 2010).

Model order choice

Illustration on audio data



Initialization

A good **initialization** of parameters (**W** and **H**) **is important for any local optimization** approach (including MU rules) due to the existence of many local minima.

Random initializations:

- initialize (nonnegative) parameters **randomly several times**;
- keep the solution with the lowest final cost.

Strcutural data-driven initializations:

- initialize **W** by **clustering** of data points **V** (Kim and Choi, 2007);
- initialize **W** by **singular value decomposition (SVD)** of data points **V** (Boutsidis and Gallopoulos, 2008);
- etc ...

Stopping criteria

How many iterations?

For any iterative optimization strategy (including MU rules) **the total number of iterations is important** and results in a tradeoff between:

- the computational load from one side, and
- the data fitting (approximation error) and model quality from the other side.

Stopping criteria (Albright et al., 2006):

- after a **fixed number of iterations**;
- once the **approximation error** (the cost) is below a pre-defined **threshold**;
- once the **approximation error relative decrease** is below a pre-defined **threshold**;
- etc ...

- Introduction
- NMF models
- Algorithms for solving NMF
- Constrained NMF schemes
  - Regularized NMF
  - Geometric approaches
- Multi-stream and cross-modal NMF schemes
- Applications
- Conclusion

Motivation

Reminder !

Problems:

- **NMF is not unique**.
- Hence **NMF is not guaranteed to extract latent components as desired** within a particular application.

Possible solution: Given the application, **impose some knowledge-based constraints** on $\mathbf{W}$, on $\mathbf{H}$, or on both $\mathbf{W}$ and $\mathbf{H}$.

- Adding constraints usually **makes the decomposition "more unique"**.
- Appropriate constrains may lead to **more suitable latent components**.

Shape-constrained NMF

Convex NMF (Ding et al., 2010)

Constrain the basis vectors $\mathbf{w}_k$ to be convex combinations of the input vectors:

Convex-NMF model

$$\mathbf{w}_k = \sum_{n=1}^{N} g_{nk}\mathbf{v}_n\,;\ g_{nk} \geq 0,\ \sum_n g_{nk} = 1$$

hence the model:

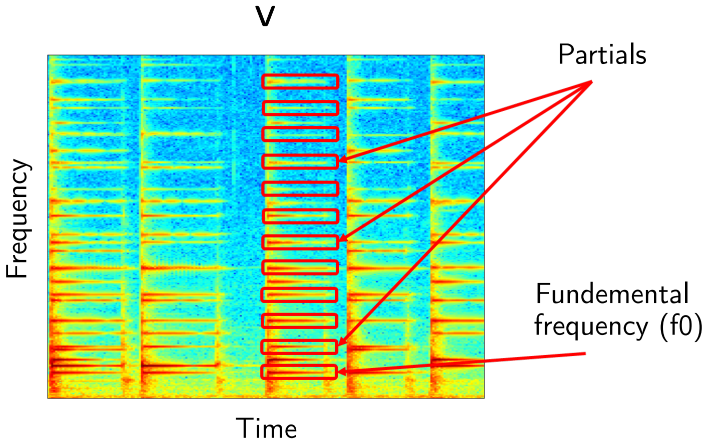
$$\mathbf{V} \approx (\mathbf{VG})\mathbf{H}\,;\ h_{kn} \geq 0$$

- Remarks:
  - vectors $\mathbf{w}_k$ can be better interpreted as **centroids**, being weighted sums of data points;
  - the solution $(\mathbf{G}, \mathbf{H})$ tends to be more sparse (Ding et al., 2010);
  - in practice, the model yields more regular solutions than NMF (see, *e.g.*, Essid and Fevotte 2013).

Shape-constrained NMF

Harmonic NMF (Vincent et al., 2008)

Many audio sound (*e.g.*, speech, harmonic music sounds, etc.) exhibit a **harmonic structure**.



Shape-constrained NMF

Harmonic NMF (Vincent et al., 2008)

Constrain the basis vectors $\mathbf{w}_k$ to be mixtures of M **pre-defined narrow-band harmonic spectra** $\mathbf{E} = \{\mathbf{e}_m\}_{m=1}^{M}$.

Harmonic NMF model

$$\mathbf{w}_k = \sum_{m=1}^{M} g_{mk}\mathbf{e}_m; \;\; g_{mk} \geq 0,$$

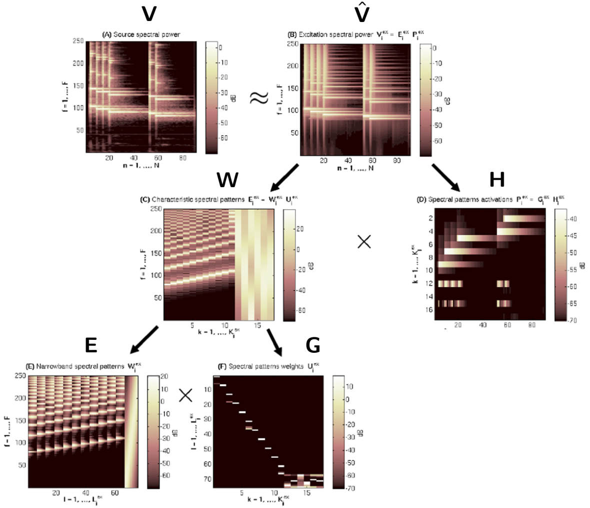
where many entries of matrix $\mathbf{G} = \{g_{mk}\}_{m,k}$ are constrained to be zero (combining any harmonic spectra together is not allowed).
Hence the model:

$$\mathbf{V} \approx (\mathbf{EG})\mathbf{H}\,; \;\; h_{kn} \geq 0\,; \;\; g_{kn} \geq 0.$$

- Remarks:
  - resulting $\mathbf{W} = \mathbf{EG}$ is always harmonic by construction;
  - ($\mathbf{G}$, $\mathbf{H}$) includes less free parameters than ($\mathbf{W}$, $\mathbf{H}$) in the unconstrained NMF.

Shape-constrained NMF

Example of harmonic NMF (Ozerov et al., 2012)

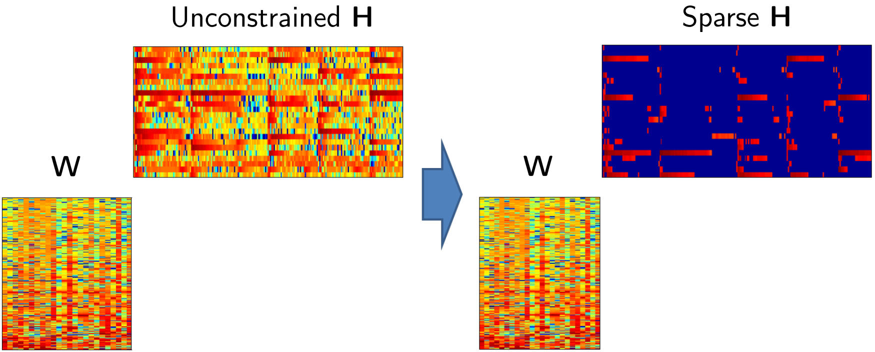


A slightly modified version of a figure from (Ozerov et al., 2012).

Sparse NMF

Sparsity constraints on **H**

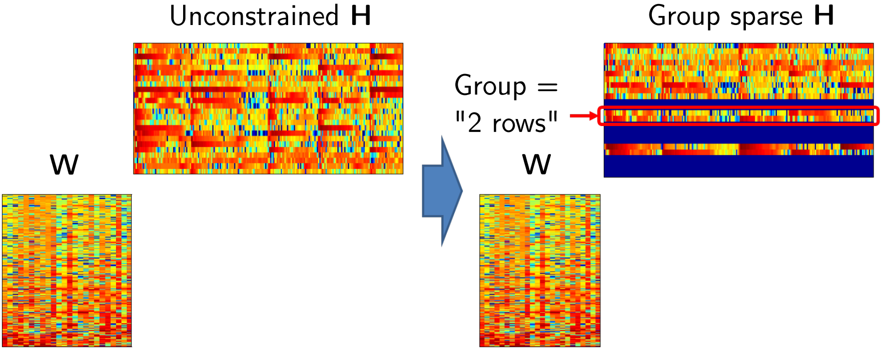
Enforcing only few non-zero entries in $\mathbf{h}_n$:



Sparse NMF

Group sparsity constraints on **H**

Enforcing only few non-zero pre-defined groups (blocks) in **H**:



Sparse NMF

Implementation

Sparsity and group sparsity constraints on $\mathbf{H}$ and/or $\mathbf{W}$ are usually implemented by adding **sparsity-inducing penalties** to the divergence to be minimized:

$$C(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{WH}) + \lambda\psi(\mathbf{H}) + \eta\phi(\mathbf{W})\,.$$

See for example:

- (Hoyer, 2004; Eggert and Korner, 2004) for **sparsity-inducing** penalties in NMF,
  and
- (A. Lefèvre et al., 2011; Sun and Mazumder, 2013; El Badawy et al., 2014) for **group sparsity-inducing** penalties in NMF.

Smooth NMF schemes

Motivation: NMF temporal activations (rows of $\mathbf{H}$) are often erratic, while we know that for some decompositions they should be rather smooth (*e.g.*, for music notes activations).

Solution: introducing smoothness constraints into NMF decomposition (Virtanen, 2007; Jia and Qian, 2009; Essid and Fevotte, 2013; Seichepine et al., 2014a), *e.g.*, by adding a **smoothness penalty** to the divergence to be minimized:

$$C(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{WH}) + \beta_s S(\mathbf{H})\,;$$

where

$$S(\mathbf{H}) = \frac{1}{2} \sum_{k=1}^{K} \sum_{n=2}^{N} |h_{kn} - h_{k(n-1)}|^p; \qquad p = 1 \textit{ or } 2$$

β_s is a regularization parameter controlling the amount of smoothing.

Smooth NMF schemes

Illustration on synthetic sequential data

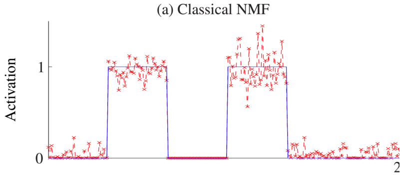
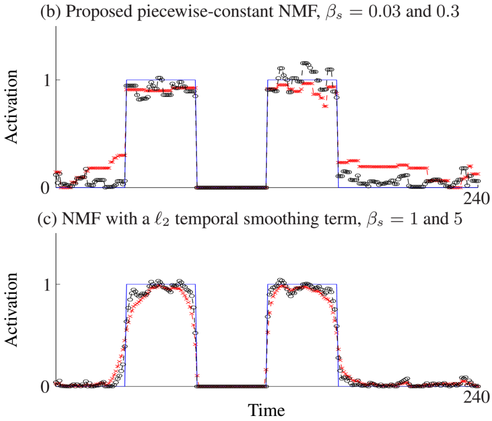


Illustration by N. Seichepine (Seichepine et al., 2014a)



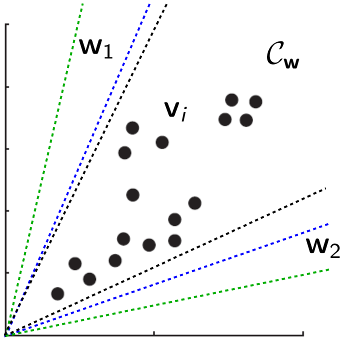
Other constraints

as mentioned in (Gillis, 2014)

- graph regularized NMF (Cai et al., 2011),
- orthogonal NMF (Choi, 2008),
- tri-NMF (Ding et al., 2006),
- projective NMF (Yang and Oja, 2010),
- minimum volume NMF (Miao and Qi, 2007),
- hierarchical NMF (Li et al., 2013),
- etc ...

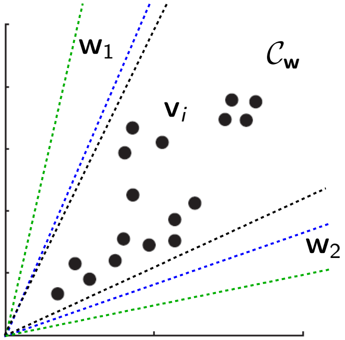
- Introduction
- NMF models
- Algorithms for solving NMF
- Constrained NMF schemes
  - Regularized NMF
  - Geometric approaches
- Multi-stream and cross-modal NMF schemes
- Applications
- Conclusion

Preliminaries

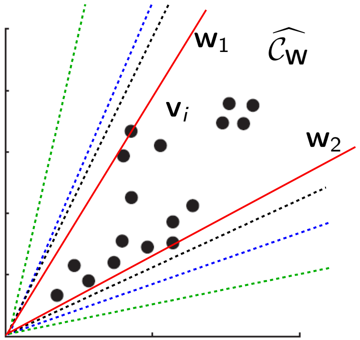


Problem: which $\mathcal{C}_{\mathbf{w}}$?

Preliminaries



Problem: which $\mathcal{C}_{\mathbf{w}}$?



Seek the **smallest cone** $\widehat{\mathcal{C}_{\mathbf{w}}}$ containing the data.

Why the smallest cone?

- Assume the data has **actually** been generated as $\mathbf{V} = \mathbf{BC}$; $\mathbf{B} \geq 0$ and $\mathbf{C} \geq 0$;
- the **exact NMF** model is then $\mathbf{W} = \mathbf{B}$ and $\mathbf{H} = \mathbf{C}$.

Klingenberg et al. lemma

If $Prob\{c_{kn} \in \mathcal{V}(0^+)\} \neq 0$,
i.e. the distribution of the activation coefficients c_{kn} is non-zero in a positive neighborhood of the origin, so that some observations may be arbitrarily close to the vertices of the generating cone,
then the smallest cone $\widehat{\mathcal{C}_{\mathbf{W}}}$ is exactly the generating cone $\widehat{\mathcal{C}_{\mathbf{B}}}$ as $N \to \infty$.

Why the smallest cone?

Example after (Klingenberg et al., 2009)

Generate data according to $\mathbf{V} = \mathbf{BC}$; such that $\mathbf{B} \geq 0$, $\mathbf{C} \geq 0$ and c_{kn} uniformly drawn in $[0, 1]$:

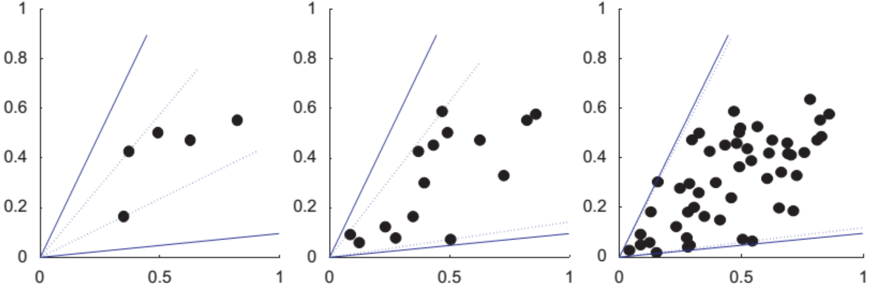


Illustration extracted from (Klingenberg et al., 2009)

Advantages

- $\mathbf{W}$ is uniquely determined (upto a permutation matrix): its columns are the vertices of the cone $\widehat{\mathcal{C}_{\mathbf{W}}}$.
- $\mathbf{H}$ can then be uniquely determined using standard nonnegative linear regression.

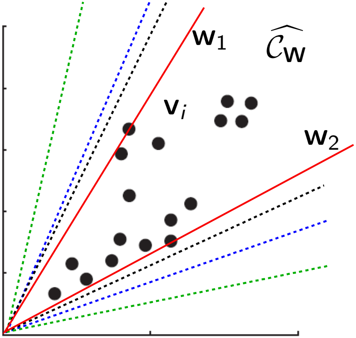
$\rightarrow$ the model becomes **identifiable**!

- Simpler algorithms can potentially be devised, significantly lowering the computational load...

Determining $\widehat{\mathcal{C}_{\mathbf{W}}}$

Preliminary

Assume (without loss of generality) that the data is scaled to unit length, *i.e.* $||\mathbf{v}_n|| = 1, \forall n$:

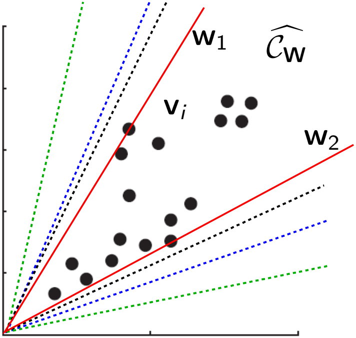


Original data

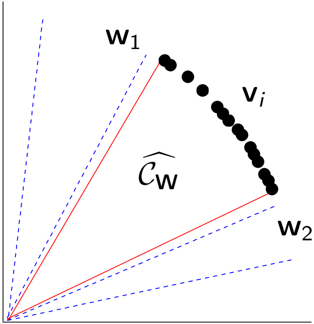
Determining $\widehat{\mathcal{C}_{\mathbf{W}}}$

Preliminary

Assume (without loss of generality) that the data is scaled to unit length, *i.e.* $||\mathbf{v}_n|| = 1, \forall n$:



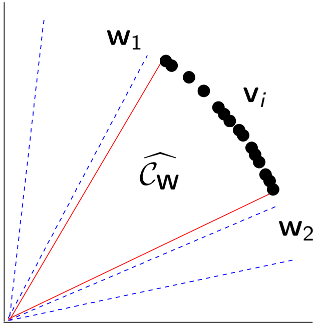
Original data



Data rescaled to unit-length

Determining $\widehat{\mathcal{C}_{\mathbf{W}}}$ with the EVA in $\mathbb{R}^2$

EVA: Extreme Vector Algorithm (Klingenberg et al., 2009)



Find the two data vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ which are the **furthest apart in an angular sense**:

$$\begin{aligned} \mathbf{w}_1, \mathbf{w}_2 &= \underset{m,n}{\operatorname{argmax}} \cos^{-1}(\mathbf{v}_m^T \mathbf{v}_n) \\ &= \underset{m,n}{\operatorname{argmin}}\, \mathbf{v}_m^T \mathbf{v}_n \end{aligned}$$

The EVA in higher dimensions
(Klingenberg et al., 2009)

Initialisation

Set $\mathbf{w}_1, \mathbf{w}_2 = \text{argmin}_{m,n} \mathbf{v}_m^T \mathbf{v}_n$: *first two vectors furthest apart*

For $i = 2 : K$ (repeat until target rank K is reached)

- Set $\mathbf{W}_i = \begin{bmatrix} \mathbf{w}_1 & \dots & \mathbf{w}_i \end{bmatrix}$
- Let $\mathbf{P}_i = \mathbf{W}_i (\mathbf{W}_i^T \mathbf{W}_i)^{-1} \mathbf{W}_i^T$: *projection onto span*$\{\mathbf{w}_1, \dots, \mathbf{w}_i\}$
- Project $\mathbf{V}$ onto current span of $\mathbf{W}_i$: $\mathbf{V}' = \mathbf{P}_i \begin{bmatrix} \mathbf{v}_1 & \dots & \mathbf{v}_n \end{bmatrix}$
- Find k such that $\mathbf{v}_k^T \mathbf{v}'_k = \mathbf{v}_k^T (\mathbf{P}_i \mathbf{v}_k) = \min_n \mathbf{v}_n^T \mathbf{v}'_n$: *the furthest in angular sense to its projection onto span*$\{\mathbf{w}_1, \dots, \mathbf{w}_i\}$.

Discussion

- The basis vectors $\mathbf{w}_k$ are assumed to be among the data:
  $\exists \mathcal{K}:$ a set of indices; $|\mathcal{K}| = K$; such that $\mathbf{W} = \mathbf{V}(:, \mathcal{K})$
- $\mathbf{V}$ is assumed to be K-separable, that is:
  $\exists \mathcal{K}$ such that $\mathbf{V} = \mathbf{V}(:, \mathcal{K})\mathbf{H}$; $\mathbf{H} \geq 0$

Discussion

- The basis vectors $\mathbf{w}_k$ are assumed to be among the data:
  $\exists\,\mathcal{K}$: a set of indices; $|\mathcal{K}| = K$; such that $\mathbf{W} = \mathbf{V}(:,\mathcal{K})$
- $\mathbf{V}$ is assumed to be K-separable, that is:
  $\exists\,\mathcal{K}$ such that $\mathbf{V} = \mathbf{V}(:,\mathcal{K})\mathbf{H}$; $\mathbf{H} \geq 0$
  in other words, $\mathbf{V}$ satisfies the **Extreme Data Property (EDP)** (Klingenberg et al., 2009):

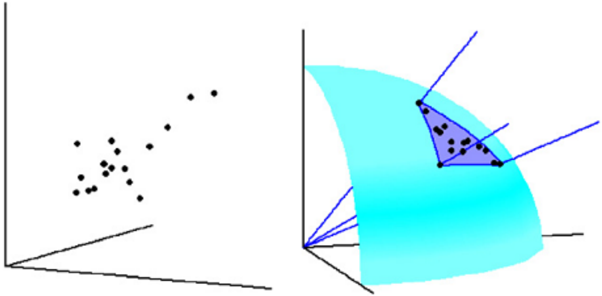


Illustration from (Klingenberg et al., 2009)

EDP satisfied

Discussion

- The basis vectors $\mathbf{w}_k$ are assumed to be among the data:
  $\exists\, \mathcal{K}$: a set of indices; $|\mathcal{K}| = K$; such that $\mathbf{W} = \mathbf{V}(:, \mathcal{K})$
- $\mathbf{V}$ is assumed to be K-separable, that is:
  $\exists\, \mathcal{K}$ such that $\mathbf{V} = \mathbf{V}(:, \mathcal{K})\mathbf{H}$; $\mathbf{H} \geq 0$
  in other words, $\mathbf{V}$ satisfies the **Extreme Data Property (EDP)** (Klingenberg et al., 2009):

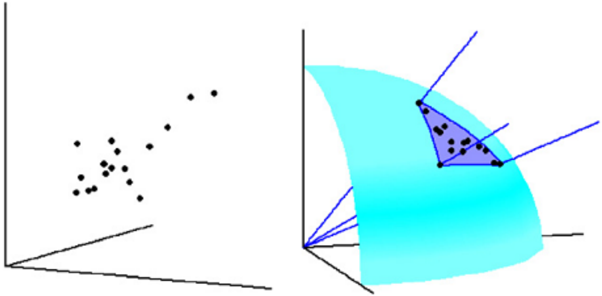


Illustration from (Klingenberg et al., 2009)

EDP satisfied

Discussion

- The basis vectors $\mathbf{w}_k$ are assumed to be among the data:
  $\exists\,\mathcal{K}$: a set of indices; $|\mathcal{K}| = K$; such that $\mathbf{W} = \mathbf{V}(:, \mathcal{K})$
- $\mathbf{V}$ is assumed to be K-separable, that is:
  $\exists\,\mathcal{K}$ such that $\mathbf{V} = \mathbf{V}(:, \mathcal{K})\mathbf{H}$; $\mathbf{H} \geq 0$

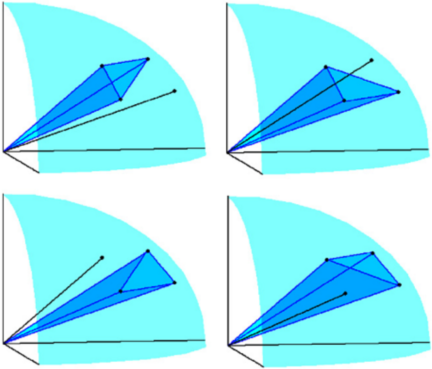


Illustration from (Klingenberg et al., 2009)

EDP not satisfied!

Discussion

- The basis vectors $\mathbf{w}_k$ are assumed to be among the data:
  $\exists\, \mathcal{K}$: a set of indices; $|\mathcal{K}| = K$; such that $\mathbf{W} = \mathbf{V}(:, \mathcal{K})$
- $\mathbf{V}$ is assumed to be K-separable, that is:
  $\exists\, \mathcal{K}$ such that $\mathbf{V} = \mathbf{V}(:, \mathcal{K})\mathbf{H}$; $\mathbf{H} \geq 0$
- The EVA is potentially very complex: its time complexity is $O(K^4)$

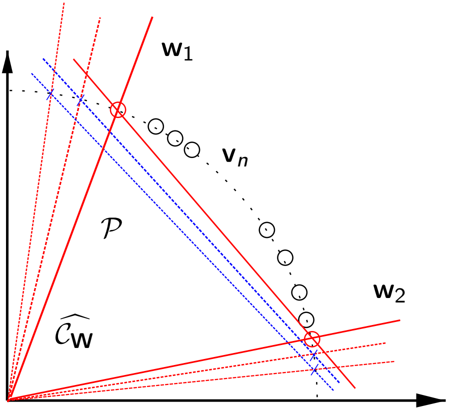
Discussion

- The basis vectors $\mathbf{w}_k$ are assumed to be among the data:
  $\exists\, \mathcal{K}$: a set of indices; $|\mathcal{K}| = K$; such that $\mathbf{W} = \mathbf{V}(:, \mathcal{K})$
- $\mathbf{V}$ is assumed to be K-separable, that is:
  $\exists\, \mathcal{K}$ such that $\mathbf{V} = \mathbf{V}(:, \mathcal{K})\mathbf{H}$; $\mathbf{H} \geq 0$
- The EVA is potentially very complex: its time complexity is $O(K^4)$

$\rightarrow$ need for alternative geometric algorithms...

Determining $\widehat{\mathcal{C}_{\mathbf{W}}}$ by a separating hyperplane

A geometric intuition

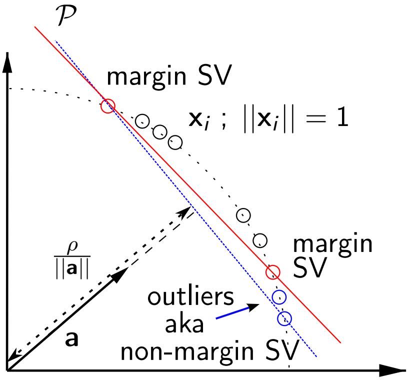


To determine $\widehat{\mathcal{C}_{\mathbf{W}}}$, find the hyperplane $\mathcal{P}$ that separates the data from the origin with **maximum margin**.

$\rightarrow$ this is the **single-class Support Vector Machine** problem!

Single-class Support Vector Machines

Handling outliers



$\mathcal{P} : \langle \mathbf{a}, \mathbf{x} \rangle - \rho = 0 \ , \ \rho \geq 0$

$\mathbf{a}$: normal vector; ρ : offset

Single-class SVM problem:

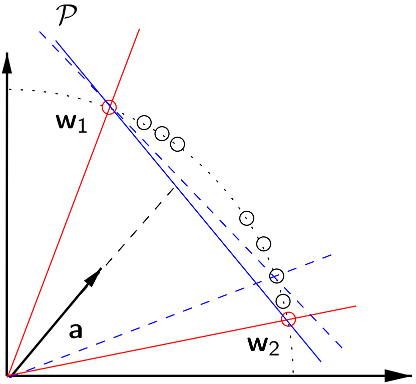
$\min_{\mathbf{a},\xi_i,\rho} \frac{1}{2}||\mathbf{a}||^2 + \frac{1}{\nu N} \sum_i \xi_i - \rho,$

s.t. $\langle \mathbf{a}, \mathbf{x}_i \rangle \geq \rho - \xi_i \ , \ \xi_i \geq 0 \ , \ \rho \geq 0;$

- $\xi_i \geq 0$: are slack variables ($1 \leq i \leq N$);
- ν : positive penalization parameter:
  - ν is an upper-bound on the fraction of **outliers**;
  - a lower-bound on the fraction of **support vectors.**

Determinig $\widehat{\mathcal{C}_{\mathbf{W}}}$ using single-class SVM
(Essid, 2012)

$\widehat{\mathcal{C}_{\mathbf{W}}}$ vertices $\mathbf{w}_k$ are merely the **margin-support vectors.**



$$\mathcal{P} : \langle \mathbf{a}, \mathbf{x} \rangle - \rho = 0$$

The SVM-NMF algorithm I
(Essid, 2012)

1. Apply single-class SVM on data $\mathbf{V} = \{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$;
2. Select margin support vectors as basis vectors $\mathbf{w}_k$;
3. Solve for $\mathbf{H}$: $\min_{\mathbf{h}_i} C(\mathbf{h}_i) = ||\mathbf{v}_i - \mathbf{W}\mathbf{h}_i||^2$ s.t. $\mathbf{h}_{k,i} \geq 0$;
   a classic **nonnegative least-squares** problem.

- **Advantages:**
  - The proposed algorithm can be straightforwardly **kernelized**, hence:
    - allowing for non-linear data decompositions;
    - incorporating prior knowledge through the use of appropriate kernels.
  - **Model order selection**: the choice of K is no longer required, it is determined from the data through a proper choice of ν.

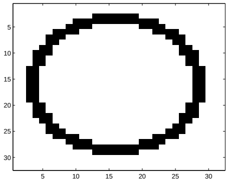
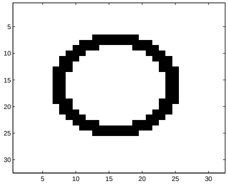
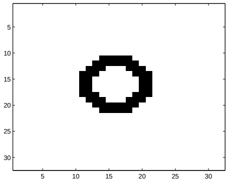
The SVM-NMF algorithm II
(Essid, 2012)

- **Advantages:**
  - **Lower complexity** compared to reference geometric approach (Extreme Vector Algorithm) which is $O(K^4)$ while SVM can be solved with $O(n)$.
  - Straightforward adaptation to **online** processing (online SVM techniques).
  - Ability to exclude **outliers**.

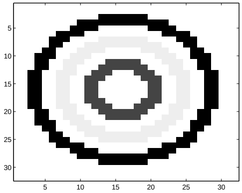
Image analysis example

Data generation

- 1500 images of size 32×32;
- generated as positive linear combinations of

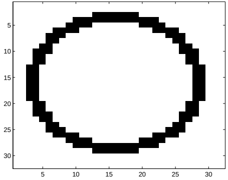
Observation example:

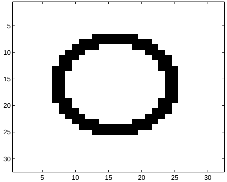


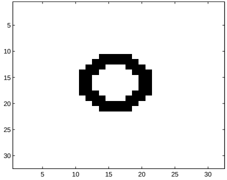
- Images as 1024-coefficient vectors
- Basis vectors stacked in a 1024×3 matrix $\mathbf{B}$
- Data generated as $\mathbf{V} = \mathbf{BC}$
- $\mathbf{C}$ drawn **uniformly** in the range $[0, 1]$

Image analysis example

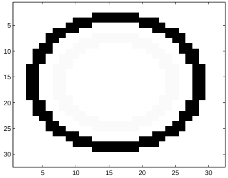
Components $\mathbf{w}_k$ found by applying SVM-NMF on $\mathbf{V} = [\mathbf{BC}, \mathbf{B}]$; $\nu = 0.001$:

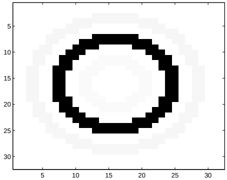






Components $\mathbf{w}_k$ found by applying SVM-NMF on $\mathbf{V} = \mathbf{BC}$; $\nu = 0.001$:





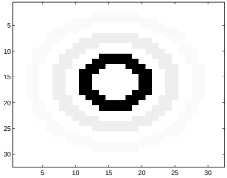
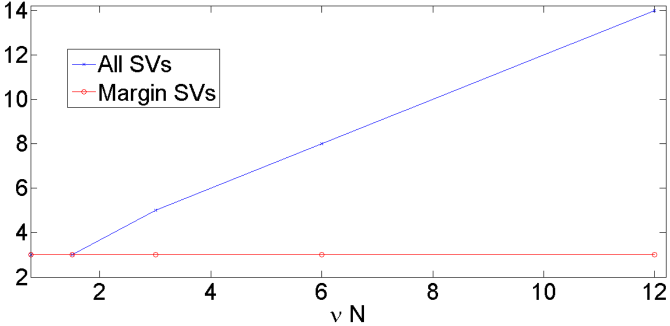


Image analysis example

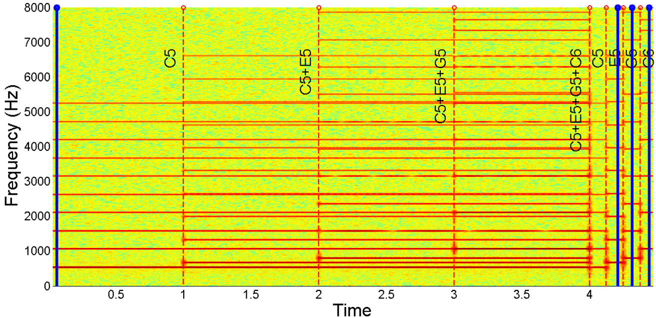
Number of support vectors and basis vectors (*i.e.* **margin support vectors**) as a function of νN.



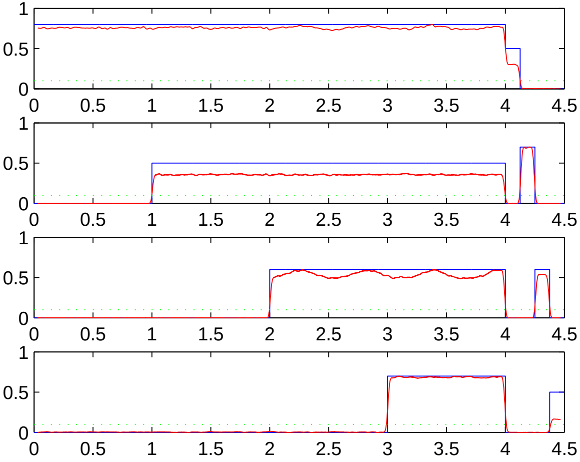
$\rightarrow$ When νN increases, outliers are created but the number of components remains fixed.

Synthetic music transcription example

- **Data generation**
  - $s(t) = \sum_m s_m(t) r_m(t) + b(t)$
  - $s_m(t) = \sum_{p=1}^{10} \frac{a_m}{p} \cos(2\pi p f_m t)$
  - $b(t)$: Gaussian noise; SNR $=$ 6dB
  - $r_m(t)$ defines the temporal activations
  - **V** formed by stacking short-term **power spectra** column-wise.



Synthetic music transcription example



Estimated activations (in red) allow for a perfect transcription.

Non-linear NMF using kernels

- $\Phi : \mathcal{V} \rightarrow \mathcal{H}$; $\mathcal{H}$: a feature space.
- $\kappa(x, y) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$, $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^F \times \mathbb{R}^F$.

▶ **Kernel-based NMF problem:**

$\Phi(\mathbf{V}) \approx \mathbf{WH}$ s.t. $w_{fk} \geq 0$ and $h_{kn} \geq 0$

▶ **Solution:**

- Determine **W** using kernel single-class SVM;
- Solve: $\min_{\mathbf{h}_n} C_\Phi(\mathbf{h}_n) = \frac{1}{2} ||\Phi(\mathbf{v}_n) - \mathbf{W}\mathbf{h}_n||^2_{\mathcal{H}}$ s.t. $h_{kn} \geq 0$.

It is easily shown, using the kernel trick, that:

$\frac{\partial C_\Phi(\mathbf{h}_n)}{\partial h_{kn}} = \sum_{k=1}^{K} h_{kn} \kappa(v_k, v_l) - \kappa(v_n, v_l)$ and $\frac{\partial^2 C_\Phi(\mathbf{h}_n)}{\partial h_{ln} \partial h_{kn}} = \kappa(v_k, v_l)$

$\rightarrow$ The **Hessian** matrix is exactly the **Gram** matrix which is **positive definite** for positive definite kernels.

Ongoing work

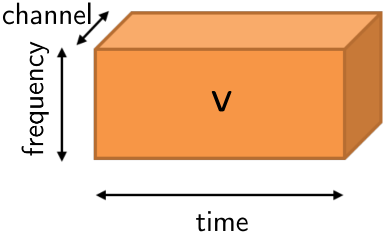
- Considering real-world applications.
- Using application-specific kernels: potential for **incorporating prior knowledge** on the data.
- Studying the impact of ν on model-order selection.

- Introduction
- NMF models
- Algorithms for solving NMF
- Constrained NMF schemes
- Multi-stream and cross-modal NMF schemes
  - NTF models
  - Co-factorisation schemes
- Applications
- Conclusion

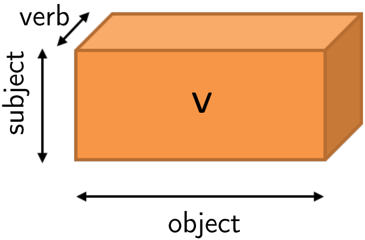
Multi-way data representations

Some data can have more **meaningful representation** using **multi-way arrays** rather than **matrices** (two-way arrays).

Multichannel audio (Févotte and Ozerov, 2010)



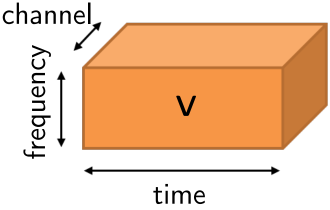
Text statistics for natural language processing (de Cruys, 2010)

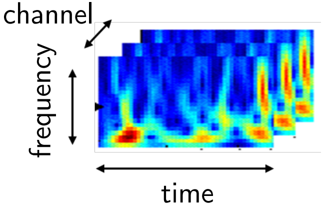


Multi-way data representations

Some data can have more **meaningful representation** using **multi-way arrays** rather than **matrices** (two-way arrays).

Electroencephalography (EEG) data (Lee et al., 2007)

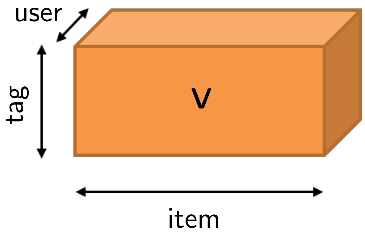




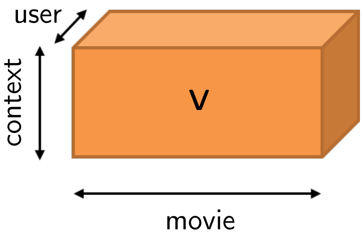
Multi-way data representations

Some data can have more **meaningful representation** using **multi-way arrays** rather than **matrices** (two-way arrays).

Tag recommendation (Rendle et al., 2009)



Context-aware collaborative filtering (Karatzoglou et al., 2010)



Definition

What do we mean by a **tensor**?

- Tensor is an L-way array or simply a dataset indexed by L indices ($L = 2$ for matrices, $L = 3$ for "boxes")

2-way array (matrix)

$\mathbf{V}$

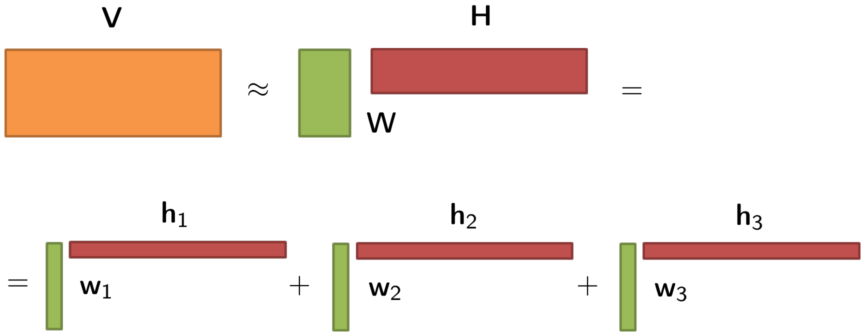
3-way array (tensor)

$\mathbf{V}$

- **NOTE**: In Physics tensors have a different meaning.

Reminder on NMF models

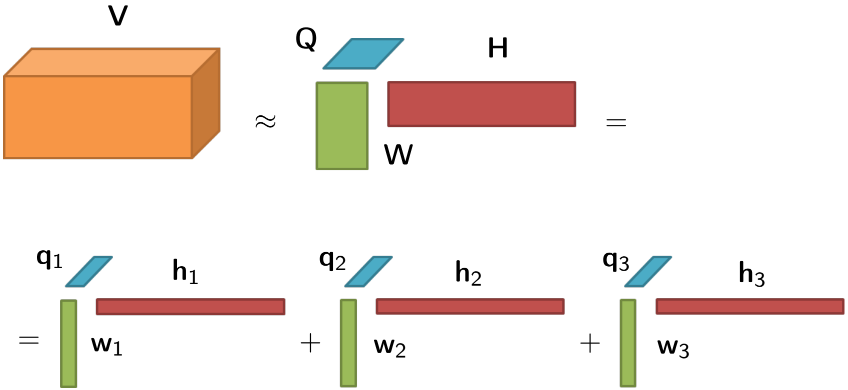
$v_{fn} \approx \sum_k w_{fk} h_{kn}, \quad \mathbf{V} = \mathbf{W}\mathbf{H}$



CANDECOMP / PARAFAC (NTF) models

(Bro, 1997)

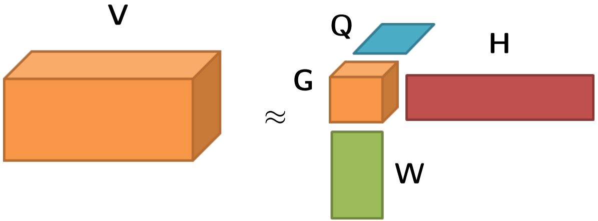
$v_{fnl} \approx \sum_k w_{fk} h_{kn} q_{lk}, \quad \mathbf{V} \approx \sum_k \mathbf{w}_k \circ \mathbf{h}_k^T \circ \mathbf{q}_k$



TUCKER3

(Kiers, 2000)

$v_{fnl} \approx \sum_{p,k,r} w_{fp} h_{kn} q_{rl} g_{pkr}$

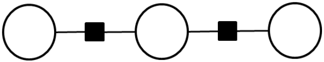


- **G** is called a **core tensor**

Factor graphs representation for NTF I
(Yilmaz and Cemgil, 2010)

Yilmaz and Cemgil 2010 proposed representing various NTF models using **factor graphs** (Loeliger, 2004), which are connected graphs:

- with cycles and squares as nodes, and
- where a vertex cannot connect two boxes or two cycles together (a bipartite graph).

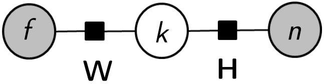


Factor graphs representation for NTF II
(Yilmaz and Cemgil, 2010)

Representing NTF models with factor graphs (Yilmaz and Cemgil, 2010):

- **Latent factors** (*e.g.*, **W** or **H**) are represented by **square nodes**.
- Latent factor **indices** are represented by **cycle nodes** (in gray for observed indices and in white for latent indices).
- A **vertex** connecting a cycle node with a square node means that the corresponding latent factor is indexed by the corresponding index.

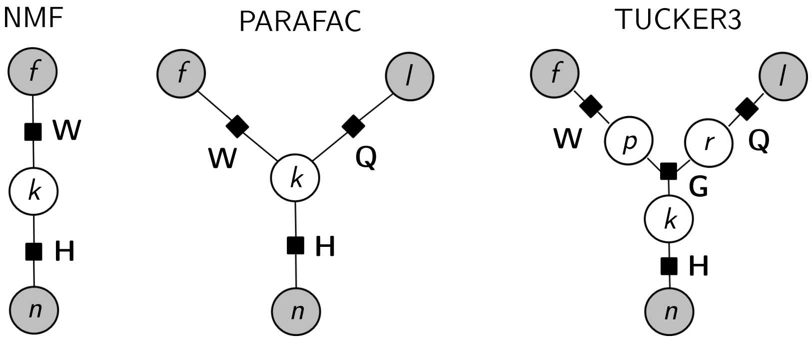
NMF $v_{fn} \approx \sum_k w_{fk} h_{kn}$



Factor graphs representation for NTF III

(Yilmaz and Cemgil, 2010)

- NMF: $v_{fn} \approx \sum_k w_{fk} h_{kn}$,
- PARAFAC: $v_{fnl} \approx \sum_k w_{fk} h_{kn} q_{lk}$,
- TUCKER3: $v_{fnl} \approx \sum_{p,k,r} w_{fp} h_{kn} q_{rl} g_{pkr}$.



Generalized NTF
(Yilmaz and Cemgil, 2010)

Yilmaz and Cemgil 2010 proposed to call **Generalized NTF** any nonnegative factorization model that can be represented this way as a **factor graph**.

Advantages of such a general and graphical representation:

- *Having a graphical representation is **always nice!***
- This representation **generalizes** many existing models (Yilmaz and Cemgil, 2010; Cemgil et al., 2011).
- Results from the factor graph theory (*e.g.*, the **sum-product algorithm** Kschischang et al. 2001) can be re-used.
- Given a graph, a corresponding optimization **algorithm** (e.g., MU rules) can be **generated automatically**.

Example of MU rules for TUCKER3 NTF

$$v_{fnl} \approx \hat{v}_{fnl} = \sum_{p,k,r} w_{fp} h_{kn} q_{rl} g_{pkr}$$

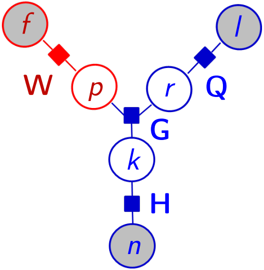
MU rules for TUCKER3 NTF with the β-divergence (one iteration):

$$h_{kn} \leftarrow h_{kn} \frac{\sum_{f,l,p,r} v_{fnl} \hat{v}_{fnl}^{\beta-2} w_{fp} q_{rl} g_{pkr}}{\sum_{f,l,p,r} \hat{v}_{fnl}^{\beta-1} w_{fp} q_{rl} g_{pkr}},$$

$$w_{fp} \leftarrow w_{fp} \frac{\sum_{n,l,k,r} v_{fnl} \hat{v}_{fnl}^{\beta-2} h_{kn} q_{rl} g_{pkr}}{\sum_{n,l,k,r} \hat{v}_{fnl}^{\beta-1} h_{kn} q_{rl} g_{pkr}},$$

$$q_{rl} \leftarrow q_{rl} \frac{\sum_{f,n,p,k} v_{fnl} \hat{v}_{fnl}^{\beta-2} w_{fp} h_{kn} g_{pkr}}{\sum_{f,n,p,k} \hat{v}_{fnl}^{\beta-1} w_{fp} h_{kn} g_{pkr}},$$

$$g_{pkr} \leftarrow g_{pkr} \frac{\sum_{f,n,l} v_{fnl} \hat{v}_{fnl}^{\beta-2} w_{fp} h_{kn} q_{rl}}{\sum_{f,n,l} \hat{v}_{fnl}^{\beta-1} w_{fp} h_{kn} q_{rl}}.$$



Surprising facts about PARAFAC
(Yilmaz and Cemgil, 2010)

There are a lot of interesting **mathematical developments** around NTF (see, *e.g.*, Lim and Comon 2010).

Surprisingly, but some results for the PARAFAC NTF (L-way arrays with $L \geq 3$) are **quantitatively different** from the results for the NMF.

Some surprising results on PARAFAC NTF:

- In contrast to the NMF, the **uniqueness conditions** for the PARAFAC NTF are **mild** (Kruskal, 1977).
- In contrast to the NMF, there exist tensors for which **rank-**R **approximations** (i.e., PARAFACs with R latent components) **do not exist** (Lim and Comon, 2010). In other words, the set of tensors with rank $\leq R$ is not guaranteed to be closed.

Applications of NTF to audio-visual content processing

Audio

- Music genre classification (Benetos and Kotropoulos, 2008)
- Source separation (FitzGerald et al., 2008; Ozerov et al., 2011a)
- Compression (Nikunen et al., 2011; Ozerov et al., 2013)

Videos and images

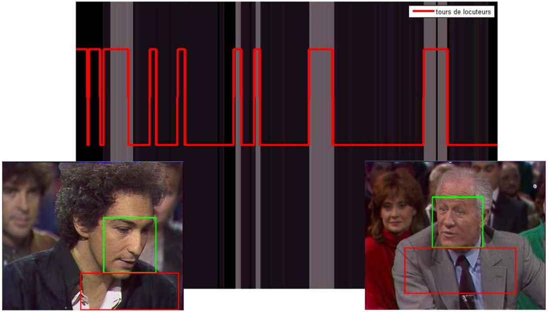
- Action recognition (Kim and Cipolla, 2009; Krausz and Bauckhage, 2010)
- Faces analysis (Shashua and Hazan, 2005)
- Image retrieval (Liu et al., 2014)

- Introduction
- NMF models
- Algorithms for solving NMF
- Constrained NMF schemes
- Multi-stream and cross-modal NMF schemes
  - NTF models
  - Co-factorisation schemes
- Applications
- Conclusion

Motivation

Multimodal speaker diarization on edited videos

- **Observation**: the audio and visual streams are related:



→ Exploit both **audio** and **visual** features to perform speaker diarization.

NMF for multimodal data analysis I

Possible approaches

- **NTF** cannot be used: features from each modality do not live in spaces of same dimensionality;

$\rightarrow$ an observation tensor can**not** be built for such data.

- Alternatively, concatenate the feature observations:

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{bmatrix} \approx \mathbf{W}\mathbf{H}$$

$\rightarrow$ same cost functions need to be used for different modalities: not always optimal.

NMF for multimodal data analysis II

Possible approaches

- Another idea: perform **co-factorisation** constraining the activations to be the same:

$$\begin{cases} \mathbf{V}_1 \approx \mathbf{W}_1\mathbf{H} \\ \mathbf{V}_2 \approx \mathbf{W}_2\mathbf{H} \end{cases} \qquad \text{solving:} \quad \min_{\mathbf{W}_1,\mathbf{W}_2,\mathbf{H}} D_1(\mathbf{V}_1, \mathbf{W}_1\mathbf{H}) + \beta_2 D_2(\mathbf{V}_2, \mathbf{W}_2\mathbf{H})$$

$\rightarrow$ does not account for possible local discrepancies across modalities:

$\rightarrow$ constrain the audio and visual data factorisations to be "**related**": temporal activations relating to these two streams of data should be **close**, not necessarily equal.

Soft nonnegative matrix co-factorisation

(Seichepine et al., 2013)

Solve the problem:

$$\min_{\mathbf{W}_1,\mathbf{H}_1,\mathbf{W}_2,\mathbf{H}_2} D_1(\mathbf{V}_1 \,|\, \mathbf{W}_1\mathbf{H}_1) + \beta_2 D_2(\mathbf{V}_2 \,|\, \mathbf{W}_2\mathbf{H}_2) + \beta_c \left\|\mathbf{H}_1 - \mathbf{H}_2\right\|_p$$

- D_1 is a **measure of fit** penalizing the reconstruction errors for the **first modality** ;
- D_2 is a **measure of fit** penalizing the reconstruction errors for the **second modality** ;
- $\|\cdot\|_p$ is a **penalization term** coupling factorizations for the first and the second modality;
- β_2 and β_c are weighting **hyperparameters**.

Soft nonnegative matrix co-factorisation

(Seichepine et al., 2013)

Solve the problem:

$$\min_{\mathbf{W}_1,\mathbf{H}_1,\mathbf{W}_2,\mathbf{H}_2} D_1(\mathbf{V}_1 \,|\, \mathbf{W}_1\mathbf{H}_1) + \beta_2 D_2(\mathbf{V}_2 \,|\, \mathbf{W}_2\mathbf{H}_2) + \beta_c \left\|\mathbf{H}_1 - \mathbf{H}_2\right\|_p$$

- **Remarks:**
  - similarly a dependency between $\mathbf{W}_1$ and $\mathbf{W}_2$ could be accounted for: $\min D(\mathbf{V}|\mathbf{WH})$ is equivalent to $\min D(\mathbf{V}^T|\mathbf{H}^T\mathbf{W}^T)$.
  - if $\mathbf{H}_1$ and $\mathbf{H}_2$ have different dimensions, the penalty term can readily ignore rows and columns of $\mathbf{H}_1$ that have no match in $\mathbf{H}_2$.

Solving the problem

- We have devised a **block-coordinate MM** algorithm solving the problem:
  - $\mathbf{H}_1$, $\mathbf{H}_2$, $\mathbf{W}_1$ and $\mathbf{W}_2$ are updated sequentially;
  - the cost function is decreased at each iteration;
- update rules have been determined for:
  - Kullback-Liebler and Itakura-Saito cost functions;
  - ℓ_2 and ℓ_1-coupling penalties;
  - ℓ_2 and ℓ_1-temporal smoothing penalties.

  see (Seichepine et al., 2014b) for more details.
- Matlab scripts are available on N. Seichepine's web page:

  http://www.telecom-paristech.fr/~seichepi

Applications

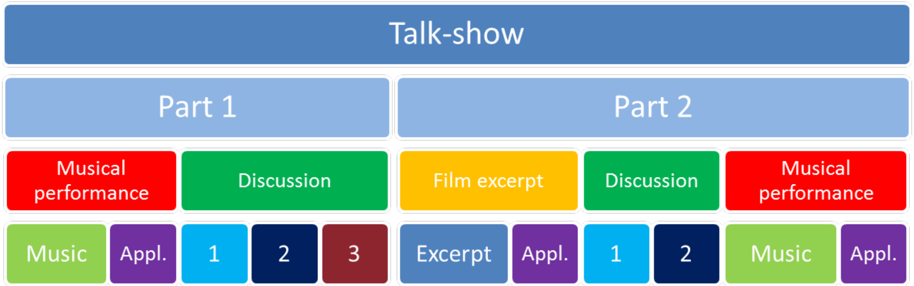
The method has been successfully applied (Seichepine et al., 2014b) to:

- **multimodal speaker diarization**;
- multi-channel musical audio source separation.

- Introduction
- NMF models
- Algorithms for solving NMF
- Constrained NMF schemes
- Multi-stream and cross-modal NMF schemes
- Applications
  - Video structuring
  - Audio source separation
- Conclusion

The video structuring problem

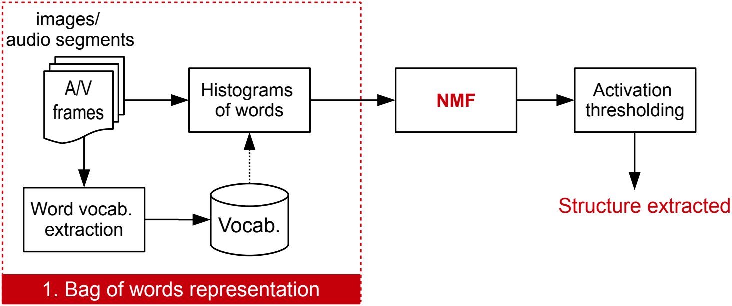
Goal: automatically extract a **temporal organization** of a document into units conveying a homogeneous type of (audio/video) content.



A generic video structuring system using NMF

Challenge: perform the task in a **non-supervised fashion**.

Proposed approach: a **generic** structuring scheme using **NMF** (Essid and Fevotte, 2013):

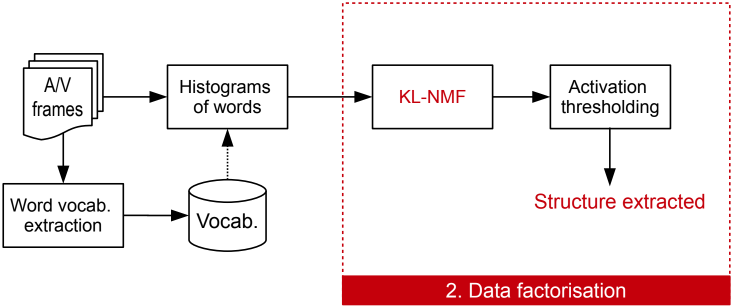


1. create a low-level (visual/audio) vocabulary and use it to extract **histogram of (visual/audio) words** from the sequence of observation frames;

A generic video structuring system using NMF

Challenge: perform the task in a **non-supervised fashion**.

Proposed approach: a **generic** structuring scheme using **NMF** (Essid and Fevotte, 2013):

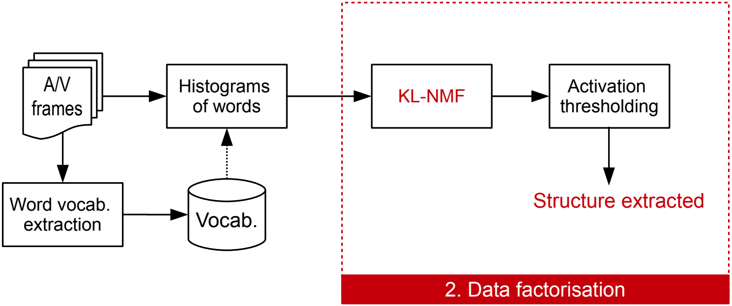


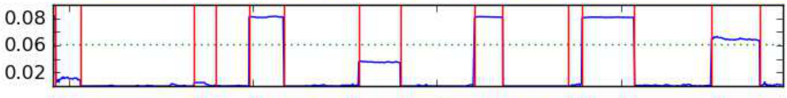
2. apply a variant of **smooth NMF** using the **Kullback-Leibler** divergence to extract **latent structuring events** and their **activations** across the duration of the document.

A generic video structuring system using NMF

Challenge: perform the task in a **non-supervised fashion**.

Proposed approach: a **generic** structuring scheme using **NMF** (Essid and Fevotte, 2013):

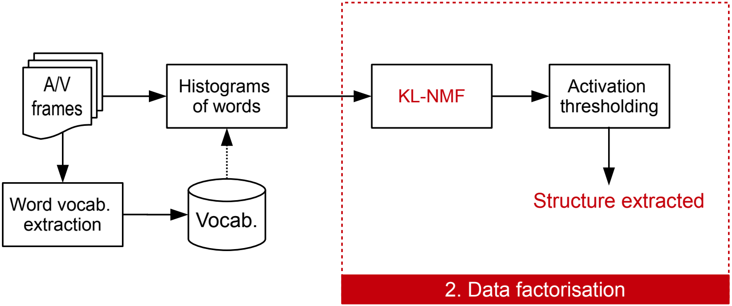




A generic video structuring system using NMF

Challenge: perform the task in a **non-supervised fashion**.

Proposed approach: a **generic** structuring scheme using **NMF** (Essid and Fevotte, 2013):



Activations should be **temporally smooth**: structuring events naturally exhibit a “certain” temporal continuity.

Applications

Onscreen person-oriented structuring

Discover the video editing structure: label the video frames as follows in a **non-supervised** fashion:

Using the **Canal9 political debates** database (Vinciarelli et al., 2009).

Applications

Speaker diarization

"Who spoke when?"

- **A notable difficulty**: handling overlapped speech segments.

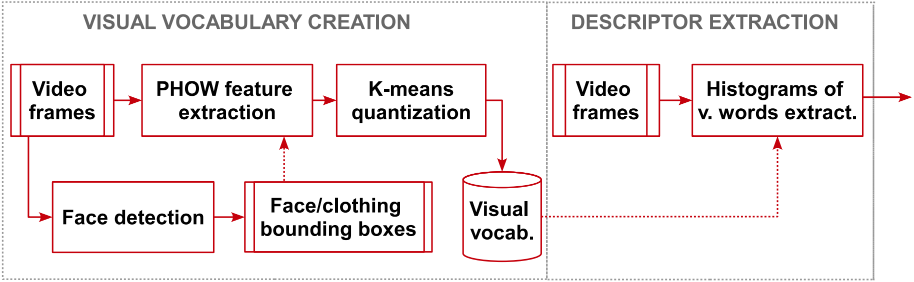
→ NMF has the potential to alleviate this issue.

Experimental validation

Canal9 political debates database (Vinciarelli et al., 2009)

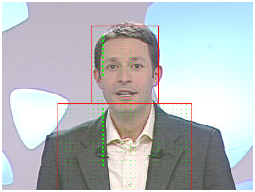
- broadcasts featuring a moderator and 2 to 4 guests;
- moderators, guest and background vary;
- 7 hours of video content: 10 minutes from each of the first 41 shows;
- 189 distinct persons; 28521 video shots.

Visual features



Visual vocabulary creation

- **PHOW** features (Bosch et al., 2007): histograms of orientation gradients over 3 scales, on 8-pixel step grid; extracted from **faces** and **clothing** regions, determined automatically for current video;
- quantization over 128 bins using K-means.



Evaluation

Reference system: ergodic Hidden Markov Models (HMM) states:

- $N_{sp} + 2$ states, N_{sp}: number of speakers;
- Gaussian-emission probabilities with full covariance matrices;
- same features as NMF system.

NMF Parameters:

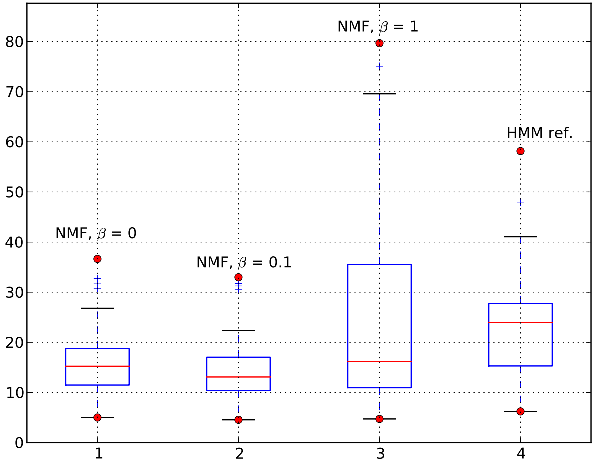
- $K = N_{sp} + 1$;
- best (in terms of cost-function value) of 10 random initializations;
- 3 values of smoothing penaltyβ_s are tested: $\beta_s \in \{0; 0.1; 1\}$.

Scoring:

- using **frame-type classification error rate**;
- computed following the **NIST** speaker diarization error procedure (performs a one-to-one mapping between groundtruth and automatically determined segment labels).

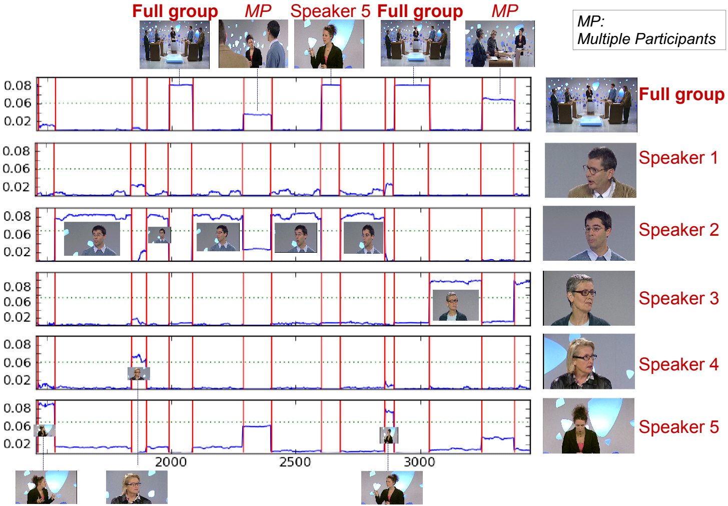
Results

Shot-type classification error rates



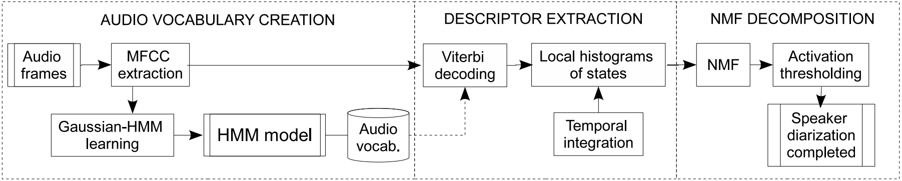
Results

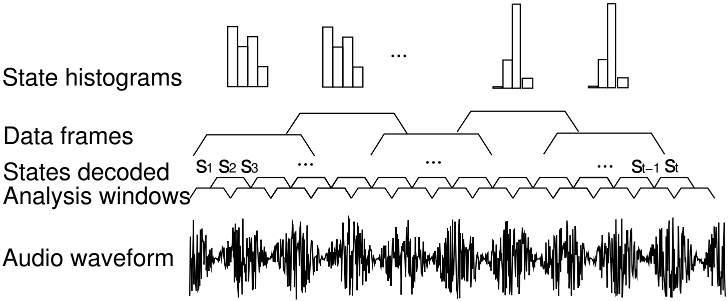
Visualising the activations



NMF for speaker diarization

(Essid and Fevotte, 2013)





Evaluation

(Seichepine et al., 2014b)

Dataset and scoring:

- using the *Canal9 political debates* videos;
- using the **NIST DER** (Diarization Error Rate).

Reference methods:

- a simple K-means applied to data matrices $\mathbf{V}$;
- a state-of-the-art GMM-based diarization system: the LIUM Speaker Diarization system (Meignier and Merlin, 2010).

NMF Parameters:

- audio: N_{sp} components; video: $N_{sp}+1$ components;
- initializations based on output of previously computed monomodal NMFs;
- $\beta_1 = 0.02$, $\beta_2 = 0.2$ and $\beta_c = 0.1$, respectively for visual, audio and coupling penalties; tuned on development data.

Results

Method	K-means	NMF	S-NMF	**CS-NMF**
Mean score	13.51	14.13	11.45	**10.24**
Mean (3 speakers)	15.86	16.37	17.40	**15.92**
Mean (4 speakers)	10.90	13.65	10.69	**9.29**
Mean (5 speakers)	12.46	12.67	7.43	**6.46**
Prop. better than K-means	/	30%	85%	85%

S-NMF: ℓ_1-smoothed NMF;
CS-NMF: ℓ_1-coupled ℓ_1-smoothed NMF.

Results

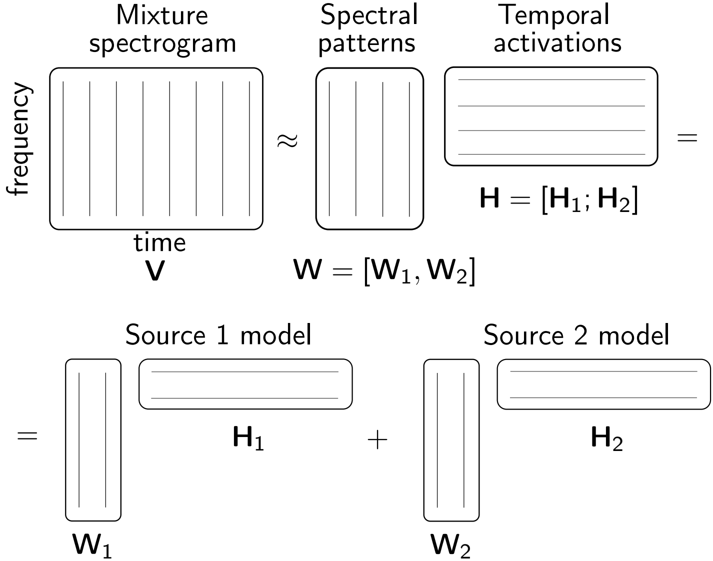
Method	K-means	NMF	S-NMF	**CS-NMF**	LIUM
Mean score	13.51	14.13	11.45	10.24	**6.87**
Mean (3 speakers)	15.86	16.37	17.40	15.92	**7.67**
Mean (4 speakers)	10.90	13.65	10.69	9.29	**8.26**
Mean (5 speakers)	12.46	12.67	7.43	6.46	**5.97**
Prop. better than K-means	/	30%	85%	85%	100%
Prop. better than LIUM	0%	0%	30%	**52%**	/

S-NMF: ℓ_1-smoothed NMF;
CS-NMF: ℓ_1-coupled ℓ_1-smoothed NMF.

- Introduction
- NMF models
- Algorithms for solving NMF
- Constrained NMF schemes
- Multi-stream and cross-modal NMF schemes
- Applications
  - Video structuring
  - Audio source separation
- Conclusion

NMF for audio source separation

Main idea



NMF for audio source separation

Details

All audio signals are represented in the **complex-values** short time Fourier transform (**STFT**) domain (a time-frequency representation).

Problem: Given a mixture of two sources

$$\mathbf{X} = \mathbf{S}_1 + \mathbf{S}_2, \qquad \mathbf{X}, \mathbf{S}_1, \mathbf{S}_2 \in \mathbb{C}^{F \times N},$$

estimate $\mathbf{S}_1$ and $\mathbf{S}_2$.

Basic approach:

- Compute an NMF decomposition $\mathbf{V} = |\mathbf{X}|^{.2} \approx \mathbf{WH} = \mathbf{W}_1\mathbf{H}_1 + \mathbf{W}_2\mathbf{H}_2$.
- Compute source estimates by Wiener filtering:

$$\widehat{\mathbf{S}}_1 = \frac{\mathbf{W}_1\mathbf{H}_1}{\mathbf{W}_1\mathbf{H}_1 + \mathbf{W}_2\mathbf{H}_2} \odot \mathbf{X}, \qquad \widehat{\mathbf{S}}_2 = \frac{\mathbf{W}_2\mathbf{H}_2}{\mathbf{W}_1\mathbf{H}_1 + \mathbf{W}_2\mathbf{H}_2} \odot \mathbf{X}.$$

NMF for audio source separation

Details

Main difficulty: How to compute the decomposition

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} = \mathbf{W}_1\mathbf{H}_1 + \mathbf{W}_2\mathbf{H}_2$$

such that $(\mathbf{W}_1, \mathbf{H}_1)$ and $(\mathbf{W}_2, \mathbf{H}_2)$ represent well the sources $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively?

One popular approach:

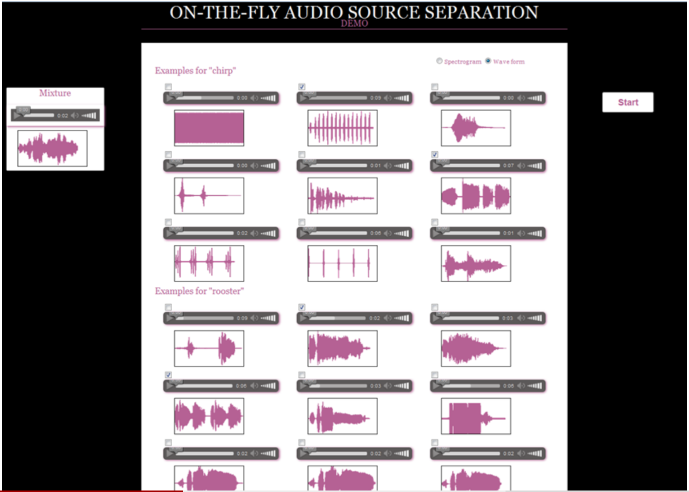
- Compute $\mathbf{W}_1$ and $\mathbf{W}_2$ from some training samples (*e.g.*, downloaded from the internet).
- Set $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2]$ and fix it during mixture decomposition.

Problem: Training samples are not always available and/or representative.

Source separation demo

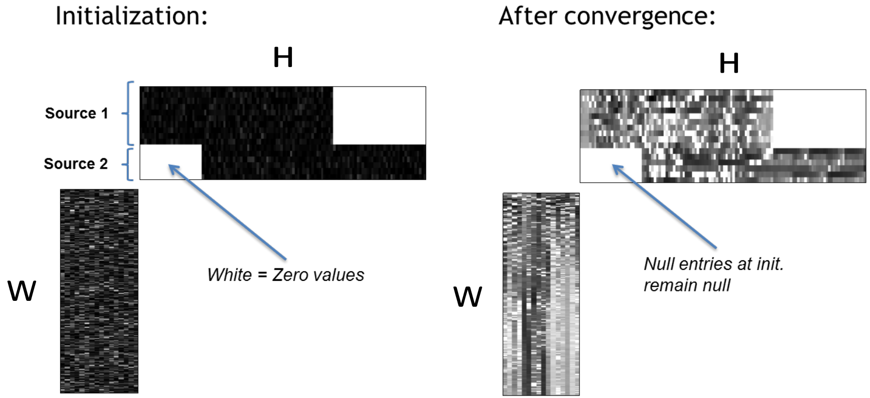
On-the-fly audio source separation (El Badawy et al., 2014)

A user queries **audio samples from the internet** to pre-train $\mathbf{W}_1$ and $\mathbf{W}_2$.



User-guided audio source separation
(Ozerov et al., 2011a; Duong et al., 2014)

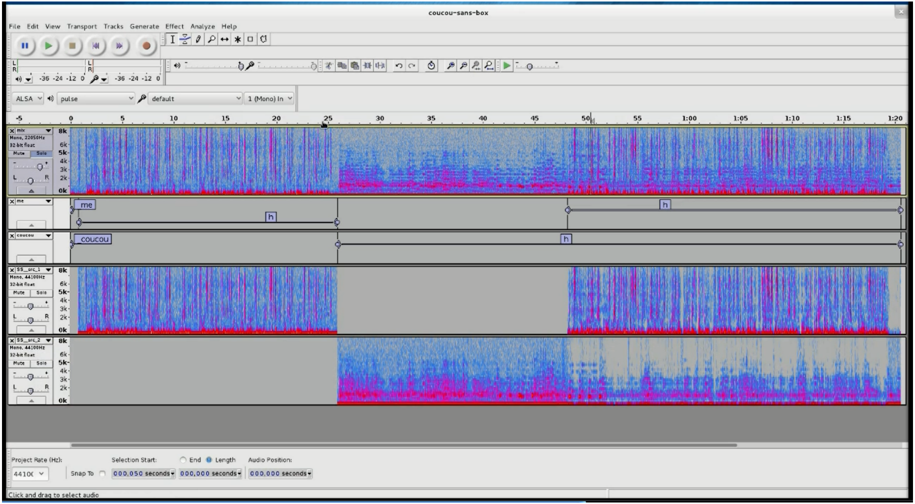
A user is simply asked to **annotate temporal segments** of source activities (active or non-active).



Due to multiplicative update rules, zero entries at the initialization stay at zero

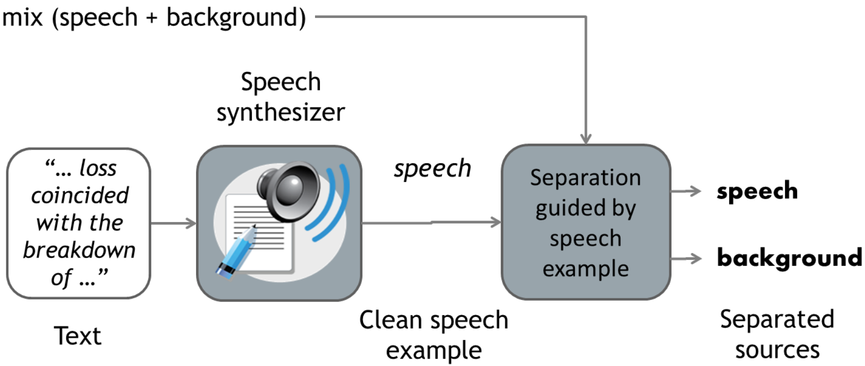
User-guided audio source separation

Demo (Ozerov et al., 2011a; Duong et al., 2014)



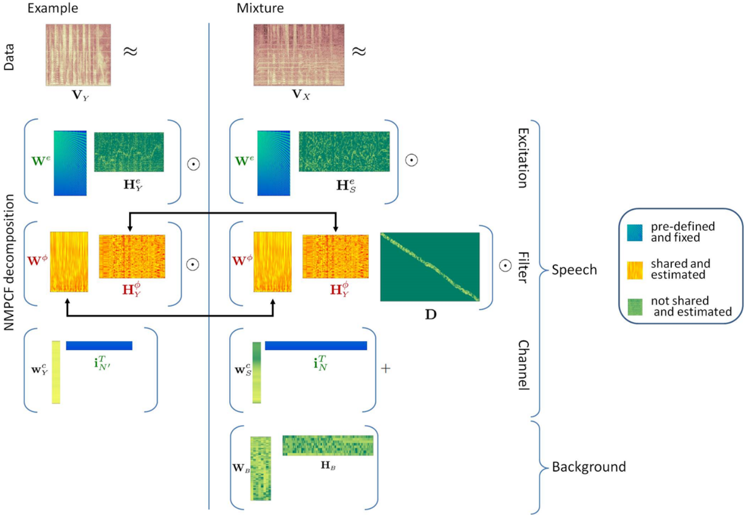
Text-informed audio source separation

General scheme (Le Magoarou et al., 2013)



Text-informed audio source separation

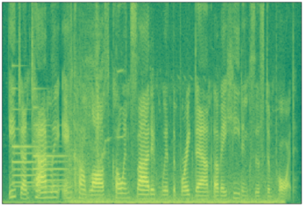
Coupled NMF-based approach (Le Magoarou et al., 2013)



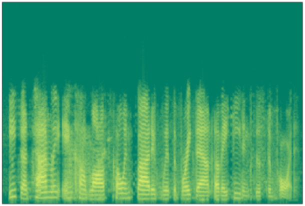
Text-informed audio source separation

Demo (Le Magoarou et al., 2013)

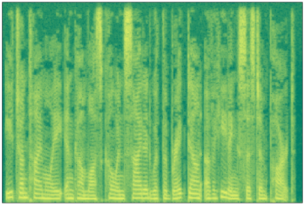
Mixture (speech + background)



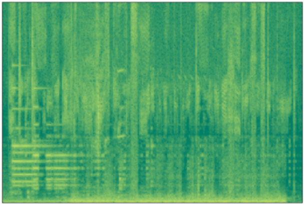
Estimated speech



Synthesized speech example



Estimated background



Take-home messages I

- NMF is a **versatile** data decomposition technique that has proven effective for **diverse applications** across **numerous disciplines**,
  - it tends to provide "meaningful" and "natural" **part-based** data representations,
  - it can be used both for feature learning, topic extraction, clustering, segmentation, source separation, coding...
- For NMF to be successful, it has to be estimated using **appropriate cost-functions** reflecting prior knowledge about the data.
- Being non-unique, NMF should **incorporate constraints** relating to the data, either though:
  - **regularized cost-functions** accounting for sparsity, shape, smoothness, cross-modal dependency constraints..., or
  - alternative formulations, *e.g*, **geometric** approaches having the potential to estimate **exact NMF** models.

Take-home messages II

- Many algorithms are available to estimate NMF, mostly alternating updates of $\mathbf{W}$ and $\mathbf{H}$; variants include:
  - **multiplicative updates:** heuristic, simple and easy to implement, but slow and instable,
  - **majorisation-minimisation:** well-founded for a variety of cost functions, stable, still slow,
  - **gradient-descent** and **Newton**: fast but unstable.
- NMF is a state-of-the-art technique for a number of audio-processing tasks (transcription, source separation...),
- it has a great potential for video (and RBG+depth) analysis tasks, especially temporal structure analysis.

Ongoing and future research

- How to properly estimate the **model-order** K?
- How to achieve **better** and **faster** "convergence"?
- How to perform **non-linear** data decompositions?
- How to handle **big data**?

A selection of NMF software

Software	Language	Main features
beta_ntf	Python	Weighted tensor decomposition, all β-divergences, MM
sklearn.decomposition.NMF	Python	ℓ_2-norm, gradient-descent, sparsity
IMM DTU NMF toolbox	Matlab	ℓ_2-norm, MM, gradient-descent, ALS
Févotte's matlab scripts	Matlab	ℓ_2-norm, KL and IS-div, MM, probabilistic
Seichepine's matlab scripts	Matlab	Soft **co-factorisation**, ℓ_2-norm, KL and IS-div, ℓ_1/ℓ_2-norm **temporal smoothing**, MM
svmnmf	Matlab	Geometric SVM-based NMF, **kernel**-based non-linear decompositions, fast
libNMF	C	ℓ_2-norm, MM, gradient-descent, ALS, multi-core, fast

Bibliography I

A. Cichocki, S. Cruces, and S. Amari. Generalized {A}lpha-{B}eta Divergences and Their Application to Robust Nonnegative Matrix Factorization. *Entropy*, 13:134–170, 2011.

A. Lefèvre, F. Bach, and C. Févotte. {I}takura-{S}aito nonnegative matrix factorization with group sparsity. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011. URL http://perso.telecom-paristech.fr/~fevotte/Proceedings/icassp11c.pdf.

S. A. Abdallah and M. D. Plumbley. Polyphonic transcription by nonnegative sparse coding of power spectra. In *Proc. 5th International Symposium Music Information Retrieval (ISMIR'04)*, pages 318–325, Barcelona, Spain, 2004.

R. Albright, J. Cox, D. Duling, A. Langville, and C. Meyer. Algorithms, initializations, and convergence for the nonnegative matrix factorization. Technical Report Math 81706, NCSU, 2006.

S. Arberet, A. Ozerov, F. Bimbot, and R. Gribonval. A tractable framework for estimating and combining spectral source models for audio source separation. *Signal Processing*, 92(8):1886–1901, 2012.

M. Arngren, M. Schmidt, and J. Larsen. Unmixing of hyperspectral images using Bayesian non-negative matrix factorization with volume prior. *Journal of Signal Processing Systems*, 65(3):479–496, 2011.

R. Badeau, N. Bertin, and E. Vincent. Stability analysis of multiplicative update algorithms and application to non-negative matrix factorization. *IEEE Transactions on Neural Networks*, 21(12):1869–1881, Dec. 2010.

E. Benetos and C. Kotropoulos. A tensor-based approach for automatic music genre classification. In *Proceedings of the European Signal Processing Conference*, Lausanne, Switzerland, 2008.

M. W. Berry and M. Browne. Email Surveillance Using Non-negative Matrix Factorization. *Computational and Mathematical Organization Theory*, 11(3):249–264, Jan. 2006. ISSN 1381-298X. doi: 10.1007/s10588-005-5380-5. URL http://link.springer.com/10.1007/s10588-005-5380-5.

M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons.

Bibliography II

N. Bertin, R. Badeau, and E. Vincent. Enforcing Harmonicity and Smoothness in Bayesian Non-Negative Matrix Factorization Applied to Polyphonic Music Transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):538–549, Mar. 2010. ISSN 1558-7916. doi: 10.1109/TASL.2010.2041381. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5410052.

V. D. Blondel, N.-D. Ho, and P. V. Dooren. Weighted non-negative matrix factorization and face feature extraction. In *Image and Vision Computing*, 2008.

A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *IEEE 11th International Conference on Computer Vision*. IEEE, 2007. URL http://www.computer.org/portal/web/csdl/doi/10.1109/ICCV.2007.4409066.

C. Boutsidis and E. Gallopoulos. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41:1350–1362, 2008.

L. M. Bregman. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.*, 7(3):210–217, 1967.

R. Bro. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38:149–171, Oct. 1997.

J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and Molecular Pattern Discovery Using Matrix Factorization. In *Proceedings of the National Academy of Sciences*, pages 4164–4169, 2004.

S. Bucak and B. Gunsel. Incremental subspace learning via non-negative matrix factorization. *Pattern Recognition*, 42(5):788–797, May 2009.

S. S. Bucak and B. Gunsel. Video Content Representation by Incremental Non-Negative Matrix Factorization. In *2007 IEEE International Conference on Image Processing*, volume 2, pages II – 113–II – 116. IEEE, 2007. ISBN 978-1-4244-1436-9. doi: 10.1109/ICIP.2007.4379105. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4379105.

D. Cai, X. He, J. Han, and T. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33:1548–1560, 2011.

Bibliography III

A. T. Cemgil. Bayesian Inference for Nonnegative Matrix Factorisation Models, Feb. 2009a. URL http://www.hindawi.com/journals/cin/2009/785152.abs.html.

A. T. Cemgil. Bayesian Inference for Nonnegative Matrix Factorisation Models. *Computational Intelligence and Neuroscience*, 2009(Article ID 785152):17 pages, 2009b.

A. T. Cemgil, U. Simsekli, and Y. C. Subakan. Probabilistic latent tensor factorization framework for audio modeling. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'11)*, 2011.

J.-C. Chen. The nonnegative rank factorizations of nonnegative matrices. 62:207–217, Nov 1984. ISSN 00243795. doi: 10.1016/0024-3795(84)90096-X. URL http://www.sciencedirect.com/science/article/pii/002437958490096X.

S. Choi. Algorithms for orthogonal nonnegative matrix factorization. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2008.

a. Cichocki and T. Rutkowski. Constrained non-Negative Matrix Factorization Method for EEG Analysis in Early Detection of Alzheimer Disease. *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, 5(4):V–893–V–896, 2006. doi: 10.1109/ICASSP.2006.1661420. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1661420.

A. Cichocki, R. Zdunek, and S. Amari. Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. In *Proc. 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06)*, pages 32–39, Charleston SC, USA, 2006.

A. Cichocki, Y.-D. K. H. Lee, and S. Choi. Non-negative matrix factorization with α-divergence. *Pattern Recognit. Lett.*, 29:1433–1440, 2008.

O. Cirakman, B. Gunsel, N. Sengor, and O. Gursoy. Key-frame based video fingerprinting by NMF. In *2010 IEEE International Conference on Image Processing*, pages 2373–2376. IEEE, Sept. 2010. ISBN 978-1-4244-7992-4. doi: 10.1109/ICIP.2010.5652649. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5652649.

Bibliography IV

M. Cooper and J. Foote. Summarizing Video using Non-Negative Similarity Matrix Factorization. In *Proc. IEEE Workshop on Multimedia Signal Processing*, volume 00, pages 2–5, 2002. ISBN 0780377141.

C. Damon, A. Liutkus, A. Gramfort, and S. Essid. Nonnegative Matrix Factorization for Single-Channel EEG Artifact Rejection. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013a.

C. Damon, A. Liutkus, A. Gramfort, and S. Essid. Nonnegative Tensor Factorization for Single-Channel EEG Artifact Rejection. In *IEEE International Workshop on Machine Learning for Signal Processing*, Southampton, UK, 2013b.

T. V. de Cruys. A non-negative tensor factorization model for selectional preference induction. *Natural Language Engineering*, 16(4):417–437, 2010.

K. Devarajan and N. Ebrahimi. Molecular pattern discovery using non-negative matrix factorization based on Renyi's information measure. In *Proceedings of the XII SCMA International Conference*, Auburn, Alabama, December 2005.

C. Ding, X. He, and H. Simon. On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. In *SIAM Data Mining Conference*, number 4, 2005. URL http://epubs.siam.org/doi/abs/10.1137/1.9781611972757.70.

C. Ding, W. P. T. Li, and H. Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proc. of the 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 126–135, 2006.

C. H. Ding, T. Li, and M. I. Jordan. Convex and Semi-Nonnegative Matrix Factorizations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(1):45–55, 2010. ISSN 0162-8828. doi: 10.1109/TPAMI.2008.277. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?isnumber=5339303&arnumber=4685898&count=16&index=4.

K. Drakakis, S. Rickard, R. de Frein, and A. Cichocki. Analysis of Financial Data using Non-Negative Matrix Factorization. *International Journal of Mathematical Sciences*, 6(2), 2007.

Bibliography V

N. Q. K. Duong, A. Ozerov, and L. Chevallier. Temporal annotation-based audio source separation using weighted nonnegative matrix factorization. In *4th IEEE International Conference on Consumer Electronics - Berlin (IEEE 2014 ICCE-Berlin)*, 2014.

J. Durrieu, A. Ozerov, and C. Févotte. Main instrument separation from stereophonic audio signals using a source/filter model. *European Signal Processing Conference (EUSIPCO)*, 2009. URL http://www.quaero.org/media/files/bibliographie/eusipco09_2.pdf.

E. Vincent, N. Bertin, and R. Badeau. Harmonic and inharmonic Nonnegative Matrix Factorization for Polyphonic Pitch transcription. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, pages 109–112, 2008. doi: 10.1109/ICASSP.2008.4517558.

J. Eggert and E. Korner. Sparse coding and NMF. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, volume 4, pages 2529–2533. IEEE, 2004. ISBN 0-7803-8359-1. doi: 10.1109/IJCNN.2004.1381036. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1381036.

S. Eguchi and Y. Kano. Robustifying maximum likelihood estimation. Technical report, Institute of Statistical Mathematics, June 2001. Research Memo. 802.

D. El Badawy, N. Q. K. Duong, and A. Ozerov. On-the-fly audio source separation. In *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, Reims, France, Sept. 2014.

S. Essid. A single-class SVM based algorithm for computing an identifiable NMF. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, 2012.

S. Essid and C. Fevotte. Decomposing the Video Editing Structure of a Talk-show using Nonnegative Matrix Factorization. In *International Conference on Image Processing (ICIP)*, Orlando, FL, USA, 2012.

S. Essid and C. Fevotte. Smooth Nonnegative Matrix Factorization for Unsupervised Audiovisual Document Structuring. *IEEE Transactions on Multimedia*, 15(2):415–425, 2013. ISSN 1520-9210. doi: 10.1109/TMM.2012.2228474.

Bibliography VI

C. Fevotte. Slides of lecture on unsupervised data decompositions. Technical report, CNRS/LTCI, Telecom ParisTech, 2012.

C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. Oct. 2010. URL http://arxiv.org/abs/1010.1763.

C. Févotte and A. Ozerov. Notes on nonnegative tensor factorization of the spectrogram for audio source separation~: statistical insights and towards self-clustering of the spatial cues. In *Proc. 7th International Symposium on Computer Music Modeling and Retrieval (CMMR)*, volume 5493 of *Lecture Notes in Computer Science*, pages 102–115, Malaga, Spain, 2010. Springer. URL http://perso.telecom-paristech.fr/~fevotte/Proceedings/cmmr10.pdf.

C. Fevotte, N. Bertin, and J.-L. Durrieu. Nonnegative Matrix Factorization with the Itakura-Saito Divergence. With Application to Music Analysis. *Neural Computation*, 21(3), Mar. 2009.

C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the {I}takura-{S}aito divergence. {W}ith application to music analysis. *Neural Computation*, 21(3):793–830, 2009. doi: 10.1162/neco.2008.04-08-771. URL http://www.tsi.enst.fr/~fevotte/Journals/neco09_is-nmf.pdf.

D. FitzGerald, M. Cranitch, and E. Coyle. Extended Nonnegative Tensor Factorisation Models for Musical Sound Source Separation. *Computational Intelligence and Neuroscience*, 2008(Article ID 872425):15 pages, 2008. doi: 10.1155/2008/872425.

H. Fujisawa and S. Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, Oct. 2008.

Y. Gao and G. Church. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics*, 21:3970–3975, 2005. doi: doi:10.1093/bioinformatics/bti653.

E. Gaussier and C. Goutte. Relation between PLSA and NMF and implications. In *Proc. 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'05)*, pages 601–602, New York, NY, USA, 2005. ACM. ISBN 1595930345. URL http://dl.acm.org/citation.cfm?id=1076148.

Bibliography VII

N. Gillis. *Regularization, Optimization, Kernels, and Support Vector Machines*, chapter The why and how of nonnegative matrix factorization. Chapman & Hall/CRC, 2014.

D. Greene, G. Cagney, N. Krogan, and P. Cunningham. Ensemble Non-negative Matrix Factorization Methods for Clustering Protein-Protein Interactions. *Bioinformatics*, 24(15):1722–1728, 2008.

R. Hennequin, B. David, and R. Badeau. Score informed audio source separation using a parametric model of non-negative spectrogram. In *International Conference on Acoustics, Speech, and Signal Processing*, 2011. URL http://hal.inria.fr/docs/00/94/52/94/PDF/hennequin_icassp2011.pdf.

T. Hofmann. Probabilistic latent semantic analysis. *Proceedings of the Fifteenth conference on Uncertainty . . .*, 1999. URL http://dl.acm.org/citation.cfm?id=2073829.

P. O. Hoyer. Non-negative Matrix Factorization with Sparseness Constraints. *The Journal of Machine Learning Research*, 5:1457–1469, Dec. 2004. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1005332.1044709.

D. R. Hunter and K. Lange. A tutorial on mm algorithms. *Amer. Stat.*, 58(1):30–37, Feb. 2004.

I. S. Dhillon and S. Sra. Generalized Nonnegative Matrix Approximations with {B}regman Divergences. *Advances in Neural Information Processing Systems (NIPS)*, 19, 2005.

M. Jeter and W. Pye. A note on nonnegative rank factorizations. *Linear Algebra and its Applications*, 38:171–173, Jun 1981. ISSN 00243795. doi: 10.1016/0024-3795(81)90018-5. URL http://www.sciencedirect.com/science/article/pii/0024379581900185.

S. Jia and Y. Qian. Constrained Nonnegative Matrix Factorization for Hyperspectral Unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 47(1):161–173, Jan. 2009. ISSN 0196-2892. doi: 10.1109/TGRS.2008.2002882. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4694061.

M. M. Kalayeh, H. Idrees, and M. Shah. NMF-KNN : Image Annotation using Weighted Multi-view Non-negative Matrix Factorization. In *CVPR*, 2014.

Bibliography VIII

A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. O. . Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering. In ACM, editor, *Proceedings of the 4th ACM Conference on Recommender Systems (RecSys'10)*, pages 79–86, New York, NY, 2010.

H. A. L. Kiers. Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics*, 14: 105–122, 2000.

J. Kim and H. Park. Sparse Nonnegative Matrix Factorization for Clustering. Technical report, Georgia Institute of Technology, 2008. URL https://smartech.gatech.edu/handle/1853/20058.

T.-K. Kim and R. Cipolla. Canonical correlation analysis of video volume tensors for action categorization and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8):1415–1428, 2009.

Y.-D. Kim and S. Choi. A Method of Initialization for Nonnegative Matrix Factorization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, volume 2, pages 537–540, Honolulu, Hawaii, 2007.

B. Klingenberg, J. Curry, and A. Dougherty. Non-negative matrix factorization: Ill-posedness and a geometric algorithm. *Pattern Recognition*, 42(5):918–928, May 2009. ISSN 0031-3203. doi: 10.1016/j.patcog.2008.08.026. URL http://linkinghub.elsevier.com/retrieve/pii/S0031320308003403http://www.sciencedirect.com/science/article/B6V14-4TCR1KR-1/2/8bedc245dc0fd9ba7487561f8df431cahttp://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V14-4TCR1KR-1&_user=771355&_coverDate=05/31/2009&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_searchStrId=1272450458&_rerunOrigin=google&_acct=C000028498&_version=1&_urlVersion=0&_userid=771355&md5=35ba172c08afbfa2b0676a0f2dca1897.

B. Krausz and C. Bauckhage. Action Recognition in Videos Using Nonnegative Tensor Factorization. In *2010 20th International Conference on Pattern Recognition*, pages 1763–1766. IEEE, Aug. 2010. ISBN 978-1-4244-7542-1. doi: 10.1109/ICPR.2010.435. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5597190.

J. B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2):95 – 138, 1977.

F. Kschischang, B. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

Bibliography IX

L. Le Magoarou, A. Ozerov, and N. Q. K. Duong. Text-informed audio source separation using nonnegative matrix partial co-factorization. In *Proc IEEE. Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2013.

D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 401: 788–791, 1999.

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural and Information Processing Systems 13*, pages 556–562, 2001.

H. Lee, Y.-D. Kim, A. Cichocki, and S. Choi. Nonnegative tensor factorization for continuous EEG classification. *International Journal of Neural Systems*, 17(4):305–317, 2007.

H. Lee, A. Cichocki, and S. Choi. Kernel nonnegative matrix factorization for spectral EEG feature extraction. *Neurocomputing*, 72(13-15):3182–3190, Aug. 2009. ISSN 09252312. doi: 10.1016/j.neucom.2009.03.005. URL `http://linkinghub.elsevier.com/retrieve/pii/S0925231209000757`.

Y. Li, D. Sima, S. V. Cauter, A. C. Sava, U. Himmelreich, Y. Pi, and S. V. Huffel. Hierarchical non-negative matrix factorization (hNMF): A tissue pattern differentiation method for glioblastoma multiforme diagnosis using MRSI. *NMR in Biomedicine*, 26:307–319, 2013.

L.-H. Lim and P. Comon. Multiarray signal processing: Tensor decomposition meets compressed sensing. *Compte-Rendus de l'Academie des Sciences, section Mecanique*, 338(6):311—320, June 2010.

A. Limem, G. Delmaire, M. Puigt, G. Roussel, and D. Courcot. Non-negative matrix factorization using weighted Beta divergence and equality constraints for industrial source apportionment. In *23rd IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2013)*, Southampton, UK, September 22–25 2013.

C.-J. Lin. On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization. *IEEE Transactions on Neural Networks*, 18:1589–1596, 2007a.

C.-J. Lin. Projected Gradient Methods for Nonnegative Matrix Factorization. *Neural Computation*, 19:2756–2779, 2007b.

J. Liu, C. Wang, J. Gao, and J. Han. Multi-view clustering via joint nonnegative matrix factorization. In *Proc. of SDM*, 2013. URL `http://epubs.siam.org/doi/abs/10.1137/1.9781611972832.28`.

Bibliography X

X. Liu, Q. Xu, S. Yan, G. W. H. Jin, and S.-W. Lee. Nonnegative tensor co-factorization and its unified solution. *IEEE Transactions on Image Processing (TIP)*, 2014.

H. A. Loeliger. An introduction to factor graphs. *IEEE Signal ProcessingMagazine*, 21:28–41, 2004.

W. Lu, W. Sun, and H. Lu. Robust watermarking based on dwt and nonnegative matrix factorization. *Computers and Electrical Engineering*, 35(1):183–188, 2009.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11(10-60), 2010. URL `http://dl.acm.org/citation.cfm?id=1756008`.

A. Masurelle, S. Essid, and G. Richard. Gesture recognition using a NMF-based representation of motion-traces extracted from depth silhouettes. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014.

S. Meignier and T. Merlin. {LIUM SPKDIARIZATION}: AN OPEN SOURCE TOOLKIT FOR DIARIZATION. In *CMU SPUD Workshop*, Texas, USA, 2010. URL `http://lium3.univ-lemans.fr/diarization/doku.php/welcome`.

P. Melville and V. Sindhwai. Recommender systems. In C. Sammut and W. G., editors, *Encyclopedia of Machine Learning*. Springer-Verlag, 2010. URL `http://dl.acm.org/citation.cfm?id=245121`.

L. Miao and H. Qi. Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *IEEE Trans. on Geoscience and Remote Sensing*, 47:765–777, 2007.

N. Mohammadiha, P. Smaragdis, and A. Leijon. Supervised and unsupervised speech enhancement using NMF. *IEEE Transactions on Audio Speech and Language Processing*, 21(10):2140–2151, Oct. 2013.

V. Monga and M. K. Mihcak. Robust and secure image hashing via non-negative matrix factorizations. *IEEE Trans. on information Forensics and Security*, 2(3):376–390, Sep. 2007.

J. Nikunen, T. Virtanen, and M. Vilermo. Multichannel audio upmixing based on non-negative tensor factorization representation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'11)*, pages 33–36, 2011.

Bibliography XI

A. Ozerov and C. Fevotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE transactions on Audio, Speech, and Language Processing*, 18(3):550–563, 2010. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5229304.

A. Ozerov, C. Févotte, and M. Charbit. Factorial scaled hidden Markov model for polyphonic audio representation and source separation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'09)*, Mohonk, NY, Oct. 18-21 2009.

A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu. Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011a. URL http://perso.telecom-paristech.fr/~fevotte/Proceedings/icassp11d.pdf.

A. Ozerov, A. Liutkus, R. Badeau, and G. Richard. Informed source separation: source coding meets source separation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'11)*, Mohonk, NY, Oct. 2011b.

A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 20(4):1118–1133, 2012.

A. Ozerov, A. Liutkus, R. Badeau, and G. Richard. Coding-Based Informed Source Separation: Nonnegative Tensor Factorization Approach. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(8):1699–1712, Aug. 2013. ISSN 1558-7916. doi: 10.1109/TASL.2013.2260153. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6508860.

P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, Jun 1994. ISSN 11804009. doi: 10.1002/env.3170050203. URL http://doi.wiley.com/10.1002/env.3170050203.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Bibliography XII

Z. Rafii, D. Sun, F. Germain, and G. Mysore. Combining modeling of singing voice and background music for automatic separation of musical mixtures. In *14th International Society for Music Information Retrieval (ISMIR)*, Curitiba, PR, Brazil, 2013.

S. Rendle, L. B. Marinho, A. Nanopoulos, and . L.S. Thieme. Learning optimal ranking with tensor factorization for tag recommendation. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 727–736, 2009.

J. L. Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama. Computational auditory induction as a missing-data model-fitting problem with Bregman divergence. *Speech Communication*, 53(5):658–676, May-June 2011.

R. Sandler and M. Lindenbaum. Nonnegative Matrix Factorization with Earth Mover's Distance Metric for Image Analysis. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1590–1602, Jan. 2011. ISSN 1939-3539. doi: 10.1109/TPAMI.2011.18. URL http://www.ncbi.nlm.nih.gov/pubmed/21263163.

M. N. Schmidt and M. Morup. Infinite non-negative matrix factorizations. In *Proc. European Signal Processing Conference (EUSIPCO)*, 2010.

M. N. Schmidt, J. Larsen, and F.-T. Hsiao. Wind Noise Reduction using Non-Negative Sparse Coding. In *2007 IEEE Workshop on Machine Learning for Signal Processing*, pages 431–436. IEEE, Aug. 2007. ISBN 978-1-4244-1565-6. doi: 10.1109/MLSP.2007.4414345. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4414345.

M. N. Schmidt, O. Winther, and L. K. Hansen. Bayesian non-negative matrix factorization. In *Independent Component Analysis and Signal Separation*, pages 540–547, 2009.

N. Seichepine, S. Essid, C. Fevotte, and O. Cappe. Soft nonnegative matrix co-factorization with application to multimodal speaker diarization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, 2013.

N. Seichepine, S. Essid, C. Fevotte, and O. Cappe. Piecewise constant nonnegative matrix factorization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014a.

Bibliography XIII

N. Seichepine, S. Essid, C. Fevotte, and O. Cappe. Soft l1 and l2 coupling of nonnegative matrix factorization problems. *submitted to IEEE transactions on signal processing*, 2014b.

A. Shashua and T. Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In *ICML 2005: Proceedings of the 22nd International Conference on Machine Learning*, pages 792–799, 2005.

L. S. R. Simon and E. Vincent. A general framework for online audio source separation. In *International conference on Latent Variable Analysis and Signal Separation*, Tel-Aviv, Israel, Mar. 2012.

J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objets and their locations in images. In *ICCV*, Beijing, China, 2005.

P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2003.

D. Soukup and I. Bajla. Robust object recognition under partial occlusions using NMF. *Computational intelligence and neuroscience*, 2008. URL http://www.hindawi.com/journals/cin/aip/857453/.

D. Sun and R. Mazumder. Non-negative matrix completion for bandwidth extension: A convex optimization approach. *Machine Learning for Signal . . .*, 2013. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6661924.

Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization with the β-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1592–1605, 2013.

M. Türkan and C. Guillemot. Image prediction based on neighbor-embedding methods. *IEEE Transactions on Image Processing*, 21(4):1885–1898, 2011.

E. Vincent, N. Bertin, and R. Badeau. Two nonnegative matrix factorization methods for polyphonic pitch transcription. In *Proc. Music Information Retrieval Evaluation eXchange (MIREX)*, 2007.

E. Vincent, N. Bertin, and R. Badeau. Harmonic and inharmonic Nonnegative Matrix Factorization for Polyphonic Pitch transcription. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 109–112. IEEE, Mar. 2008. ISBN 978-1-4244-1483-3. doi: 10.1109/ICASSP.2008.4517558. URL http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=4517558.

Bibliography XIV

E. Vincent, N. Bertin, and R. Badeau. Adaptive Harmonic Spectral Decomposition for Multiple Pitch Estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):528–537, Mar. 2010. ISSN 1558-7916. doi: 10.1109/TASL.2009.2034186. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5282583.

A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. Canal9: A database of political debates for analysis of social interactions. In *IEEE International Workshop on Social Signal Processing*, Amsterdam, 2009. Ieee. ISBN 978-1-4244-4800-5. doi: 10.1109/ACII.2009.5349466. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5349466.

T. Virtanen. Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1066–1074, 2007.

T. Virtanen and A. Cemgil. Mixtures of gamma priors for non-negative matrix factorization based speech separation. In *Independent Component Analysis and Signal Separation*. Springer-Verlag, 2009. URL http://link.springer.com/chapter/10.1007/978-3-642-00599-2_81.

F. Wang, T. Li, X. Wang, S. Zhu, and C. Ding. Community discovery using nonnegative matrix factorization. *Data Mining and Knowledge Discovery*, 22(3):493–521, July 2010. ISSN 1384-5810. doi: 10.1007/s10618-010-0181-y. URL http://link.springer.com/10.1007/s10618-010-0181-y.

K. Wilson, B. Raj, and P. Smaragdis. Regularized non-negative matrix factorization with temporal dependencies for speech denoising. In *INTERSPEECH*, 2008. URL http://www.cs.illinois.edu/~paris/Paris_Smaragdis_page/Paris_Smaragdis_Publications_files/wilson-is2008.pdf.

W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03*, page 267, New York, New York, USA, July 2003. ACM Press. ISBN 1581136463. doi: 10.1145/860435.860485. URL http://dl.acm.org/citation.cfm?id=860435.860485.

Z. Yang and E. Oja. Linear and nonlinear projective nonnegative matrix factorization. *IEEE Trans. on Neural Networks*, 21:734–749, 2010.

Bibliography XV

Z. Yang and E. Oja. Unified Development of Multiplicative Algorithms for Linear and Quadratic Nonnegative Matrix Factorization. *IEEE Trans. Neural Networks*, 22(12):1878–1891, 2011. doi: http://dx.doi.org/10.1109/TNN.2011.2170094.

K. Yilmaz and A. T. Cemgil. Probabilistic latent tensor factorisation. In *Proc. of International Conference on Latent Variable analysis and Signal Separation*, pages 346–353, 2010.

K. Y. Yilmaz, A. T. Cemgil, and U. Simsekli. Generalized coupled tensor factorization. In *NIPS*, 2011.

R. Zdunek and A. Cichocki. Non-negative matrix factorization with quasi-Newton optimization. In *Eighth International Conference on Artificial Intelligence and Soft Computing (ICAISC)*, pages 870–879, 2006.

M. Zetlaoui, M. Feinberg, P. Verger, and S. Clémençon. Extraction of food consumption systems by non-negative matrix factorization (NMF) for the assessment of food choices. Technical report, Arxiv, 2010. URL http://hal.archives-ouvertes.fr/docs/00/48/47/94/PDF/NMF_food.pdf.