An introduction to Nonnegative Matrix Factorisation

Slim ESSID

Telecom ParisTech

June 2015

Credits

Some illustrations, slides and demos are reproduced courtesy of:

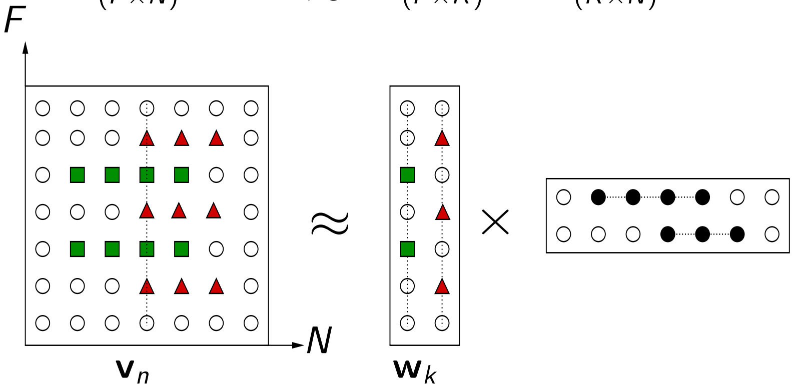
- A. Ozerov,
- C. Févotte,
- N. Seichepine,
- R. Hennequin,
- F. Vallet,
- A. Liutkus.

- Introduction
- NMF models
- Algorithms for solving NMF
- Applications
- Conclusion

Explaining data by factorisation

General formulation

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

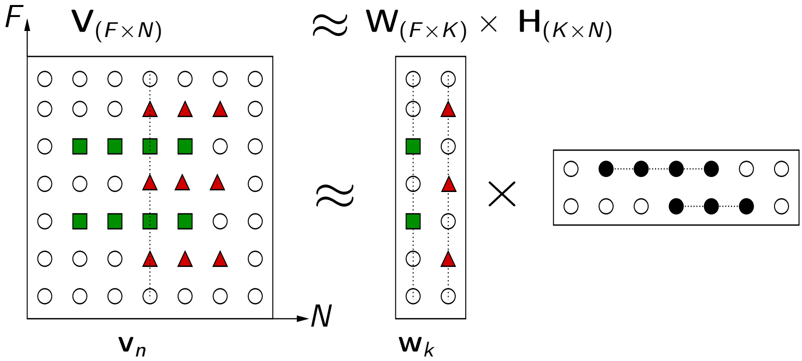


$$\mathbf{v}_n \approx \sum_{k=1}^{K} h_{kn}\mathbf{w}_k$$

Illustration by C. Févotte

Explaining data by factorisation

General formulation



data matrix

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

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

Illustration by C. Févotte

Data is often nonnegative by nature[1]

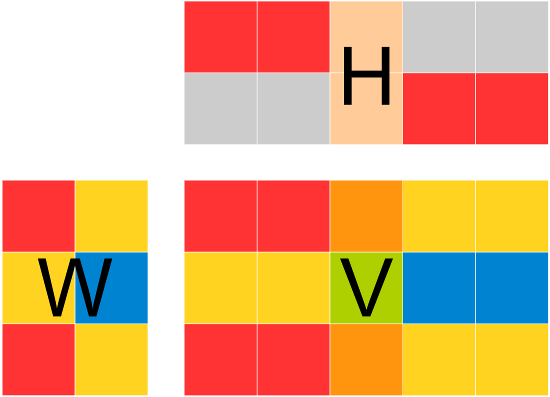
- pixel intensities;
- amplitude spectra;
- occurrence counts;
- food or energy consumption;
- user scores;
- stock market values;
- ...

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

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

The Nonnegative Matrix Factorisation model

NMF provides an unsupervised linear representation of the data:



$$\mathbf{V} \approx \mathbf{WH};$$

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

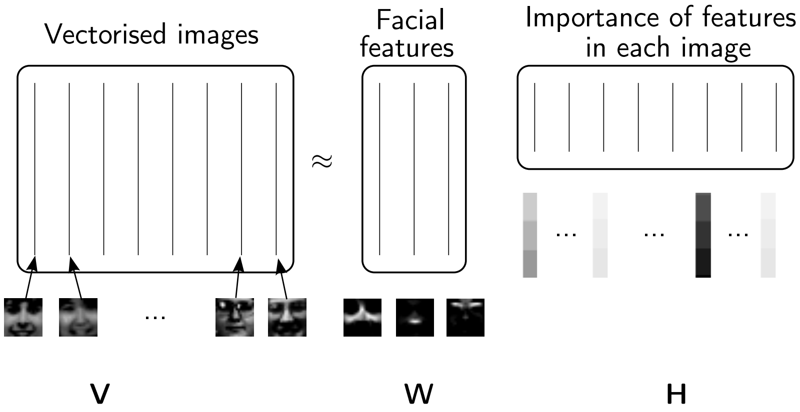
Illustration by N. Seichepine

Explaining face images by NMF[2]

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

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

Explaining face images by NMF
Method



NMF outputs

Image example

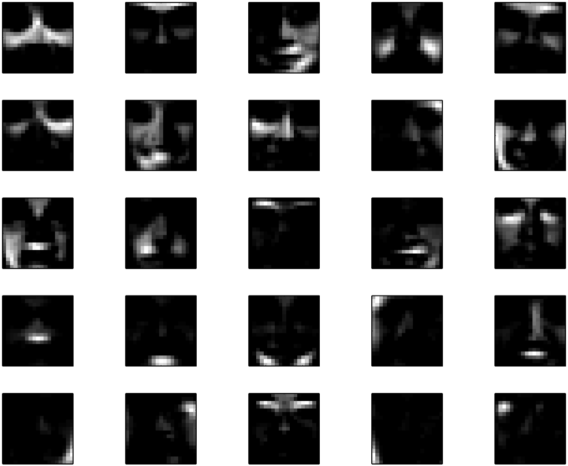


Illustration by C. Févotte

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

- $\mathbf{H}$: 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 $\mathbf{w}_k$.

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)

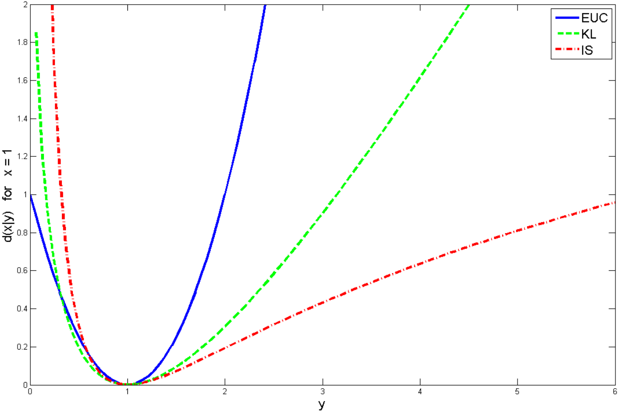
$$d_{KL}(x|y) = x \log \frac{x}{y} - x + y$$

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[3]

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

The IS divergence is **scale-invariant** $\rightarrow$ it provides higher accuracy in the representation of data with large dynamic range (*e.g.* audio spectra).

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

Weighted NMF

Conventional NMF optimization criterion:

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

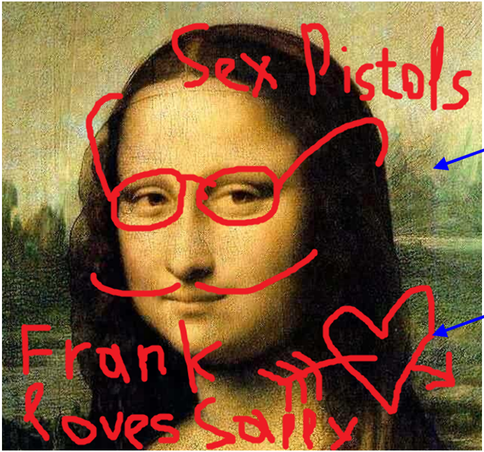
Weighted NMF optimization criterion:

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

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

Weighted NMF application example I

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



Observed value

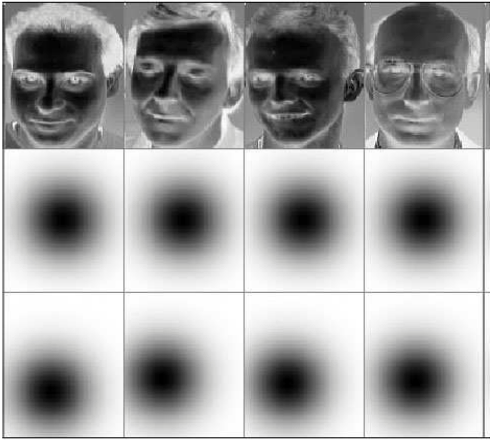
$b_{fn} = 1$

Missing value

$b_{fn} = 0$

Weighted NMF application example II

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



Data **V**

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

Image-centered weights

Face-centered weights

- Introduction
- NMF models
- Algorithms for solving NMF
  - Preliminaries
  - Difficulties in NMF
  - Multiplicative update rules
- Applications
- Conclusion

Optimization problem

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;
- the solution is **not unique**...

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

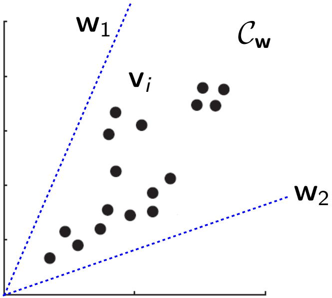
In particular, $\mathbf{Q}$ can be any **nonnegative generalised permutation matrix**; *e.g.*, in $\mathbb{R}^3$:

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

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

Geometric interpretation and ill-posedness

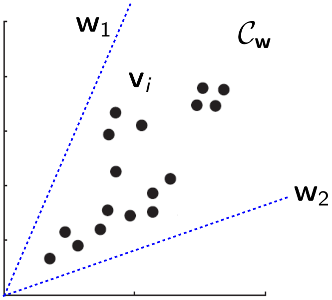
NMF assumes the data is well described by a **simplicial convex cone** $\mathcal{C}_{\mathbf{w}}$ generated by the columns of $\mathbf{W}$:

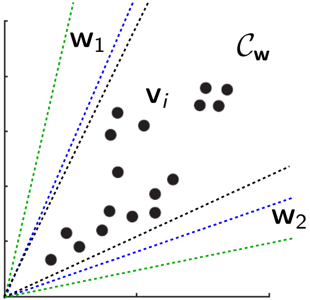


$$\mathcal{C}_{\mathbf{w}} = \left\{ \sum_{k=1}^{K} \lambda_k \mathbf{w}_k ;\ \lambda_k \geq 0 \right\}$$

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 $\mathbf{W}$:



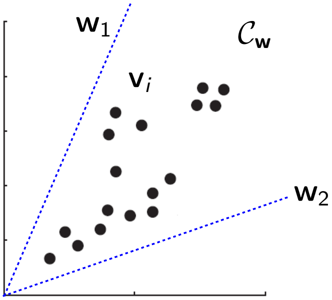


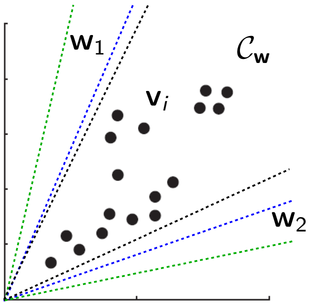
$$\mathcal{C}_{\mathbf{w}} = \left\{ \sum_{k=1}^{K} \lambda_k \mathbf{w}_k ;\ \lambda_k \geq 0 \right\}$$

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.

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 $\mathbf{H}$ (similarly for $\mathbf{W}$) 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

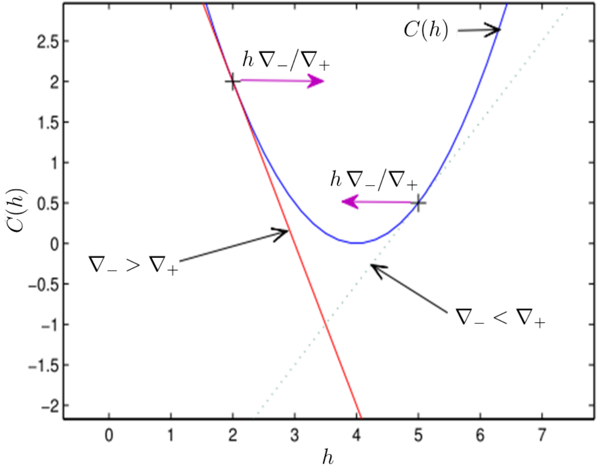
$$\nabla_{h_{kn}} C(\boldsymbol{\theta}) = \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_{+} - \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_{-} \ ,$$

and the summands are both nonnegative.

NOTE: The nonnegativity of $\mathbf{W}$ and $\mathbf{H}$ is guaranteed by construction.

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:**

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

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 $C(s)$.

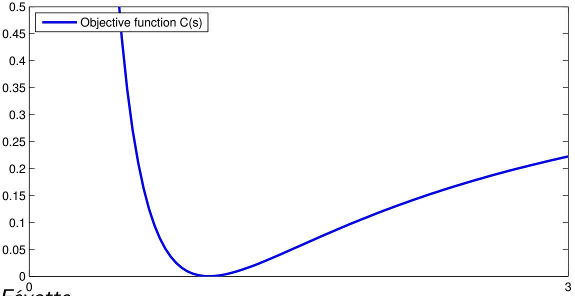


Illustration by C. Févotte

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 $C(s)$.

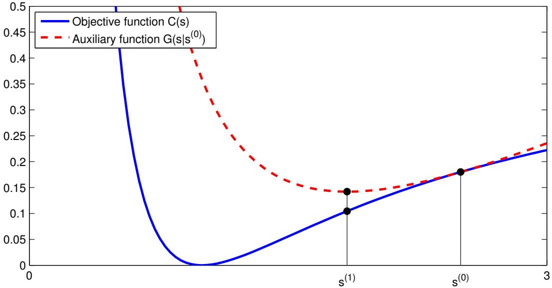


Illustration by C. Févotte

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 $C(s)$.

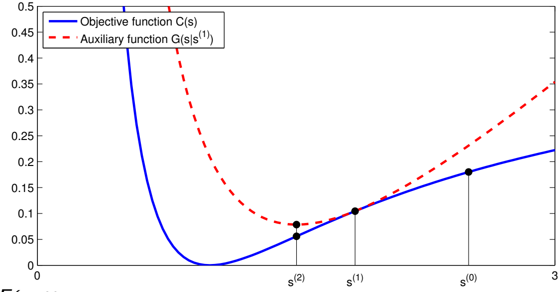


Illustration by C. Févotte

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 $C(s)$.

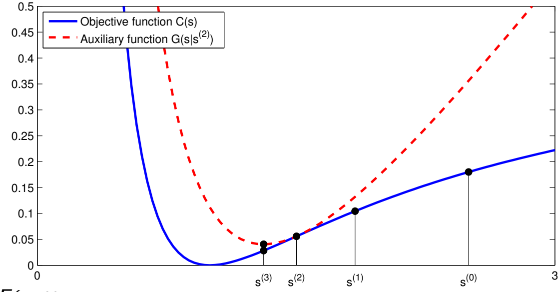


Illustration by C. Févotte

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 $C(s)$.

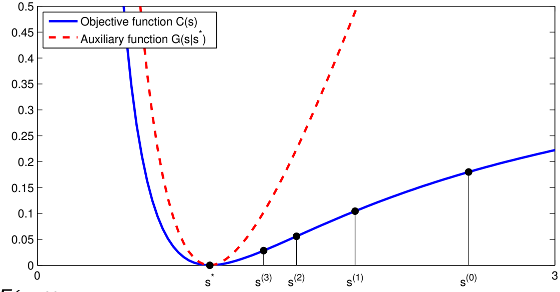


Illustration by C. Févotte

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:

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

Summary

Multiplicative Update rules:

Advantages:

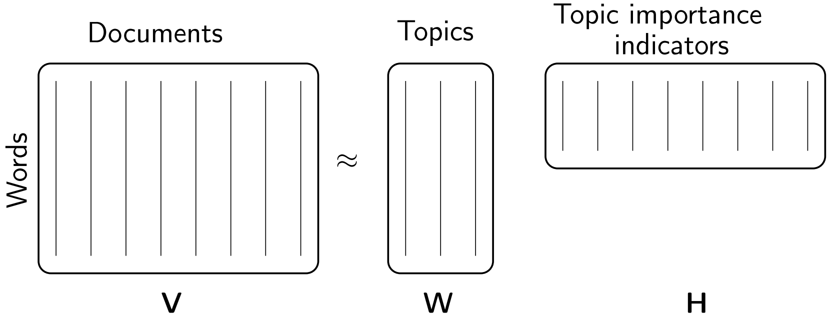
- easy to implement;
- non-negativity of $\mathbf{W}$ and $\mathbf{H}$ is guaranteed.

Drawbacks:

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

Topics recovery

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



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

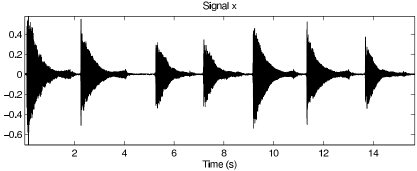
Topics described by most frequent words in each dictionary element $\mathbf{W}_k$.

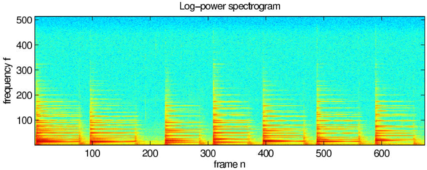
- Introduction
- NMF models
- Algorithms for solving NMF
- Applications
  - Text analysis
  - Music transcription
  - Video structuring
- Conclusion

NMF-based music transcription

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

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

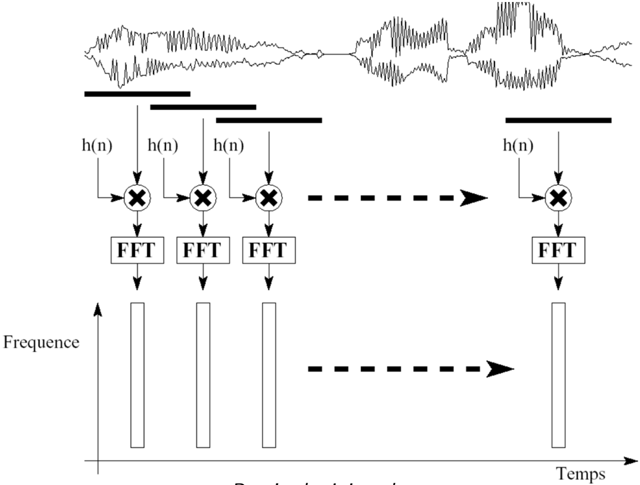




Three representations of the data.

Spectral analysis

Short-Term Fourier Transform (STFT)

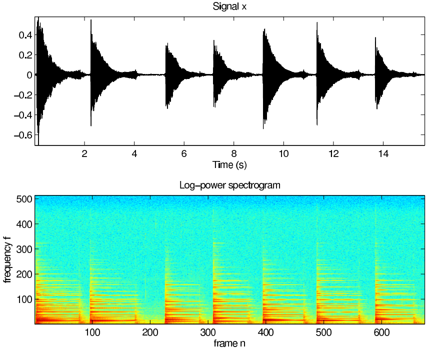


Drawing by J. Laroche

NMF-based music transcription demo

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

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

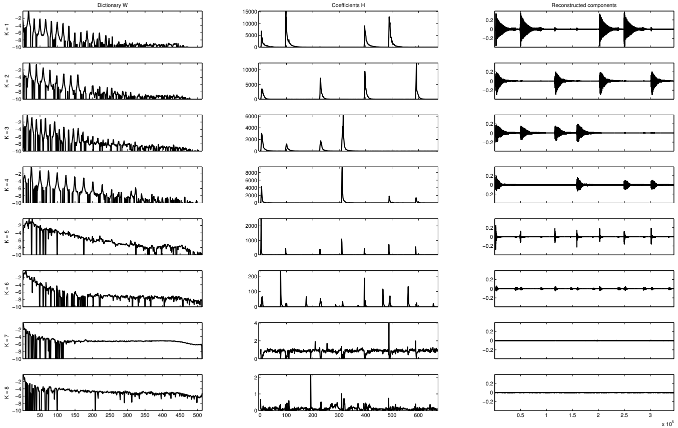


Three representations of the data.

Music transcription demo

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

NMF decomposition with $K = 8$



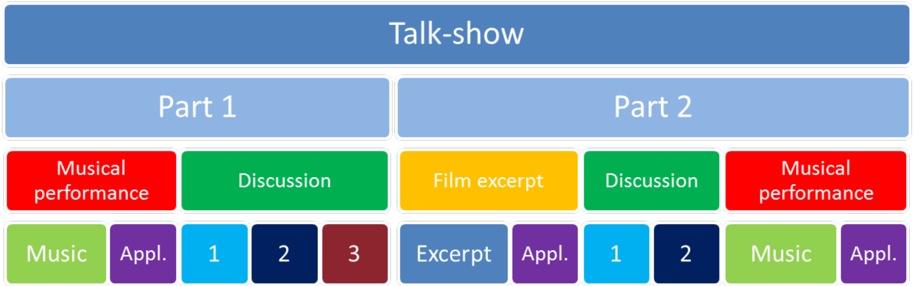
Pitch estimates: 65.0 68.0 61.0 72.0 0 0 0 0

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

- Introduction
- NMF models
- Algorithms for solving NMF
- Applications
  - Text analysis
  - Music transcription
  - Video structuring
- Conclusion

The video structuring problem

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



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)

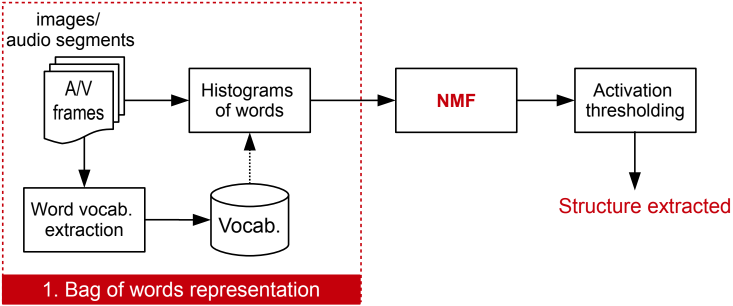
"Who spoke when?"

illustration by N. Seichepine

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):

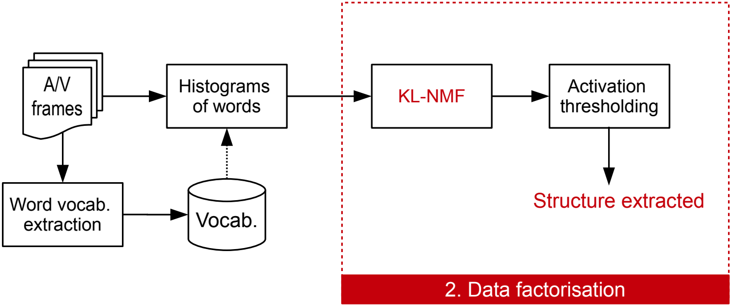


1. create a low-level (visual/audio) vocabulary and use it to extract **histogram of (visual/audio) words** from the sequence of observation frames;

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):

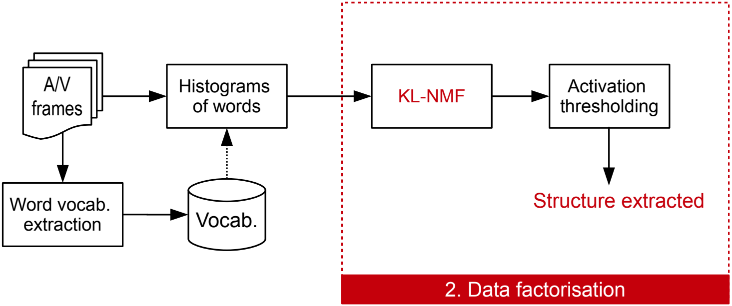


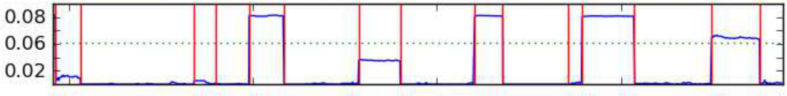
2. apply a variant of **smooth NMF** using the **Kullback-Leibler** divergence to extract **latent structuring events** and their **activations** across the duration of the document.

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):

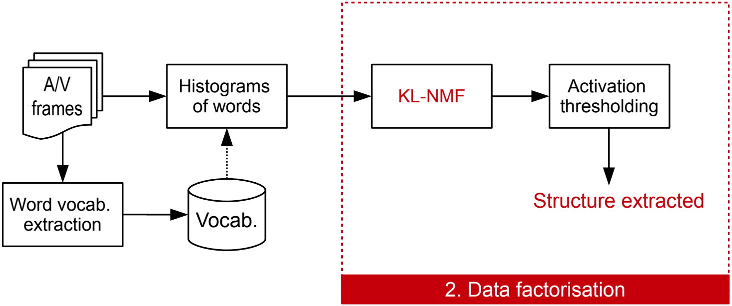




A generic video structuring system using NMF

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.

Smooth KL-NMF

Using the Kullback-Leibler (KL) divergence as a measure of fit

Given histogram data (whose columns are frame-wise descriptors), we seek a factorization $\mathbf{V} \approx \mathbf{WH}$; $w_{fk} \geq 0$; $h_{kn} \geq 0$ that minimises

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

- $D(\mathbf{V}|\mathbf{WH}) = \sum_{fn} d_{KL}(v_{fn}|\sum_k w_{fk} h_{kn})$: **fit-to-data term** such that $d_{KL}(x|y) = x \log \frac{x}{y} - x + y$;
- $S(H)$ is a **regularisation** term that controls the **temporal smoothness** of the activation coefficients:

$$S(H) = \frac{1}{2} \sum_{k=1}^{K} \sum_{n=2}^{N} (h_{kn} - h_{k(n-1)})^2.$$

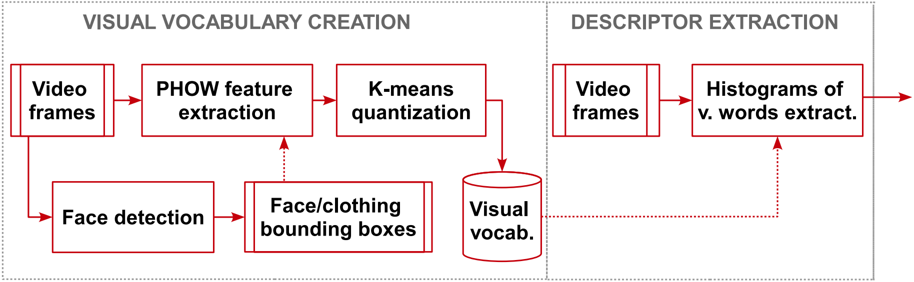
Applications

Onscreen person-oriented structuring

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

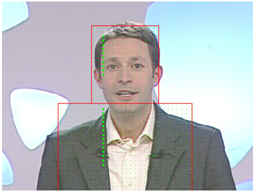
Using the **Canal9 political debates** database (Vinciarelli et al., 2009).

Visual features



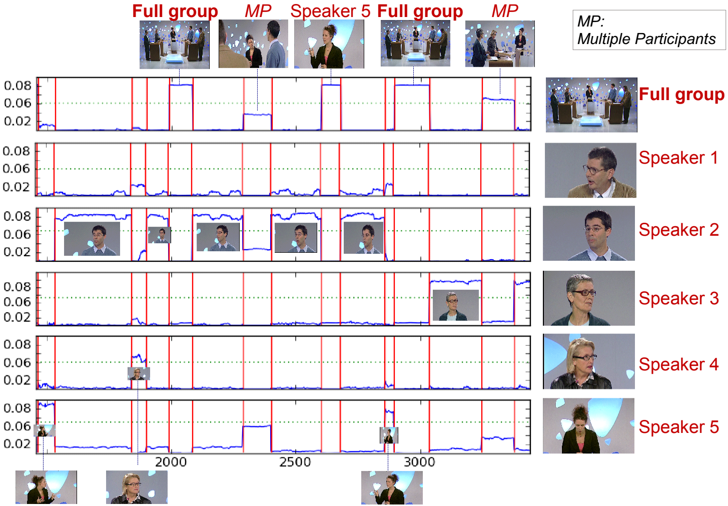
Visual vocabulary creation

- **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.



Results

Visualising the activations



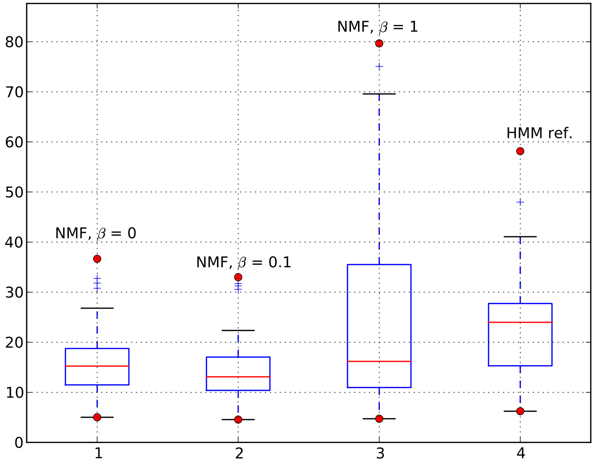
Experimental validation

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.

Results

Shot-type classification error rates



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.

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 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

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.

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.

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 Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

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

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.

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.

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 II

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.

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.