An introduction to NMF

followed by:

A Quasi-Newton algorithm on the orthogonal manifold for NMF with transform learning

Pierre Ablin
Parietal

Joint work with: D.Fagot, H.Wendt, A.Gramfort and C.Févotte

https://arxiv.org/pdf/1811.02225.pdf

Parietal presentation, 2018

Introduction to NMF

Non-negative Matrix Factorization

Matrix factorization technique, just like:

- Dictionnary learning
- Principal Component Analysis
- Independent Component Analysis
- ...

NMF: the problem

Let $V \in \mathbb{R}^{p \times n}$ a matrix of **positive** entries.
The **rank-k NMF** of V consists in finding $W \in \mathbb{R}^{p \times k}$ and $H \in \mathbb{R}^{k \times n}$ of **positive** entries such that (Lee and Seung, 1999):

$$\boxed{V \simeq WH}$$

- Used with $k < \min(n, p)$ to obtain a low dimensional representation of V.
- Lifts some of the usual indeterminacy of the factorization : $WH = (WM^{-1})(MH)\ \forall M \in \mathbb{R}^{k \times k}$ invertible.
- Only scale and order indeterminacy remains.

Applications of NMF

Applied to data that are instrinsically nonnegative:

- Spectrograms: astronomy (Blanton and Roweis, 2007), music signal processing(Smaragdis and Brown, 2003), neuroscience (Rutkowski et al., 2007), $\cdots$
- Gene expression matrix in biology (Devarajan, 2008)
- Document-term matrix in text mining (Arora et al., 2013)

Algorithms for NMF

NMF as an optimization problem: find W, H solution of :

$$\text{minimize } d(V||WH) \text{ s.t } W \geq 0, H \geq 0$$

Several choices for d:

- Frobenius: $d(V||\hat{V}) = ||V - \hat{V}||_F^2$
- Kullback-Leibler divergence:

$$d(V||\hat{V}) = \sum_{i,j} V_{ij} \log(\frac{V_{ij}}{\hat{V}_{ij}}) + \hat{V}_{ij} - V_{ij}$$

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

$$d(V||\hat{V}) = \sum_{i,j} \frac{V_{ij}}{\hat{V}_{ij}} - \log(\frac{V_{ij}}{\hat{V}_{ij}}) - 1$$

Multiplicative update rules

$$\text{minimize } d(V||WH) \text{ s.t } W \geq 0, H \geq 0$$

Alternate optimization: Fix W, and update H, then fix H and update W.

Safe multiplicative update rules, e.g. for Itakura-Saito divergence ($\odot$ is element-wise multiplication):

$$H \leftarrow H \odot \frac{W^\top((WH)^{\odot -2} \odot V)}{W^\top(WH)^{\odot -1}}$$

$$W \leftarrow W \odot \frac{((WH)^{\odot -2} \odot V)H^\top}{(WH)^{\odot -1}H^\top}$$

Safe = one iteration decreases the cost function.

- What happens to the iterations if the factorization is perfect?
- Equivalent to alternate diagonaly rescaled gradient descent

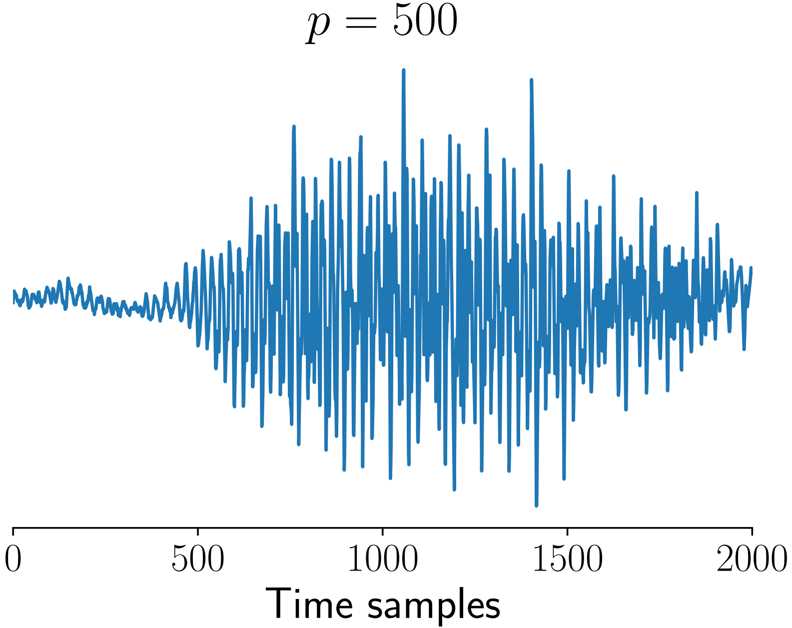
NMF applied to music processing

NMF for audio signal processing

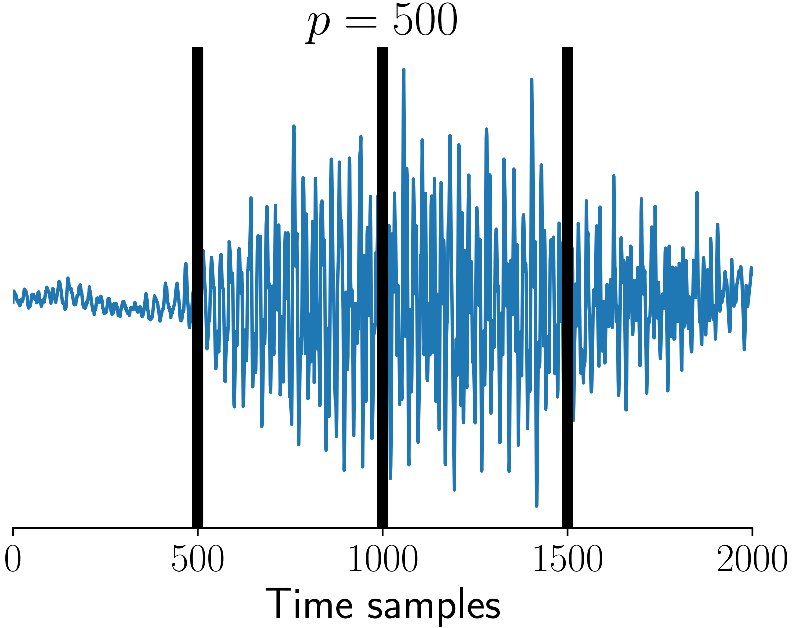
NMF is usually applied to the **spectrogram** of a song.

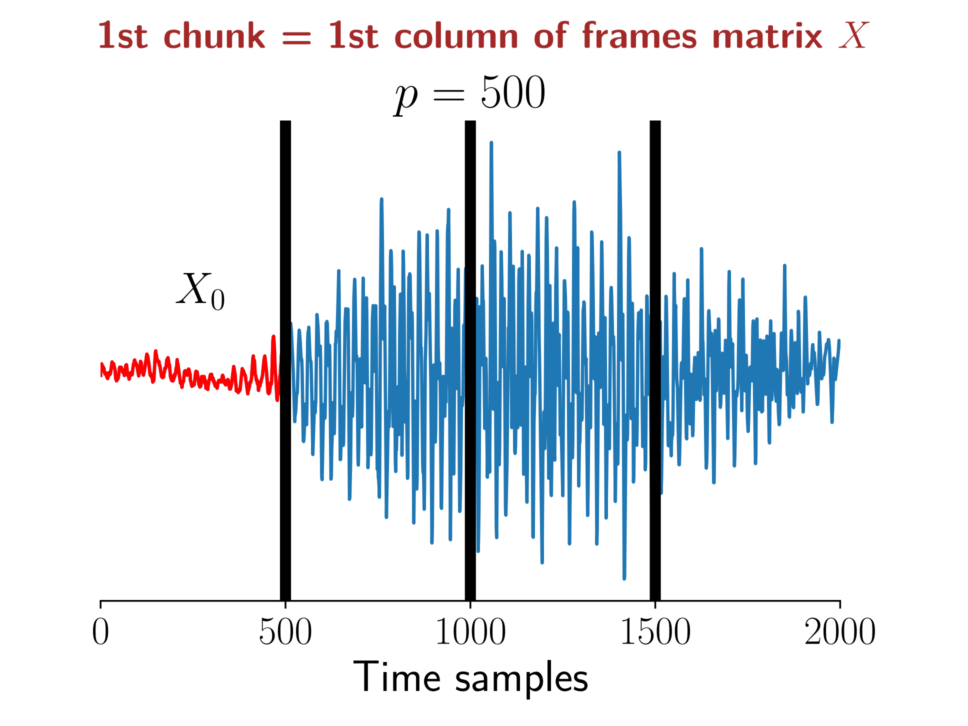
- Signal s of T samples, sampled at f_s.
- Cut it in n frames of size p
- Yields a **frames** matrix $X \in \mathbb{R}^{p \times n}$

Toy example on 2000 samples with $p = 500$

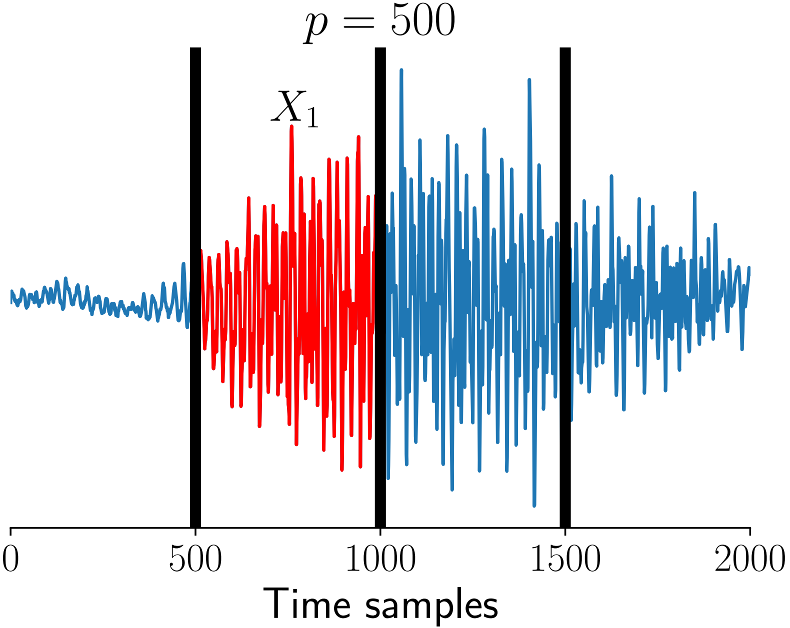


Cut the signal in chunks of size p




1st chunk = 1st column of frames matrix X
$p = 500$
X_0
0
500
1000
1500
2000
Time samples

Repeat n times to have a $p \times n$ matrix



The spectrogram is obtained by taking the DCT of X

Φ^{dct} Discrete Cosine Transform matrix of size p:

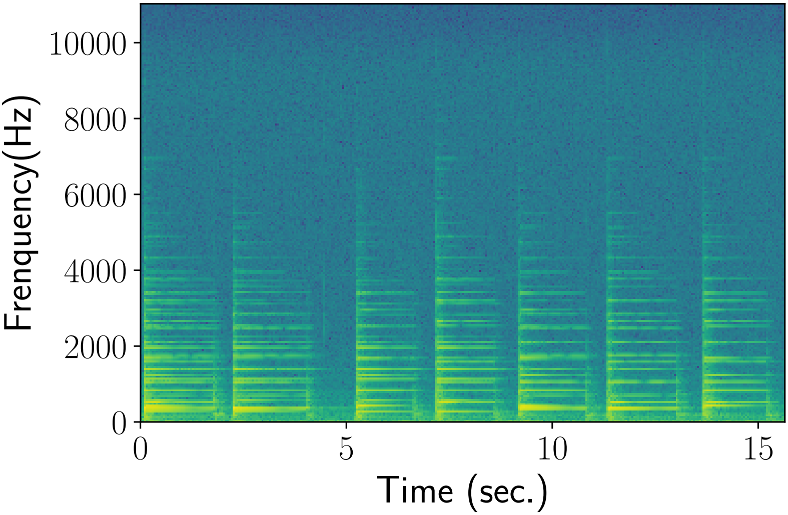
$$\Phi^{\mathsf{dct}}_{ij} = \sqrt{\frac{2}{p}} \cdot \cos[\frac{\pi}{p}(i + \frac{1}{2})(j + \frac{1}{2})], \ \ 0 \leq i, j \leq p - 1$$

It is an **orthogonal** matrix: $\boxed{\Phi\Phi^\top = I_p}$

- The spectrogram is then $V = (\Phi^{\mathsf{dct}} X)^{\odot 2}$

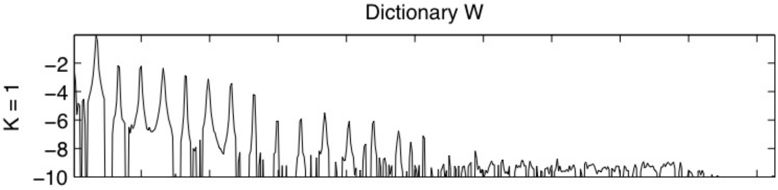
It corresponds to the concatenation of the power spectral densities for each selected frame.

Spectrogram of some piano chords

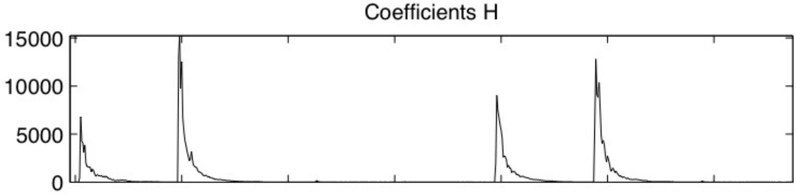


NMF on spectrograms

$V \simeq WH$, $W \in \mathbb{R}^{p \times k}$, $H \in \mathbb{R}^{k \times n}$.



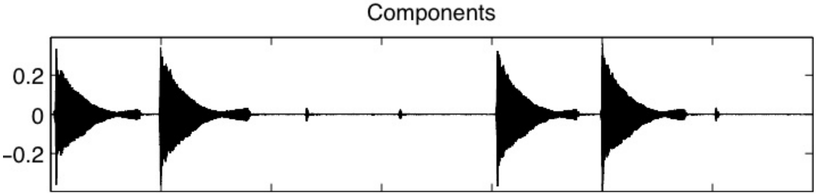
The columns of W correspond to spectral profiles



The rows of H correspond to temporal activations

NMF on spectrograms

$V \simeq WH$, $W \in \mathbb{R}^{p \times k}$, $H \in \mathbb{R}^{k \times n}$.



Recovery of the signal corresponding to a specific column i of W /row of H is possible using Wiener filtering (Févotte et al., 2009):

$$X^i = \Phi^\top \left(\frac{W_{:,i} H_{i,:}}{WH} \odot \Phi X \right)$$

- It isolates single notes in the simple piano case.
- More generally, it is an important tool for musical unsupervised source separation

Transform learning for NMF

Transform learning

Traditional NMF for audio signal processing:

$$\text{minimize } d((\Phi^{\text{dct}}X)^{\odot 2}||WH) \text{ s.t } W \geq 0, H \geq 0$$

Transform learning (Fagot et al., 2018):

$$\text{minimize } \mathcal{C}(\Phi, W, H) = d((\Phi X)^{\odot 2}||WH)$$
$$\text{s.t } W \geq 0, H \geq 0, \Phi\Phi^\top = I_p$$

- Alternate optimization in Φ, W and H.
- Regular multiplicative updates for W, H.
- For Φ: optimization on the orthogonal manifold (Absil et al., 2009) ♡

Optimization on the orthogonal manifold

Litterature methods for TL-NMF:

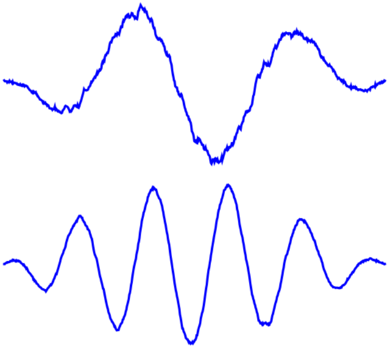
- Projected gradient (Fagot et al., 2018): $\Phi \leftarrow \Pi(\Phi - \eta G)$, where G is the gradient of $\mathcal{C}$ w.r.t. Φ and Π is the projection on the orthogonal manifold.
- Jacobi angles, similar to coordinate descent (Wendt et al., 2018): $\Phi \leftarrow J\Phi$ where J is a Jacobi rotation:

$$J = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & 0 & \\ & & \cos(\theta) & \cdots & -\sin(\theta) & \\ & & \vdots & \ddots & \vdots & \\ & & \sin(\theta) & \cdots & \cos(\theta) & \\ & 0 & & & \ddots & \\ & & & & & 1 \end{bmatrix}$$

Takes 1 day to converge on a regular music track. 1 min for standard NMF. ☹

But transform learning is useful !

The learned transform Φ captures the frequencies of the signals, and it obtains better results than NMF for some source separation tasks.



Method	SDR (dB)		SIR (dB)		SAR (dB)	
SNR = -10 dB	$\hat{y}_{\text{sp}}$	$\hat{y}_{\text{no}}$	$\hat{y}_{\text{sp}}$	$\hat{y}_{\text{no}}$	$\hat{y}_{\text{sp}}$	$\hat{y}_{\text{no}}$
Baseline	-9.50	10.00	-9.50	10.00	∞	∞
IS-NMF	-6.75	6.82	-5.00	**13.95**	**4.12**	7.93
TL-NMF	**1.73**	**12.29**	**13.44**	13.33	2.22	**19.20**
SNR = 0 dB	$\hat{y}_{\text{sp}}$	$\hat{y}_{\text{no}}$	$\hat{y}_{\text{sp}}$	$\hat{y}_{\text{no}}$	$\hat{y}_{\text{sp}}$	$\hat{y}_{\text{no}}$
Baseline	0.10	0.08	0.10	0.08	∞	∞
IS-NMF	1.73	0.69	3.06	5.32	**9.30**	3.65
TL-NMF	**6.50**	**5.81**	**12.11**	**9.16**	8.16	**9.00**

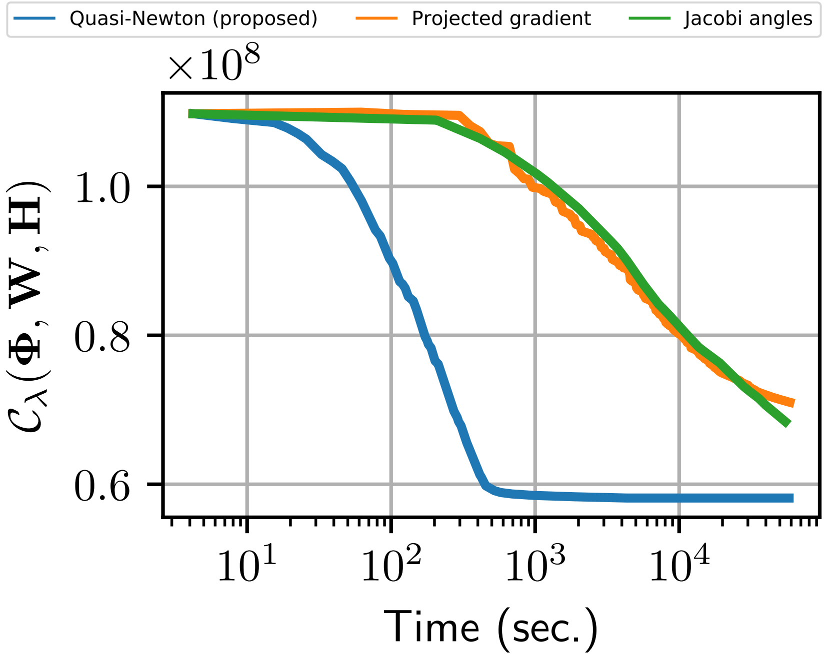
Table 1. Source separation performance.

Our contribution

Faster optimization:

- Optimize directly on the manifold using matrix exponential: $\Phi \leftarrow \exp(\mathcal{E})\Phi$ with $\mathcal{E} + \mathcal{E}^\top = 0$ enforces orthogonality
- Use a sparse approximation of the Hessian of $\mathcal{C}$ to obtain a quasi-Newton method

Results: from one day to 10 min



Thanks for your attention!

Online code:
https://github.com/pierreablin/tlnmf

Bibliography I

Absil, P.-A., Mahony, R., and Sepulchre, R. (2009). *Optimization algorithms on matrix manifolds*. Princeton University Press.

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. (2013). A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pages 280–288.

Blanton, M. R. and Roweis, S. (2007). K-corrections and filter transformations in the ultraviolet, optical, and near-infrared. *The Astronomical Journal*, 133(2):734.

Devarajan, K. (2008). Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS computational biology*, 4(7):e1000029.

Fagot, D., Wendt, H., and Févotte, C. (2018). Nonnegative matrix factorization with transform learning. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2431–2435.

Bibliography II

Févotte, C., Bertin, N., and Durrieu, J.-L. (2009). Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation*, 21(3):793–830.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Rutkowski, T. M., Zdunek, R., and Cichocki, A. (2007). Multichannel eeg brain activity pattern analysis in time–frequency domain with nonnegative matrix factorization support. In *International Congress Series*, volume 1301, pages 266–269. Elsevier.

Smaragdis, P. and Brown, J. C. (2003). Non-negative matrix factorization for polyphonic music transcription. In *IEEE workshop on applications of signal processing to audio and acoustics*, volume 3, pages 177–180. New York.

Bibliography III

Wendt, H., Fagot, D., and Févotte, C. (2018). Jacobi algorithm for nonnegative matrix factorization with transform learning. In *Proc. European Signal Processing Conference (EUSIPCO)*.