Online Non-negative Matrix Factorization as a Tool in Data Processing

Palina Salanevich

Email: p.salanevich@uu.nl

Utrecht University

June 10, 2022

1st Workshop on AI and Mathematics

Plan

- What is dictionary learning and non-negative matrix factorization?
  - data-driven representations
  - interpretability
- Applications in audio enhancement
- Applications in EEG data processing

Plan

- What is dictionary learning and non-negative matrix factorization?
  - data-driven representations
  - interpretability
- Applications in audio enhancement
  Based on the joint work with A. Sack (UCLA), M. Perlmutter (UCLA), and W. Jiang (USTC); D. Needell (UCLA)
- Applications in EEG data processing

Plan

- What is dictionary learning and non-negative matrix factorization?
  - data-driven representations
  - interpretability

- Applications in audio enhancement
  Based on the joint work with A. Sack (UCLA), M. Perlmutter (UCLA), and W. Jiang (USTC); D. Needell (UCLA)

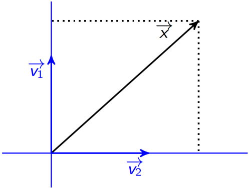
- Applications in EEG data processing
  Based on the joint work with H. Lyu (Wisconsin-Madison), Ch. Huang (UCLA), J. Li (UCLA), and D. Needell (UCLA)

Data representation

Basis - an economic representation

$$V = \{v_i\}_{i=1}^d \subset \mathbb{R}^d, \ \ \mathrm{span}(V) = \mathbb{R}^d;$$

$$x \mapsto \{< x, v_i >\}_{i=1}^d$$

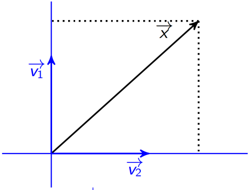


Data representation

Basis - an economic representation

$$V = \{v_i\}_{i=1}^{d} \subset \mathbb{R}^d, \quad \mathrm{span}(V) = \mathbb{R}^d;$$

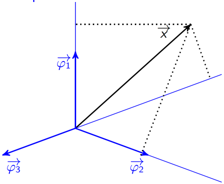
$$x \mapsto \{< x, v_i >\}_{i=1}^{d}$$



Frame - a stable representation

$$\Phi = \{\varphi_i\}_{i=1}^{N} \subset \mathbb{R}^d, \ \mathrm{span}(\Phi) = \mathbb{R}^d \ (N > d);$$

$$x \mapsto \{< x, \varphi_i >\}_{i=1}^{N}$$

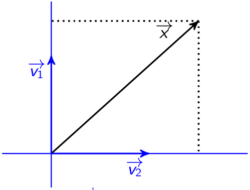


Data representation

Basis - an economic representation

$$V = \{v_i\}_{i=1}^d \subset \mathbb{R}^d, \quad \mathrm{span}(V) = \mathbb{R}^d;$$

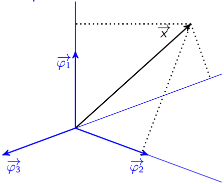
$$x \mapsto \{< x, v_i >\}_{i=1}^d$$



Frame - a stable representation

$$\Phi = \{\varphi_i\}_{i=1}^N \subset \mathbb{R}^d, \; \mathrm{span}(\Phi) = \mathbb{R}^d \; (N > d);$$

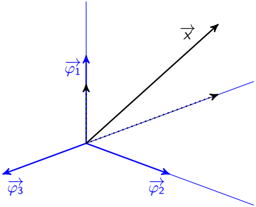
$$x \mapsto \{< x, \varphi_i >\}_{i=1}^N$$



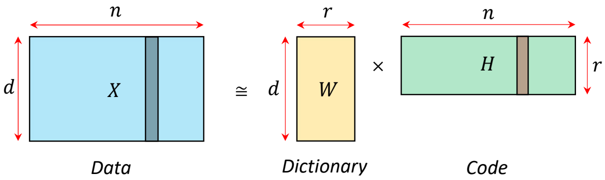
Dictionary - a data-driven representation

$$W = \{w_i\}_{i=1}^r \subset \mathbb{R}^d;$$

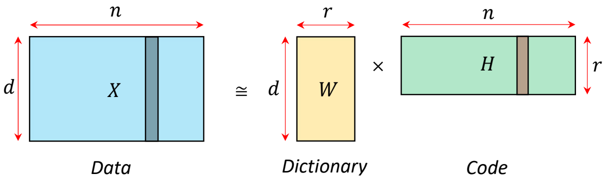
$$x \mapsto \{h_i\}_{i=1}^r, \text{ s.t. } \; x \approx \sum_{i=1}^r h_i w_i$$



Non-negative matrix factorization

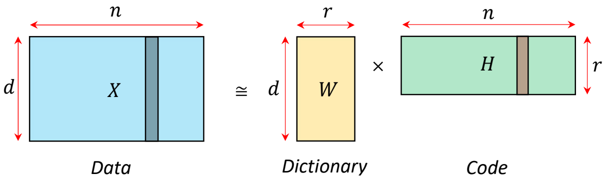


Non-negative matrix factorization



Idea: dictionary atoms should represent additive features, without cancellations

Non-negative matrix factorization



Idea: dictionary atoms should represent additive features, without cancellations

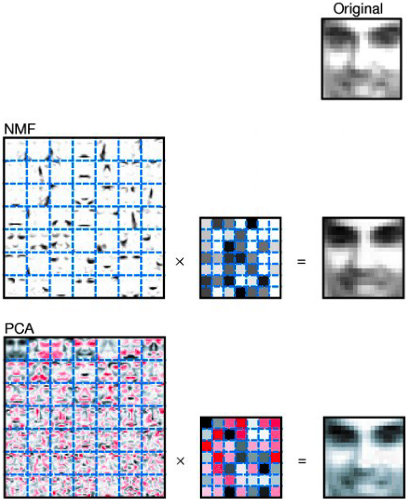
- Non-negative matrix factorization:

$$\min_{\substack{W \in \mathbb{R}^{d \times r}_{\geq 0} \\ H \in \mathbb{R}^{r \times n}_{\geq 0}}} \|X - WH\|_F$$

- Additive dictionary learning:

$$\min_{\substack{W \in \mathbb{R}^{d \times r} \\ H \in \mathbb{R}^{r \times n}_{\geq 0}}} \|X - WH\|_F$$

Non-negative matrix factorization: illustrative example



Data set: pictures of people's faces

NMF: data is represented as a non-negative linear combination of dictionary atoms, which thus represent "additive parts" of data
(e.g., eyes, nose, mouth).

PCA: Due to cancellation between eigenvectors, each 'eigenface' does not have to represent parts of a face

Non-negative matrix factorization: algorithm

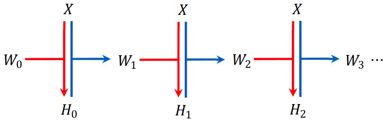
NMF optimization problem: $\min_{\substack{W \in \mathbb{R}_{\geq 0}^{d \times r} \\ H \in \mathbb{R}_{\geq 0}^{r \times n}}} \|X - WH\|_F$

Non-negative matrix factorization: algorithm

NMF optimization problem: $\min_{\substack{W \in \mathbb{R}^{d\times r}_{\geq 0} \\ H \in \mathbb{R}^{r\times n}_{\geq 0}}} \|X - WH\|_F$

Block coordinate descent: iteratively

1. Fix W and solve $\min_{H \in \mathbb{R}^{r\times n}_{\geq 0}} \|X - WH\|_F$
2. Fix H and solve $\min_{W \in \mathbb{R}^{d\times r}_{\geq 0}} \|X - WH\|_F$

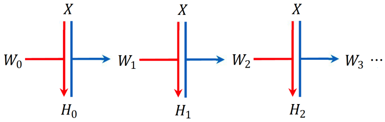


Non-negative matrix factorization: algorithm

NMF optimization problem: $\min_{\substack{W\in\mathbb{R}^{d\times r}_{\geq 0}\\ H\in\mathbb{R}^{r\times n}_{\geq 0}}} \|X - WH\|_F$

Block coordinate descent: iteratively

1. Fix W and solve $\min_{H\in\mathbb{R}^{r\times n}_{\geq 0}} \|X - WH\|_F$
2. Fix H and solve $\min_{W\in\mathbb{R}^{d\times r}_{\geq 0}} \|X - WH\|_F$



Multiplicative Update:

$$H_{ij} \leftarrow H_{ij}\frac{(W^TX)_{ij}}{(W^TWX)_{ij}}, \qquad W_{ij} \leftarrow W_{ij}\frac{(XH^T)_{ij}}{(XHH^T)_{ij}}$$

D. D. Lee and H. S. Seung

Learning the parts of objects by non-negative matrix factorization, *Nature*, vol. 401, no. 6755, p. 788, 1999.

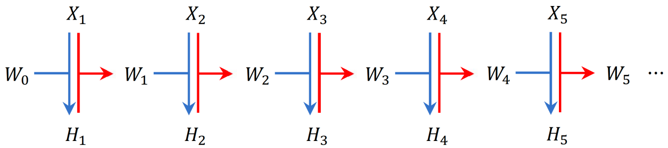
Online non-negative matrix factorization

Question: Suppose the columns of X are randomly drawn from the data set $\mathcal{X}$. Can we learn a dictionary that efficiently describes all elements of $\mathcal{X}$?

Online non-negative matrix factorization

Question: Suppose the columns of X are randomly drawn from the data set $\mathcal{X}$. Can we learn a dictionary that efficiently describes all elements of $\mathcal{X}$?

Online Non-negative Matrix Factorization (ONMF):
Learn the dictionary W from a sequence of input matrices $(X_t)_{t\in\mathbb{N}}$.



Goal: construct a sequence $(W_t, H_t)_{t\in\mathbb{N}}$ such that (almost surely)

$$\|X_t - W_{t-1}H_t\|_F^2 \to_{t\to\infty} \min_{\substack{W\in\mathbb{R}_{\geq 0}^{d\times r}\\ H\in\mathbb{R}_{\geq 0}^{r\times n}}} \mathbb{E}\left(\|X - WH\|_F^2\right)$$

Online non-negative matrix factorization: algorithm

1. (Sparse) code matrix update: $H_t = \arg \min_{H \in \mathbb{R}_{\geq 0}^{r \times n}} \|X_t - W_{t-1} H\|_F^2 + \alpha \|H\|_1$
2. Aggregation of the past information:

$$A_t = \frac{1}{t}\left((t-1)A_{t-1} + H_t H_t^T\right), \qquad B_t = \frac{1}{t}\left((t-1)B_{t-1} + H_t X_t^T\right)$$

3. Dictionary matrix update: $W_t = \arg \min_{W \in \mathbb{R}_{\geq 0}^{d \times r}} \frac{1}{2} \operatorname{Tr}(W A_t W_t^T) - \operatorname{Tr}(B_t W)$

Online non-negative matrix factorization: algorithm

1. (Sparse) code matrix update: $H_t = \arg\min_{H \in \mathbb{R}^{r\times n}_{\geq 0}} \|X_t - W_{t-1}H\|_F^2 + \alpha\|H\|_1$
2. Aggregation of the past information:

$$A_t = \frac{1}{t}\left((t-1)A_{t-1} + H_t H_t^T\right), \qquad B_t = \frac{1}{t}\left((t-1)B_{t-1} + H_t X_t^T\right)$$

3. Dictionary matrix update: $W_t = \arg\min_{W \in \mathbb{R}^{d\times r}_{\geq 0}} \frac{1}{2}\mathrm{Tr}(W A_t W_t^T) - \mathrm{Tr}(B_t W)$

Convergence guarantees:

- i.i.d. $(X_t)_{t\in\mathbb{N}}$;

J. Mairal, F. Bach, J. Ponce, and G. Sapiro

Online learning for matrix factorization and sparse coding, *Journal of Machine Learning Research*, 11 (2010).

Online non-negative matrix factorization: algorithm

1. (Sparse) code matrix update: $H_t = \arg\min_{H \in \mathbb{R}_{\geq 0}^{r \times n}} \|X_t - W_{t-1}H\|_F^2 + \alpha\|H\|_1$

2. Aggregation of the past information:

$$A_t = \frac{1}{t}\left((t-1)A_{t-1} + H_t H_t^T\right), \qquad B_t = \frac{1}{t}\left((t-1)B_{t-1} + H_t X_t^T\right)$$

3. Dictionary matrix update: $W_t = \arg\min_{W \in \mathbb{R}_{\geq 0}^{d \times r}} \frac{1}{2}\mathrm{Tr}(WA_tW_t^T) - \mathrm{Tr}(B_tW)$

Convergence guarantees:

- i.i.d. $(X_t)_{t\in\mathbb{N}}$;
- irreducible Markov chain $(X_t)_{t\in\mathbb{N}}$.

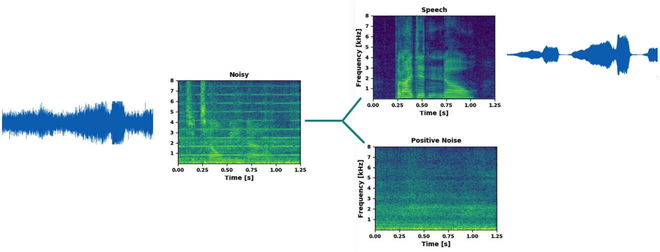
J. Mairal, F. Bach, J. Ponce, and G. Sapiro

Online learning for matrix factorization and sparse coding, *Journal of Machine Learning Research*, 11 (2010).

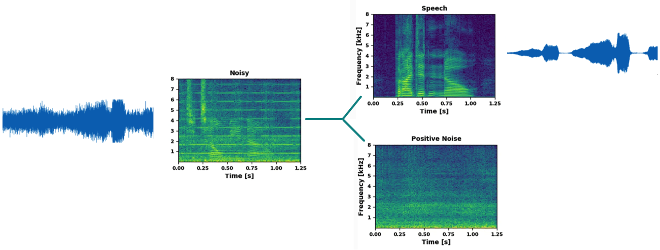
H. Lyu, D. Needell, and L. Balzano

Online matrix factorization for Markovian data and applications to network dictionary learning, *arXiv:1911.01931* (2019).

Problem: audio enhancement



Problem: audio enhancement



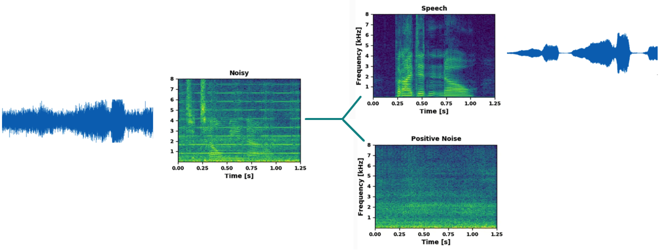
A. Lefevre, F. Bach, and C. Févotte

Online algorithms for non-negative matrix factorization with the Itakura-Saito divergence, *WASPAA*, 2011.

C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller

Real-time speech separation by semi-supervised non-negative matrix factorization, *LVA/ICA*, 2012.

Problem: audio enhancement



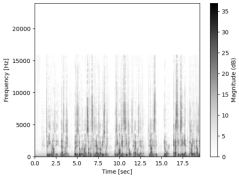
Advantages of ONMF over NMF methods:

- Memory efficiency and parallelization. Can be adapted to streaming audio.
- Uses regularized loss function leading to better performance and theoretical convergence guarantees.

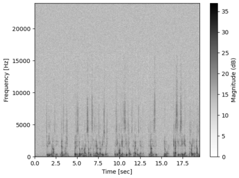
A. Lefevre, F. Bach, and C. Févotte
Online algorithms for non-negative matrix factorization with the Itakura-Saito divergence, *WASPAA*, 2011.

C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller
Real-time speech separation by semi-supervised non-negative matrix factorization, *LVA/ICA*, 2012.

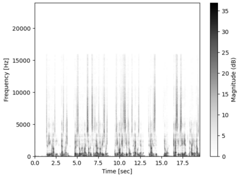
Problem: audio enhancement



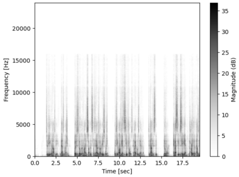
a Original clean speech



b Original noisy speech



c NMF-based denoising



d ONMF-based denoising

Method	SDR	SIR	SAR
NMF	19.43	31.42	19.72
ONMF	22.70	53.45	22.70
ORIGINAL	9.75	9.76	37.41

Performance measures with artificial noise.

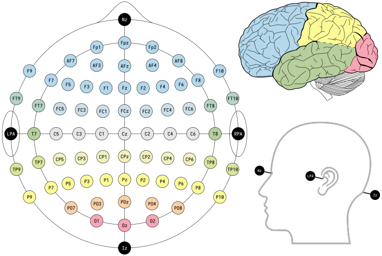
Method	SDR	SIR	SAR
NMF	9.46	13.90	11.63
ONMF	10.41	13.11	13.95
ORIGINAL	5.91	5.91	286.50

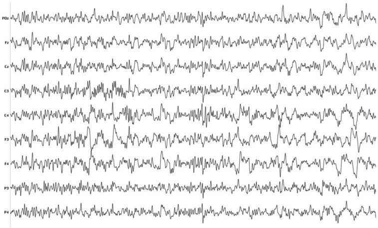
Performance measures with real-world noise.

A. Sack, W. Jiang, M. Perlmutter, P. Salanevich, and D. Needell

On audio enhancement via online non-negative matrix factorization, *56th Annual Conference on Information Sciences and Systems (CISS)*, 2022.

Electroencephalogram (EEG) measures the neurons electro-physiological activity that is accessible on the surface of the scalp.

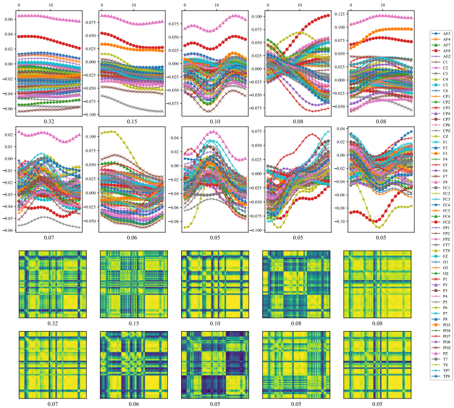




Problem: Determine functional connections between different brain regions (important, e.g., for diagnostics).

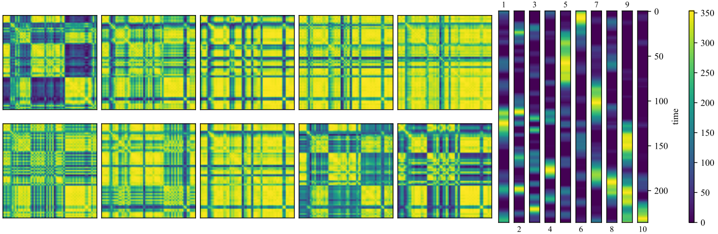
Idea: Use correlation between signals from different detectors to determine functional dependencies.

Correlation matrix via ONMF



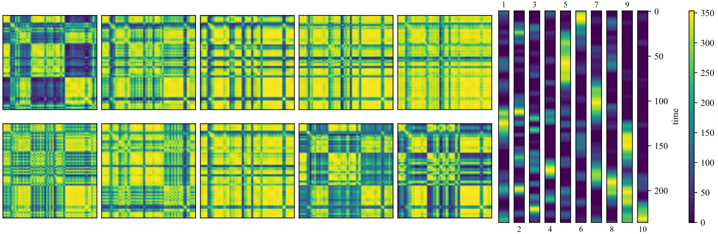
Temporal dictionary of $r = 10$ atoms for $k = 20$-step evolution in the EEG signal. Any k-step joint evolution of all 61-sensor signals are approximated by a non-negative combination of these atoms, given by the learned code matrix H.

Problem: EEG data processing

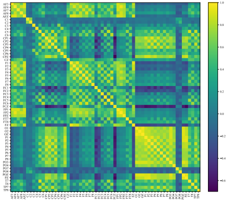


Dictionary-based correlation matrices and their time evolution.

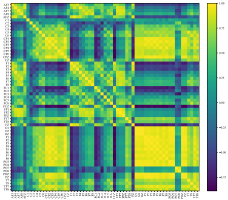
Problem: EEG data processing



Dictionary-based correlation matrices and their time evolution.



Pearson correlation matrix.



ONMF correlation matrix.

EEG data processing: ongoing research

Question 1. *Can ONMF effectively parse event-related neural responses into their underlying neural components?*

We aim to use graph-based regularization to obtain dictionary atoms that are

1. reliable across subjects
2. interpretable in parsing the mixed responses into underlying neural processes

EEG data processing: ongoing research

Question 1. *Can ONMF effectively parse event-related neural responses into their underlying neural components?*

We aim to use graph-based regularization to obtain dictionary atoms that are

1. reliable across subjects
2. interpretable in parsing the mixed responses into underlying neural processes

Question 2. *Can ONMF improve upon ICA in denoising of EEG data?*

ICA fails when data contains non-stationary noise (e.g., muscle movement or heart artifacts in EEG-fMRI data), while NMF works effectively with such data.

Thank You for Your Attention!