Nonnegative Matrix Factorization, Wasserstein metric, source separation

Andersen Ang

ECS, Uni. Southampton, UK
andersen.ang@soton.ac.uk
Homepage angms.science

Version: June 28, 2023
First draft: June 21, 2023

Distance-based methods in Machine Learning
University College London
27-28 June 2023

Content

Joint work with

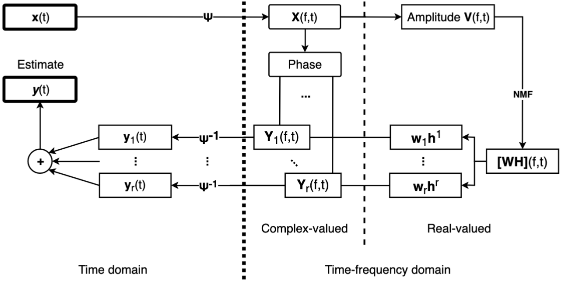
Xinwen Ding @ U.Waterloo, CA

Giang Tran @ U.Waterloo, CA

Steve Vavasis @ U.Waterloo, CA

Overview: Single-channel blind source separation (BSS) of audio

- Typical BSS pipeline



https://angms.science/doc/NMF/20201202iTWIST_12_page_slide.pdf

- Single-channel: one measurement
- blind: no prior knowledge
- source separation: de-mixing
- audio: time series

- This talk: modify this pipeline
- Content
  - Power spectrum representation of audio
  - Time-frequency transform
  - Random estimation of kernel
  - Nonnegative Matrix Factorization (NMF)
  - Separable NMF and SPA
  - Spectrum misalignment
  - Wasserstien distance

Single-channel blind source separation

- **Given**: $x(\tau) = \sum_{k=1}^{K} s_k(\tau)$ single observation in $\mathbb{R}^{\boldsymbol{T}}$

  $\tau \in [0, \boldsymbol{T}]$ time domain

- $s_k(\tau), k = 1, 2, ..., K$ $\underbrace{K}_{\text{unknown}}$ $\underbrace{\text{sources}}_{\text{unknown}}$

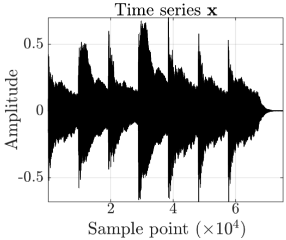
- **Goal**: recover all K sources $s_k(\tau)$ from single measurement $x(\tau)$ under-determined problem

- Blind $=$ no prior knowledge on s, K

- What is known
  - $\boldsymbol{T}$: the duration of the time series
  - $x(\tau)$: the observed time series

Representation of time series

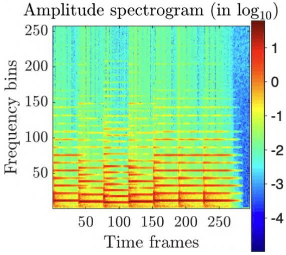
- $x(\tau) \in \mathbb{R}^{\boldsymbol{T}} \xrightarrow[\text{sampling}]{\text{discretize}} \boldsymbol{x} \in \mathbb{R}^L \xrightarrow{\psi} \boldsymbol{X}[f,t] \in \mathbb{C}^{F\times T}$
  - $\boldsymbol{x} \in \mathbb{R}^L$: vector of L elements
  - $\boldsymbol{X}[f,t]$: time-frequency content of $\boldsymbol{x}$ at (f-Hz, t-second)
  - $f \in [0,F]$, there are F frequency bins (y-coordinate)
  - $t \in [0,T]$, there are T time frame (x-coordinate)
  - $x(\tau)$ with $(\cdot)$ is continuous, $\boldsymbol{x}[t]$ with $[\cdot]$ is discrete



- ψ DSTFT (Discrete Short-time Fourier Transform) $\mathbb{R}^L \to \mathbb{C}^{F\times(T+1)}$

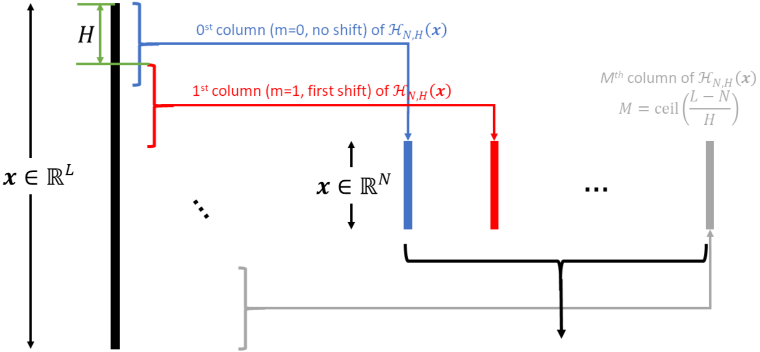
$$\boldsymbol{X}[f,t] \;:=\; \sum_{n=0}^{N-1} \boldsymbol{w}[n]\boldsymbol{x}[n+tH]\exp\Big[-i\frac{2\pi}{N}fn\Big]. \qquad (\psi)$$

  - N number of short-time intervals
  - $n \in [0, N-1]$ is interval index
  - $\big[\boldsymbol{x}[0+tH], \boldsymbol{x}[1+tH], \dots, \boldsymbol{x}[N-1+tH]\big]$ is a segment of $\boldsymbol{x} \in \mathbb{R}^L$
  - $H \in [0, L]$ hop size, a shift parameter
  - $\boldsymbol{w} \in \mathbb{R}^N$: $\big[w[0], w[1], \dots, w[N-1]\big]$ a window function
  - $t \in [0,T]$ time frame and $T = \lfloor \frac{L-N}{H} \rfloor$ is the max frame
  - $f \in [0,F]$ frequency bin, $F = N-1$ and $f = \lfloor \frac{N}{2} \rfloor$ is Shannon-Nyqusit frequency



Picture of DSTFT

Hankelization of $\mathbf{x}$



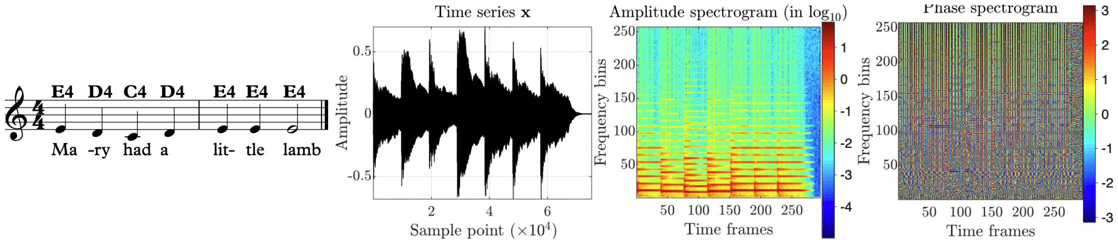
$\mathcal{H}_{N,H}(\boldsymbol{x}) \in \mathbb{R}^{N \times M+1}$: a "Hankel matrix" of $\boldsymbol{x}$,
with shift parameter H and segment length N

https://angms.science/doc/SP/SP_STFT.pdf

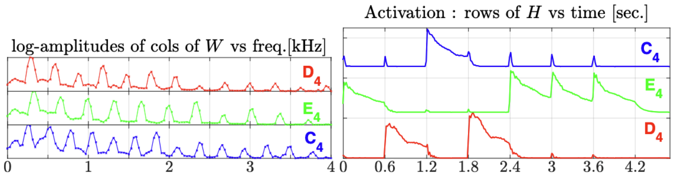
Power spectrogram and decomposition

- Suppose we have a complex spectrogram $\boldsymbol{X}[f,t] \in \mathbb{C}^{F\times T}$
- Convert the complex $\boldsymbol{X}$ to real power spectrogram / amplitude spectrogram / magnitude spectrogram

$$z = re^{i\theta} \quad \Longrightarrow \quad \boldsymbol{X}[f,t] = \underbrace{\big|\boldsymbol{X}[f,t]\big|}_{\boldsymbol{V}} e^{\angle i\boldsymbol{\Theta}[f,t]}.$$

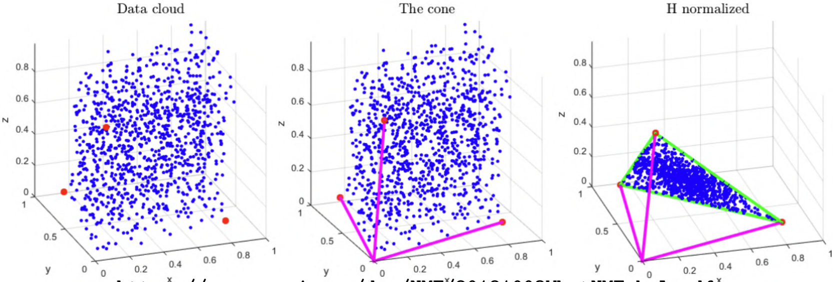


$$\boldsymbol{V} \overset{NMF}{=} \boldsymbol{W}\boldsymbol{H}$$



Nonnegative Matrix Factorization

- Given $\boldsymbol{V} \in \mathbb{R}_+^{m \times n}$, find $\boldsymbol{W} \in \mathbb{R}_+^{m \times r}$ and $\boldsymbol{H} \in \mathbb{R}_+^{r \times n}$ such that $\boldsymbol{V} = \boldsymbol{W}\boldsymbol{H}$
  - A linear algebra problem, earliest apperance in chemistry in 1960s — See Sect1.4 in Gillis 2020[1]
  - A NP-hard problem — Vavasis 07[2]
  - A nonsmooth nonconvex biconvex optimization problem — many works
- Conic geometry



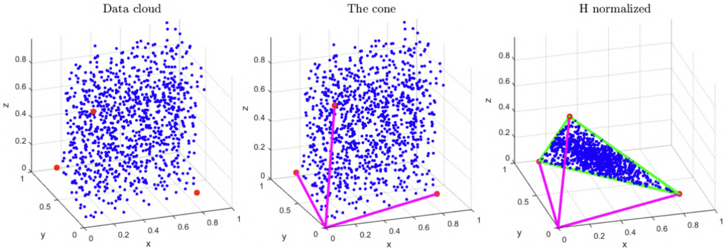
https://angms.science/doc/NMF/20181003WhatNMF_kul.pdf

---

[1] Nicolas Gillis, Nonnegative Matrix Factorization, SIAM, 2020

[2] Steve Vavasis, On the complexity of nonnegative matrix factorization, SIAM J OPT, 2007

Separable NMF



SPA (Successive Projection alg.)

$R = V$

$J = \{\}$

For i $= 1 : k$

$\quad j = \underset{j}{\text{argmax}}\ f(R_{:j})$ $\qquad f = ||\cdot||_2^2$

$\quad J = J \cup \{j\}$

$\quad H = \underset{y \in Y}{\text{argmin}}\ g(V, V_{:J}\, y)$ $\qquad g = ||A - B||_F^2$

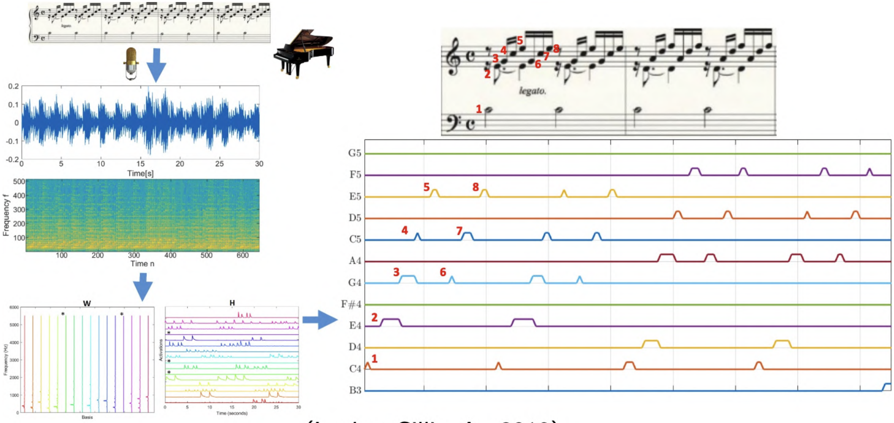
$\quad R = R - V_{:J}\, H$

- Separable NMF: $\boldsymbol{W}$ are certain columns of $\boldsymbol{V}$

$$\boldsymbol{V} = \boldsymbol{W}\boldsymbol{H} = \boldsymbol{V}_{:J}[\boldsymbol{I}_r \ \ \boldsymbol{H}']\boldsymbol{\Pi}_n.$$

  - $\boldsymbol{W}$ comes from r columns of $\boldsymbol{V}$, labelled by an r-set J.
  - $\boldsymbol{\Pi}_n$ is column permutation
  - $\boldsymbol{I}_r$ is r-order identity matrix
  - $\boldsymbol{H}' \in \mathbb{R}^{r \times (n-r)}$
- SPA (Successive Projection Algorithm)
  - find column with largest norm
  - projects out such column from the residual data matrix

NMF works quite well on (simple) audio ...



(Leplat, Gillis, A., 2019)

- Fast algorithm
- Identifiability / solution of is unique, even the problem is nonconvex
- Rank selection power?

  Leplat et al., Blind Audio Source Separation with Minimum-Volume Beta-Divergence NMF, IEEE TSP, 2020

Two challenges

1. It is expensive to obtain a spectrogram
   - STFT is expensive: $\sim \mathcal{O}(N^3)$ cost ($N = \#$short intervals)
   - $\longrightarrow$ Treat STFT as a kernel process, approximate it by randomization
2. Spectrum misalignment on more complicated audio
   - Inharmonicity[3], an unavoidable physical phenomenon
   - $\longrightarrow$ use Wasserstien metric to allow spectrum shifting

[3] Chris Murray, *Musical String Inharmonicity*, https://publicwebuploads.uwec.edu/documents/Musical-string-inharmonicity-Chris-Murray.pdf

Randomization: idea

- Observation: STFT is a dot product

$$\boldsymbol{X}[f,t] = \sum_{n=0}^{N-1} \boldsymbol{w}[n]\boldsymbol{x}[n+tH] \exp\Big[-i\frac{2\pi}{N}fn\Big] \qquad \longrightarrow \qquad \boldsymbol{X}[\cdot,t] = \Big\langle \boldsymbol{x}[n+tH],\ \underbrace{\boldsymbol{w}[n]\exp\Big[-i\frac{2\pi}{N}fn\Big]}_{\text{“nonlinear kernel”}} \Big\rangle$$

  This is giving a hint on kernel estimation.

- We treat STFT as a nonlinear kernel and we approximate the power spectrum $\boldsymbol{V}$

$$\boldsymbol{V}[f,t] \ = \ |\boldsymbol{X}[f,t]| \ \approx \ \sum_{ij} S_{ij}\sin\omega_i t_j + C_{ij}\cos\omega_i t_j \ = \ \sum_{ij}\begin{bmatrix}\sin\omega_i t_j & \cos\omega_i t_j\end{bmatrix}\begin{bmatrix}S_{ij}\\ C_{ij}\end{bmatrix}$$

  - sine-cosine is because $\underbrace{\exp\Big[-i\frac{2\pi}{N}fn\Big]}_{\sin,\cos}$
  - we work on $\boldsymbol{V}$ instead of $\boldsymbol{X}$ because
    - we don't want to deal with complex numbers / phase
    - standard NMF works on $\mathbb{R}$ not $\mathbb{C}$

- Why it will work: Rahimi-Recht's random feature[4]

[4] Rahimi and Recht, *Random Features for Large-Scale Kernel Machines*, NIPS, 2007

Randomization: procedure

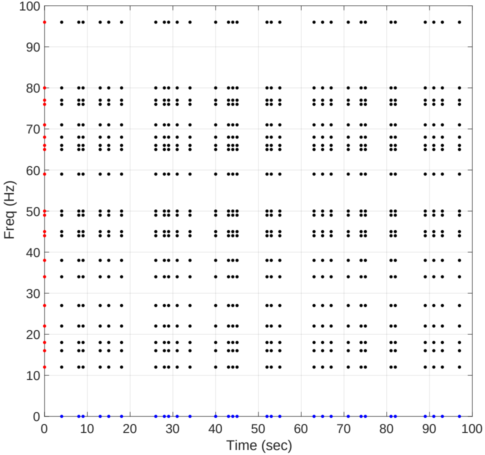
- Step 1. Randomization on frequency
  - Randomly pick N_1 frequencies f_i such that $f_i \sim \mathcal{U}[0, F]$.
  - Let $\omega = 2\pi f$, construct a frequency-basis matrix $\boldsymbol{A} = \begin{bmatrix} \sin\omega_1 t & \cos\omega_1 t \\ \vdots & \vdots \\ \sin\omega_{N_1} t & \cos\omega_{N_1} t \end{bmatrix} \in \mathbb{R}^{N_1 \times 2}$
- Step 2. Randomization on time
  - Divide the time domain $[0, T]$ into N_2 disjoint time windows of the same length.
  - Uniformly sample M time points in, t_j, $j \in [M]$ each time window.
  - Extract signal $\boldsymbol{x}_j := \boldsymbol{x}(t_j)$ associated to the time points t_j.
  - For each time window j, solve

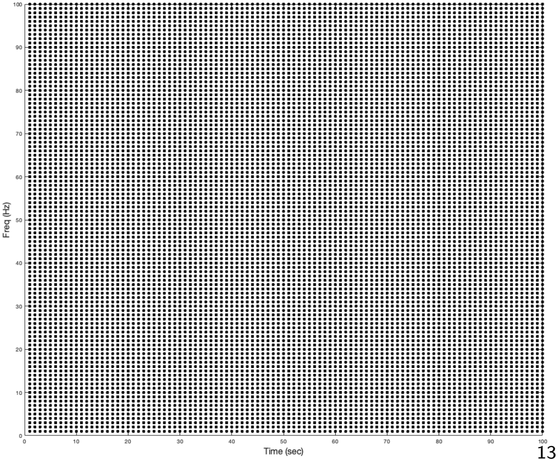
$$\boldsymbol{y}_j^* := \underset{\boldsymbol{y} \in \mathbb{R}^{2 \times N_1}}{\operatorname{argmin}} \; \|\boldsymbol{y}\|_1 \text{ s.t. } \|\boldsymbol{A}\boldsymbol{y} - \boldsymbol{x}_j\|_2 \leq \sigma,$$

  where $\boldsymbol{y}_j^* = \begin{bmatrix} S_{:j}^* \\ C_{:j}^* \end{bmatrix}$ is the sparse sine-cosine coefficient that makes $\boldsymbol{x}_j$ best match $\boldsymbol{A}$
  - The j^{th} column (jth time frame) of the estimated power spectrogram $\hat{\boldsymbol{V}}$ is $\hat{V}[:, j] = \sqrt{S_{:j}^2 + C_{:j}^2}$.
  - For the whole $\hat{\boldsymbol{V}}$ across all time point, the whole problem on is nonsmooth non-proximable convex.
- Here we are estimating the power spectrogram $\implies$ we do not need to deal with phase.

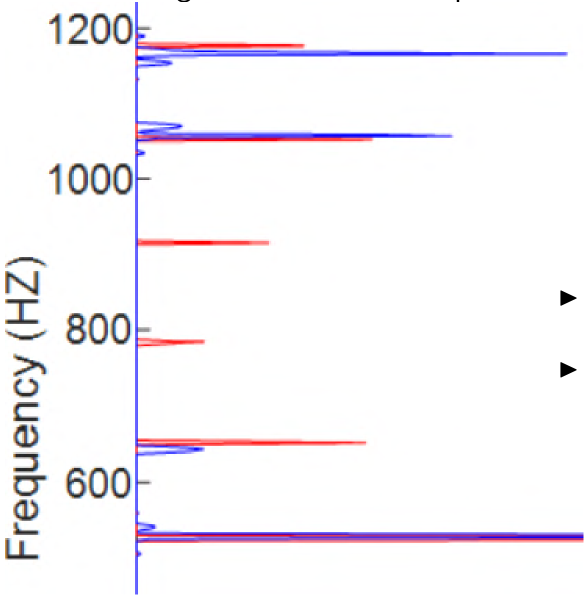
Illustration: 600 time-freq points vs 10000 time-freq points

- $F = 100$Hz, frequency resolution $\Delta f = 1$Hz
- $T = 100$Second with a temporal resolution Δt of 1second
- Random $N_1 = 20$ frequencies: $12,\ 16,\ 18,\ 22,\ 27,\ 34,\ 38,\ 44,\ 45,\ 49,\ 50,\ 59,\ 65,\ 66,\ 68,\ 71,\ 76,\ 77,\ 80,\ 96$
- We divide $[0, T]$ into $N_2 = 10$ time windows (each 10 second).
- In each time window we randomly pick $M = 3$ time points.





Spectrum misalignment on more complicated audio: inharmonics



$$\mu \frac{\partial^2 y}{\partial t^2} = T \frac{\partial^2 y}{\partial x^2} - ESK^2 \frac{\partial^4 y}{\partial x^4}$$

- E: Young's modulus (string's resistance to deform)
- Wave equation for ideal string $E = 0$ (string deforms without effort)

Wasserstein distance / OT distance

- 1-dimensional discrete Wasserstein distance

$$d_C(\boldsymbol{x},\boldsymbol{y}) := \|\boldsymbol{C}(\boldsymbol{x}-\boldsymbol{y})\|_1 = \Bigg\| \underbrace{\begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ 1 & 1 & 1 & \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}}_{\text{1-dim OT cost matrix}} (\boldsymbol{x}-\boldsymbol{y}) \Bigg\|_1 .$$

do not confuse the $\boldsymbol{C}$ here with the $C_{:j}$ in the previous slide, they are different things

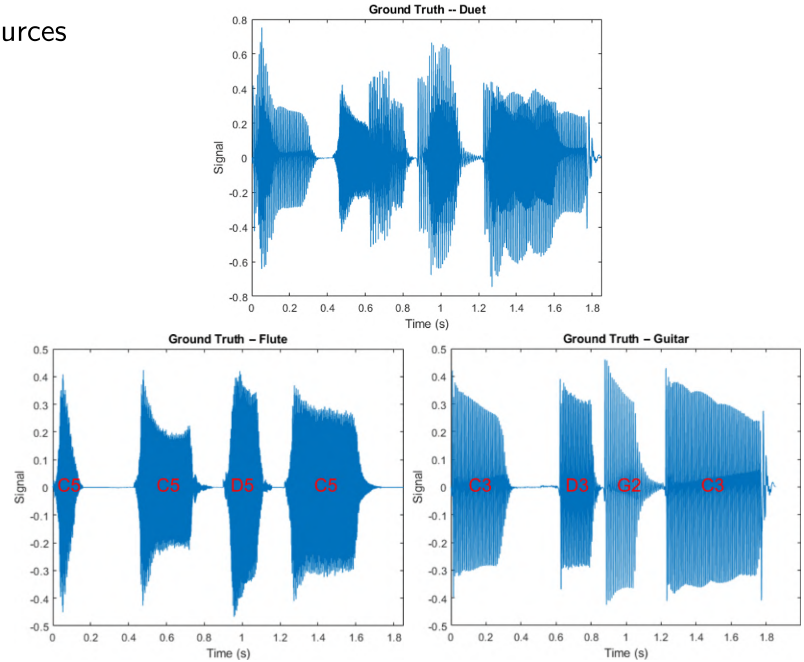
- Ideas
  - SPA in Wass-distance
  - Transform the data via the Wasserstein cost matrix $\boldsymbol{C}$
  - Why Wass-distance: $\underbrace{\text{holistic comparison fitting}}_{\text{allow misalignment}}$ vs $\underbrace{\text{element-wise comparison fitting}}_{\text{does not allow misalignment}}$

- Wasserstein-NMF is not a new idea — e.g., Flammy 2016[5]

  Our approach differs in
  - NMF vs separable NMF
  - OT divergence solved via linear program vs nonlinear problem
  - Different transport matrix $\boldsymbol{C}$: different $\boldsymbol{C}$ and also different dimension for OT
  - Semi-supervised (pre-define $\boldsymbol{W}$ as a comb) vs unsupervised

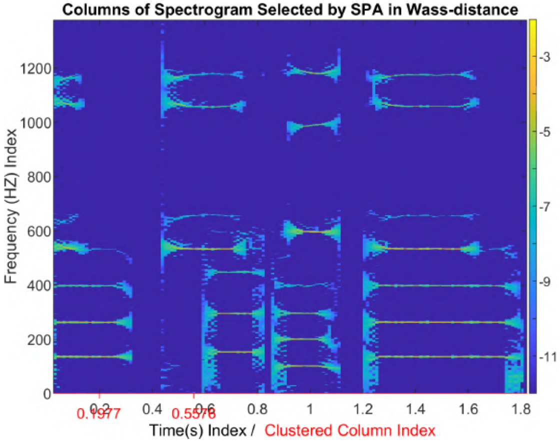
[5] Flamary et al., Optimal spectral transportation with application to music transcription, NIPS2016

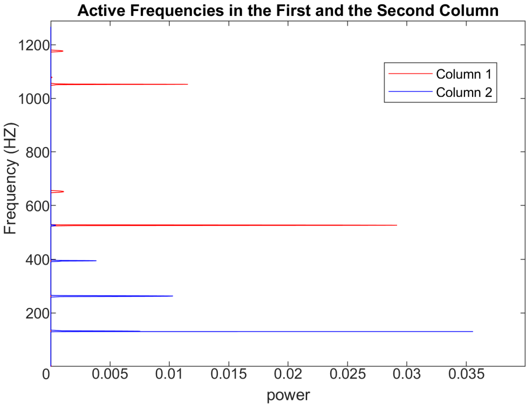
Example: 5 sources



Columns of spectrogram selected by SPA in Wass-distance

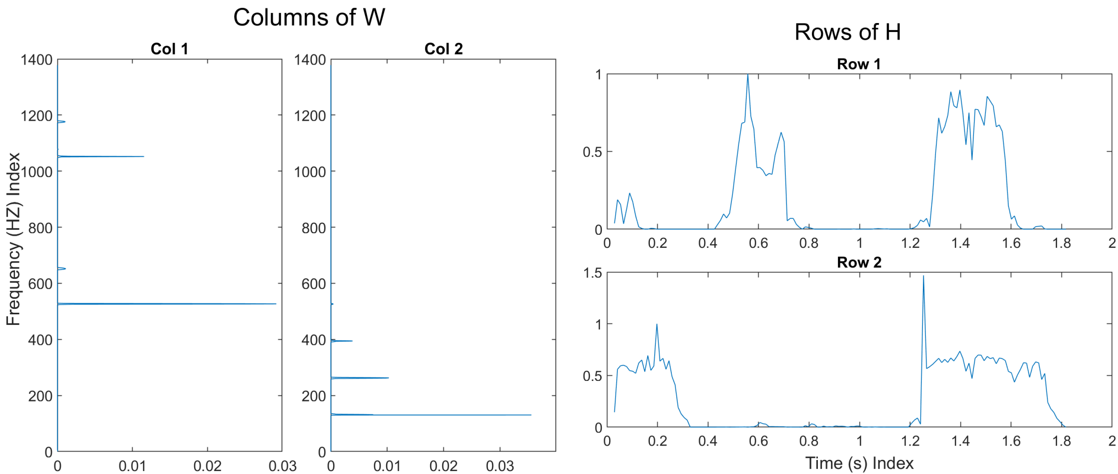
- In this example, we let SPA with Wass-distance select **2** features (i.e. two columns of the spectrogram)
- SPA with Wass-distance captures the solo periods of both instruments.



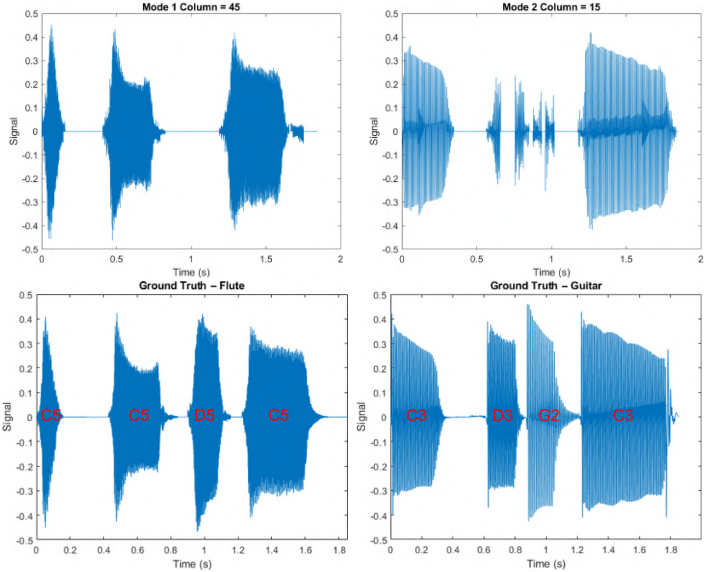


There are 5 sources: C5, D5, C3, D3, G2 and here we are demonstrating 2 features

Plots of $\boldsymbol{W}$ and $\boldsymbol{H}$

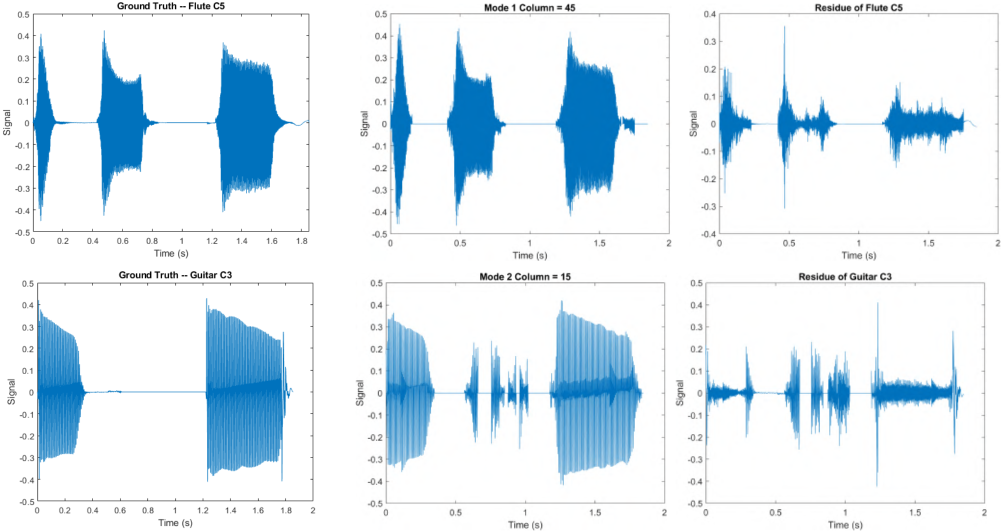


Reconstructed modes (C5 and C3)



There are 5 sources: C5, D5, C3, D3, G2 and here we are demonstrating 2 features

Reconstructed modes (C5 and C3)



There are 5 sources: C5, D5, C3, D3, G2 and here we are demonstrating 2 features

Last page - summary

Blind Source separation
Power spectrum
Nonnegative Matrix Factorization
Separable NMF
Random kernel estimation
Spectrum misalignment
Wasserstein distance

End of document