Expectation-Maximization Algorithms for Itakura-Saito Nonnegative Matrix Factorization

Paul Magron, Tuomas Virtanen

Laboratory of Signal Processing, Tampere University of Technology, Finland

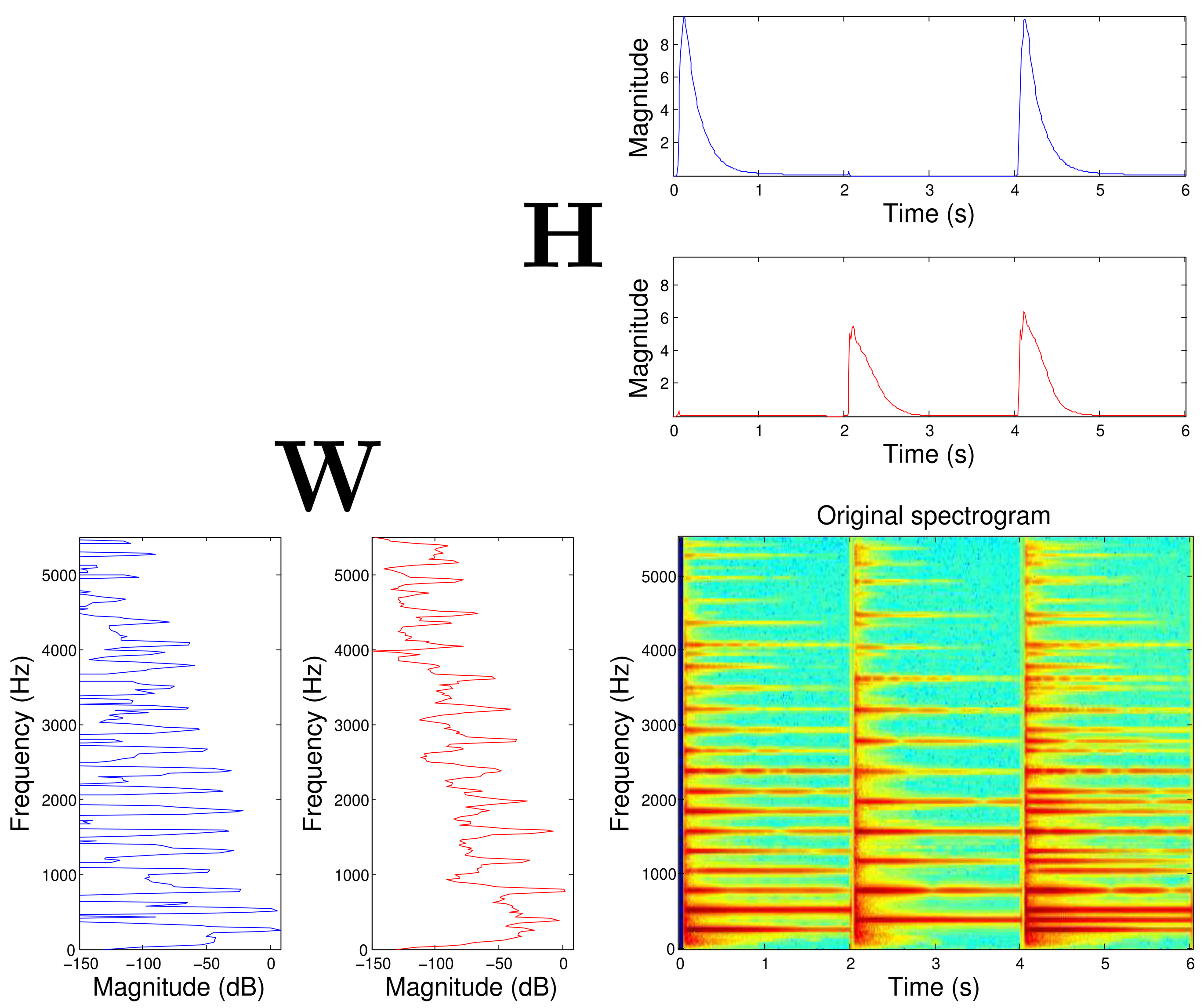
Motivation

- ISNMF is popular for audio source separation.
- Estimation: maximum-likelihood (ML) or the SAGE variant of EM, which performs worse than ML.

$\rightarrow$ What about **alternative EM algorithms**?

Baseline ISNMF [1]

- STFT mixture: $\mathbf{X} = \sum_j \mathbf{S}_j$
- Gaussian sources: $s_{j,ft} \sim \mathcal{N}(0, v_{j,ft})$
- NMF variances: $\mathbf{V}_j = \mathbf{W}_j\mathbf{H}_j$



- ML estimation:

$$\max p(\mathbf{X}; \mathbf{W}, \mathbf{H}) \Leftrightarrow \min D_{\text{IS}}(\mathbf{V}, \mathbf{WH}) \text{ with } \mathbf{V} = |\mathbf{X}|^{\odot 2}$$

- Minimization of $D_{\text{IS}} \rightarrow$ multiplicative update rules:

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{([\mathbf{WH}]^{\odot -2} \odot \mathbf{V})\mathbf{H}^T}{[\mathbf{WH}]^{\odot -1}\mathbf{H}^T} \text{ and } \mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T([\mathbf{WH}]^{\odot -2} \odot \mathbf{V})}{\mathbf{W}^T[\mathbf{WH}]^{\odot -1}}$$

EM framework

Instead of the likelihood, maximize:

$$\mathcal{Q} = \int p(\mathbf{Z}|\mathbf{X}; \Theta') \log p(\mathbf{X}, \mathbf{Z}; \mathbf{W}, \mathbf{H}) d\mathbf{Z}$$

- Θ' = current estimates of the parameters;
- $\mathbf{Z}$ = latent variables (sources or rank-1 NMF components).

Alternate between:

- E-step: compute $\mathcal{Q}$ given Θ';
- M-step: maximize $\mathcal{Q}$ it with respect to $\mathbf{W}/\mathbf{H}$

⚠ **Computing the joint posterior $p(\mathbf{Z}|\mathbf{X}; \Theta')$ is difficult.**

The **SAGE variant** [1] approximates $\mathcal{Q}$:

- ☺ No need to compute the joint posterior distribution.
- ☹ Updates are sequential.

Proposed algorithms

Key-idea: we can compute the joint posterior.

- J sources, but $J-1$ latent variables: $\mathbf{Z} = [\mathbf{S}_1, ..., \mathbf{S}_{J-1}]$.
- Last source $\mathbf{S}_J$ = a degree of freedom.
- The joint posterior becomes non-degenerate.

$$\max \mathcal{Q} \Leftrightarrow \min \sum_j D_{\text{IS}}(\mathbf{P}_j, \mathbf{W}_j\mathbf{H}_j)$$

where $\mathbf{P}_j$ is the posterior power of $\mathbf{S}_j$.

- Minimization of $D_{\text{IS}} \rightarrow$ multiplicative update rules:

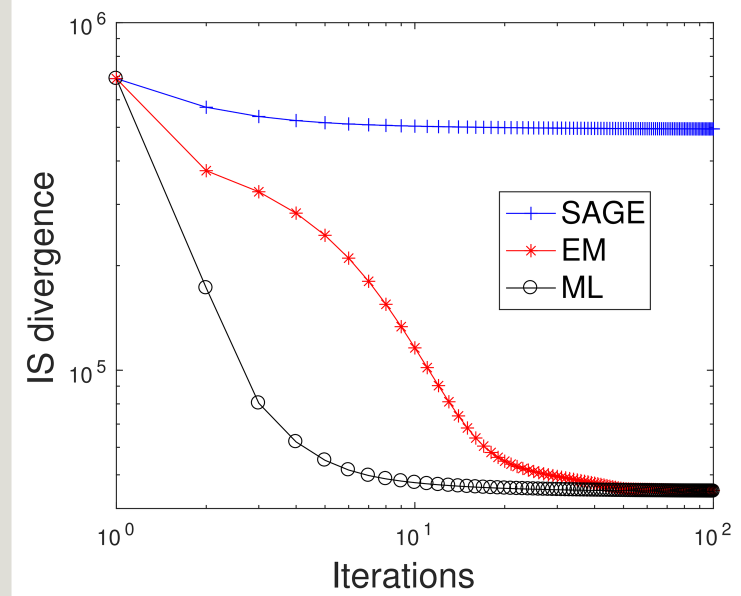
  $\mathbf{V}$, $\mathbf{W}$ and $\mathbf{H}$ are replaced by $\mathbf{P}_j$, $\mathbf{W}_j$ and $\mathbf{H}_j$.
- ☺ Updates can be done in parallel.

Other algorithms: different sets of latent variables:

- SAGE variant with $\mathbf{Z}$ = sources;
- EM with $\mathbf{Z}$ = rank-1 NMF components.

Speech separation experiments

- Two speakers (male and female) from the GRID corpus.
- For each speaker, 100 sentences (= a sequence of six words):
  - 90 = training material for learning the dictionaries $\mathbf{W}_j$;
  - 10 = test material for performing the separation.
- Separation quality: signal to distortion, interference and artifact ratios (SDR, SIR and SAR).



	SDR	SIR	SAR	Time
ML	7.0	**15.4**	7.8	3.6
SAGE	2.4	7.0	5.1	25.3
EM	**7.1**	15.1	**8.0**	6.7

- SAGE: time consuming, poor convergence and separation.
- EM outperforms SAGE and competes with ML: slightly more interference but less artifacts / overall distortion.

Conclusion

Novel EM algorithms for ISNMF that outperform SAGE and compete with ML

Applications: estimate more sophisticated models with non-tractable likelihood, e.g. [2].

References

[1] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis", in Neural Computation, March 2009.

[2] P. Magron and T. Virtanen, "Complex ISNMF: a phase-aware model for monaural audio source separation", to be published in the IEEE Trans. on Audio, Speech, and Language Processing, 2018.

https://github.com/magronp/em-isnmf