Décompositions tensorielles non-négatives du spectrogramme multicanal pour la séparation de sources musicales

Cédric Févotte

Laboratoire Lagrange

Journée ISIS "Décompositions tensorielles et applications"
Janvier 2013

Travaux en collaboration avec Alexey Ozerov (Technicolor, Rennes)

Outline

Generalities about nonnegative matrix factorization (NMF)

Nonnegative tensor decomposition for multichannel audio source separation
- CP decomposition
- Multichannel NMF

Audio results
- SiSEC 2008
- User-guided separation

Nonnegative matrix factorization (NMF)

Given a *nonnegative* matrix $\mathbf{V}$ of dimensions $F \times N$, NMF is the problem of finding a factorization

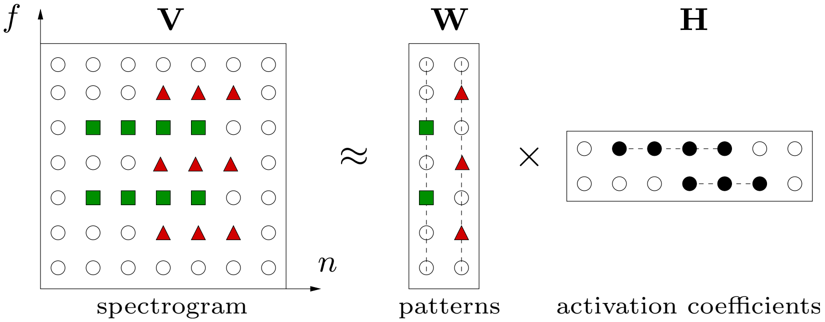
$$\mathbf{V} \approx \mathbf{W}\mathbf{H}$$

where $\mathbf{W}$ and $\mathbf{H}$ are *nonnegative* matrices of dimensions $F \times K$ and $K \times N$, respectively.

Early work by Paatero and Tapper (1994), landmark paper in *Nature* by Lee and Seung (1999).

NMF and music signal processing

NMF applied to the spectrogram, for source separation & transcription (Smaragdis and Brown, 2003)



NMF as a constrained minimization problem

Minimize a measure of fit between data $\mathbf{V}$ and model $\mathbf{WH}$, subject to nonnegativity of $\mathbf{W}$ and $\mathbf{H}$:

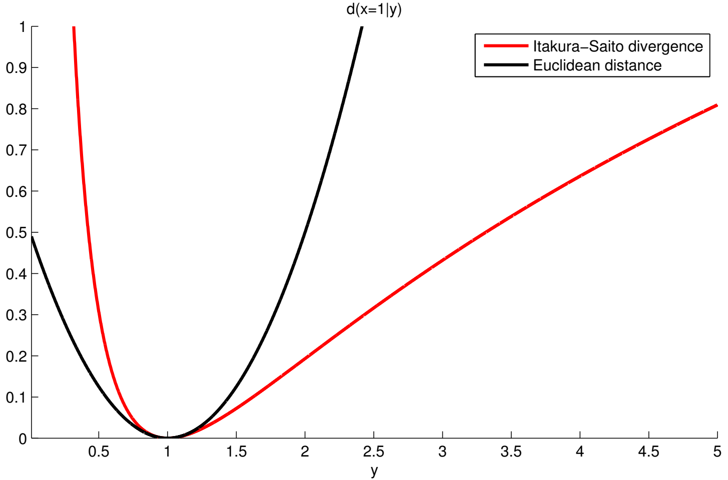
$$\min_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} D(\mathbf{V}|\mathbf{WH}) = \sum_{fn} d([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn})$$

where $d(x|y)$ is a scalar cost function.

Itakura-Saito NMF

(Févotte, Bertin, and Durrieu, 2009)

Itakura-Saito divergence: $d_{IS}(x|y) = \frac{x}{y} - \log \frac{x}{y} - 1$



Itakura-Saito NMF: inference in a generative model

(Févotte, Bertin, and Durrieu, 2009)

Let $\mathbf{X} = \{x_{fn}\}$ be the (complex-valued) STFT of the signal. Assume

$$x_{fn} = \sum_{k=1}^{K} c_{kfn}$$

$$c_{kfn} \sim \mathcal{N}_c(0, w_{fk} h_{kn})$$

and the components $c_{1fn}, \ldots, c_{Kfn}$ are independent given $\mathbf{W}$ and $\mathbf{H}$.

Itakura-Saito NMF: inference in a generative model

(Févotte, Bertin, and Durrieu, 2009)

Let $\mathbf{X} = \{x_{fn}\}$ be the (complex-valued) STFT of the signal. Assume

$$x_{fn} = \sum_{k=1}^{K} c_{kfn}$$

$$c_{kfn} \sim \mathcal{N}_c(0, w_{fk} h_{kn})$$

and the components $c_{1fn}, \ldots, c_{Kfn}$ are independent given $\mathbf{W}$ and $\mathbf{H}$. Then

$$-\log p(\mathbf{X}|\mathbf{W},\mathbf{H}) = D_{IS}(|\mathbf{X}|^2|\mathbf{WH}) \quad + cst.$$

Additivity assumed in the STFT domain. Phase is preserved in the model, though in a noninformative way (uniform distribution).

Related work by Benaroya et al. (2003); Parry and Essa (2007)

What about multichannel data ?

- NMF is suitable for single-channel data.
- Music is usually available in multichannel (at least stereo).
- Factorizing the channel spectrograms separately is suboptimal.
- The channel spectrograms form the slices of tensor.
- Use adequate tensor decompositions !

Outline

Generalities about nonnegative matrix factorization (NMF)

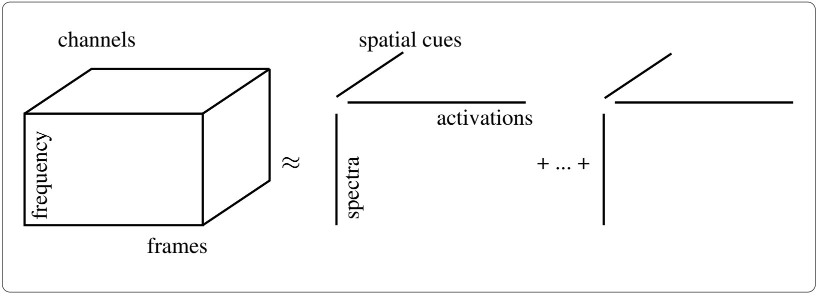
Nonnegative tensor decomposition for multichannel audio source separation

- CP decomposition
- Multichannel NMF

Audio results

- SiSEC 2008
- User-guided separation

CP decomposition
Principles



Considered for multichannel source separation and/or transcription by (FitzGerald et al., 2005, 2008; Parry and Essa, 2006; Févotte and Ozerov, 2010).

$$\min_{\mathbf{W},\mathbf{H},\mathbf{Q}} \sum_{ifn} d(v_{ifn} | \sum_k q_{ik} w_{fk} h_{nk})$$

CP decomposition

Limitations

- Underlies a linear instantaneous mixing model.
- One source is usually made of several rank-1 components: manual grouping of the spatial cues is required.
- Estimation is statistically not optimal in the standard linear instantaneous mixing model.

CP decomposition

Limitations

- Underlies a linear instantaneous mixing model.
- One source is usually made of several rank-1 components: manual grouping of the spatial cues is required.
- Estimation is statistically not optimal in the standard linear instantaneous mixing model.

IS-NMF

$$x_{fn} = \sum_{k=1}^{K} c_{kfn} \quad \text{with} \quad c_{kfn} \sim \mathcal{N}_c(0, w_{fk} h_{nk})$$

CP decomposition

Limitations

- Underlies a linear instantaneous mixing model.
- One source is usually made of several rank-1 components: manual grouping of the spatial cues is required.
- Estimation is statistically not optimal in the standard linear instantaneous mixing model.

CP IS-NMF

$$x_{ifn} = \sum_{k=1}^{K} \sqrt{q_{ik}}\, c_{kfn}^{(i)} \quad \text{with} \quad c_{kfn}^{(i)} \sim \mathcal{N}_c(0, w_{fk} h_{nk})$$

CP decomposition

Limitations

- Underlies a linear instantaneous mixing model.
- One source is usually made of several rank-1 components: manual grouping of the spatial cues is required.
- Estimation is statistically not optimal in the standard linear instantaneous mixing model.

DESIRED

$$x_{ifn} = \sum_{k=1}^{K} \sqrt{q_{ik}}\, c_{kfn} \quad \text{with} \quad c_{kfn} \sim \mathcal{N}_c(0, w_{fk} h_{nk})$$

(point source)

CP decomposition

Limitations

- Underlies a linear instantaneous mixing model.
- One source is usually made of several rank-1 components: manual grouping of the spatial cues is required.
- Estimation is statistically not optimal in the standard linear instantaneous mixing model.

DESIRED

$$x_{ifn} \;=\; \sum_{k=1}^{K} \; a_{ik} \; c_{kfn} \quad \text{with} \quad c_{kfn} \sim \mathcal{N}_c(0, w_{fk} h_{nk})$$

(point source + real mixing coefficients)

CP decomposition

Limitations

- Underlies a linear instantaneous mixing model.
- One source is usually made of several rank-1 components: manual grouping of the spatial cues is required.
- Estimation is statistically not optimal in the standard linear instantaneous mixing model.

DESIRED

$$x_{ifn} = \sum_{k=1}^{K} a_{ikf}\, c_{kfn} \quad \text{with} \quad c_{kfn} \sim \mathcal{N}_c(0, w_{fk} h_{nk})$$

(point source + real mixing coefficients + convolution)

CP decomposition
Limitations

- Underlies a linear instantaneous mixing model.
- One source is usually made of several rank-1 components: manual grouping of the spatial cues is required.
- Estimation is statistically not optimal in the standard linear instantaneous mixing model.

DESIRED

$$x_{ifn} = \sum_{k=1}^{K} a_{ikf}\, c_{kfn} \quad \text{with} \quad c_{kfn} \sim \mathcal{N}_c(0, w_{fk} h_{nk})$$

(point source + real mixing coefficients + convolution)

+ grouping: $a_{ikf} = a_{ij_kf}$ (J sources)

CP decomposition

Limitations

- Underlies a linear instantaneous mixing model.
- One source is usually made of several rank-1 components: manual grouping of the spatial cues is required.
- Estimation is statistically not optimal in the standard linear instantaneous mixing model.

DESIRED

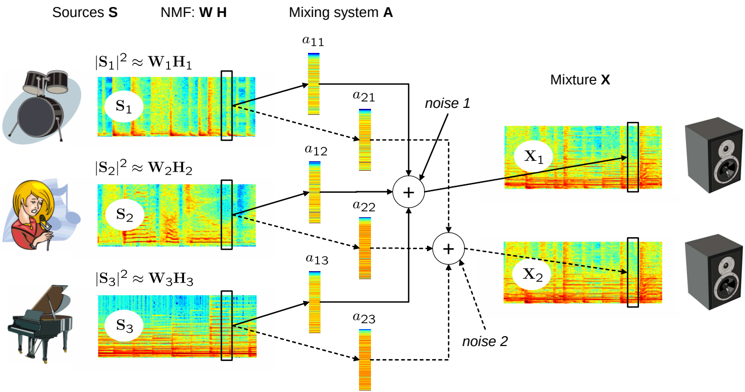
$$x_{ifn} = \sum_{k=1}^{K} {\color{red}a_{ikf}}\, c_{kfn} \quad \text{with} \quad c_{kfn} \sim \mathcal{N}_c(0, w_{fk} h_{nk})$$

(point source + real mixing coefficients + convolution)

+ grouping: $a_{i{\color{red}k}f} = a_{i{\color{red}j_k}f}$ (J sources)

= **MULTICHANNEL NMF**

Multichannel NMF
(Ozerov and Févotte, 2010)



Multichannel NMF problem: Estimate **W**, **H** and **A** from **X**

Multichannel NMF (ctd)
(Ozerov and Févotte, 2010)

Model:

$$x_{ifn} = \sum_j a_{ijf} s_{jfn}$$

$$s_{jfn} \sim \mathcal{N}_c(0, [\mathbf{W}_j \mathbf{H}_j]_{fn})$$

Maximum likelihood estimation:

$$\min_{\mathbf{W},\mathbf{H},\mathbf{A}} -\log p(\mathbf{X}|\mathbf{W}, \mathbf{H}, \mathbf{A})$$

Possible with an EM algorithm that uses the sources $\{s_{jfn}\}$ as latent data.

Outline

Generalities about nonnegative matrix factorization (NMF)

Nonnegative tensor decomposition for multichannel audio source separation

- CP decomposition
- Multichannel NMF

Audio results

- SiSEC 2008
- User-guided separation

SiSEC 2008 results

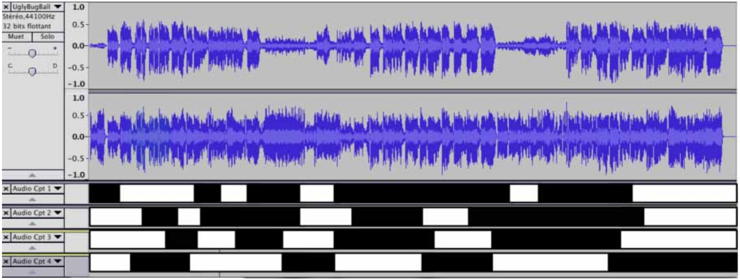
Best scores on task “Under-determined speech and music mixtures” at the 2008 Signal Separation Evaluation Campaign.

http://www.irisa.fr/metiss/SiSEC08/SiSEC_underdetermined/test_eval.html

User-guided multichannel IS-NMF

(Ozerov, Févotte, Blouet, and Durrieu, 2011)

- The decomposition is "guided" by the operator: source activation time-codes are input to the separation system.
- The temporal segmentation is reflected in the form of zeros in **H** when a source is silent.



References I

L. Benaroya, R. Gribonval, and F. Bimbot. Non negative sparse representation for Wiener based source separation with a single sensor. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 613–616, Hong Kong, 2003.

C. Févotte and A. Ozerov. Notes on nonnegative tensor factorization of the spectrogram for audio source separation : statistical insights and towards self-clustering of the spatial cues. In S. Ystad, M. Aramaki, R. Kronland-Martinet, and K. Jensen, editors, *Proc. 7th International Symposium on Computer Music Modeling and Retrieval (CMMR)*, volume 6684 of *Lecture Notes in Computer Science*, pages 102–115, Málaga, Spain, 2010., 2010. Springer. URL `http://perso.telecom-paristech.fr/~fevotte/Proceedings/cmmr10.pdf`. Long paper.

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, Mar. 2009. doi: 10.1162/neco.2008.04-08-771. URL `http://www.tsi.enst.fr/~fevotte/Journals/neco09_is-nmf.pdf`.

D. FitzGerald, M. Cranitch, and E. Coyle. Non-negative tensor factorisation for sound source separation. In *Proc. of the Irish Signals and Systems Conference*, Dublin, Ireland, Sep. 2005.

References II

D. FitzGerald, M. Cranitch, and E. Coyle. Extended nonnegative tensor factorisation models for musical sound source separation. *Computational Intelligence and Neuroscience*, 2008(Article ID 872425):15 pages, 2008. doi: 10.1155/2008/872425.

D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):550–563, Mar. 2010. doi: 10.1109/TASL.2009.2031510. URL http://www.tsi.enst.fr/~fevotte/Journals/ieee_asl_multinmf.pdf.

A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu. Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011. URL http://perso.telecom-paristech.fr/~fevotte/Proceedings/icassp11d.pdf.

P. Paatero and U. Tapper. Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5: 111–126, 1994.

References III

R. M. Parry and I. Essa. Phase-aware non-negative spectrogram factorization. In *Proc. International Conference on Independent Component Analysis and Signal Separation (ICA)*, pages 536–543, London, UK, Sep. 2007.

R. M. Parry and I. A. Essa. Estimating the spatial position of spectral components in audio. In *Proc. 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 666–673, Charleston SC, USA, Mar. 2006.

P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)*, Oct. 2003.