Matrix Co-Factorisation and Applications to Music Analysis

Slim Essid [1]

Abstract

This presentation will very briefly introduce the matrix co-factorisation paradigm, especially *nonnegative matrix co-factorisation* and discuss its applications to music analysis, in particular, multiview audio source separation in music.

1. Introduction

Matrix co-factorisation methods perform two (or more) factorisations in parallel, which are linked in a particular way. They have proven useful for various *multiview* data analysis settings, that is settings where observations are obtained from multiple sensors or interfaces, each sensor/interface contributing a particular *view* of the data. Instances of this include *multichannel* audio data, as acquired by microphone arrays, or more generally, *multimodal* data, *i.e.* heterogeneous data that involves two or more *modalities* such as the audio, textual or visual modalities in video recordings or web content.

In particular, these methods are well suited to the analysis of music as it is by essence a multimodal artefact that can be sensed in a variety of ways: music is materialized in the head of a composer, or a trained musician reading a musical-score; it is translated into sound and motion in a musician's gestures or a dancer's movements and steps; it becomes visual art when it is illustrated by disc cover designs or transformed into an audiovisual production; not to mention its textual dimension that encapsulates not only the lyrics (in sung music) and editorial meta-data, but also social web content such as user-tags, reviews, ratings, etc.

This presentation will very briefly introduce the matrix co-factorisation paradigm, discussing its applications to music analysis, especially multiview audio source separation in music. The attention will be on *nonnegative matrix co-factorisation*, which offers improved interpretability and has proven its superiority for numerous musical audio analysis tasks.

[1]LTCI, Télécom ParisTech, Université Paris Saclay. Correspondence to: Slim Essid <slim.essid@telecom-paristech.fr>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

2. Co-factorisation methods

Matrix factorisation techniques, especially their nonnegative variants (Lee & Seung, 1999), can be profitably used to extract meaningful representations for the data being analysed. When dealing with *multichannel* data—*i.e.* with data views of the same nature (*e.g.* multichannel audio or images)—observations from multiple channels may be assembled in *multi-way arrays*, *i.e. tensors*, before being modelled by tensor factorisation methods (Kolda & Bader, 2009; Cichocki et al., 2008; Yilmaz & Cemgil, 2010).

In contrast to the previous setting, data from different modalities usually live in feature spaces of completely different topology and dimensionality (think of audio as opposed to images of a video), preventing the possibility of "naturally" representing them by the same tensor. In this case, one may resort to *co-factorisation* techniques, that is techniques performing two (or more) factorisations in parallel, which are linked in a particular way. Because of the different nature of the modalities, this link has usually to be characterized through dependencies between the expansion coefficients, a.k.a activations, in cross-modal correspondence, and unlikely through dependencies between dictionary elements of different modalities.

Assuming that appropriate nonnegative features have been extracted at the same rate from the two modalities being analysed[1]—say the audio and images of a video—so that two observation matrices $\mathbf{V}_1 \in \mathbb{R}_+^{K_1 \times N}$ and $\mathbf{V}_2 \in \mathbb{R}_+^{K_2 \times N}$, assembled by stacking the observations column-wise, are available for the audio and visual modalities, one may seek a model $(\mathbf{W}_1, \mathbf{W}_2, \mathbf{H})$ such that:

$$\begin{cases} \mathbf{V}_1 \approx \mathbf{W}_1 \mathbf{H} \\ \mathbf{V}_2 \approx \mathbf{W}_2 \mathbf{H} \\ \mathbf{W}_1 \geq 0,\ \mathbf{W}_2 \geq 0,\ \mathbf{H} \geq 0\,; \end{cases} \tag{1}$$

in such a way that the activations in $\mathbf{H}$ be the same for both modalities. This is referred to as *hard co-factorisation*, an approach that has been followed in a number of works (see *e.g.* Fitzgerald et al. (2009); Yoo & Choi (2011); Yokoya et al. (2012)). Clearly, such an approach is limited in that

[1]To simplify, we consider the case of two modalities, but clearly the methods described here can be straightforwardly generalized to more than two data views by considering the relevant pairwise associations.

it does not account for possible local discrepancies across the modalities. This happens for example, in video analysis scenarios, when there is a mismatch between the audio and the images information, say because of a visual occlusion. Our *soft co-factorisation* model (Seichepine et al., 2014) stems from that limitation: it merely encourages the activations corresponding to each modality to be close, as opposed to equal, according to:

$$\begin{cases} \mathbf{V}_1 \approx \mathbf{W}_1\mathbf{H}_1 \\ \mathbf{V}_2 \approx \mathbf{W}_2\mathbf{H}_2 \\ \mathbf{H}_1 \approx \mathbf{H}_2 \\ \mathbf{W}_1 \geq 0,\, \mathbf{W}_2 \geq 0,\, \mathbf{H}_1 \geq 0,\, \mathbf{H}_2 \geq 0. \end{cases} \tag{2}$$

The model (2) is estimated by solving the following optimization problem:

$$\begin{cases} \min\limits_{\boldsymbol{\theta}} C_c(\boldsymbol{\theta}) \;;\; \boldsymbol{\theta} \triangleq (\mathbf{W}_1, \mathbf{H}_1, \mathbf{W}_2, \mathbf{H}_2) \\ \mathbf{W}_1 \geq 0,\, \mathbf{W}_2 \geq 0,\, \mathbf{H}_1 \geq 0,\, \mathbf{H}_2 \geq 0\,; \end{cases} \tag{3}$$

$$C_c(\boldsymbol{\theta}) \triangleq D_1(\mathbf{V}_1 \,|\, \mathbf{W}_1\mathbf{H}_1) + \gamma D_2(\mathbf{V}_2 \,|\, \mathbf{W}_2\mathbf{H}_2) + \delta P\left(\mathbf{H}_1, \mathbf{H}_2\right) \;; \tag{4}$$

where:

- $D_1(.\,|\,.)$ and $D_2(.\,|\,.)$ are the measures of fit respectively relating to the first and second views; note that they may be chosen to be different divergences, each well suited to the corresponding feature space;
- $P\left(.,.\right)$ is a penalty on the difference between (properly rescaled) activation values occurring at the same instant; it can for instance be chosen to be the ℓ_1 or ℓ_2-norm of the difference between the rescaled activations;
- γ and δ are regularization parameters controlling, respectively, the relative importance of each modality and the coupling penalty.

We have devised stable algorithms to solve this problem for different choices of coupling penalties and measures of fit, as well as temporal smoothing penalties, using the majorisation-minimisation approach. For more details, the interested reader is referred to (Seichepine et al., 2014).

3. Applications

The soft co-factorisation scheme has proven effective for various tasks (Seichepine et al., 2014; 2013; Sedighin et al., 2017; Parekh et al., 2017a;b) and is believed to hold promise for various multiview music analysis tasks, for instance when considering jointly:

- **Audio and visuals** in music video analysis tasks, especially in live performance videos, where the visual information can be very valuable for tasks as diverse as musical instrument recognition, source separation, melody/singing voice extraction, beat and downbeat estimation, etc. In fact, we have successfully applied soft matrix co-factorisation to multichannel (Seichepine et al., 2014) and multimodal musical audio source separation (Parekh et al., 2017a;b). In the last two works, the task considered is the separation of musical instrument sources in multimodal recordings of string quartets (Marchini et al., 2014), including audio, visual and motion-capture data. Audio source separation in this type of ensembles is known to be very challenging, hence leveraging motion features obtained from visual data turns out to be very useful for the task, based on the assumption that a set of audio activations would be "similar" to the velocity of sound-producing motion (Parekh et al., 2017b).
- **Crowd data and music**, possibly represented by audio, scores and/or lyrics, for autotagging or music recommendation tasks, where possibly item–user matrices could be processed jointly with item–musical features matrices.
- **User data and music**, as part of relevance feedback schemes, where the feedback could be either explicit, *i.e.* textual, or implicit, *e.g.* physiological, for example in settings where the user would be equipped with ECG (electrocardiographic), EMG (electromyographic) and/or EEG (electroencephalographic) sensors as they listen to the music.

4. Conclusion

Matrix co-factorisation proves to be a versatile multi-view data analysis technique that encompasses a diversity of highly expressive models. In particular, simple regularisation schemes can be deployed for the analysis of multimodal data so as to take advantage of the dependencies that exist between the data views being analysed. Our soft co-factorisation scheme goes along this line by flexibly binding together the related factors across concurrent modalities. It can additionally accommodate local regularity requirements when processing temporal sequences, through smoothing penalties.

Acknowledgements

The author warmly thanks collaborators, both PhD students, faculty colleagues and senior researchers, who have taken part in one or more aspects of the research discussed here, namely, Cédric Févotte, Nicolas Seichepine, Olivier Cappé, Sanjeel Parekh, Gaël Richard, Alexey Ozerov, Ngoc Q. K. Duong and Patrick Perez.

References

Cichocki, Andrzej, Zdunek, Raphael, and Amari, Shun-ichi. Nonnegative Matrix and Tensor Factorization. *IEEE Signal Process Mag*, 25(1):142–145, 2008. ISSN 1053-5888.

Fitzgerald, Derry, Cranitch, Matt, and Coyle, Eugene. Using tensor factorisation models to separate drums from polyphonic music. In *Proc Int Conf Digit Audio Eff*, 2009.

Kolda, T. G. and Bader, B. W. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

Lee, D. Daniel and Seung, Sebastian. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–91, oct 1999. ISSN 0028-0836. doi: 10.1038/44565.

Marchini, Marco, Ramirez, Rafael, Papiotis, Panos, and Maestre, Esteban. The sense of ensemble: a machine learning approach to expressive performance modelling in string quartets. *Journal of New Music Research*, 43 (3):303–317, 2014.

Parekh, Sanjeel, Essid, Slim, Ozerov, Alexey, Duong, Ngoc Q. K., Pérez, Patrick, and Richard, Gaël. Motion informed audio source separation. In *Under Review*, 2017a.

Parekh, Sanjeel, Essid, Slim, Ozerov, Alexey, Duong, Quang-Khanh-Ngoc, Perez, Patrick, and Richard, Gael. Guiding audio source separation by video object information. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Orleans, USA, October 2017b. Accepted.

Sedighin, F., Babaie-Zadeh, M., Rivet, B., and Jutten, C. Multimodal soft nonnegative matrix co-factorization for convolutive source separation. *IEEE Transactions on Signal Processing*, 65(12):3179–3190, June 2017. ISSN 1053-587X. doi: 10.1109/TSP.2017.2679692.

Seichepine, N, Essid, Slim, Févotte, Cédric, and Cappe, O. Soft nonnegative matrix co-factorization with application to multimodal speaker diarization. In *Proc IEEE Int Conf Acoust Speech Signal Process*, Vancouver, 2013.

Seichepine, Nicolas, Essid, Slim, Fevotte, Cedric, and Cappe, Olivier. Soft nonnegative matrix co-factorization. *IEEE Trans Signal Process*, PP(99), 2014. ISSN 1053-587X.

Yilmaz, K. and Cemgil, A. T. Probabilistic latent tensor factorisation. In *Proc Int Conf Latent Var Anal Signal Sep*, pp. 346–353, 2010.

Yokoya, Naoto, Yairi, Takehisa, and Iwasaki, Akira. Coupled Nonnegative Matrix Factorization Unmixing for Hyperspectral and Multispectral Data Fusion. *IEEE Trans Geosci Remote Sens*, 50(2):528–537, February 2012.

Yoo, Jiho and Choi, Seungjin. Matrix co-factorization on compressed sensing. In *Proc Int Joint Conf Artif Intell*, 2011.