Extraction of Temporal Patterns in Multi-rate and Multi-modal Datasets

Antoine Liutkus, Umut Şimşekli, Taylan Cemgil

Extraction of Temporal Patterns in Multi-rate and Multi-modal Datasets

Antoine Liutkus[1], Umut Şimşekli[2], and A. Taylan Cemgil[2]

[1] Inria, Speech processing team, LORIA,
Universit de Lorraine, Villers-ls-Nancy, France
antoine.liutkus@inria.fr
[2] Boğaziçi University, Dept. of Computer Engineering, İstanbul, Turkey
{umut.simsekli, taylan.cemgil}@boun.edu.tr

Abstract. We focus on the problem of analyzing corpora composed of irregularly sampled (multi-rate) heterogeneous temporal data. We propose a novel convolutive multi-rate factorization model for extracting multi-modal patterns from such multi-rate data. Our model builds up on previously proposed multi-view (coupled) nonnegative matrix factorization techniques, and extends them by accounting for heterogeneous sample rates and enabling the patterns to have a duration. We illustrate the proposed methodology on the joint study of audiovisual data for speech analysis.

Keywords: Coupled factorization, Multi-rate data analysis

1 Introduction

The last decade has witnessed a rapid growth in the size of available data. Thanks to the current technological infrastructure, massive amounts of data are continuously produced and the cost of storing this massive data gets cheaper everyday. This growth in the size of the data has brought new scientific challenges.

One major challenge is handling the data-heterogeneity. Data are often collected in different modalities (e.g., audio, video, text, etc.) at different time instances. Combining different but related data can improve estimation and prediction performance drastically, provided the different modes of the data contain sufficiently rich information and a proper model is established for jointly modeling these modes.

Various research fields have focused on the data-heterogeneity problem, such as transfer learning [9], multiple-kernel learning [3], and coupled factorizations [14]. Each of these fields has different application-specific objectives (such as increasing classification acuracy or separation performance) and therefore approach the problem from slightly different perspectives. A common theme in these works is modeling a collection of observed matrices $\{V_n\}_{n=1}^N$ by using a factorization model:

$$V_n(l,t) \approx \hat{V}_n(l,t) = \sum_k W_n(l,k)H(k,t). \quad (1)$$

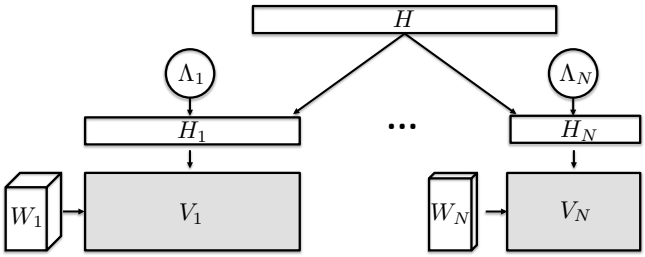

Fig. 1. Illustration of the proposed model (MULTICONV). The blocks represent the matrices and the tensors that appear in the model. The shaded blocks are observed, whereas the other ones are latent. The arrows visualize the dependency structure.

Here, V_n denotes the different modes of data, where each V_n is modeled as the product of a dictionary matrix W_n and an activation matrix H. In this modeling strategy, each mode n has its own dictionary W_n but their corresponding activations are shared among all modes, making the overall model coupled.

When different modes of the data contain temporal information, alternative factorization models can be proposed [12, 2]. In this study, we will consider the non-negative matrix factor deconvolution (NMFD) model [12], where the temporal information is incorporated through convolution:

$$V_n(l,t) \approx \hat{V}_n(l,t) = \sum_{k,p} W_n(l,k,p) H(k,p-t). \tag{2}$$

Here, the dictionary tensors W_n have temporal axis (p) that enables the dictionaries to encapsulate temporal information.

This modeling strategy has yielded many practical applications, when there is only one observed matrix (i.e., $N = 1$). However, when there are multiple observed matrices, this model requires all modalities V_n to be synchronized temporally. However, in practice, different modes of the data are often collected with different technologies. Therefore, they are usually sampled at different sampling rates, which we call 'multi-rate' data. In this study, we propose a novel convolutive factorization model that is able to model multi-rate multi-modal data. In the sequel, we will describe the model in detail and present a practical inference algorithm to estimate the parameters of the model. We illustrate the proposed method on the joint decomposition of audiovisual data for speech analysis.

2 The MULTICONV Model

In this section, we describe our model in detail. We assume that we observe N matrices $\{V_n\}_{n=1}^N$ with nonnegative entries, each one of them being of size $L_n \times T_n$, where L_n and T_n are the dimensions of each sample and the number of samples for modality n, respectively. For instance, V_1 can be the magnitude or power spectrogram of audio data, where each column might contain the spectrum

of a single audio frame, and V_2 can be video data where each column contains the vectorized version of an image. Our objective is to jointly model different modalities $\{V_n\}_n$ when their sampling rates are different, yielding possibly different number of samples T_n. Without loss of generality, let us assume that the modalities are sorted by decreasing number of samples, so that $T_1 \geq \cdots \geq T_N$. Finally, let $T_0 \geq T_1$ be an arbitrary integer, corresponding to the number of samples in some absolute *time reference*, where sampling is regular and achieved at a high precision. For concision, a sample index t of modality n will be written $t \in \mathbb{T}_n$. Note that, even though we assume all the observed data to be matrices, it is straightforward to extend the model where any V_n can be a tensor.

We will model each V_n by using an NMFD model. In order to accurately model patterns with a temporal structure, they will be taken as lasting P_n samples in modality n. A typical choice for P_n in the case of constant sampling rates is to enforce patterns to have the same absolute duration through different modalities, picking an arbitrary P_0 as the *absolute duration* of the patterns, and then choosing:

$$\forall n, P_n = \left\lceil \frac{P_0 T_n}{T_0} \right\rceil, \tag{3}$$

where $\lceil \cdot \rceil$ is the ceiling function.

The first important issue we face is to establish a temporal correspondence between the samples observed through the different modalities. The difficulty on this point is that not only the different sampling rates may be different, they may also be varying over time or even be irregular. In full generality, we introduce a *link tensor* Λ_n for each modality, of dimension $T_n \times P_n \times T_0 \times T_0$. In essence, $\Lambda_n(t, p, \tau, \tau')$ is high whenever time instants t and $t-p$ in $\mathbb{T}_n$ correspond to reference samples τ and $\tau - \tau'$ in $\mathbb{T}_0$. For instance, assume that all sampling frequencies are constant and equal. Then, we can pick $\Lambda_n(t, p, \tau, \tau') = \delta(t, \tau)\,\delta(p, \tau')$ with $\delta(t, t') = 1$ iff $t = t'$ and 0 otherwise. If sampling frequencies f_n are constant but unequal, we can for instance pick:

$$\Lambda_n(t, p, \tau, \tau') \propto \delta(\lfloor f_1 t \rceil, \lfloor f_n \tau \rceil)\,\delta(\lfloor f_1 p \rceil, \lfloor f_n \tau' \rceil), \quad \forall n \tag{4}$$

where $\lfloor \cdot \rceil$ is the rounding function and $\propto$ denotes equality up to a normalizing constant. Indeed, we assume in the sequel that Λ_n is normalized so that:

$$\forall t, p \in \mathbb{T}_n, \sum_{\tau, \tau' \in \mathbb{T}_0} \Lambda_n(t, p, \tau, \tau') = 1, \quad \forall n \tag{5}$$

With the link tensors Λ_n in hand, we can describe the actual model that decomposes the observations as the sum of only a few multi-modal patterns. For this purpose, we introduce a latent activation matrix $H(k, t)$, with fixed dimension $K \times T_0$, i.e. with the resolution of the reference time line $\mathbb{T}_0$. Finally, observation V_n is modeled as the superposition of the K patterns, activated over

time through convolution, given as follows:

$$V_n\left(l,t\right) \approx \hat{V}_n\left(l,t\right) = \sum_{k=0}^{K-1} \sum_{p=0}^{P_n-1} W_n\left(l,k,p\right) \underbrace{\sum_{\tau,\tau' \in \mathbb{T}_0} \Lambda_n\left(t,p,\tau,\tau'\right) H\left(k,\tau-\tau'\right)}_{H_n(k,t-p)}. \tag{6}$$

where H_n is the temporally adjusted activations for each modality V_n. Fig.1 illustrates the model.

3 Inference

Once we observe $\{V_n\}_n$, our aim is to estimate the parameters $\Theta = \{\{W_n\}_n, H\}$ and thus to find the multi-modal patterns as well as the way they are activated over time that best permits to account for the observed data $\{V_n\}_n$. For this purpose, we choose the parameters that minimize a cost function $C\left(\Theta\right)$, which is taken as the sum of a data-fit $J_V\left(\Theta\right)$ and a regularization term $\Psi_H\left(\Theta\right)$ for the activations H:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}}\ C\left(\Theta\right) = J_V\left(\Theta\right) + \Psi_H\left(\Theta\right). \tag{7}$$

In our setup, the data-fit criterion in (7) is taken as the sum over all the data of a scalar (element-wise) cost-function:

$$J_V\left(\Theta\right) = \sum_n \lambda_n \left[\sum_{l,t}^{L_n,T_n} d_n\left(V_n\left(l,t\right) \| \hat{V}_n\left(l,t\right)\right) \right],$$

where $d_n\left(v \| \hat{v}\right)$ is the particular cost function used for modality n and $\lambda_n > 0$ is a scalar indicating the global importance of a good fit for modality n. It assesses the similarity between one of the elements v from the observation and the corresponding element $\hat{v}$ from the model (6).

We allow for the cost-function to differ from one modality to another, mainly because observation noise may have strongly different physical origins depending on the modality. In this work, we will assume that d_n belongs to the family of β-divergences, that is defined as follows:

$$d_\beta(v||\hat{v}) = \frac{v^\beta}{\beta(\beta-1)} - \frac{v\hat{v}^{\beta-1}}{\beta-1} + \frac{\hat{v}^\beta}{\beta} \tag{8}$$

This divergence is the squared Euclidean distance for $\beta = 2$, and it can be extended by continuity at $\beta = 1$ and $\beta = 0$ to coincide with the Kullback-Leibler and Itakura-Saito divergences, respectively.

The term $\Psi_H\left(\Theta\right)$ in (7) permits to enforce some additional constraints concerning the activations H. In our context, *sparsity* is relevant and means that we expect most activations to be close to 0, and only occasionally to bear a significant magnitude. Sparse regularization for NMF has been the topic of many studies [4, 11, 7] and here, we pick the ℓ_1 norm over H as a sparsity-enforcing criterion [5].

3.1 Multiplicative updates

To update the parameters Θ so as to minimize a given cost function $C(\Theta)$, such as $C = J_V + \Psi_H$, we adopt a Majoration-Equalization approach, through Multiplicative Updates (MU) that was first presented in [6] and whose proofs for convergence were recently given in [1] for the β-divergence J_V with $\beta \in [0, 2]$. The MU methodology may be described as follows. First, we compute the derivative of $C(\Theta)$ with respect to any one Θ_i of the parameters and then express it as the difference of two nonnegative terms:

$$\frac{\partial C}{\partial \Theta_i}(\theta) = \underbrace{G_+^i(\theta)}_{\geq 0} - \underbrace{G_-^i(\theta)}_{\geq 0}. \tag{9}$$

In our case, this is easily done for the data fit and regularization functions J_V and Ψ_H chosen here. Then, instead of adopting a classical gradient descent for Θ_i, we update it multiplicatively through:

$$\Theta_i \leftarrow \Theta_i \frac{G_-^i(\Theta_i)}{G_+^i(\Theta_i)}. \tag{10}$$

When applied to $C = J_V + \Psi_H$, (10) leads to the following multiplicative updates for the parameter Θ_i:

$$\Theta_i \leftarrow \Theta_i \frac{\sum_{n,l,t} \lambda_n \hat{V}_n(l,t)^{\beta_n - 2} V_n(l,t) \frac{\partial \hat{V}_n(l,t)}{\partial \Theta_i} + \nabla_{\Theta_i}^- \Psi_H(\Theta)}{\sum_{n,l,t} \lambda_n \hat{V}_n(l,t)^{\beta_n - 1} \frac{\partial \hat{V}_n(l,t)}{\partial \Theta_i} + \nabla_{\Theta_i}^+ \Psi_H(\Theta)}. \tag{11}$$

These update rules are applied iteratively until convergence.

4 Experiments

4.1 Dataset and experimental setup

In this paper, we apply the MULTICONV model to the extraction of multimodal patterns in the MRI-TIMIT database [8]. This corpus features real time Magnetic Resonance Imaging (rtMRI) data along with the corresponding audio, for 10 different speakers, 5 males and 5 females, each one of them recorded while uttering 460 sentences from the MOCHA-TIMIT database [13].

This corpus hence consists of synchronized rtMRI and audio recordings. The rtMRI (video) has an image resolution of 68×68, with a sampling rate of 23 frames per second. Each image corresponds to a mid-sagittal slice of a speaker. The corresponding audio is sampled at 20kHz. For analysis, the audio was split into frames of 128 samples (6.5ms), with an overlap of 50% between adjacent frames. The resulting Short-Term Fourier Transform (STFT) has hence a framerate of 312 frames per second. One of the excerpt of the database is depicted on figure 2.

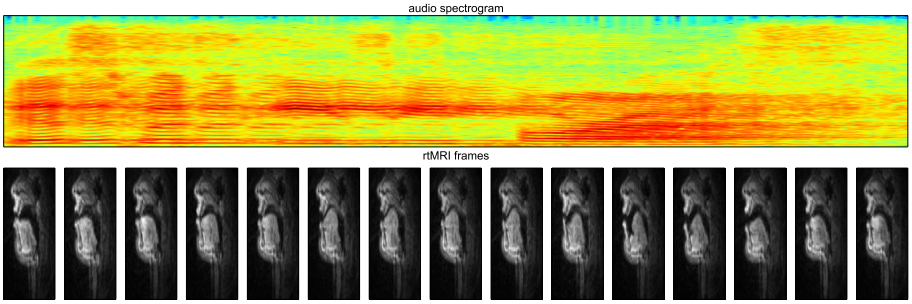

Fig. 2. Excerpt of the MRI-TIMIT database. Speaker F3, 'It was easy for us". Audio is sampled at 20kHz, rtMRI at 23 frames per second.

Our objective here is to extract meaningful articulatory audio-visual patterns, or *primitives*, from the MRI-TIMIT database by using the MULTICONV model. In this respect, the current study goes further in the direction undertaken by the pioneering work presented in [10], that focused on the same objective, but exploited the rtMRI modality only. Furthermore, while [10] also integrated sparsity constraints in the activations, these constraints were slightly different than those considered here and do not lead to straightforward multiplicative updates.

4.2 Results

The MULTICONV model was fitted on the MRI-TIMIT data for speaker F1 (sentences 1-25), using the Kullback-Leibler divergence both for the audio and rtMRI data. We chose to estimate $K = 25$ patterns having a duration of about 380ms. After two hundred iterations, we recover both the patterns and their underlying activation vectors. For the audio modality, the template is a spectrogram, while it is a short video for the rtMRI modality. Nine patterns are displayed in figure 3, where the average of the rtMRI modality is represented, for conciseness.

Interestingly, the joint convolutive modeling clearly isolates some parts of the rtMRI data, thus automatically locating the main places of articulation. In figure 3 for instance, we clearly see that the lips are identified as moving together, thus being important parts of the same pattern. However, the main interest of the MULTICONV model is to also automatically relate these places of articulation with a corresponding audio spectrogram, even if the sampling frequencies of these modalities are very different.

Notwithstanding the interest of performing such an unsupervised analysis of multi-modal data, the qualitative use of these results by professional linguists is made difficult by the lack of phonological information. Indeed, it seems natural to associate each pattern to a phoneme, and learning the MULTICONV model would then amount to estimating the best rtMRI and associated spectrogram for

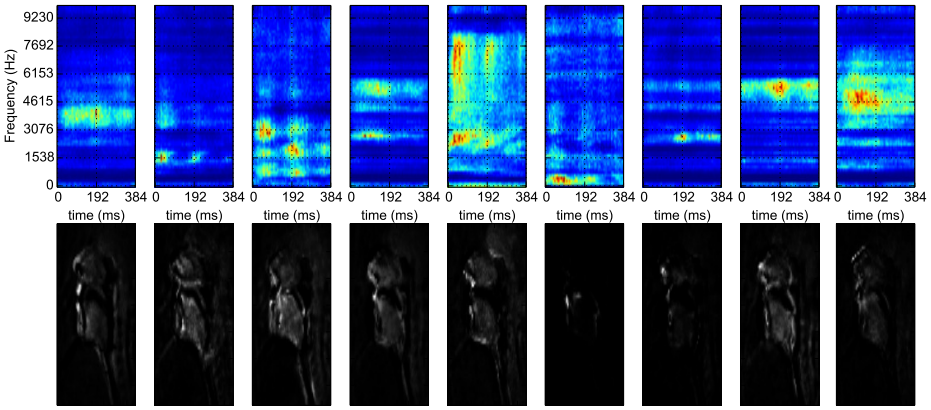

Fig. 3. Nine multi-modal patterns learned with the MULTICONV model on the MRI-TIMIT dataset.

each phoneme. To achieve this, we simply need to make use of a transcription and adapt the model so that the activation of each pattern-phoneme is zero except at the beginning of all occurrences of the phoneme. We leave this for future work.

5 Conclusion

Multi-view data analysis is concerned with corpora composed of heterogeneous items from different modalities, such as audio, video, images or text. In most studies, either all modalities are assumed perfectly synchronized or the phenomenon under study is intrinsically non-temporal, such as images or textual documents. In those cases, a joint analysis in effect often boils down to data concatenation. In this paper, we have proposed the MULTICONV model, to extract multi-modal patterns from the joint analysis of data-streams that exhibit different or even non-constant frame-rates, and that capture different aspects of the same phenomenon. This model builds on previously proposed multi-view Nonnegative Matrix Factorization techniques (NMF), but significantly extends them by both accounting for heterogeneous sample rates and by enabling patterns to have a duration, which proves fundamental in the study of datasets that are relative to temporal phenomena. In practice, we propose a convolutive multi-rate NMF model, where temporal patterns are activated simultaneously over the different modalities through a shared underlying activation stream. The multi-rate problem is addressed by the incorporation of a reference time scale, which subsumes many different sampling scenarios and permits to bind the modalities together. We illustrated the proposed methodology with a preliminary analysis of an audio-visual corpus of speech data.

References

1. Févotte, C., Idier, J.: Algorithms for nonnegative matrix factorization with the beta-divergence. Neural Computation 23(9), 2421–2456 (Sep 2011)
2. Févotte, C., Le Roux, J., Hershey, J.R.: Non-negative dynamical system with application to speech and audio. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. pp. 3158–3162 (2013)
3. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. Journal of Machine Learning Research 12(Jul), 2211–2268 (2011)
4. Joder, C., Weninger, F., Virette, D., Schuller, B.: A comparative study on sparsity penalties for nmf-based speech separation: Beyond lp-norms. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. pp. 858–862. IEEE (2013)
5. Le Roux, J., Weninger, F., Hershey, J.: Sparse NMF – half-baked or well done? Tech. Rep. TR2015-023, Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA (Mar 2015)
6. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems (NIPS). vol. 13, pp. 556–562. The MIT Press (Apr 2001)
7. Lefèvre, A., Bach, F., Févotte, C.: Itakura-saito nonnegative matrix factorization with group sparsity. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Prague, Czech Republic (May 2011)
8. Narayanan, S., Bresch, E., Ghosh, P., Goldstein, L., Katsamanis, A., Kim, Y., Lammert, A., Proctor, M., Ramanarayanan, V., Zhu, Y.: A multimodal real-time mri articulatory corpus for speech research. In: INTERSPEECH. pp. 837–840 (2011)
9. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. on Knowl. and Data Eng. (2010)
10. Ramanarayanan, V., Katsamanis, A., Narayanan, S.: Automatic Data-Driven Learning of Articulatory Primitives from Real-Time MRI Data Using Convolutive NMF with Sparseness Constraints. In: INTERSPEECH. pp. 61–64. ISCA (2011)
11. Smaragdis, P., Shashanka, M., Raj, B., Mysore, G.J.: Probabilistic factorization of non-negative data with entropic co-occurrence constraints. In: ICA '09: Proc. of the 8th Int. Conf. on Independent Component Analysis and Signal Separation (2009)
12. Smaragdis, P.: Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In: Independent Component Analysis and Blind Signal Separation, pp. 494–499. Springer (2004)
13. Wrench, A.: A multi-channel/multi-speaker articulatory database for continuous speech recognition research. Phonus. 5, 1–13 (2000)
14. Yilmaz, Y.K., Cemgil, A.T., Simsekli, U.: Generalised coupled tensor factorisation. In: NIPS (2011)