CONVOLUTIVE NON-NEGATIVE MATRIX FACTORISATION WITH A SPARSENESS CONSTRAINT

Paul D. O'Grady *Barak A. Pearlmutter*

Hamilton Institute
National University of Ireland Maynooth,
Co. Kildare, Ireland.

ABSTRACT

Discovering a representation which allows auditory data to be parsimoniously represented is useful for many machine learning and signal processing tasks. Such a representation can be constructed by Non-negative Matrix Factorisation (NMF), a method for finding parts-based representations of non-negative data. We present an extension to NMF that is convolutive and includes a sparseness constraint. In combination with a spectral magnitude transform, this method discovers auditory objects and their associated sparse activation patterns.

1. INTRODUCTION

A preliminary step in many data analysis tasks is to find a suitable representation of the data. Typically, methods exploit the latent structure in the data. For example, ICA [1] reduces the redundancy of the data by projecting the data onto its independent components, which can be discovered by maximising a statistical measure such as independence [2] or non-Gaussianity [3].

Non-negative Matrix Factorisation (NMF) approximately decomposes a non-negative matrix $\mathbf{V}$ into a product of two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ [4, 5]. NMF is a parts-based approach that does not make a statistical assumption about the data. Instead, it assumes that for the domain at hand, negative numbers would be physically meaningless. The lack of statistical assumptions makes it difficult to prove that NMF will give correct decompositions, although it has been shown geometrically that NMF provides a correct decomposition for some classes of images [6].

Data that contains negative components, for example audio, must be transformed into a non-negative form before NMF can be applied. Here, we use a magnitude spectrogram. Spectrograms have been used in audio analysis for many years [7] and in combination with NMF have been applied to a variety of problems such as speech separation [8–10] and automatic transcription of music [11].

Supported by Higher Education Authority of Ireland (An tÚdarás Um Ard-Oideachas) and Science Foundation Ireland grant 00/PI.1/C067.

In this paper we combine a previous convolutive extension of NMF [9] that is able to identify auditory objects with time-varying spectra with a sparseness constraint [12] and apply the resulting algorithm to the analysis of audio data. The paper is structured as follows. Section 2 presents NMF and discusses its performance using experiments on synthetic data. Convolutive NMF and its advantages over conventional NMF are discussed in Section 3. In Section 4 we add an additional sparseness constraint to the convolutive NMF objective and discuss an experiment on musical data.

2. NON-NEGATIVE MATRIX FACTORISATION

NMF is a linear non-negative approximate factorisation, and is formulated as follows. Given a non-negative $M \times N$ matrix $\mathbf{V} \in \mathbb{R}^{\geq 0, M \times N}$ the goal is to approximate $\mathbf{V}$ as a product of two non-negative matrices $\mathbf{W} \in \mathbb{R}^{\geq 0, M \times R}$ and $\mathbf{H} \in \mathbb{R}^{\geq 0, R \times N}$

$$\mathbf{V} \approx \mathbf{W} \cdot \mathbf{H} \tag{1}$$

where $R \leq M$, such that the reconstruction error is minimised. Two NMF algorithms were introduced by Lee and Seung [4, 13], each optimising its own measure of the quality of the approximation. The first cost function presented is the Euclidean distance between $\mathbf{V}$ and $\mathbf{WH}$, the second is a generalised version of the Kullback-Leibler divergence. We will use the latter

$$D(\mathbf{V} \| \mathbf{W}, \mathbf{H}) = \left\| \mathbf{V} \otimes \log \frac{\mathbf{V}}{\mathbf{W} \cdot \mathbf{H}} - \mathbf{V} + \mathbf{W} \cdot \mathbf{H} \right\| \tag{2}$$

where $\otimes$ denotes an element-wise (also known as Hadamard or Schur product) multiplication, and division is also element-wise. NMF can now be written as an optimisation problem.

$$\min_{\mathbf{W}, \mathbf{H}} D(\mathbf{V} \| \mathbf{W}, \mathbf{H}) \qquad \mathbf{W}, \mathbf{H} \geq 0$$

The above objective is convex in $\mathbf{W}$ and $\mathbf{H}$ individually but not together. Therefore algorithms usually alternate updates of $\mathbf{W}$ and $\mathbf{H}$. The objective can be minimised using a diagonally rescaled gradient descent algorithm [4], which leads to multiplicative updates

$$\mathbf{H} = \mathbf{H} \otimes \frac{\mathbf{W}^{\mathsf{T}} \cdot \frac{\mathbf{V}}{\mathbf{W}\cdot\mathbf{H}}}{\mathbf{W}^{\mathsf{T}} \cdot \mathbf{1}} \qquad \mathbf{W} = \mathbf{W} \otimes \frac{\frac{\mathbf{V}}{\mathbf{W}\cdot\mathbf{H}} \cdot \mathbf{H}^{\mathsf{T}}}{\mathbf{1} \cdot \mathbf{H}^{\mathsf{T}}} \qquad (3)$$

where $\mathbf{1}$ is an $M \times N$ matrix with all its elements equal to unity. As the algorithm iterates, the factors converge to a local optimum of Eq. 2.

The parameter R, which is the number of columns in $\mathbf{W}$ and rows in $\mathbf{H}$, specifies the rank of the approximation. If $R < M$ then $\mathbf{W}$ is under-determined and NMF reveals low-rank features of the data. The columns of $\mathbf{W}$ will contain the basis for the data while the rows of $\mathbf{H}$ will contain activation patterns for each basis. The selection of an appropriate value for R usually requires prior knowledge, and is important to obtaining a satisfactory decomposition.

2.1. NMF applied on audio spectra

To illustrate the application of NMF on audio data consider the example shown in Figure 1. The signal under consideration is composed of two band-limited noise bursts with magnitude spectra constant over time. The first burst is centred around 2 kHz and occurs four times, while the second burst is centred around 6 kHz and occurs three times. The signal's spectrogram is an $M \times N$ matrix $\mathbf{V}$ with magnitude information for M frequency bins at N time intervals. NMF is applied to $\mathbf{V}$ with $R = 2$ and the resultant factors shown. In this example both the frequency spectra of the bursts (columns of $\mathbf{W}$) and their activations in time (rows of $\mathbf{H}$) have been identified. Therefore, this decomposition has successfully revealed the structure of $\mathbf{V}$ by correctly describing its constituent elements in both the frequency and time domains.

Now consider the example presented in Figure 2. Here, the signal under consideration is composed of two auditory objects that have differing frequency sweeps over time. The first object is centred around 2 kHz and the second object is centred around 6 kHz, each occurring four times. NMF is applied to the data with the same parameters as above and the factors are shown. It is evident from the columns of $\mathbf{W}$ that the identified spectra contain frequency components that are centred around both 2 kHz and 6 kHz. Thus, NMF fails to identify the spectra of each object and instead discovers objects that are a combination of both. The reasons for this are that the spectra of the auditory objects evolve over time and that NMF is not expressive enough to reveal this temporal structure. Therefore, in order to reveal a correct decomposition, the expressive properties of NMF need to be extended to consider the evolution of each object's spectrum.

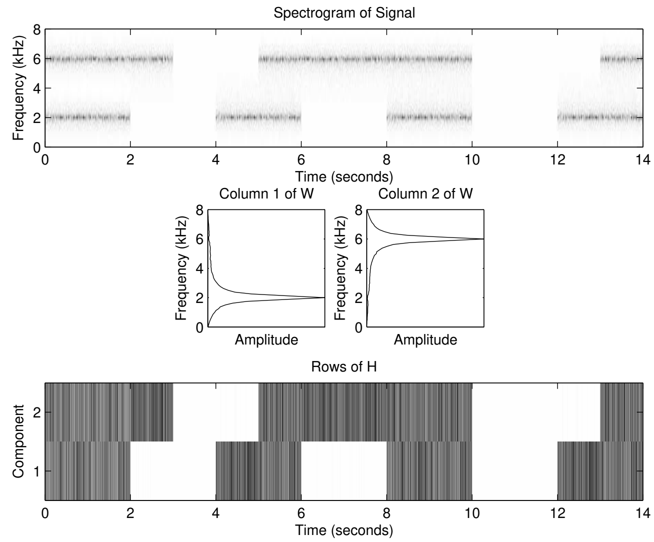

Fig. 1. Spectrogram of a signal composed of band-limited noise bursts, and its factors obtained by NMF.

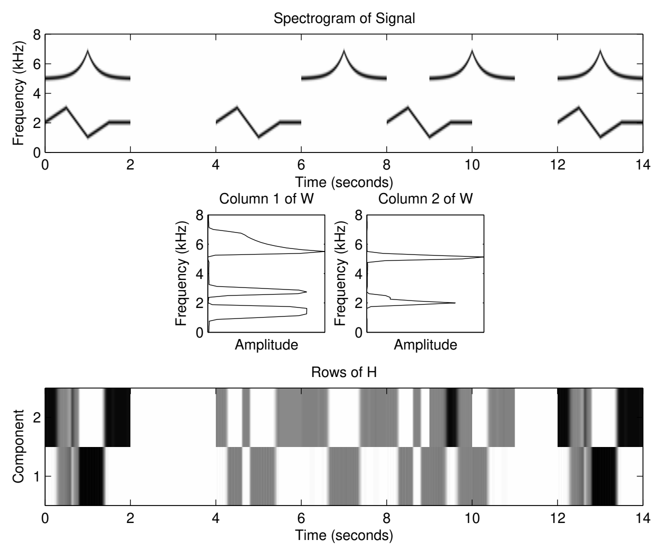

Fig. 2. Spectrogram of a signal composed of auditory objects with time-varying spectra, and its factors obtained by NMF.

3. CONVOLUTIVE NMF

Often the temporal relationship between multiple observations over nearby intervals of time are discovered using a convolutive generative model. Such a model has previously been used to extend ICA [14] and NMF [9], the latter constituting the algorithm we review in this section. In conventional NMF each object is described by its spectrum and corresponding activation in time, while in convolutive NMF each object has a sequence of successive spectra and corresponding activation pattern across time. The model of Eq. 1 is extended to the convolutive case

$$\mathbf{V} \approx \sum_{t=0}^{T-1} \mathbf{W}_t \cdot \overset{t\rightarrow}{\mathbf{H}}$$

where $\mathbf{V} \in \mathbb{R}^{\geq 0, M \times N}$ is the input to be decomposed, $\mathbf{W}_t \in \mathbb{R}^{\geq 0, M \times R}$ and $\mathbf{H} \in \mathbb{R}^{\geq 0, R \times N}$ are its two factors, and T is the length of each spectrum sequence. The i-th column of $\mathbf{W}_t$ describes the spectrum of the i-th object t time steps after the object has begun. The $\overset{i\rightarrow}{(\cdot)}$ denotes a column shift operator that moves its argument i places to the right; as each column is shifted off to the right the leftmost columns are zero filled. Conversely, the $\overset{\leftarrow i}{(\cdot)}$ operator shifts columns off to the left, with zero filling on the right.

Using the previously presented framework for NMF, the new cost function for the convolutive generative model is

$$D(\mathbf{V}\|\mathbf{\Lambda}) = \left\| \mathbf{V} \otimes \log \frac{\mathbf{V}}{\mathbf{\Lambda}} - \mathbf{V} + \mathbf{\Lambda} \right\| \tag{4}$$

where $\mathbf{\Lambda}$ is the approximation to $\mathbf{V}$ and is defined as

$$\mathbf{\Lambda} = \sum_{t=0}^{T-1} \mathbf{W}_t \cdot \overset{t\rightarrow}{\mathbf{H}}$$

This new cost function can be viewed as a set of T conventional NMF operations that are summed to produce the final result. Consequently, as opposed to updating two matrices ($\mathbf{W}$ and $\mathbf{H}$) as in conventional NMF, $T+1$ matrices require an update ($\mathbf{W}_0, \ldots, \mathbf{W}_{T-1}$ and $\mathbf{H}$). The resultant convolutive NMF update equations are

$$\mathbf{H} = \mathbf{H} \otimes \frac{\mathbf{W}_t^{\mathsf{T}} \cdot \overset{\leftarrow t}{[\frac{\mathbf{V}}{\mathbf{\Lambda}}]}}{\mathbf{W}_t^{\mathsf{T}} \cdot \mathbf{1}} \qquad \mathbf{W}_t = \mathbf{W}_t \otimes \frac{\frac{\mathbf{V}}{\mathbf{\Lambda}} \cdot \overset{t\rightarrow}{\mathbf{H}}^{\mathsf{T}}}{\mathbf{1} \cdot \overset{t\rightarrow}{\mathbf{H}}^{\mathsf{T}}} \tag{5}$$

where $\mathbf{H}$ is updated to the average result of its updates for all t. When $T = 1$ this reduces to conventional NMF (Eq. 3).

3.1. Convolutive NMF applied on audio spectra

We have shown that conventional NMF reveals a correct decomposition for auditory objects with constant spectra, but fails for objects that exhibit time-varying spectra. Let us now consider the application of convolutive NMF to this example. The performance of the algorithm now depends on two parameters R and T, where T must be larger than the temporal extent of each object. Convolutive NMF is applied to the data with $R = 2$ and $T = 2$ seconds, and the resultant factors are presented in Figure 3. It is evident from the spectral sequences obtained (i-th column of $\mathbf{W}_t$, for $t = 0, 1, \ldots, T-1$) that the time-varying spectra of each object has been revealed and that the rows of $\mathbf{H}$ identify the start of each object. Therefore, the decomposition has successfully revealed the structure of $\mathbf{V}$ by correctly describing the spectral evolution of each object and its position in time.

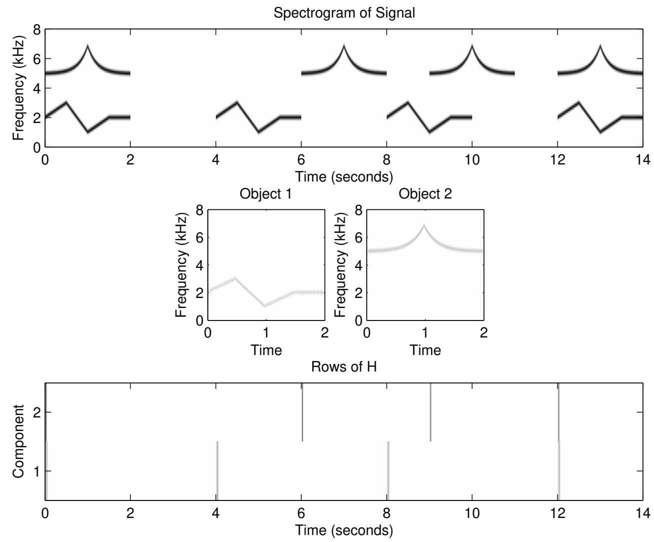

Fig. 3. Spectrogram of a signal composed of auditory objects with time-varying spectra, and its factors obtained by convolutive NMF.

4. CONVOLUTIVE NMF WITH ADDITIONAL CONSTRAINTS

For some tasks it may be advantageous to perform NMF with additional constraints placed on either $\mathbf{W}$ or $\mathbf{H}$. One increasingly popular and powerful constraint is that the rows of $\mathbf{H}$ have a parsimonious activation pattern for the basis contained in the columns of $\mathbf{W}$. This is the so called *Sparseness Constraint* [15, 16]. A signal is said to be sparse when it is zero or nearly zero more than might be expected from its variance. Such a signal has a probability density function or distribution of values with a sharper peak at zero and fatter tails than a Gaussian. A standard sparse distribution is the Laplacian distribution ($p(c) \propto \exp -|c|$). The advantage of a sparse signal representation is that the probability of two or more activation patterns being active simultaneously is low. Thus, sparse representations lend themselves to good separability [17]. Although convolutive NMF produces activation patterns that tend to be sparse, the addition of the sparseness constraint on $\mathbf{H}$ provides a means of trading off the sparseness of the representation against accurate reconstruction.

The most widely used method for multi-objective optimisation is the weighted sum method. This method creates an aggregate objective function by multiplying each constituent cost function by a weighting factor and summing the weighted costs. Combining our reconstruction cost function (Eq. 4) with a sparseness constraint on $\mathbf{H}$ results in the following objective function

$$G(\mathbf{V}\|\mathbf{\Lambda}) = D(\mathbf{V}\|\mathbf{\Lambda}) + \lambda \sum_{ij} \mathbf{H}_{ij} \tag{6}$$

The left term of the objective function corresponds to NMF, while the right term is an additional constraint on $\mathbf{H}$ that enforces sparsity by minimising the L_1-norm of its elements [18]. The parameter λ controls the trade off between sparseness and accurate reconstruction.

This objective creates a new problem: the right term is a strictly increasing function of the absolute value of its argument, so it is possible that the objective can be decreased by scaling $\mathbf{W}$ up and $\mathbf{H}$ down ($\mathbf{W} \mapsto \alpha\mathbf{W}$ and $\mathbf{H} \mapsto (1/\alpha)\mathbf{H}$, with $\alpha > 1$). This situation does not alter the left term in the objective function, but will cause the right term to decrease, resulting in the elements of $\mathbf{W}$ growing without bound and $\mathbf{H}$ tending toward zero. Consequently, the solution arrived at by the optimisation algorithm is not influenced by the right term of the objective function and the resultant $\mathbf{H}$ matrix is not sparse. Therefore another constraint needs to be introduced. This is achieved by fixing the L_2-norm of the i-th object of $\mathbf{W}$ to unity, over all $t = 0, 1, \ldots, T-1$. This appropriately constrains the scale of the elements in $\mathbf{W}$ and $\mathbf{H}$.

4.1. New update rules

The classic NMF update rules [4] implement gradient descent and our new updates will also follow this approach. First we consider the gradient descent update for $\mathbf{H}$,

$$\mathbf{H} = \mathbf{H} + \eta_{\mathbf{H}} \nabla_{\mathbf{H}} G(\mathbf{V} \| \mathbf{\Lambda})$$

Taking the gradient of Eq. 6 with respect to $\mathbf{H}$ gives

$$\nabla_{\mathbf{H}} G(\mathbf{V} \| \mathbf{\Lambda}) = \mathbf{W}_t^{\mathsf{T}} \cdot \overset{\leftarrow t}{\left[\frac{\mathbf{V}}{\mathbf{\Lambda}} \right]} - \mathbf{W}_t^{\mathsf{T}} \cdot \mathbf{1} + \lambda \cdot \mathbf{1}$$

Diagonally rescaling the variables [4, 12] and setting the learning rate to

$$\eta_{\mathbf{H}} = \frac{\mathbf{H}}{\mathbf{W}_t^{\mathsf{T}} \cdot \mathbf{1} + \lambda \cdot \mathbf{1}}$$

gives the new multiplicative update rule for $\mathbf{H}$

$$\mathbf{H} = \mathbf{H} \otimes \frac{\mathbf{W}_t^{\mathsf{T}} \cdot \overset{\leftarrow t}{[\frac{\mathbf{V}}{\mathbf{\Lambda}}]}}{\mathbf{W}_t^{\mathsf{T}} \cdot \mathbf{1} + \lambda \cdot \mathbf{1}} \tag{7}$$

Similarly, we derive a new update for the $\mathbf{W}_t$

$$\mathbf{W}_t = \mathbf{W}_t + \eta_{\mathbf{W}} \nabla_{\mathbf{W}_t} G(\mathbf{V} \| \mathbf{\Lambda})$$

where the gradient of Eq. 6 with respect to $\mathbf{W}_t$ is

$$\nabla_{\mathbf{W}_t} G(\mathbf{V} \| \mathbf{\Lambda}) = \frac{\mathbf{V}}{\mathbf{\Lambda}} \cdot \overset{t\rightarrow}{\mathbf{H}}^{\mathsf{T}} - \mathbf{1} \cdot \overset{t\rightarrow}{\mathbf{H}}^{\mathsf{T}}$$

The additional unit norm constraint on $\mathbf{W}_t$ complicates the update rule and impedes the discovery of a suitable form for $\eta_{\mathbf{W}}$ that would result in a multiplicative update [12], resulting in the following additive update

$$\mathbf{W}_t = \mathbf{W}_t + \eta_{\mathbf{W}} \left[\frac{\mathbf{V}}{\mathbf{\Lambda}} \cdot \overset{t\rightarrow}{\mathbf{H}}^{\mathsf{T}} - \mathbf{1} \cdot \overset{t\rightarrow}{\mathbf{H}}^{\mathsf{T}} \right] \tag{8}$$

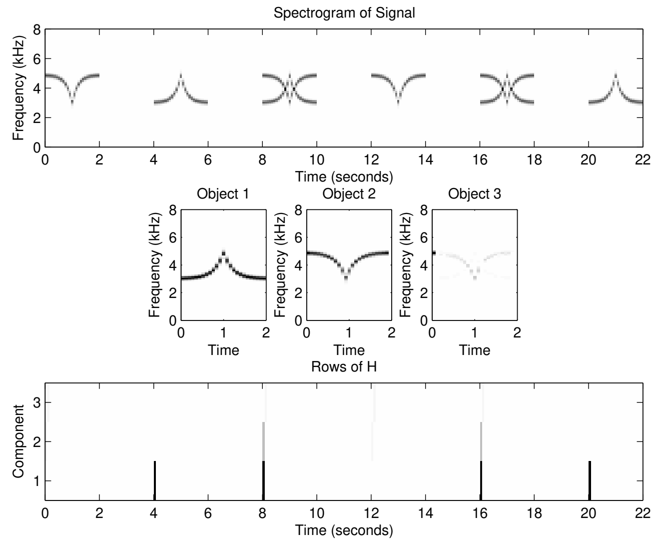

Fig. 4. Spectrogram of a signal composed of an over-complete basis, and its factors obtained by convolutive NMF.

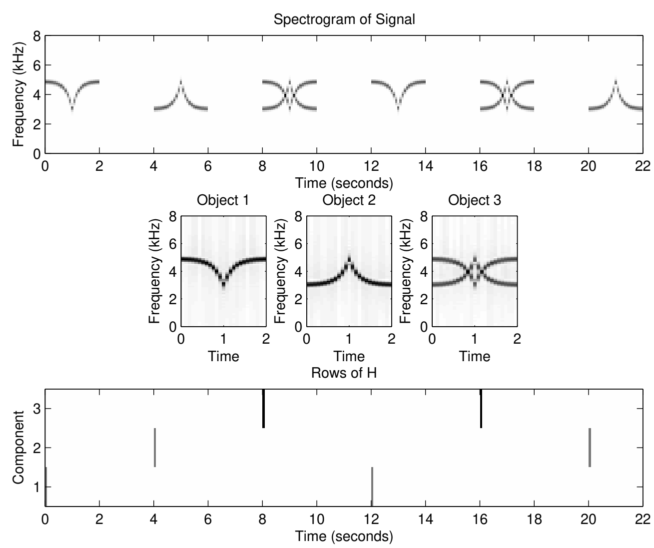

Fig. 5. Spectrogram of a signal composed of an over-complete basis, and its factors obtained by sparse convolutive NMF.

After this update, any negative values in $\mathbf{W}_t$ are set to zero (non-negativity constraint), and each $\mathbf{W}_t$ is rescaled to unit L_2-norm. As long as $\eta_{\mathbf{W}}$ and $\eta_{\mathbf{H}}$ are sufficiently small these updates should reduce Eq. 6.

4.2. Sparse Convolutive NMF applied on audio spectra

An interesting property of the sparseness constraint is that it enables the discovery of an over-complete basis, *i.e.*, a basis that contains more basis functions than are necessary to span

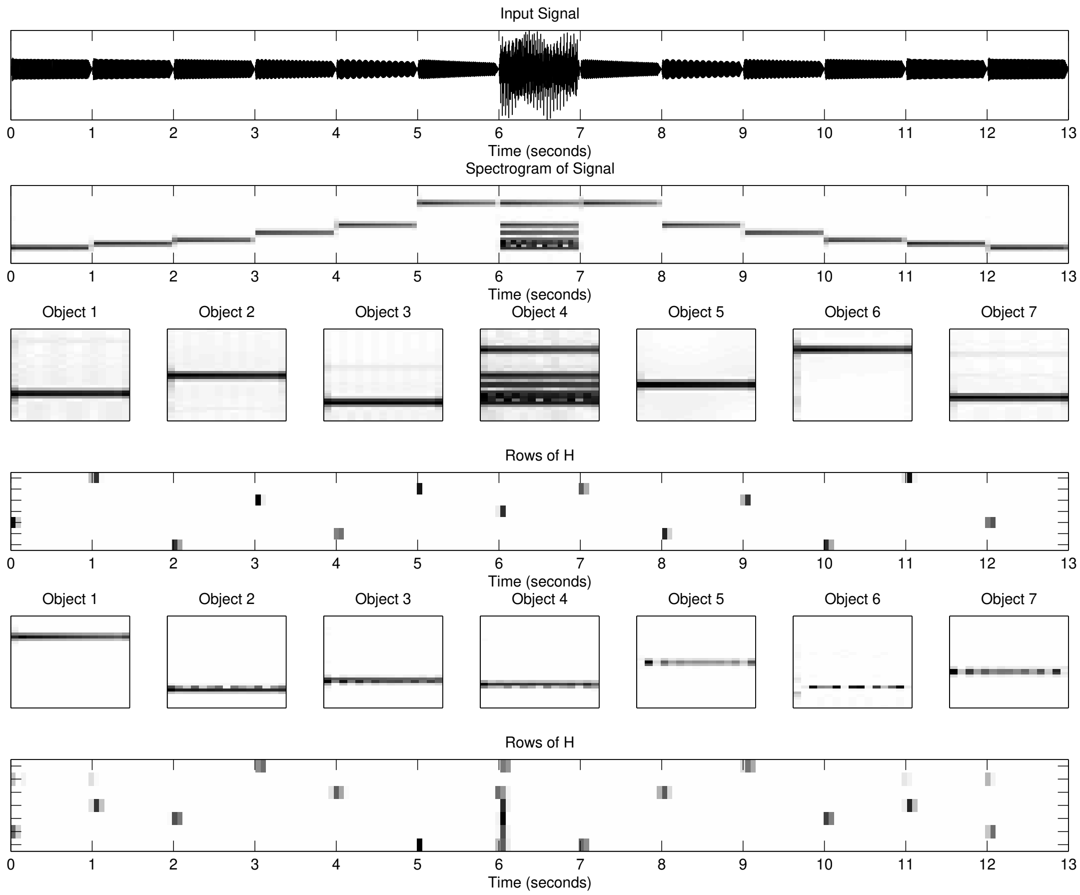

Fig. 6. Music waveform and its associated spectrogram along with its factors obtained by sparse convolutive NMF (rows 3 & 4) and conventional convolutive NMF (rows 5 & 6).

the projection space.

To illustrate the performance of convolutive NMF on data generated from an over-complete basis consider the example presented in Figure 4. The signal under consideration is composed of three auditory objects each occurring twice, where the first object is an exponentially decreasing then increasing frequency sweep centred around 4 kHz, the second object is the reverse of the first, and the third object is a combination of the first two. Convolutive NMF is applied to the data with $R = 3$ and $T = 2$ seconds, and the resultant factors are presented. It is evident from the results that only the first two auditory objects are identified. This is because the third object can be expressed in terms of the first two and the signal can be described by using just the first two objects. Thus, convolutive NMF achieves its optimum with just the first two linearly independent objects without the need for an over-complete representation.

When the sparseness constraint is included, the existence of an over-complete representation helps minimise the objective, allowing for a sparser description of the signal. Sparse convolutive NMF applied to the same signal (Figure 5) identifies all three objects and their associated activation patterns, successfully revealing the over-complete basis used to generate the signal.

4.3. Sparse Convolutive NMF applied on music

We now compare these algorithms on a simple music example, consisting of rudimentary synthesised guitar notes. Each note produces only its fundamental frequency. The arrangement is simple, composed of three sections: the six notes of a G chord are played individually in descending order; all six notes of the chord are played simultaneously; and each note is played in reverse order. Each note is played for one second, and the frequencies of the notes are 98.00 Hz (G), 123.47 Hz (B), 146.83 Hz (D), 196.00 Hz (G), 246.94 Hz (B) and 392.00 Hz (G).

Both sparse convolutive NMF and convolutive NMF are applied to the music and the resultant factors are presented in Figure 6. It is evident from the spectrogram that the music can be represented by an over-complete representation consisting of each individual note and the chord. Convolutive NMF is applied with $R = 7$, $T = 1$ seconds and the resultant factors are presented in rows 5 & 6. As can be seen from the activation pattern, the algorithm has failed to represent the chord as an individual auditory object and instead represents it as a combination of notes. Sparse convolutive NMF is applied with the same parameters, where the additional parameter λ is selected on an *ad hoc* basis. The resultant factors are presented in rows 3 & 4. Here, it is evident that an over-complete representation is discovered in which the chord is represented as an individual auditory object.

5. CONCLUSIONS

In this paper we have presented a sparse convolutive version of NMF that effectively discovers a sparse parts-based representation for non-negative data. This method extends the convolutive NMF objective by including a sparseness constraint on the activation patterns, enabling the discovery of over-complete representations. We have reviewed how the expressive properties of NMF can be improved by reformulation of the problem in a convolutive framework, and shown that the addition of a sparseness constraint can lead to the discovery of appropriate over-complete representations in music.

References

[1] P. Comon, "Independent component analysis: A new concept," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[2] Anthony J. Bell and Terrence J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neu. Comp.*, vol. 7, no. 6, pp. 1129–1159, 1995.

[3] Aapo Hyvärinen and Erkki Oja, "A fast fixed-point algorithm for independent component analysis," *Neu. Comp.*, vol. 9, no. 7, pp. 1483–1492, Oct. 1997.

[4] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Adv. in Neu. Info. Proc. Sys. 13*. 2001, pp. 556–562, MIT Press.

[5] P. Paatero and U. Tapper, "Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, pp. 111–126, 1994.

[6] David Donoho and Victoria Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?," in *Adv. in Neu. Info. Proc. Sys. 16*. 2004, MIT Press.

[7] Ralph K. Potter, George A. Kopp, and Harriet C. Green, *Visible Speech*, D. Van Nostrand Company, 1947.

[8] Tuomas Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *in Proceedings of the International Computer Music Conference (ICMC 2003)*, 2003.

[9] Paris Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Fifth International Conference on Independent Component Analysis*, Granada, Spain, Sept. 22–24 2004, LNCS 3195, pp. 494–499, Springer-Verlag.

[10] D. FitzGerald, M. Cranitch, and E. Coyle, "Sound source separation using shifted non-negative tensor factorisation," in *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

[11] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, 2004, pp. 318–325.

[12] Patrik O. Hoyer, "Non-negative sparse coding," in *IEEE Workshop on Neural Networks for Signal Processing*, 2002.

[13] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects with nonnegative matric factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[14] R. H. Lambert, *Multichannel Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures*, Ph.D. thesis, Univ. of Southern California, 1996.

[15] B. A. Olshausen and D. J. Field, "Sparse coding of sensory inputs," *Curr Opin Neurobiol*, vol. 14, no. 4, pp. 481–487, 2004.

[16] D. J. Field, "What is the goal of sensory coding?," *Neural Computation*, vol. 6, pp. 559–601, 1994.

[17] Michael Zibulevsky and Barak A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neu. Comp.*, vol. 13, no. 4, pp. 863–882, Apr. 2001.

[18] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.