Sparse Non-negative Tensor Factor Double Deconvolution (SNTF2D) for multi channel time-frequency analysis

Morten Mørup and **Mikkel N. Schmidt**
Informatics and Mathematical Modelling
Technical University of Denmark
Richard Petersens Plads, Building 321
2800 Kgs Lyngby
email: {mm,mns}@imm.dtu.dk

Editor:

Abstract

We recently introduced two algorithms for sparse non-negative matrix factor 2-D deconvolution (SNMF2D) (Mørup and Schmidt, 2006) that are useful for single channel source separation (Schmidt and Mørup, 2006a) and music transcription (Schmidt and Mørup, 2006b). We here extend this approach to the analysis of the log-frequency spectrograms of a multichannel recording. The model proposed forms a non-negative tensor factor 2-D deconvolution (NTF2D) based on the parallel factor (PARAFAC) model. Two algorithms are given for NTF2D; one based on least squares the other on Kullback-Leibler divergence minimization. Both algorithms are extended to give sparse decompositions. The algorithms are demonstrated to successfully identify the components of both artificially generated as well as real stereo music.

Keywords: Non-negative Matrix Factorization (NMF), PARAFAC, Sparse Coding, SNMF2D, SNTF2D.

1. Introduction

We recently proposed the non-negative matrix factor 2-D deconvolution (NMF2D) model extending the regular non-negative matrix factorization (NMF) model to a 2-dimensional convolution of the non negative matrices $\mathbf{W}^\tau \in \mathbb{R}^{F\times D}$ and $\mathbf{H}^\phi \in \mathbb{R}^{T\times D}$, that is

$$\mathbf{V}_{f,t} \approx \mathbf{\Lambda}_{f,t} = \sum_{\tau,\phi} \mathbf{W}^\tau_{f-\phi,d}\mathbf{H}^\phi_{t-\tau,d}, \quad \text{where} \quad \tau \in \{0,1,\ldots,\Upsilon\} \text{ and } \phi \in \{0,1,\ldots,\Phi\}. \tag{1}$$

The model can be considered an extension of the non-negative matrix factor deconvolution (NMFD) independently proposed by Smaragdis (2004), Eggert et al. (2004) and FitzGerald et al. (2005b) corresponding to either $\phi = \{0\}$ or $\tau = \{0\}$.

The NMF2D model has proven useful in the analysis of the log-frequency spectrogram $\mathbf{V}$ of a signal of mixed musical instruments. Here the ϕ-th notes played by the instruments are captured by $\mathbf{H}^\phi$ while the frequency structure, i.e. the harmonics of the instruments at time lag τ are captured by $\mathbf{W}^\tau$. As a result, the change in pitch of an instrument corresponds to a vertical shift in the log-frequency spectrogram captured by the ϕ shifts while each instruments is assumed to have a fixed temporal frequency structure captured by the τ shifts, see also Schmidt and Mørup (2006b) for details.

Often music is not solely represented by the spectrogram of one single channel but by several channels, i.e. by several microphones or the two stereo signals in stereo recordings. The NMF2D model can only handle such data by either analyzing each channels separately or unfolding the extra channel modality onto one of the existing modalities to form an analyzable matrix. However, unfolding can, to some extent, hamper interpretation but, more importantly, potentially dismiss

modality specific information by mixing information in a given modality with the more or less arbitrarily chosen modalities that it has been unfolded to. Rather than unfolding, we will extend the NMF2D model to handle spectral data of more than one channel. The model proposed turns out to be a 2-D convolutive PARAFAC model, i.e. a non-negative tensor factor 2-D deconvolution (NTF2D).

The paper is structured as follows: First, the NTF2D model is introduced and two algorithms ensured to converge are given. These algorithms are extended to form sparse decompositions in order to handle ambiguity between the factors in $\mathbf{W}$ and $\mathbf{H}$ and to improve interpretability of the components. This is followed by a demonstration of the ability of the algorithms to identify the components of synthetic data. Finally, we demonstrate how the algorithms also correctly identify the components of real stereo music. The algorithms can be downloaded from *www2.imm.dtu.dk/pubdb/views/edoc_download.php/4652/zip/imm4652.zip*. To illustrate the NTF2D algorithm we presently put it in the framework of music analysis. However, the algorithm is in general useful when a fixed translated 2-D structure is present in the data.

2. Method

Consider the signal $\mathbf{V} \in \mathbb{R}^{C \times F \times T}$ being a three way array where $\mathbf{V}_{c,f,t}$ denotes the spectral coefficients at channel c at frequency f and time t. In the following we will assume that the frequency harmonics of each instrument given in $\mathbf{W}^{\tau}$ and each note played given in $\mathbf{H}^{\phi}$ is the same regardless of the channels analyzed. We will further assume that each channel has an instantaneous linear mix of these signatures given by $\mathbf{D} \in \mathbb{R}^{C \times D}$, i.e. we will for convenience assume all frequencies of a given instrument to be mixed with same strength. From these assumptions the log-frequency spectrogram can be approximated as

$$\mathbf{V}_{c,f,t} \approx \mathbf{\Lambda}_{c,f,t} = \sum_{d,\tau,\phi} \mathbf{D}_{c,d} \mathbf{W}^{\tau}_{f-\phi,d} \mathbf{H}^{\phi}_{t-\tau,d}. \tag{2}$$

Consequently, $\mathbf{H}^{\phi}_{t,d}$ represents the degree in which the ϕ-th note is present at time t in instrument d. $\mathbf{W}^{\tau}_{f,d}$ is the harmonical structure at lag τ at frequency f of the d-th instrument and $\mathbf{D}_{c,d}$ is the degree in which instrument d is present in channel c. Notice, if $\tau = \{0\}$ and $\phi = \{0\}$ this model becomes the conventional PARAFAC model (Welling and Weber, 2001) as proposed for the analysis of sound signals by Parry and Essa (2006) and FitzGerald et al. (2005a) whereas the single convolutive model recently proposed by FitzGerald and Coyle (2006) corresponds to $\tau = \{0\}$. Consequently, the NTF2D model forms a PARAFAC model that is convolutive in two of the three modalities, i.e. convolutive in the time and frequency domain. While the instantaneous mixing into the channels in general is a rough assumption it becomes reasonable when considering the time-frequency representation. Here each time-frequency point in the spectrogram is an average of the frequency activity over the time window used for the representation. Delays present between the channels are, in general, far less than the extend of this time frame.

Define the Khatri-Rao product $\mathbf{A} \odot \mathbf{B} = [\mathbf{A}_1 \otimes \mathbf{B}_1 \ ... \ \mathbf{A}_F \otimes \mathbf{B}_F]$ and the matricizing operation, i.e. $\mathbf{V}_{(1)} = \mathbf{V}^{C \times F \cdot T}$, $\mathbf{V}_{(2)} = \mathbf{V}^{F \times C \cdot T}$ and $\mathbf{V}_{(3)} = \mathbf{V}^{T \times C \cdot F}$. Let further $\overset{\downarrow q}{\mathbf{A}}$ and $\overset{\uparrow p}{\mathbf{A}}$ denotes the upward and downward shift operator on the matrix $\mathbf{A}$ given by shifting and zero padding the rows of $\mathbf{A}$, i.e.:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}, \quad \overset{\downarrow 2}{\mathbf{A}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 2 & 3 \end{pmatrix}, \quad \overset{\uparrow 1}{\mathbf{A}} = \begin{pmatrix} 4 & 5 & 6 \\ 7 & 8 & 9 \\ 0 & 0 & 0 \end{pmatrix}$$

The NTF2D model can then be formulated as the following three equivalent approximations:

$$\mathbf{V}_{(1)} \approx \mathbf{\Lambda}_{(1)} = \mathbf{D}(\sum_{\tau,\phi} \overset{\downarrow\tau}{\mathbf{H}}{}^{\phi} \odot \overset{\downarrow\phi}{\mathbf{W}}{}^{\tau})^T, \tag{3}$$

$$\mathbf{V}_{(2)} \approx \mathbf{\Lambda}_{(2)} = \sum_{\tau,\phi} \overset{\downarrow\phi}{\mathbf{W}}{}^{\tau} \ (\overset{\downarrow\tau}{\mathbf{H}}{}^{\phi} \odot \mathbf{D})^T, \tag{4}$$

$$\mathbf{V}_{(3)} \approx \mathbf{\Lambda}_{(3)} = \sum_{\tau,\phi} \overset{\downarrow\tau}{\mathbf{H}}{}^{\phi} \ (\overset{\downarrow\phi}{\mathbf{W}}{}^{\tau} \odot \mathbf{D})^T. \tag{5}$$

We will give two algorithms to estimate $\mathbf{D}$, $\mathbf{W}$ and $\mathbf{H}$ —one based on least squares (LS) and the other on Kullbach-Leibler (KL) divergence minimization, forming the following three equivalent minimizations

$$C_{LS} = \frac{1}{2}||\mathbf{V}_{(1)} - \mathbf{\Lambda}_{(1)}||_F^2 = \frac{1}{2}||\mathbf{V}_{(2)} - \mathbf{\Lambda}_{(2)}||_F^2 = \frac{1}{2}||\mathbf{V}_{(3)} - \mathbf{\Lambda}_{(3)}||_F^2 \tag{6}$$

$$\text{where} \quad ||\mathbf{A}^{I\times J} - \mathbf{B}^{I\times J}||_F^2 = \sum_{i,j}(\mathbf{A}_{i,j} - \mathbf{B}_{i,j})^2, \tag{7}$$

$$C_{KL} = D(\mathbf{V}_{(1)}|\mathbf{\Lambda}_{(1)}) = D(\mathbf{V}_{(2)}|\mathbf{\Lambda}_{(2)}) = D(\mathbf{V}_{(3)}|\mathbf{\Lambda}_{(3)}) \tag{8}$$

$$\text{where} \quad D(\mathbf{A}^{I\times J}|\mathbf{B}^{I\times J}) = \sum_{i,j} \mathbf{A}_{i,j} \log \frac{\mathbf{A}_{i,j}}{\mathbf{B}_{i,j}} - \mathbf{A}_{i,j} + \mathbf{B}_{i,j}. \tag{9}$$

However, 2-D convolutive models suffer from ambiguity between $\mathbf{W}$ and $\mathbf{H}$ (Mørup and Schmidt, 2006). Consequently, the harmonics of the components can be captured in both $\mathbf{H}$ and $\mathbf{W}$. Furthermore, when including many ϕ and τ shifts the number of free parameters of the model can become larger than the number of data points available, i.e. the representation can become overcomplete. As a result, constraints in the form of sparseness have proven useful (Mørup and Schmidt, 2006). Consequently, we impose the sparseness cost $C_{Sparse}(\mathbf{H})$ to restrict $\mathbf{H}$ to be sparse in order for the harmonic frequency structure of the instruments to be solely present in $\mathbf{W}$. $C_{Sparse}(\mathbf{H})$ can be any function with positive derivative (Mørup and Schmidt, 2006), we will in the present analysis use the $L_1 - norm$ since this corresponds to a probability density which is highly peaked at zero and have heavy tails hence form a sparse representation (Hoyer, 2002):

$$C_{Sparse}(\mathbf{H}) = \beta||\mathbf{H}||_1 = \beta \sum_{j,\phi,d} \mathbf{H}_{t,d}^{\phi} \tag{10}$$

Adding this penalty to the existing cost functions, β becomes the weight of sparseness to the reconstruction error. This sparseness constraint is, however, easily minimized letting the components in $\mathbf{H}$ go to zero while letting the corresponding components in $\mathbf{W}$ and $\mathbf{D}$ go to infinity. Consequently, we impose extra constraints of unit Frobenius-norm to the components in $\mathbf{W}$ and $\mathbf{D}$, i.e. $\|\mathbf{W}_d\|_F = 1$, $\|\mathbf{D}_d\|_F = 1$ where $\mathbf{W}_d = \mathbf{W}_{:,d}^{:}$, $\mathbf{D}_d = \mathbf{D}_{:,d}$ and : is the MATLAB shorthand notation denoting all elements of the given modality. As proposed for conventional NMF by Eggert and Korner (2004) we reformulate the reconstruction to be invariant of this normalization:

$$\tilde{\mathbf{\Lambda}}_{c,f,t} = \sum_{\tau,\phi,d} \frac{\mathbf{W}_{f-\phi,d}^{\tau}}{\|\mathbf{W}_d\|_F} \frac{\mathbf{D}_{c,d}}{\|\mathbf{D}_d\|_F} \mathbf{H}_{t-\tau,d}^{\phi}. \tag{11}$$

Consequently, the p-th of the three equivalent cost functions in Equation 6 and 8 using this reconstruction also become invariant of the normalization:

$$C_{SparseLS} = \frac{1}{2}||\mathbf{V}_{(p)} - \tilde{\mathbf{\Lambda}}_{(p)}||_F^2 + C_{Sparse}(\mathbf{H}) \tag{12}$$

$$C_{SparseKL} = D(\mathbf{V}_{(p)}|\tilde{\mathbf{\Lambda}}_{(p)}) + C_{Sparse}(\mathbf{H}). \tag{13}$$

The cost functions given in Equations (6) and (8) and including sparseness in Equation 12 and 13 were all differentiated with respect to given elements in $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{D}$. The parameters were then updated using a gradient based search with a step size giving multiplicative updates (see Mørup and Schmidt (2006) as well as Lee and Seung (2000) for details of this approach). The algorithms are given in Table 1 and 2. Here $A \bullet B$ denotes element-wise multiplication and $\frac{A}{B}$ denotes element-wise division. Furthermore, $diag(\mathbf{a})$ is a square matrix containing the elements in the vector $\mathbf{a}$ along the diagonal while $\mathbf{1}$ is a matrix of ones.

NTF2D/SNTF2D Least squares

1. Initialize $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{D}$ randomly.
2. $\mathbf{\Lambda}_{(1)} = \mathbf{D}(\sum_\tau \sum_\phi \overset{\downarrow\tau}{\mathbf{H}}^\phi \odot \overset{\downarrow\phi}{\mathbf{W}}^\tau)^T$
3. $\mathbf{D} \leftarrow \mathbf{D} \bullet \frac{\mathbf{V}_{(1)}\mathbf{Z} + \mathbf{D}diag(\mathbf{1}((\mathbf{D}\mathbf{Z}^\mathbf{T}\mathbf{Z}) \bullet \mathbf{D}))}{\mathbf{D}\mathbf{Z}^\mathbf{T}\mathbf{Z} + \mathbf{D}diag(\mathbf{1}((\mathbf{V}_{(1)}\mathbf{Z}) \bullet \mathbf{D}))}$ where $\mathbf{Z} = (\sum_\tau \sum_\phi \overset{\downarrow\phi}{\mathbf{W}}^\tau \odot \overset{\downarrow\tau}{\mathbf{H}}^\phi)$
4. $\mathbf{D}^\tau_{k,d} = \frac{\mathbf{D}^\tau_{k,d}}{\|\mathbf{D}_d\|_F}$, $\mathbf{\Lambda}_{(2)} = \sum_\tau \sum_\phi \overset{\downarrow\phi}{\mathbf{W}}^\tau (\overset{\downarrow\tau}{\mathbf{H}}^\phi \odot \mathbf{D})^T$
5. $\mathbf{W}^\tau \leftarrow \mathbf{W}^\tau \bullet \frac{\sum_\phi \overset{\uparrow\phi}{\mathbf{V}}_{(2)} (\overset{\downarrow\tau}{\mathbf{H}}^\phi \odot \mathbf{D}) + \mathbf{W}^\tau diag(\mathbf{1} \sum_\tau (\overset{\uparrow\phi}{\mathbf{\Lambda}}_{(2)} (\overset{\downarrow\tau}{\mathbf{H}}^\phi \odot \mathbf{D})) \bullet \mathbf{W}^\tau)}{\sum_\phi \overset{\uparrow\phi}{\mathbf{\Lambda}}_{(2)} (\overset{\downarrow\tau}{\mathbf{H}}^\phi \odot \mathbf{D}) + \mathbf{W}^\tau diag(\mathbf{1} \sum_\tau (\overset{\uparrow\phi}{\mathbf{V}}_{(2)} (\overset{\downarrow\tau}{\mathbf{H}}^\phi \odot \mathbf{D})) \bullet \mathbf{W}^\tau)}$
6. $\mathbf{W}^\tau_{i,d} = \frac{\mathbf{W}^\tau_{i,d}}{\|\mathbf{W}_d\|_F}$, $\mathbf{\Lambda}_{(3)} = \sum_\tau \sum_\phi \overset{\downarrow\tau}{\mathbf{H}}^\phi (\overset{\downarrow\phi}{\mathbf{W}}^\tau \odot \mathbf{D})^T$
7. $\mathbf{H}^\phi \leftarrow \mathbf{H}^\phi \bullet \frac{\sum_\tau \overset{\uparrow\tau}{\mathbf{V}}_{(3)} (\overset{\downarrow\phi}{\mathbf{W}}^\tau \odot \mathbf{D})}{\sum_\tau \overset{\uparrow\tau}{\mathbf{\Lambda}}_{(3)} (\overset{\downarrow\phi}{\mathbf{W}}^\tau \odot \mathbf{D}) + \beta \frac{\partial C_{Sparse}(\mathbf{H})}{\partial \mathbf{H}^\phi}}$
8. Repeat from step 2 until convergence.

Table 1: Algorithm for NTF2D/SNTF2D by Least Squares. The algorithm is given for SNTF2D but by omitting everything in gray the corresponding algorithm without sparseness constraint, i.e NTF2D is achieved. The convergence was in the present analysis set to a maximum of 250 iterations or when the relative change in the cost function was less than 10^{-6}

According to Equation (3) the updates can be transformed into the framework of regular matrix analysis. Consequently, the convergence of $\mathbf{W}^\tau$ is given by replacing $\mathbf{H}^\phi$ with $\mathbf{Z} = \mathbf{H}^\phi \odot \mathbf{D}$ and $\mathbf{V}$ with $\mathbf{V}_{(2)}$ in the proof of the $\mathbf{W}^\tau$ update given by Mørup and Schmidt (2006) while in the proof of the convergence of $\mathbf{H}^\phi$ replacing $\mathbf{W}^\tau$ with $\mathbf{Z} = \mathbf{W}^\tau \odot \mathbf{D}$ and $\mathbf{V}$ with $\mathbf{V}_{(3)}$. The convergence of the $\mathbf{D}$ update follows straight from the proof of the regular NMF updates given by Lee and Seung (2000): Noticing $\mathbf{V}_{(1)} \approx \mathbf{\Lambda}_{(1)} = \mathbf{D}(\sum_{\tau,\phi} \overset{\downarrow\tau}{\mathbf{H}}^\phi \odot \overset{\downarrow\phi}{\mathbf{W}}^\tau)^T$, and defining $\mathbf{Z} = (\sum_{\tau,\phi} \overset{\downarrow\tau}{\mathbf{H}}^\phi \odot \overset{\downarrow\phi}{\mathbf{W}}^\tau)^T$, this becomes the conventional NMF, i.e. $\mathbf{\Lambda}_{(1)} = \mathbf{D}\mathbf{Z}$. While the convergence of the updates including sparsity for conventional NMF (Eggert and Korner, 2004) and the SNMF2D (Mørup and Schmidt, 2006) has not been proved, they were all conjectured convergent. Although extensively analyzed, we never experienced divergence in any of the updates of $\mathbf{H}^\phi$ and $\mathbf{W}^\tau$ nor $\mathbf{D}$ in the two SNTF2D algorithms. Consequently, we conjecture that also the algorithms including sparsity are convergent.

NTF2D/SNTF2D KL-divergence

1. Initialize $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{D}$ randomly.
2. $\mathbf{\Lambda}_{(1)} = \mathbf{D}(\sum_\tau \sum_\phi \overset{\downarrow\tau}{\mathbf{H}}^{\phi} \odot \overset{\downarrow\phi}{\mathbf{W}}^{\tau})^T$
3. $\mathbf{D} \leftarrow \mathbf{D} \bullet \frac{\frac{\mathbf{V}_{(1)}}{\mathbf{DZ^T}} + \mathbf{D}diag(\mathbf{1} \cdot ((\mathbf{1Z}) \bullet D))}{\mathbf{1Z} + \mathbf{D}diag(\mathbf{1} \cdot (\frac{\mathbf{V}_{(1)}}{\mathbf{DZ}^T}\mathbf{Z} \bullet \mathbf{D}))}$ where $\mathbf{Z} = (\sum_\tau \sum_\phi \overset{\downarrow\phi}{\mathbf{W}}^{\tau} \odot \overset{\downarrow\tau}{\mathbf{H}}^{\phi})$
4. $\mathbf{D}^\tau_{k,d} = \frac{\mathbf{D}^\tau_{k,d}}{\|\mathbf{D}_d\|_F}$, $\mathbf{\Lambda}_{(2)} = \sum_\tau \sum_\phi \overset{\downarrow\phi}{\mathbf{W}}^{\tau} (\overset{\downarrow\tau}{\mathbf{H}}^{\phi} \odot \mathbf{D})^T$
5. $\mathbf{W}^\tau \leftarrow \mathbf{W}^\tau \bullet \frac{\sum_\phi \left(\frac{\overset{\uparrow\phi}{\mathbf{V}_{(2)}}}{\mathbf{\Lambda}_{(2)}}\right)(\overset{\downarrow\tau}{\mathbf{H}}^{\phi} \odot \mathbf{D}) + \mathbf{W}^\tau diag(\mathbf{1} \cdot \sum_\tau (\mathbf{1}(\overset{\downarrow\tau}{\mathbf{H}}^{\phi} \odot \mathbf{D})) \bullet \mathbf{W}^\tau)}{\sum_\phi \mathbf{1}\overset{\downarrow\tau}{\mathbf{H}}^{\phi} + \mathbf{W}^\tau diag(\mathbf{1} \sum_\tau (\left(\frac{\overset{\uparrow\phi}{\mathbf{V}_{(2)}}}{\mathbf{\Lambda}_{(2)}}\right)(\overset{\downarrow\tau}{\mathbf{H}}^{\phi} \odot \mathbf{D}^{(i)}) \bullet \mathbf{W}^\tau)}$
6. $\mathbf{W}^\tau_{i,d} = \frac{\mathbf{W}^\tau_{i,d}}{\|\mathbf{W}_d\|_F}$, $\mathbf{\Lambda}_{(3)} = \sum_\tau \sum_\phi \overset{\downarrow\tau}{\mathbf{H}}^{\phi} (\overset{\downarrow\phi}{\mathbf{W}}^{\tau} \odot \mathbf{D})^T$
7. $\mathbf{H}^\phi \leftarrow \mathbf{H}^\phi \bullet \frac{\sum_\tau \left(\frac{\overset{\uparrow\tau}{\mathbf{V}_{(3)}}}{\mathbf{\Lambda}_{(3)}}\right)(\overset{\downarrow\phi}{\mathbf{W}}^{\tau} \odot \mathbf{D})}{\sum_\tau \mathbf{1}(\overset{\downarrow\phi}{\mathbf{W}}^{\tau} \odot \mathbf{D}) + \beta \frac{\partial C_{Sparse}(\mathbf{H})}{\partial \mathbf{H}^\phi}}$
8. Repeat from step 2 until convergence.

Table 2: Algorithm for NTF2D/SNTF2D by KL-divergence minimization. The algorithm is given for SNTF2D but by omitting everything in gray the corresponding algorithm without sparseness constraint, i.e. NTF2D, is achieved.

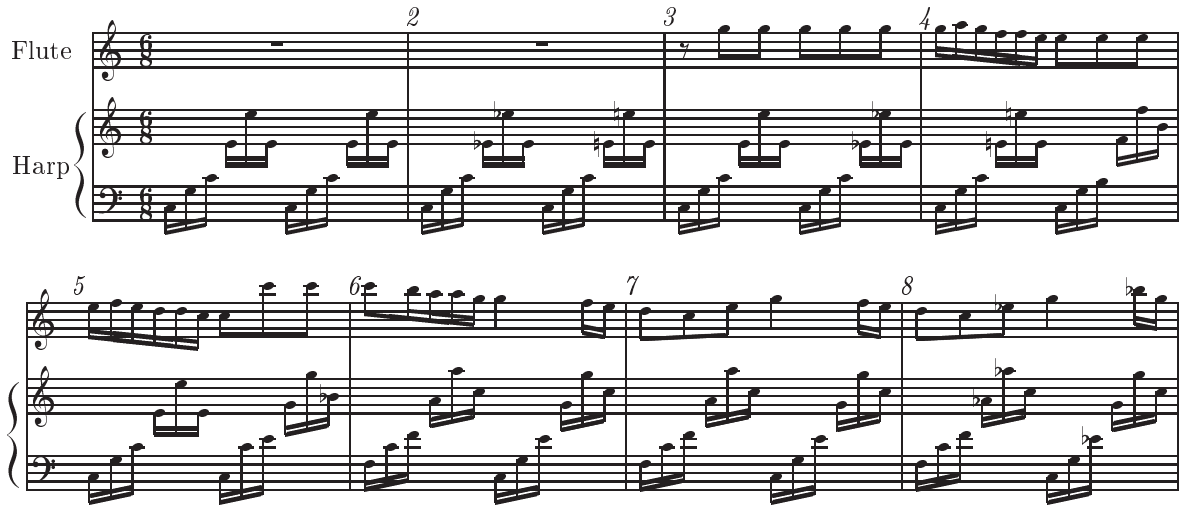

Figure 1: The first eight bars of "The Fog is Lifting" by Carl Nielsen

3. Results

The algorithms were tested on an artificial generated data set simulating a harp and flute playing "The Fog is Lifting" by Carl Nielsen, see Figure 2 (here for illustrative purposes disregarding correct fundamental frequency of the two instruments to reduce number of ϕ shifts required to cover all notes). The score of the music can be seen on Figure 1. The data set was generated having 175 frequency bins covering from 50 Hz to 8kHz corresponding to 24 bins per octave. The distance between each time point was one third of a 16th note. Consequently, $\mathbf{W}$ had seven lags, i.e. $\tau = \{0, 1, 2, \ldots, 6\}$ corresponding to a time signature covering the duration of slightly more than an 8th note. The scores were represented in $\mathbf{H}$ where $\phi = \{0, 1, 2, \ldots, 72\}$ thereby covering 3 octaves. The instruments were mixed in each channel such that the harp was dominant in channel 1 whereas the flute was dominant in channel 2. Notice, the position of the scores in $\mathbf{H}$ can be compensated by a counter change in the pitch of the frequency signature in $\mathbf{W}$ while the onset of the frequency structure in $\mathbf{W}$ can be compensated by a change in onset of the score in $\mathbf{H}$ (Schmidt and Mørup, 2006b; Mørup and Schmidt, 2006). Consequently, in the following the geometric mean of the notes in $\mathbf{H}$ was set to be at the center of all the ϕ shifts present in $\mathbf{H}$ while the geometric mean of the frequency structure was set to be at the center of all the τ shifts.

The algorithms were also tested on a real recording of "The Fog is Lifting" by Carl Nielsen (Jensen and Johansen). We sampled the music at 44.1 kHz and analyzed it by the short time Fourier transform with a 8192 point Hanning windowed FFT with 50% overlap. This gave us 283 FFT slices. We grouped the spectrogram bins into 210 logarithmically spaced frequency bins in the range of 50 Hz to 22 kHz with 24 bins per octave, which corresponds to twice the resolution of the equal tempered musical scale. To cover the duration of an eight note played we chose τ to be $\tau = \{0, 1, 2, ..., 9\}$ while $\phi = \{0, 1, 2, ..., 82\}$ covering 3.5 octaves, i.e. slightly more than the range of all the notes played. The results obtained analyzing the absolute values of these spectrograms is shown in Figure 3.

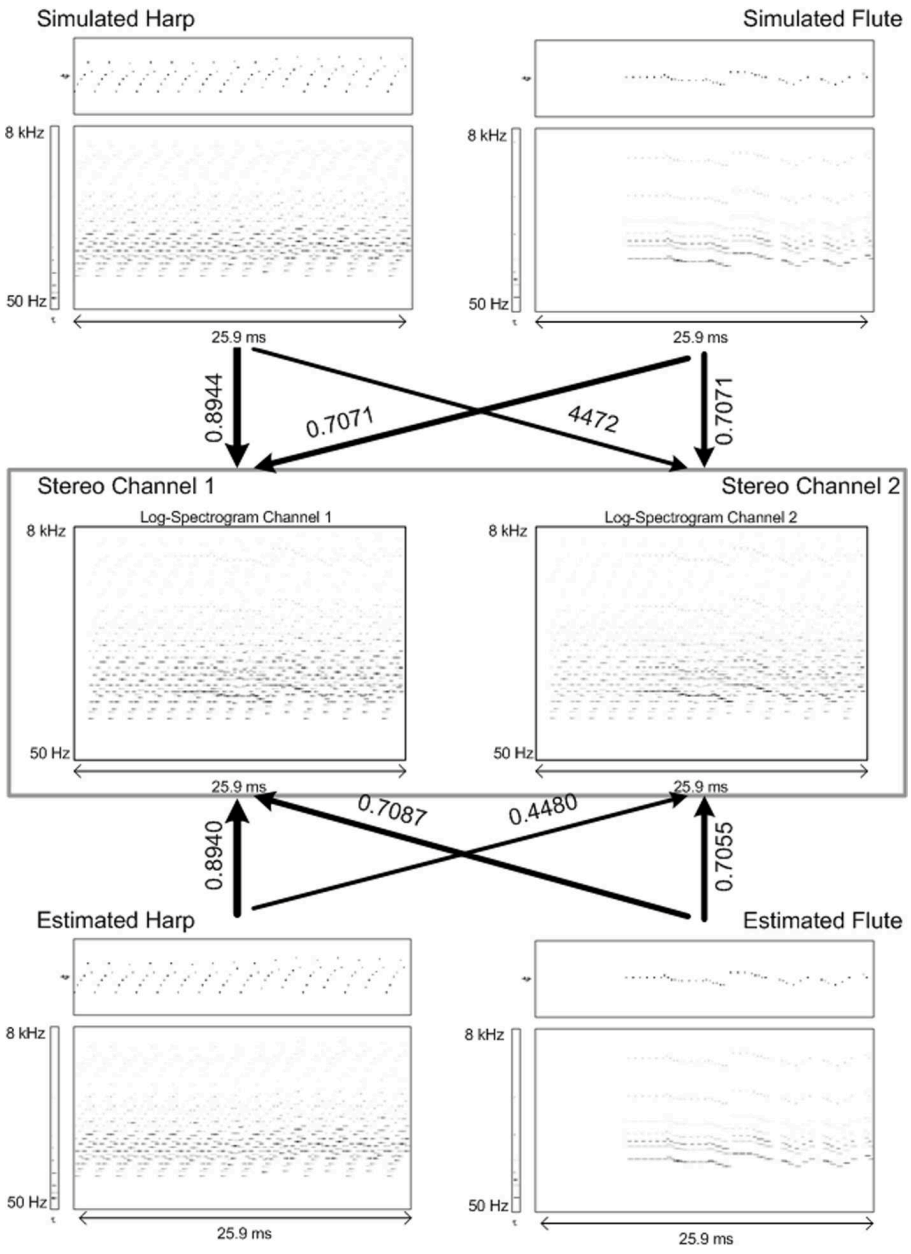

Figure 2: **Top figures:** Left panel; the artificially generated signature of a harp **W** along with the scores played given by **H** and the corresponding time-frequency signature arising from convolving **W** and **H**. The mixing in the two recording channels is given by the arrows. Right panel; the corresponding signatures for the flute. **Middle figures:** Time-frequency plot of the two channels generated from mixing the time-frequency signatures of both instruments. **Bottom figures:** The estimated signatures of the harp and flute found by the SNTF2D algorithm, here shown using LS-minimization (the KL algorithm gave similar results). The algorithms recovered more than 99% of the variance in the original data. From the decomposition it can be clearly seen that the scores **H** are perfectly recovered as well as the mixing in the channels **D** and the harmonic structure of each instrument **W**. To resolve ambiguity between **W** and **H**, β was set to 0.1 while the data was in the range $[0; 0.66]$.

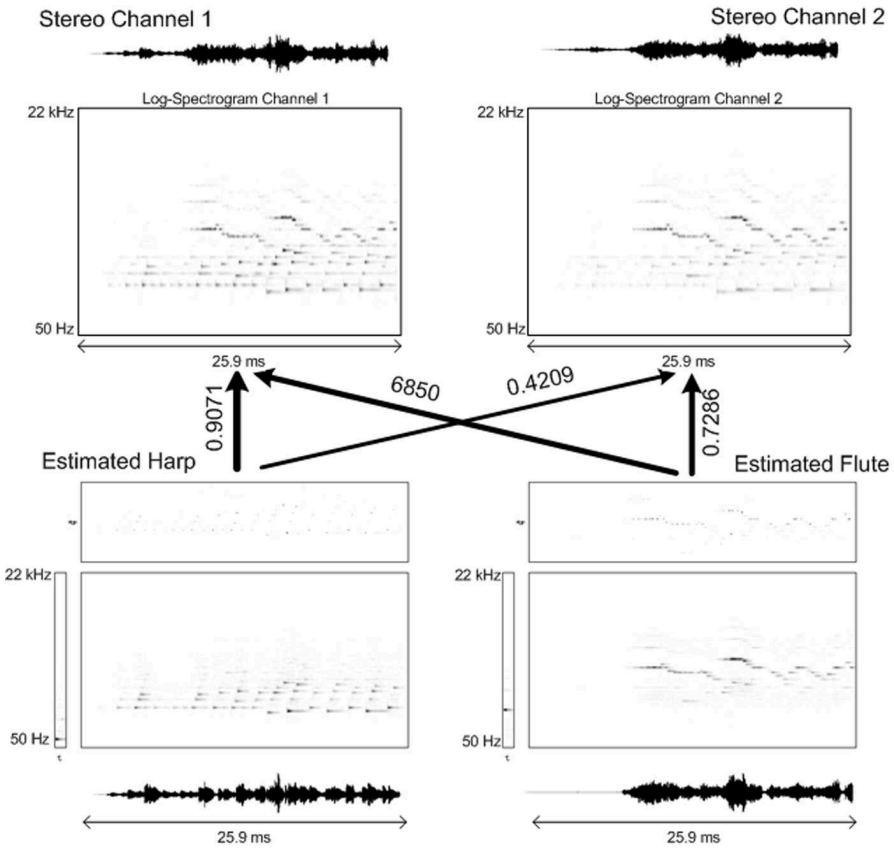

Figure 3: SNTF2D analysis of real stereo music here shown for LS-minimization (again the KL-minimization gave similar results). **Top images**: The log-frequency spectrogram and raw signal of each of the two stereo channels. **Bottom images**: The components found when decomposing the spectrogram. The first component mainly captures the harp while the second component which the flute and harp components have been mixed in the two stereo channels. Underneath the components are given the raw instrument signals found by spectral masking. The two components accounted for 86.9 % of the variance in the two log-spectrograms. To resolve ambiguity between $\mathbf{W}$ and $\mathbf{H}$ $\beta = 50$ while the data was in the range $[0; 222]$.

4. Discussion

The developed algorithms successfully captured the components of the artificially generated log-spectrogram of music. For ease of interpretability of the components the results have only been shown for the SNTF2D algorithms.

From the stereo music of the true recording of "The Fog is Lifting" (Jensen and Johansen) it was seen that the method separated well the spectrograms into two components corresponding mainly to the harp and flute respectively. From the signatures found the scores of each instrument could be read from $\mathbf{H}$ and the signature of the instrument from $\mathbf{W}$. Consequently, the SNTF2D algorithms work well for music transcription performing better than the single channel SNMF2D analysis (Schmidt and Mørup, 2006b) as the information from several channels are incorporated while including only a few extra model parameters. Furthermore, the algorithms can be used for sound separation as indicated by the reconstructed signals found by using the time-frequency signatures of each estimated instrument to perform spectral masking in the channel the instrument was the most present. Rather than evaluating the statistical properties of the raw time signals to separate the sources as could have been done by an ICA algorithm (Hyvarinen et al., 2001) or convolutive ICA algorithm (Parra et al., 1998), the SNTF2D uses prior knowledge namely the presence of harmonical structures in the log spectrogram to search for systematic patterns through the spectrograms. Consequently, the SNTF2D better models the data when the signals indeed can be assumed formed by such patterns. It is our strong belief that the SNTF2D algorithms will be useful for the analysis of other sound signals such as speech and noise when such patterns are present.

From the mixing of the components found by the model the degree in which each component is present in the channels can be estimated. Although, it is not in general correct to assume the mixing to be constant over frequencies, the linear mixing presently used is easy to implement and we believe it to be a reasonable approximation. Furthermore, the mixing $\mathbf{D}$ of the sources to the channels found by the model can be used to estimate the location of the sources when combined with information of the position at which each channels recorded the sounds. As for the SNMF2D model the assumptions of same harmonic structure across pitch for a given instrument is a rough assumption (Schmidt and Mørup, 2006b), however within a limited range this is likely to hold. Nevertheless, this is probably the main reason why the harp and flute wasn't perfectly recovered from the real music by the algorithms.

The algorithms developed are an extension of the PARAFAC model to include double convolutive mixtures. Consequently, the algorithms devised here gives both a single convolutive mixture, i.e. either $\phi = \{0\}$ or $\tau = \{0\}$ as proposed by FitzGerald and Coyle (2006) and a double convolutive mixture, i.e. $\phi \neq \{0\}$, $\tau \neq \{0\}$. These algorithms are all presently proved to converge when no sparsity is imposed. Notice, that if both ϕ and τ are zero the SNTF2D algorithms becomes a sparse PARAFAC model. Furthermore, the developed model can easily be extended to include more modalities and also to incorporate convolutive mixtures in these extra modalities, i.e. a model that is 3-D convolutive, 4-D convolutive etc. Consequently, the framework used here is generalizable to a wide range of higher order data analysis. Furthermore, the 2-D deconvolution represents the data as fixed translation invariant 2-D structures. Consequently, the algorithms proposed is useful in general when data can be represented as such structures.

Let Υ be the number of τ shifts, Φ be the number of ϕ shifts and D the number of components. The free parameters in the double convolutive model is given by $(C + F\Upsilon + T\Phi)D$ while the amount of data points is CFT. However, the data could have been analyzed by concatenating the time-frequency signatures of each channels using SNMF2D. This would have given $(F\Upsilon + CT\Phi)D >> (C + F\Upsilon + T\Phi + K)D$ free parameters. Consequently, the NTF2D is likely to be less overcomplete when operating with many lags of τ and ϕ. Furthermore, the PARAFAC model is, contrary to factor analysis, in general unique (Kruskal, 1977; Sidiropoulos and Bro, 2000). Consequently, having the analysis in the framework of the PARAFAC model improves the uniqueness properties of the components found. This is achieved through a more restricted model here assuming the time-

frequency signatures of the underlying components to be instantaneously, linearly mixed in the channels.

The above algorithms were developed under non-negativity constraints. This was the case since the amplitude of the spectrogram is positive and the components assumed additive, i.e. no cancellation of components within the spectrogram. Although algorithms could be developed to implement other assumptions the algorithms developed here are fast and easy to implement. One drawback of the sparse algorithms is that the choice of sparseness penalty β is not obvious while still influencing the solutions found.

5. Conclusion

We developed two algorithms for NTF2D with non-negative constraints and showed how they were useful in the analysis of multi-channel sound signals. While the algorithms without sparseness constraints were proven to converge we conjectured the sparse algorithms to converge. The algorithms were able to both correctly identify the components of artificially generated data as well as real music. MATLAB implementations of the algorithms can be download from ($www2.imm.dtu.dk/pubdb/views/edoc\_download.php/4652/zip/imm4652.zip$).

References

J. Eggert and E. Korner. Sparse coding and nmf. In *Neural Networks*, volume 4, pages 2529–2533, 2004.

J. Eggert, H. Wersing, and E. Korner. Transformation-invariant representation and nmf. In *Neural Networks*, volume 4, pages 2535– 2539, 2004.

D. FitzGerald and E. Coyle. Sound source separation using shifted non-negative tensor factorisation. In *ICASSP2006*, 2006.

D. FitzGerald, M. Cranitch, and E. Coyle. Non-negative tensor factorisation for sound source separation. In *proceedings of Irish Signals and Systems Conference*, pages 8–12, 2005a.

Derry FitzGerald, Matt Granitch, and Eugene Coyle. Shifted non-negative matrix factorisation for sound source separation. In *Proceedings of the IEEE conference on Statistics in Signal Processing*, 2005b.

P.O. Hoyer. Non-negative sparse coding. *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, pages 557–565, 2002.

A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons., 2001.

Thomas Jensen and Benedikte Johansen. Tåken letter (the fog is lifting) for flute and harp composed by carl nielsen. Naxos.

J.B Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl.*, 18:95–138, 1977.

Daniel D Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2000.

M. Mørup and M. N. Schmidt. Sparse nonnegative matrix factor 2-D deconvolution. Technical report, Institute for Mathematical Modelling, Technical University of Denmark, 2006.

L. Parra, C. Spence, and B. De Vries. Convolutive blind source separation based on multiple decorrelation. *Neural Networks for Signal Processing VIII, 1998. Proceedings of the 1998 IEEE Signal Processing Society Workshop*, pages 23 –32, 1998.

R. Parry, Mitchell and Irfan Essa. Estimating the spatial position of spectral components in audio. In *proceedings ICA2006*, pages 666– 673, 2006.

M. N. Schmidt and M. Mørup. Nonnegative matrix factor 2-D deconvolution for blind single channel source separation. In *ICA2006*, pages 700–707, 2006a.

M.N. Schmidt and M. Mørup. Sparse non-negative matrix factor 2-d deconvolution for automatic transcription of polyphonic music. Technical report, Institute for Mathematical Modelling, Technical University of Denmark, 2006b.

Nicholas D. Sidiropoulos and Rasmus Bro. On the uniqueness of multilinear decomposition of n-way arrays. *Journal of Chemometrics*, 14:229–239, 2000.

Paris Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sourses from monophonic inputs. *International Symposium on Independent Component Analysis and Blind Source Separation (ICA)*, 3195:494, sep 2004.

Max Welling and Markus Weber. Positive tensor factorization. *Pattern Recogn. Lett.*, 22(12):1255–1261, 2001. ISSN 0167-8655.