CUSTOM SIZED NON-NEGATIVE MATRIX FACTOR DECONVOLUTION FOR SOUND SOURCE SEPARATION

Julian M. Becker and Christian Rohlfing

Institut für Nachrichtentechnik
RWTH Aachen University
D-52056 Aachen, Germany

ABSTRACT

Non-negative Matrix Factorization (NMF) is frequently used for audio source separation. One downside of the NMF is, that it is not able to capture temporal structure of sound events. NMF splits these events into different components. In this paper we present an extension to NMF, which is capable of representing sound events with temporal structure in only one component. We also present an algorithm, which uses this method efficiently. We show that this algorithm leads to a more compact factorization (i.e. with less components) compared to NMF, without losing separation quality.

Index Terms— NMF, NMFD, Sound Source Separation

1. INTRODUCTION

NMF was introduced by Paatero [1], but only became popular after Lee and Seung published efficient algorithms [2]. NMF is frequently used in audio source separation, e.g. [3, 4], because it is able to factorize audio signals into a specified number of components which correspond to sound events. These events can be assigned to the original sources by clustering. For this step, it is necessary to chose a suitable number of components. If it is chosen too high, the clustering gets more difficult. One way of determining this number for NMF is e.g. Automatic Relevance Determination (ARD) [5].

Some sound events, especially percussive ones, have a temporal structure, that cannot be captured by NMF. These events are partitioned into different components, making the clustering more difficult. Smaragdis proposed an extension to NMF [6], Non-negative Matrix Factor Deconvolution (NMFD), which is able to represent events with temporal structure in only one component. In this paper, we show why NMFD fails in separating sources for some kinds of mixtures and present an extension to NMFD, which is more robust. We show, how this method can be used together with NMF with ARD to perform a separation of mixtures with less components.

The paper is structured as follows: In Section 2, we describe the NMFD and explain, how it can be used for audio source separation. In Section 3, we propose our extension to NMFD. In Section 4 we describe how this method can be used in a source separation algorithm and show, that it results in a factorization with a reduced number of components. Finally, in Section 6, we give our conclusions.

2. NON-NEGATIVE MATRIX FACTOR DECONVOLUTION

NMFD approximates a matrix $\mathbf{X}$ of size $K \times N$ by

$$\mathbf{X} \approx \tilde{\mathbf{X}} = \sum_{\vartheta=0}^{\theta-1} \mathbf{B}_\vartheta \overset{\rightarrow\vartheta}{\mathbf{G}}, \tag{1}$$

where the $\rightarrow l$ operator shifts the columns of the matrix l spots to the right. The newly created columns at the left are filled with zeros. $\mathbf{B}_\vartheta$ is of size $K \times I$ and $\mathbf{G}$ is a matrix of size $I \times N$. I is a user defined parameter, corresponding to the number of components in which the NMFD separates the matrix. $\mathbf{B}$ can be interpreted as basis matrices $\mathbf{B}_i$ of size θ, with corresponding activation vectors $\mathbf{G}_i$, where the index i denotes the ith component. For $\theta = 1$ the basis matrices $\mathbf{B}_i$ become basis vectors. This case is equivalent to the NMF.

$\mathbf{B}$ and $\mathbf{G}$ are iteratively calculated by minimizing a distance function between $\mathbf{X}$ and $\tilde{\mathbf{X}}$. Lee and Seung [2] introduced multiplicative update rules for NMF for the squared Euclidean (SE) distance and for Kullback Leibler (KL) divergence, resulting in convergence to a local minimum of the distance function. We will use the update rules for KL divergence for ease of notation. However, our method can be applied aquivalently for other distance functions, such as SE distance or Itakura Saito (IS) distance. The update rules for NMFD for the KL divergence proposed by Smaragdis [6] are

$$\mathbf{G} \leftarrow \mathbf{G} \otimes \left(\sum_{\vartheta=0}^{\theta-1} \mathbf{B}_\vartheta^{\mathbf{T}} \left(\frac{\mathbf{X}}{\tilde{\mathbf{X}}} \right)^{\leftarrow\vartheta} \right) \oslash \left(\sum_{\vartheta=0}^{\theta-1} \mathbf{B}_\vartheta^{\mathbf{T}} \mathbf{1} \right) \tag{2}$$

and

$$\mathbf{B}_\vartheta \leftarrow \mathbf{B}_\vartheta \otimes \left(\left(\frac{\mathbf{X}}{\tilde{\mathbf{X}}} \right) \overset{\rightarrow\vartheta^{\mathbf{T}}}{\mathbf{G}} \right) \oslash \left(\mathbf{1}\mathbf{G}^{\mathbf{T}} \right), \tag{3}$$

where $\otimes$ and $\oslash$ denote elementwise multiplication resp. division. $\mathbf{1}$ is a $K \times N$ matrix with all elements set to one. For

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

$\theta = 1$ these update rules transform to the NMF update rules proposed by Lee and Seung [2]. The convergence to a local minimum of the distance function leads to a dependency on the initializations of the matrices $\mathbf{B}$ and $\mathbf{G}$. For different initializations the NMFD can produce different factorizations.

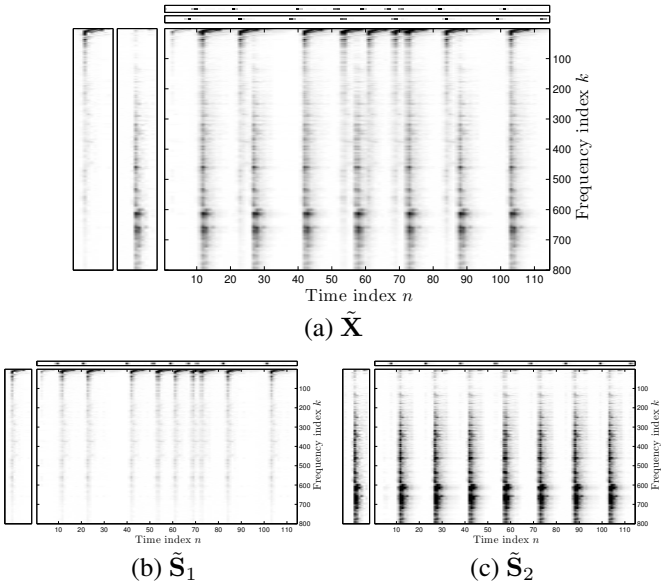
2.1. NMF with Automatic Relevance Determination

The number of components I is an important parameter regarding separation quality. If it is chosen too small, there are not enough components to approximate $\mathbf{X}$ correctly. If it is chosen too high, the components are overseparated and clustering gets difficult. Its optimal choice depends on $\mathbf{X}$.
A method to estimate this parameter during runtime of NMF is Automatic Relevance Determination (ARD). ARD places priors on $\mathbf{B}_i$ and $\mathbf{G}_i$, which are dependent on an irrelevance parameter β_i. Together with $\mathbf{B}$ and $\mathbf{G}$, β is iteratively updated, maximizing the posterior of the parameters given the data. Thus, some of the β_i converge to a large theoretical upper bound, corresponding to irrelevant components. Tan and Févotte proposed update rules for NMF with ARD, using the KL divergence [7]. Recently, generalized update rules using the β-divergence were proposed [5].

2.2. NMFD for Sound Source Separation

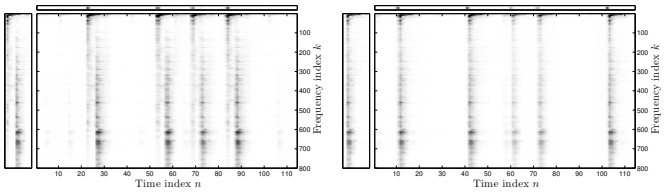
When applied to the magnitude spectrogram of an audio signal, NMFD can be used for source separation. In this case the $K \times \theta$ basis matrices $\mathbf{B}_i$ can be interpreted as spectral bases that are multiplied with temporal activation vectors $\mathbf{G}_i$. The size θ of the basis matrices defines how much temporal structure can be captured by each component. The separated matrices $\tilde{\mathbf{S}}_i$ can be interpreted as spectrograms of the separated acoustical events. Figure 1 shows an example for source separation of two drum sources by NMFD. Figure 1(a) shows the approximated spectrogram $\tilde{\mathbf{X}}$. The two basis matrices $\mathbf{B}_i$ on the left side contain the spectral shape of the two sources. The activation vectors $\mathbf{G}_i$ at the top show, at which time these sources are active. Figure 1(b) shows the approximated first source $\tilde{\mathbf{S}}_1$ and the basis vector $\mathbf{B}_1$ (left) and gain vector $\mathbf{G}_1$ (top). For this separation, we used $\theta = 10$, because this is the size of the largest temporal structure in this mixture.

In this example, the separation with NMFD works well. However, the fact, that the update rules only converge to a local minimum of the distance function, can lead to different results for different initializations of $\mathbf{B}$ and $\mathbf{G}$. While the temporal shape of the basis matrices $\mathbf{B}_i$ helps capturing temporal structure of the sources, it also makes the separation more difficult. A component with large basis matrices might be able to represent more than one note, leading to a worse separation. Figure 2 illustrates this problem. NMFD is used on the same example as before, but this time with another random initialisation. $\tilde{\mathbf{S}}_1$ and $\tilde{\mathbf{S}}_2$ do not correspond to the original sources, but are mixtures of different parts of the sources. The separation of the sources fails. Using more random initializations



(a) $\tilde{\mathbf{X}}$ (b) $\tilde{\mathbf{S}}_1$ (c) $\tilde{\mathbf{S}}_2$

Fig. 1. Example of successfull source separation with NMFD for a mixture of sources with temporal structure.



(a) $\tilde{\mathbf{S}}_1$ (b) $\tilde{\mathbf{S}}_2$

Fig. 2. Example of unsuccessfull source separation with NMFD for a mixture of sources with temporal structure.

showed, that this problem appears for about 25% of random initializations for this example. Because of this problem, the NMFD leads to worse separation results than NMF over a representative testset, as we showed in [8].

3. CUSTOM SIZED NON-NEGATIVE MATRIX FACTOR DECONVOLUTION

The examples in Section 2.2 show, that the temporal shape of $\mathbf{B}_i$ is helpful, but can lead to problems. We conclude, that it would be helpful to give the $\mathbf{B}_i$ a temporal shape that is as large as necessary, but as small as possible. Different sources behave differently and hence need a different size for the spectral bases. Therefore, it does not make sense to have a fixed size θ for the spectral bases of all the components, but different sizes θ_i for each component.
We propose a custom sized NMFD (CSNMFD), where the spectral bases of each component have different sizes. We

define this method componentwise as

$$\mathbf{X} \approx \tilde{\mathbf{X}} = \sum_{i=1}^{I} \tilde{\mathbf{S}}_i = \sum_{i=1}^{I} \sum_{\vartheta=0}^{\theta_i - 1} \mathbf{B}_{i,\vartheta} \overrightarrow{\mathbf{G}}_i^{\vartheta}, \tag{4}$$

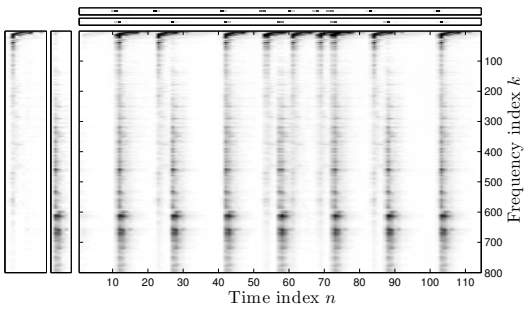
where θ is a vector of lenght I, with θ_i being the size of the spectral base $\mathbf{B}_i$ of the ith component. The componentwise update rules for the KL divergence are

$$\mathbf{G}_i \leftarrow \mathbf{G}_i \otimes \left(\sum_{\vartheta=0}^{\theta_i - 1} {\mathbf{B}_{i,\vartheta}}^{\mathbf{T}} \left(\frac{\mathbf{X}}{\tilde{\mathbf{X}}} \right)^{\overleftarrow{\vartheta}} \right) \oslash \left(\sum_{\vartheta=0}^{\theta_i - 1} {\mathbf{B}_{i,\vartheta}}^{\mathbf{T}} \mathbf{1} \right) \tag{5}$$

and

$$\mathbf{B}_{i,\vartheta} \leftarrow \mathbf{B}_{i,\vartheta} \otimes \left(\left(\frac{\mathbf{X}}{\tilde{\mathbf{X}}} \right) \overrightarrow{\mathbf{G}}_i^{\vartheta \mathbf{T}} \right) \oslash \left(\mathbf{1} \mathbf{G}_i^{\mathbf{T}} \right) \tag{6}$$

Although the update rules are presented componentwise, they should be calculated blockwise for blocks with the same size for $\mathbf{B}_i$ for reasons of efficiency. In fact, the updates of $\mathbf{B}_\vartheta$ can be calculated in one block, exactly as in NMFD (see Eq. 3).



(a) $\tilde{\mathbf{X}}$

Fig. 3. Example of source separation with CSNMFD for a mixture of sources with temporal structure.

3.1. CSNMFD for Sound Source Separation

CSNMFD provides the possibility to adapt the size of $\mathbf{B}_i$ for each component, giving the bases enough space to correctly approximate the sources, while minimizing the risk of bad separation quality due to too large spectral bases. Figure 3 shows the result of source separation using CSNMFD on the same example as before (note the different size of the matrices $\mathbf{B}_i$). For this example we used θ_1=10 and θ_2=4. The spectral bases correspond to the original sources, hence the separation is successful. In contrary to the NMFD no bad separations occured for different random initializations. The results for the CSNMFD have a comparable quality as the correct separations with NMFD.

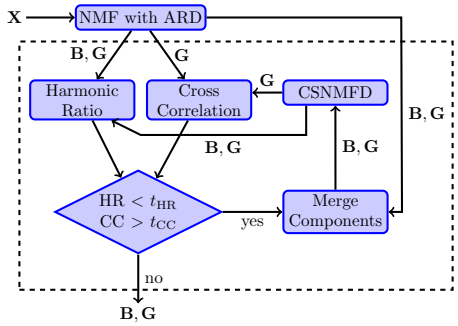


Fig. 4. Signal flow of the proposed algorithm.

4. COMPONENT SIZE ADAPTION USING CSNMFD

CSNMFD is able to capture temporal structure, that NMF cannot capture, while being more stable in results than NMFD. However, NMF, NMFD and CSNMFD require different previous knowledge. NMF only needs information about the number of components I (which can be solved by using NMF with ARD), NMFD also needs information about the temporal structures in the mixture to chose a suitable value for θ. CSNMFD even needs information about the temporal structure of all components. In the following, we present an algorithm using CSNMFD, which first gets a rough estimation of the necessary number of components I by using ARD and then reduces this number while simultaneously determining the temporals structure of the components.

Figure 4 shows the signal flow of the proposed algorithm. We start by performing NMF with ARD, resulting in a factorization with a suitable number of components. Acoustical events with a temporal structure, that cannot be represented by NMF, will be split into different components. The activation vectors $\mathbf{G}_i$ of these components have a high cross correlation (CC, see e.g. [9]). Thus, we calculate the CC of all $\mathbf{G}_i$ to find these components. As the lag of the highest correlation will be small, a maximum lag of $l_{\max}$ for CC is allowed. The problem exists for percussive sources, thus we also calculate the harmonic ratio (HR, see e.g. [10]) to make sure that no harmonic components are chosen. If the maximum of the CC exceeds a threshold t_{CC} and the HR of the components falls below a threshold t_{HR}, the components are merged. This is done by first calculating the spectrograms $\tilde{\mathbf{S}}_m$ and $\tilde{\mathbf{S}}_n$ of these components. Now NMFD with one component is performed on the spectrogram $\tilde{\mathbf{S}}_{m+n} = \tilde{\mathbf{S}}_m + \tilde{\mathbf{S}}_n$. The size θ_{m+n} of the basis matrix $\mathbf{B}_{m+n}$ is chosen corresponding to the lag, at which the CC between components m and n shows its maximum. Now the original $\mathbf{B}_m$ and $\mathbf{G}_m$ are replaced by $\mathbf{B}_{m+n}$ resp. $\mathbf{G}_{m+n}$ and θ_m is set to θ_{m+n}. At this point, the new basis matrices can be of different sizes, so NMF transforms to CSNMFD. $\mathbf{B}_n$ and $\mathbf{G}_n$ are deleted. After all components are merged, CSNMFD is performed to adjust the components. This procedure is repeated until no components are merged.

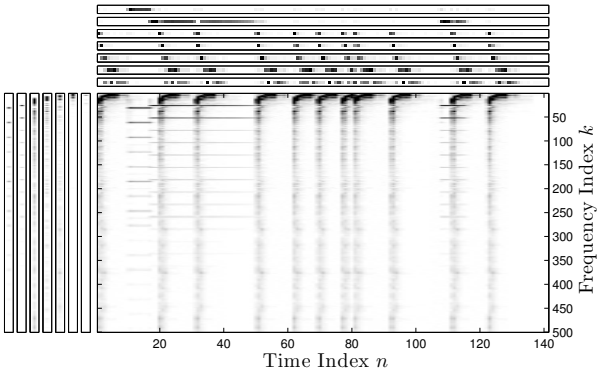


Fig. 5. Result of NMF with ARD for a mixture of saxophone and bass drum.

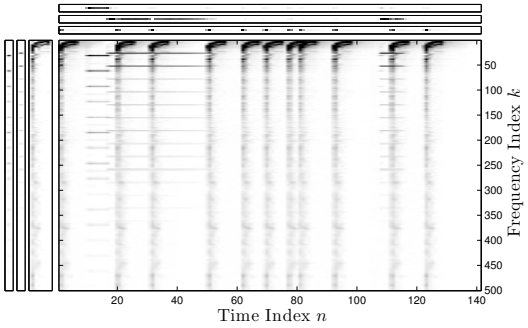


Fig. 6. Result of the algorithm using CSNMFD for a mixture of saxophone and bass drum.

5. EXPERIMENTAL RESULTS

Figure 5 shows one exemplary result of NMF with ARD and Figure 6 shows the result for the algorithm using CSNMFD for the same example. NMF with ARD needs five components to represent the bass drum. The proposed algorithm is able to represent it in one component. After optimal clustering, the SDR is 18.5 dB for both of these examples.

We performed the algorithms on 300 two-source-mixtures, that were mixed from 25 samples that we extracted from the QUASI database [11] with intial values for I in the range from 10 to 30. The thresholds were set to $t_{\mathrm{CC}} = 0.75$ and $t_{\mathrm{HR}} = 0.6$ and $l_{\max} = 3$. We used a non-blind clustering algorithm, maximizing the SDR (see [3]), to prevent our results from being deteriorated by the performance of a clustering algorithm. We performed the experiment for NMF using SE distance, KL divergence and IS distance. As mentioned in Section 2.2, NMFD leads to worse separation as NMF, therefore we only compare our method to NMF. Figure 7 shows the results for KL divergence, which led to the best results. It can be observed, that the quality in SDR is comparable for both methods, however, the proposed method needs less components. To prove the benefit of the reduced number of components for clustering, we performed the algorithm with the blind clustering algorithm proposed in [4], resulting in a 0.48 dB higher SDR for our method compared to ARD-NMF.

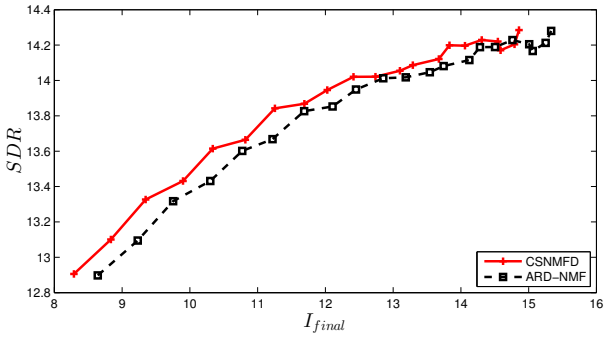


Fig. 7. NMF with ARD compared to the proposed algorithm. I_{final} is the average resulting number of components.

	SE distance	KL divergence	IS distance
ΔSDR	-0.15	0.01	0.03
ΔI_{final}	-1.66	-0.47	-0.16

Table 1. Average difference of resulting number of components I_{final} and separation quality SDR between ARD-NMF and the proposed algorithm for different distance functions.

Table 1 shows the difference of resulting number of components I_{final} and separation quality SDR between ARD-NMF and the proposed algorithm for the different distance functions for an intial number of components of $I = 30$. For KL divergence and IS distance the separation quality in SDR is comparable, however, the proposed method reduces the number of resulting components. For the SE distance the number of components is reduced even more, but the separation quality is slightly lower than for ARD-NMF.

6. CONCLUSIONS

In this paper we described the problems that arise when using NMFD for sound source separation of mixtures with temporal structure. We proposed a new method, CSNMFD, which provides the possibility to adapt the size of the spectral basis matrices to different components. We showed exemplarily, that this method is able to solve these typical problems.

We described, how CSNMFD can be used for source separation and showed, that it results in a factorization with less components. Besides this, the method provides additional advantageous information that a clustering algorithm could use, e.g. information about the temporal structure and the harmonicity of the components.

7. REFERENCES

[1] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[2] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2000, pp. 556–562.

[3] T. Virtanen, "Monaural Sound Source Separation by Nonnegative matrix Factorization With Temporal Continuity and Sparseness Criteria," in *IEEE Transactions on Audio, Speech, and Language Processing*. IEEE, 2000, vol. 3, pp. 1066–1074.

[4] M. Spiertz and V. Gnann, "Beta Divergence for Clustering in Monaural Blind Source Separation," in *128th AES Convention*. AES, 2010.

[5] V. Y. F. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization with the β-divergence," 2012.

[6] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Independent Component Analysis and Blind Signal Separation*. 2004, vol. 3195 of *Lecture Notes in Computer Science*, pp. 494–499, Springer Berlin Heidelberg.

[7] V. Y. F. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization," in *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*, 2009.

[8] J. M. Becker and V. Gnann, "Comparing Separation Quality of Nonnegative Matrix Factorization and Nonnegative Matrix Factor 2D Deconvolution in Audio Source Separation Tasks," in *133th AES Convention*. AES, 2012.

[9] J. R. Ohm, *Multimedia communication technology: Representation, transmission and identification of multimedia signals*, Springer, 2004.

[10] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 audio and beyond: Audio content indexing and retrieval*, John Wiley & Sons, 2006.

[11] "Quasi database - a musical audio signal database for source separation," `http://www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/`.