Complex SVD Initialization for NMF Source Separation on Audio Spectrograms

Julian Becker[1], Matthias Menzel[1], Christian Rohlfing[1]

[1] *Institut für Nachrichtentechnik, RWTH Aachen, 52056 Aachen, Deutschland, Email: becker@ient.rwth-aachen.de*

Introduction

Nonnegative Matrix Factorization (NMF) is an approximative low-rank matrix factorization which is frequently applied for source separation of audio signals (see e.g. [1]). The quality of source separation algorithms using NMF strongly depends on the initialization of the NMF. Very often, random values are used for initialization. Several other initialization strategies have been developed, with the aim to find better initial estimates, thus leading to a better resulting factorization. Most of these deterministic initialization methods use singular value decomposition (SVD). In this paper we introduce a new initialization scheme for audio source separation, based on complex SVD. We also evaluate several different state-of-the-art initializations in an audio source separation environment. We analyze the effect of the different methods on different kinds of mixtures and show, that our simple but efficient method leads to better results than other SVD-based initializations.

Nonnegative Matrix Factorization

NMF approximates a nonnegative matrix $\mathbf{X}$ of size $K \times N$ by a product of two nonnegative matrices $\mathbf{B}$ and $\mathbf{G}$,

$$\mathbf{X} \approx \tilde{\mathbf{X}} = \mathbf{B}\mathbf{G}, \tag{1}$$

with $\mathbf{B}$ of size $K \times I$ and $\mathbf{G}$ of size $I \times N$. I is a user defined parameter, which is usually chosen to be smaller than K and N. For better interpretation of the result of NMF, Equation 1 can be rewritten in the following way:

$$\mathbf{X}(k,n) \approx \tilde{\mathbf{X}}(k,n) = \sum_{i=1}^{I} \tilde{\mathbf{F}}(k,n,i) \tag{2}$$

with

$$\tilde{\mathbf{F}}(k,n,i) = \mathbf{B}(k,i)\mathbf{G}(i,n). \tag{3}$$

This can be interpreted as a decomposition of $\mathbf{X}$ into I rank-one matrizes $\tilde{\mathbf{F}}_i$ of size $K \times N$. Each matrix $\tilde{\mathbf{F}}(:,:,i)$ is calculated by the multiplication of a basis vector $\mathbf{B}(:,i)$ with an activation vector $\mathbf{G}(i,:)$, where $\mathbf{B}(:,i)$ denotes the ith column of $\mathbf{B}$.

The matrices $\mathbf{B}$ and $\mathbf{G}$ are iteratively calculated by minimizing an adequat distance function between $\mathbf{X}$ and $\tilde{\mathbf{X}}$. Commonly used distance functions are the Euclidean distance, the Kullback-Leibler (KL) divergence and the Itakura-Saito (IS) distance. Lee and Seung [2] introduced efficient multiplicative update rules for the square of the Euclidean distance as well as for the KL divergence, resulting in convergence to a local minimum of the distance function. The proposed update rules for the KL divergence are

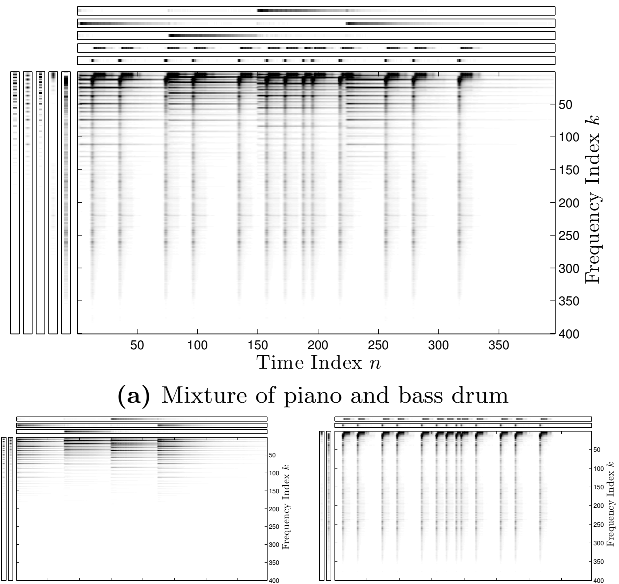
$$\mathbf{G} \leftarrow \mathbf{G} \otimes \frac{\mathbf{B}^{\mathbf{T}} \left(\frac{\mathbf{X}}{\tilde{\mathbf{X}}}\right)}{\mathbf{B}^{\mathbf{T}}\mathbf{1}} \tag{4}$$

and

$$\mathbf{B} \leftarrow \mathbf{B} \otimes \frac{\left(\frac{\mathbf{X}}{\tilde{\mathbf{X}}}\right)\mathbf{G}^{\mathbf{T}}}{\mathbf{1}\mathbf{G}^{\mathbf{T}}} \tag{5}$$

where $\mathbf{1}$ is a $K \times N$ matrix with all elements set to one and $\otimes$ denotes an elementwise multiplication. The divisions are also elementwise.

When applied to the magnitude $\mathbf{X}$ of a complex spectrogram $\underline{\mathbf{X}}$ of an audio signal, the NMF can be used for audio source separation. In this case the $K \times 1$ basis vectors $\mathbf{B}(:,i)$ can be interpreted as spectral bases that are multiplied with the temporal gain vectors $\mathbf{G}(i,:)$. Figure 1 examplarily shows source separation using NMF for a mixture spectrogram of piano and bass drum. The basis vectors $\mathbf{B}(:,i)$ (left) contain the spectral structures of the notes played by the piano and of the bass drum. The temporal gain vectors $\mathbf{G}(i,:)$ (top) contain the corresponding temporal gains.



(a) Mixture of piano and bass drum

(b) Separated piano **(c)** Separated bass drum

Figure 1: Example for source separation with NMF on a mixture of piano and bass drum.

For separation, only the first three vectors, which correspond to the components belonging to the piano, are used to reconstruct the piano spectrogram, the last two vectors are used to reconstruct the drum spectrogram.

Initializations

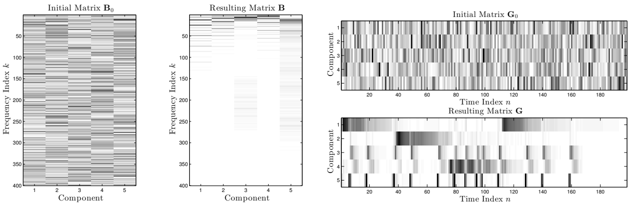
Since the multiplicative update rules only converge to a local minimum of the cost function, the results of the NMF highly depend of the initial values of the matrizes **B** and **G**. In the following, we will give a brief overview over commonly used initialization schemes.

Random Initialization

The matrizes **B** and **G** are very frequently initialized with random values. Usually the absolute values of a zero-mean normal distribution are used, however, other random distributions are possible.

The initialization has the advantage of being very easy to implement. It also has a low computational complexity, making it interesting for time-critical applications. One of the downsides of this initialization is, that there is no physical or mathematical motivation behind it. It is also a disadvantage, that it is not deterministic, different random initializations lead to different results, making it difficult to compare results.

Figure 2 shows an example of the initial and the resulting matrizes $\mathbf{B}_0, \mathbf{B}$ and $\mathbf{G}_0, \mathbf{G}$ with random initialization for the audio signal from Figure 1.



(a) Spectral basis matrix **B** **(b)** Temporal gain matrix **G**

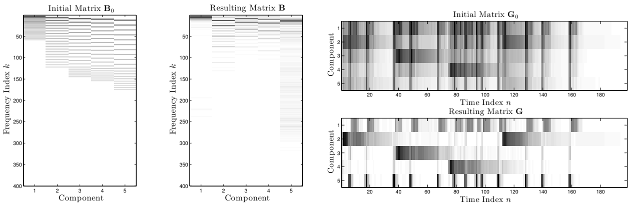
Figure 2: Initial and resulting matrizes $\mathbf{B}_0, \mathbf{B}$ and $\mathbf{G}_0, \mathbf{G}$ for random initialization.

Semantic Initialization

Under the assumption that the audio mixture contains harmonic sources, a semantic initialization [3] can be used. For this initialization, the matrix **B** is initialized with the spectral bases of the 88 notes of the piano. The matrix **G** is initialized by calculating the correlation of the corresponding components of **B** with each time frame of the spectrogram **X**. Since the number of components I is usually lower than 88, the components are deleted step by step until the desired number of I is reached. This is done after each NMF iteration by deleting the component with the lowest energy.

The semantic initialization is clearly motivated and can be expected to lead to good separation results for harmonic sources. On the other side, it is optimized only for harmonic sources and might have problems with other kinds of signals. Another downside is, that it is necessary to adapt I while performing the NMF, which makes it necessary to modify the NMF implementation.

Figure 3 shows an example of the initial and the resulting matrizes $\mathbf{B}_0, \mathbf{B}$ and $\mathbf{G}_0, \mathbf{G}$ with semantic initialization for the audio signal from Figure 1.



(a) Spectral basis matrix **B** **(b)** Temporal gain matrix **G**

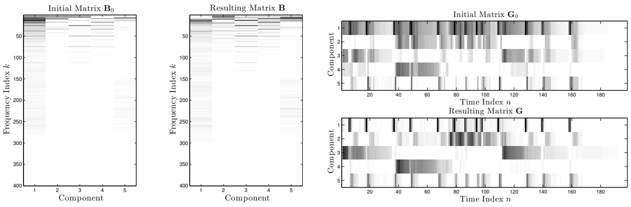
Figure 3: Initial and resulting matrizes $\mathbf{B}_0, \mathbf{B}$ and $\mathbf{G}_0, \mathbf{G}$ for semantic initialization.

SVD based Initializations

Several SVD based initializations have been proposed in the past [4, 5, 6]. Performing an SVD on a magnitude spectrogram **X** results in three matrizes **U**, **Σ** and **V**, with the property $\mathbf{X} = \mathbf{U\Sigma V}^*$. While **Σ** contains the singular values of **X**, the columns of **U** and **V** contain the left- and right-singular vectors of **X**. For an audio spectrogram, these vectors can be interpreted as spectral bases and temporal gains and thus be used as initialization for the NMF. Only the parts of **U** and **V** corresponding to the I highest singular values, are used as initialization for **B** and **G**. The negative entries of the matrizes have to be replaced with nonnegative values to make it a useful initialization for NMF. Other, more complex initializations using SVD (e.g. NNDSVD [6]) have been proposed.

Initializations using SVD are deterministic and generalized, meaning that they are not specifically designed for a special kind of signals. The problem with these initializations is the question how to treat negative values in **U** and **V**. Replacing them with zero leads to the smallest initial reconstruction error, however, it leads to problems in the NMF because of the multiplicative update rules of NMF. Using other non-zero values (e.g. absolute values) avoids this problem, but leads to a higher reconstruction error.

Figure 4 shows an example of the initial and the resulting matrizes $\mathbf{B}_0, \mathbf{B}$ and $\mathbf{G}_0, \mathbf{G}$ with NNDSVD initialization for the audio signal from Figure 1.



(a) Spectral basis matrix **B** **(b)** Temporal gain matrix **G**

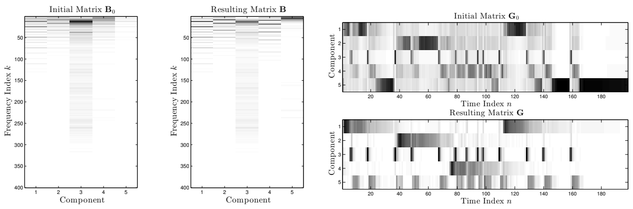
Figure 4: Initial and resulting matrizes $\mathbf{B}_0, \mathbf{B}$ and $\mathbf{G}_0, \mathbf{G}$ for NNDSVD initialization.

Fuzzy C-Means Initialization

The Fuzzy C-Means clustering algorithm can be used to initialize NMF [7]. Each time frame of **X** is interpreted as one data point. These data points are clustered into I clusters. The cluster centers can be used as initialization of **B**, while the partitioning matrix can be used as initialization for $\mathbf{G}$.

The Fuzzy C-Means initialization is deterministic and generalized. A disadvantage of this initialization is, that it only separates time frames. This means, that temporal overlapping sources might be represented in one component. It can be expected, that this might even deteriorate the separation results instead of improving them.

Figure 5 shows an example of the initial and the resulting matrizes $\mathbf{B}_0, \mathbf{B}$ and $\mathbf{G}_0, \mathbf{G}$ with Fuzzy C-Means initialization for the audio signal from Figure 1.



(a) Spectral basis matrix $\mathbf{B}$ **(b)** Temporal gain matrix $\mathbf{G}$

Figure 5: Initial and resulting matrizes $\mathbf{B}_0, \mathbf{B}$ and $\mathbf{G}_0, \mathbf{G}$ for the Fuzzy C-Means initialization.

Complex SVD Initialization

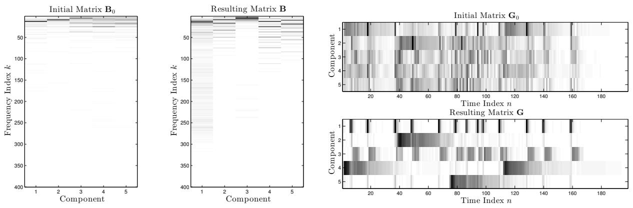
We propose to use an SVD on the complex spectrogram $\underline{\mathbf{X}} = \underline{\mathbf{U}}\mathbf{\Sigma}\underline{\mathbf{V}}^*$, instead of using the magnitude spectrogram. The parts of the matrizes $\underline{\mathbf{U}}$ and $\underline{\mathbf{V}}$ corresponding to the I highest singular values are used as initialization for $\mathbf{B}$ and $\mathbf{G}$. Since $\underline{\mathbf{U}}$ and $\underline{\mathbf{V}}$ are complex, only the magnitude is used,

$$\mathbf{B} = |\underline{\mathbf{U}}(:, 1:I)| \tag{6}$$

$$\mathbf{G} = |\underline{\mathbf{V}}(:, 1:I)^*|. \tag{7}$$

Compared to other SVD-based initialization methods, this approach has several advantages: While the other approaches perform the SVD on the magnitude spectrogram, thus neglecting phase information, the SVD of the complex spectrogram factorizes the complex spectrogram, which is the exact representation of the audio signal. Also, the problem of how to treat negative SVD values is solved, since using the abolute values does not lead to the same problems as replacing negative values with zeros.

Figure 6 shows an example of the initial and the resulting matrizes $\mathbf{B}_0, \mathbf{B}$ and $\mathbf{G}_0, \mathbf{G}$ with complex SVD initialization for the audio signal from Figure 1.



(a) Spectral basis matrix $\mathbf{B}$ **(b)** Temporal gain matrix $\mathbf{G}$

Figure 6: Initial and resulting matrizes $\mathbf{B}_0, \mathbf{B}$ and $\mathbf{G}_0, \mathbf{G}$ for the complex SVD initialization.

Evaluation

For evaluation we performed a source separation algorithm using NMF on a database of 60 different sources, containing harmonic instruments, percussive instruments, speech, vocals and noise. The 60 sources were mixed to every possible two-source mixture, resulting in 1770 mixtures. For the NMF updates we used KL-divergence as cost function and performed 300 NMF iterations.

We performed the algorithm with seven different initializations: We used the random, the semantic and the Fuzzy C-Means initialization as described. For the SVD based initializations, we used the state-of-the-art initialization method NNDSVD [6] with replacing the negative values with zeros (denoted NNDSVD) or with the absolute values (denoted NNDSVDa). We also used a basic SVD approach as described above, replacing negative values with the absolute value (denoted SVD). Finally, we also used the proposed initialization, performing an SVD on the complex spectrogram.

To evaluate the quality of the different initializations we used three different measures:

1. The reconstruction error after initialization. It can be expected, that a good initialization has a lower reconstruction error already at the beginning of the NMF. This could lead to a lower reconstruction error after performing the NMF.

2. The final reconstruction error. This is a commonly used measure to evaluate the quality of an initialization. A good initialization should lead to a local minimum of the cost function. If the local minimum is low, the initialization can be considered to be a good initial choice.

3. The signal to distortion ratio (SDR). Since we are interested in source separation, the most important property that we expect from a good initialization is, that it leads to a good separation quality. This might not necessarily be identical to the initialization with the lowest reconstruction error.

Figure 7 shows the distance of the approximated matrix with the initial matrizes $\mathbf{B}_0$ and $\mathbf{G}_0$, $\tilde{\mathbf{X}}_0 = \mathbf{B}_0 \cdot \mathbf{G}_0$ to the original matrix $\mathbf{X}$.

It can be observed, that the different initializations behave differently. The NNDSVDa seems to increase linearly. This can be explained with the replacement of negative entries with absolute values, which causes reconstruction errors. The number of datapoints, where this replacement takes place increases linearly with the number of components. The same holds for the SVD, but with a smaller impact on the reconstruction error. This problem is prevented when using the complex SVD. The NNDSVD initialization, replacing negative values with zero, has the lowest initial reconstruction error for all tested numbers of I. Figure 8 shows the distance of the approximated matrix $\tilde{\mathbf{X}} = \mathbf{B} \cdot \mathbf{G}$ after NMF to the original matrix $\mathbf{X}$.

The behaviour is very different to the one of the initial reconstruction error, showing that the initial reconstruc-

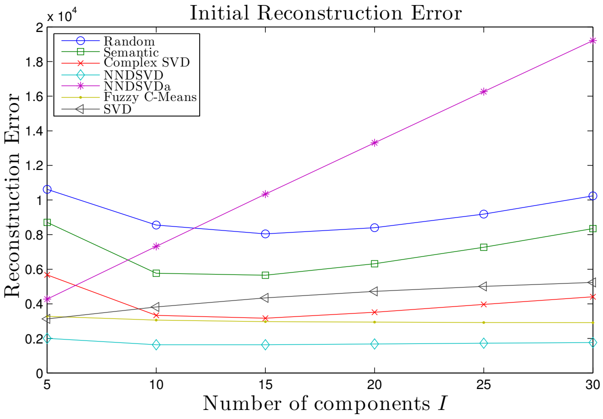


Figure 7: Initial reconstruction error for different initializations.

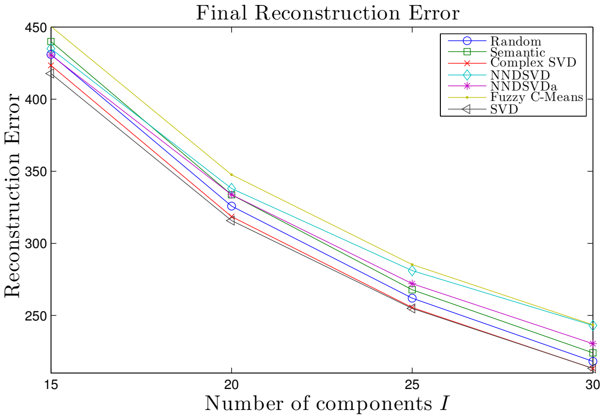


Figure 8: Final reconstruction error after NMF for different initializations

tion error is not a good indicator for a good convergence. NNDSVD, the method with the lowest initial reconstruction error, has one of the highest reconstruction errors after NMF. The SVD and the complex SVD initialization lead to the lowest final reconstruction error. Figure 9 shows the resulting separation quality for the different initializations. It should be noted, that the final reconstruction error is not a good indicator for separation quality. This means, that for an application aiming on source separation, the reconstruction error is not a reliable basis for deciding on the initialization. While the semantic initialization leads to the best separation results, it should be noted, that the used data set contains a lot of harmonic sources, favoring this initialization. Comparing the SVD-based intializations, the complex SVD leads to the best separation results.

Conclusion

We compared several state-of-the-art initialization methods for NMF in a source separation environment. We evaluated the results in terms of reconstruction error as well, as separation quality. The results indicate, that a lower reconstruction error does not necessarily lead to better source separation.

We also proposed a new initialization using an SVD on a complex audio spectrogram. The results show, that this

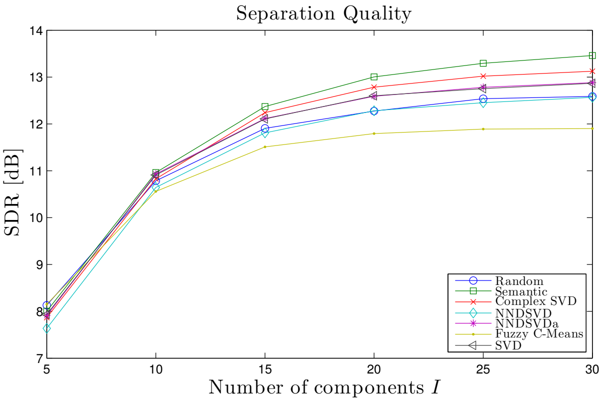


Figure 9: Separation quality for different initializations.

method leads to a very low reconstruction error as well as to a good separation quality. The separation quality was better than for all other SVD-based initializations. Only the semantic initialization resulted in a better separation quality, however, this initialization is specifically optimized for harmonic signals, while the complex SVD can be used on any kinds of signals.

We conclude, that the proposed intialization is a good alternative to existing initializations in terms of reconstruction error as well as in terms of separation quality.

References

[1] Wang, B. and Plumbley, M. D., Investigating single-channel audio source separation methods based on non-negative matrix factorization. In Proc. ICA Research Network International Workshop (pp. 17-20). 2006.

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

[3] Bertin, N., Badeau, R. and Richard, G., Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP. 2007.

[4] Tangirala, A. K., Kanodia, J., and Shah, S. L., Non-negative matrix factorization for detection and diagnosis of plantwide oscillations. Industrial & engineering chemistry research, 46(3), 801-817. 2007.

[5] Langville, A. N., Meyer, C. D., Albright, R., Cox, J., and Duling, D., Initializations for the nonnegative matrix factorization. Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 23-26). 2006.

[6] Boutsidis, C. and Gallopoulos, E., SVD based initialization: A head start for nonnegative matrix factorization. Pattern Recognition. 2008.

[7] Rezaei, M. and Boostani, R., An efficient initialization method for nonnegative matrix factorization. Journal of Applied Science. 2011.