Component-Adaptive Priors for NMF

Julian M. Becker and Christian Rohlfing

Institut für Nachrichtentechnik, RWTH Aachen University,
52056 Aachen, Germany
{becker,rohlfing}@ient.rwth-aachen.de
http://www.ient.rwth-aachen.de

Abstract. Additional priors for nonnegative matrix factorization (NMF) are a powerful way of adapting NMF to specific tasks, such as for example audio source separation. For this application, priors supporting sparseness or temporal continuity have been proposed. However, these priors are not helpful for all kinds of signals and should therefore only be used when needed. For some mixtures, only some components of the mixtures should be supported by these priors. We present an easy, but efficient method of adapting priors to different components. We show, that the separation results are improved, while the computational complexity is even slightly reduced. We also show, that our method is a helpful modification for the combination of different priors.

Keywords: NMF, audio source separation, temporal continuity, sparseness

1 Introduction

In the past, several extensions to nonnegative matrix factorization (NMF) have been proposed to adapt it to the task of source separation. Some extensions use convolutive bases instead of multiplicative ones [9, 10], others introduce additonal constraints such as sparsity [6, 13], temporal continuity [13] or spectral continuity [2]. An overview over different versions of NMF can be found in [4]. We focus on NMF with additional constraints. Most existing methods ([2, 6, 13]) add these constraints to all of the components of the NMF, weighting them equally. However, these priors often only make sense for some of the components. Recent work on only applying the constraints to some components either require an additional training step [8] or are restricted to specific tasks, such as separation of harmonic and percussive sources [3]. We propose a way of individually adapting priors to the components, so that they are used stronger on the components where they are more helpful. Our method can be used on any kind of mixture and does not require training or prior information about the sources. The paper is structured as follows: In Section 2, we provide basic information about NMF for source separation. In Section 3 we describe NMF with additional temporal continuity criterion as an example for additional priors. We then introduce our method of adapting this prior to the NMF components and generalize this method for other priors in Section 4. In Section 5 we evaluate our method by experiment, before closing the paper with our conclusions in Section 6.

2 NMF for Monaural Source Separation

We assume an audio mixture $\mathbf{x}$ in time domain, consisting of M sources $\mathbf{s}_m$. $\underline{\mathbf{X}}$ is the complex valued result of the short time Fourier transform (STFT) of $\mathbf{x}$. For source separation, NMF can be applied to the magnitude spectrogram $\mathbf{X} = |\underline{\mathbf{X}}|$. NMF approximates a matrix $\mathbf{X} \in \mathbb{R}_+^{K \times N}$ by a product of two matrices $\mathbf{B}$ and $\mathbf{G}$ as $\mathbf{X} \approx \tilde{\mathbf{X}} = \mathbf{BG},\ \mathbf{B} \in \mathbb{R}_+^{K \times I},\ \mathbf{G} \in \mathbb{R}_+^{I \times N}$. I defines the number of NMF components. $\mathbf{B}$ and $\mathbf{G}$ are iteratively calculated, minimizing a cost term $c(\mathbf{B}, \mathbf{G})$ between $\mathbf{X}$ and $\tilde{\mathbf{X}}$. Usually $c(\mathbf{B}, \mathbf{G})$ only consists of a reconstruction term $c_r(\mathbf{B}, \mathbf{G})$, where commonly used terms are the Euclidean distance, the Kullback-Leibler (KL) divergence and the Itakura-Saito (IS) distance. Lee and Seung [7] introduced multiplicative update rules for the squared Euclidean distance as well as for the KL divergence, resulting in convergence to a local minimum of the cost term. These update rules can be calculated using the gradient of $c(\mathbf{B}, \mathbf{G})$ with respect to $\mathbf{B}$, $\nabla_{\mathbf{B}} c(\mathbf{B}, \mathbf{G}) = \nabla_{\mathbf{B}}^{+} c(\mathbf{B}, \mathbf{G}) - \nabla_{\mathbf{B}}^{-} c(\mathbf{B}, \mathbf{G})$, where $\nabla_{\mathbf{B}}^{+} c(\mathbf{B}, \mathbf{G})$ and $\nabla_{\mathbf{B}}^{-} c(\mathbf{B}, \mathbf{G})$ are elementwise nonnegative terms, and the equivalently defined gradient with respect to $\mathbf{G}$, $\nabla_{\mathbf{G}} c(\mathbf{B}, \mathbf{G})$. The update rules are

$$\mathbf{B} \leftarrow \mathbf{B} \otimes \frac{\nabla_{\mathbf{B}}^{-} c(\mathbf{B}, \mathbf{G})}{\nabla_{\mathbf{B}}^{+} c(\mathbf{B}, \mathbf{G})} \quad \text{and} \quad \mathbf{G} \leftarrow \mathbf{G} \otimes \frac{\nabla_{\mathbf{G}}^{-} c(\mathbf{B}, \mathbf{G})}{\nabla_{\mathbf{G}}^{+} c(\mathbf{B}, \mathbf{G})}, \tag{1}$$

where $\otimes$ denotes elementwise multiplication and the divisions are also elementwise. For the methods presented in this paper, this generalized formulation of the update rules is sufficient. Exact update rules for KL-divergence and squared Euclidean distance can be found in [7] and for the IS-distance in [4]. Fig. 1 shows the factorization of a spectrogram of a mixture of a harmonic and a percussive source with NMF with $I = 2$ with the resulting matrices $\mathbf{B}$ (on the left) and $\mathbf{G}$ (on top). The columns of $\mathbf{B}$ capture the spectral shape of the acoustical events and can be interpreted as spectral bases. The rows of $\mathbf{G}$ can be interpreted as temporal activations. After perfoming NMF, phase information and finer structures of the spectrograms are restored using a filtering step, which is usually done by Wiener filtering (see e.g. [5, 11]). If I is higher than the number of sources M, the components have to be assigned to the sources in a clustering step, resulting in M spectrograms corresponding to the estimated sources. Finally, these spectrograms are transformed back to time domain by inverse STFT.

3 NMF with Temporal Continuity

An example for an additonal prior for NMF is the temporal continuity, proposed by Virtanen [13] to prevent incorrect factorizations. In the factorization in Fig. 1 it can be observed, that temporal gaps appear in the activation vector of the harmonic note where it is overlapped by the percussive tone. To prevent this, Virtanen proposed to add a temporal continuity term $c_t(\mathbf{G})$ to the reconstruction error term $c_r(\mathbf{B}, \mathbf{G})$, resulting in a cost function $c(\mathbf{B}, \mathbf{G}) = c_r(\mathbf{B}, \mathbf{G}) + \alpha_t c_t(\mathbf{G})$, where α_t is a weight to adjust the influence of $c_t(\mathbf{G})$. For $\alpha_t = 0$ this model

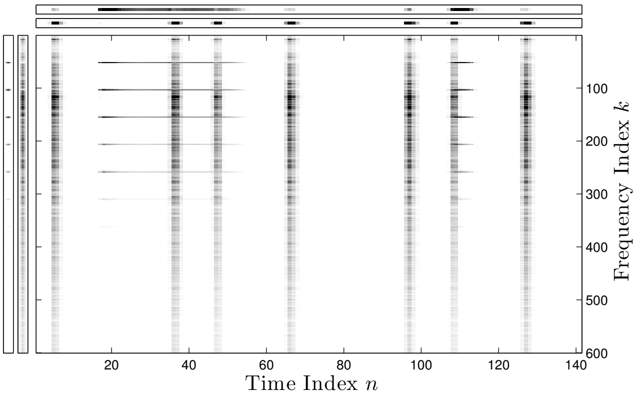


Fig. 1. Factorization of an audio mixture using NMF.

equals the standard NMF. The update rule for $\mathbf{G}$ transforms to

$$\mathbf{G} \leftarrow \mathbf{G} \otimes \frac{\nabla_{\mathbf{G}}^{-} c_r(\mathbf{B}, \mathbf{G}) + \alpha_t \nabla_{\mathbf{G}}^{-} c_t(\mathbf{G})}{\nabla_{\mathbf{G}}^{+} c_r(\mathbf{B}, \mathbf{G}) + \alpha_t \nabla_{\mathbf{G}}^{+} c_t(\mathbf{G})}, \tag{2}$$

while the update term for $\mathbf{B}$ stays the same as in Eq. (1). Virtanen proposes a temporal squared difference (TSD) cost term,

$$c_t(\mathbf{G}) = \sum_{i=1}^{I} \frac{1}{\sigma_i^2} \sum_{n=2}^{N} (g_{i,n} - g_{i,n-1})^2, \tag{3}$$

with $\sigma_i = \sqrt{(1/N) \sum_{n=1}^{N} g_{i,n}^2}$ being the standard deviation of each row of $\mathbf{G}$ and $g_{i,n}$ denoting one element of the matrix $\mathbf{G}$ at indizes i and n. The negative and positive gradient terms of this cost function are

$$[\nabla_{\mathbf{G}}^{-} c_t(\mathbf{G})]_{i,n} = \frac{2N(g_{i,n-1} + g_{i,n+1})}{\sum_{l=1}^{N} g_{i,l}^2} + \frac{2N g_{i,n} \sum_{l=2}^{N} (g_{i,l} - g_{i,l-1})^2}{\left(\sum_{l=1}^{N} g_{i,l}^2\right)^2} \tag{4}$$

and

$$[\nabla_{\mathbf{G}}^{+} c_t(\mathbf{G})]_{i,n} = \frac{4N g_{i,n}}{\sum_{l=1}^{N} g_{i,l}^2}. \tag{5}$$

4 Component-Adaptive Priors

The assumption of temporal continuity does not hold for all signals. In the example in Fig. 1 it only holds for the first (harmonic) component, whereas the second (percussive) component is not continuous in time. In fact, percussive sources usually have an impulse-like behaviour in time. Therefore, using the additional cost term is not advisable for the second component, as it deteriorates the resulting temporal activation vector. The percussive component is being smeared in time when using the temporal continuity prior. However, Virtanen only proposed a method to either use the temporal continuity term on all or no components.

In [3] it was proposed to only use priors on some components. Priors supporting harmonic structures are used on one half of the components, priors for percussive structures on the other half, assuming that harmonic and percussive structures will then automatically develop in the corresponding components. This approach has several disadvantages: First of all, it is not possible to use a structured initialization (e.g. SVD) with this approach, since these initializations already define which structures will develop in which component. Secondly, different mixtures might need a different number of percussive or harmonic components, which is not considered in this approach. The method also has the downside of only being applicable for mixtures of harmonic and percussive sources.

In the following, we will introduce a method for adapting the temporal continuity prior in a way that it is used stronger for components that need an additional temporal continuity term, while it is used weaker for components that do not. With this method, structured initializations are possible and the priors are automatically adapted to the different components, solving the problem of the fixed harmonic and percussive components and making it applicable to any kind of mixture. We also decribe, how this method can be used for other priors.

4.1 Finding Harmonic Components

Temporal continuity is mostly only desirable for harmonic components. To adapt the prior to the components, harmonic components have to be identified. Performing harmonic/percussive classification has two downsides: First of all, harmonic and percussive signals are not perfectly distinguishable. A harmonic signal might have a percussive onset, or a percussive signal a harmonic decay. This makes classification difficult and might lead to additional errors. Secondly, this step produces additional computational complexity, which might be unwanted.

An easier way is to use the cost term $c_t(\mathbf{G})$. For undistorted signals we expect it to be low for signals that are continous in time (e.g. harmonic) and high for others (e.g. percussive). Assuming that separation distortions are small, we can use this term as information about the behaviour of the different components.

4.2 Adapting the Prior

With this motivation, it seems reasonable to adapt the prior by multiplying the gradient terms $\nabla^-_{\mathbf{G}} c_t(\mathbf{G})$ and $\nabla^+_{\mathbf{G}} c_t(\mathbf{G})$ in Eq. (2) with a factor $1/c_t(\mathbf{G})$. Thus, the effect of the additional cost term is amplified for components with a low value of c_t (e.g. harmonic components) compared to components with high c_t. The update rule for $\mathbf{G}$ transforms to

$$\mathbf{G} \leftarrow \mathbf{G} \otimes \frac{\nabla^-_{\mathbf{G}} c_r(\mathbf{B},\mathbf{G}) + \frac{\alpha_t \nabla^-_{\mathbf{G}} c_t(\mathbf{G})}{c_t(\mathbf{G})}}{\nabla^+_{\mathbf{G}} c_r(\mathbf{B},\mathbf{G}) + \frac{\alpha_t \nabla^+_{\mathbf{G}} c_t(\mathbf{G})}{c_t(\mathbf{G})}}. \tag{6}$$

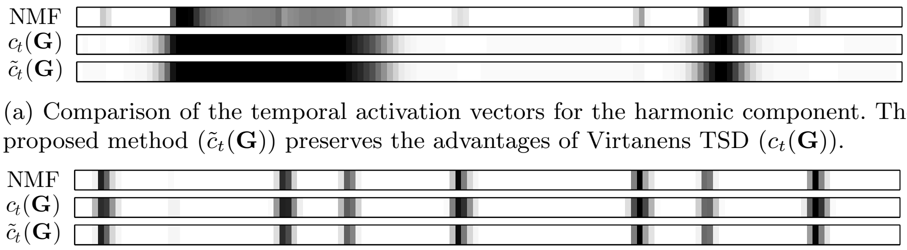
The gradient terms for the TSD prior, weighted with this factor are

$$\left[\frac{\nabla_{\mathbf{G}}^{-} c_t(\mathbf{G})}{c_t(\mathbf{G})}\right]_{i,n} = \frac{2(g_{i,n-1} + g_{i,n+1})}{\sum_{l=2}^{N}(g_{i,l} - g_{i,l-1})^2} + \frac{2g_{i,n}}{\sum_{l=1}^{N} g_{i,l}^2} \tag{7}$$

and

$$\left[\frac{\nabla_{\mathbf{G}}^{+} c_t(\mathbf{G})}{c_t(\mathbf{G})}\right]_{i,n} = \frac{4g_{i,n}}{\sum_{l=2}^{N}(g_{i,l} - g_{i,l-1})^2}. \tag{8}$$

Note, that Eqs. (7) and (8) are less computationally complex than Eqs. (4) and (5) because of a reduced number of multiplications.
Fig. 2 shows a comparison of the temporal activations for factorization of the same example as in Fig. 1 with standard NMF, Virtanens TSD ($\alpha_t = 250$) and the presented adaptive prior ($\alpha_t = 431$, the equivalent value to $\alpha_t = 250$ for Virtanens TSD with respect to the harmonic component). Fig. 2a shows the resulting temporal activations for the harmonic component. The positive effect of the temporal continuity is preserved with the proposed method, the gaps in the temporal activation of the harmonic component are avoided. However, the negative effect, the temporal smearing of the percussive component (Fig. 2b), is reduced compared to the factorization with NMF with Virtanens TSD.



(a) Comparison of the temporal activation vectors for the harmonic component. The proposed method ($\tilde{c}_t(\mathbf{G})$) preserves the advantages of Virtanens TSD ($c_t(\mathbf{G})$).

(b) Comparison of the temporal activation vectors for the percussive component. The proposed method ($\tilde{c}_t(\mathbf{G})$) reduces the temporal smearing of Virtanens TSD ($c_t(\mathbf{G})$).

Fig. 2. Comparison of temporal activation vectors $\mathbf{G}$ of standard NMF, NMF with Virtanens temporal continuity ($c_t(\mathbf{G})$, α_t=250) and the proposed method ($\tilde{c}_t(\mathbf{G})$,α_t=431).

4.3 Mathematical Description and Generalization

Comparing Eqs. (2) and (6) it can be observed, that the proposed method can be interpreted as using a new cost function $\tilde{c}_t$ with $\nabla_{\mathbf{G}}\tilde{c}_t(\mathbf{G}) = \frac{\nabla_{\mathbf{G}} c_t(\mathbf{G})}{c_t(\mathbf{G})}$. With the properties of the natural logarithm, it is obvious, that $\tilde{c}_t(\mathbf{G}) = \ln(c_t(\mathbf{G}))$. Thus, we can generalize our model for any cost function c with the requirement, that the prior only makes sense for components with a low value of c. Then, using $\ln(c)$ as prior leads to an adaption of the cost function c to the components.

5 Experimental Results

We performed source separation as described in Sec. 2. To evaluate the separation quality of the NMF without being affected by errors of a clustering algorithm, we used a non-blind clustering with knowledge of the original signals, as described in [13]. As measure for separation quality, we used the signal-to-distortion ratio (SDR), signal-to-inference ratio (SIR) and signal-to-artifacts ratio (SAR), as proposed in [12]. All given values are averaged over the complete testset.
We compared our method to standard NMF and to NMF with Virtanens TSD. To show, that our method is applicable to other priors and that it is beneficial for the combination of different priors, we also tested it on a combination of the TSD prior and a sparseness prior on **B**. Sparse spectral basis vectors can be assumed for some components (e.g. harmonic notes), but not for all (e.g. impulse-like or noisy components). Therefore, the prior should be used stronger on the components, that are relatively sparse and the requirement for our method is fullfilled. We used the same sparseness prior as was used in [13], with the difference, that we used the prior on **B** and not on **G**. The cost function of this prior was

$$c_s(\mathbf{B}) = \sum_{i=1}^{I} \frac{1}{\sqrt{(1/K)\sum_{l=1}^{K} b_{l,i}^2}} \sum_{k=2}^{K} b_{k,i} \tag{9}$$

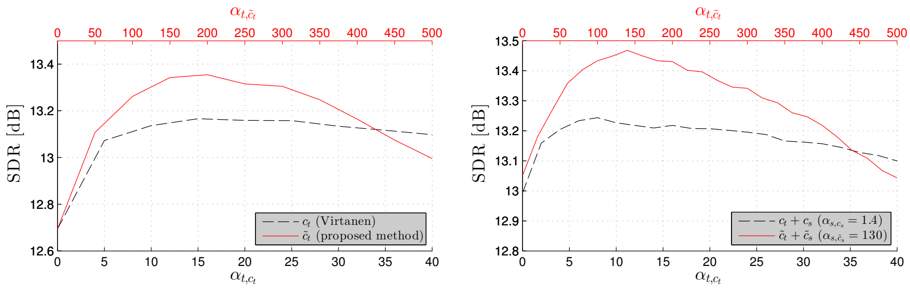
We compared a combination of c_t and c_s to one of $\tilde{c}_t$ and $\tilde{c}_s = \ln(c_s)$, using the proposed adaptive weighting. We chose $\alpha_{s,c_s} = 1.4$ and $\alpha_{s,\tilde{c}_s} = 130$, the weights with the best separation quality, for the spectral priors.

5.1 Testset & Setup

The testset consists of 60 audio signals, including harmonic and percussive signals, speech, vocals and noise, each being sampled with 44.1 kHz. These signals were mixed in every possible two-source combination, resulting in 1770 mixtures. The testset is identical to the one used in [11].
For the STFT, we used a window size of $s_w = 2^{12}$ and a hop size of $s_h = 2^{11}$ samples. I was set to 20 for every mixture, since this had shown to be a suitable number of components for this testset. As reconstruction error term, we used KL-divergence, as this produced the best separation results. We initialized the NMF by performing an SVD on the complex spectrogram $\underline{\mathbf{X}}$ as proposed in [1].

5.2 Results

A comparison of the SDR using NMF with Virtanens TSD and our adaptive version of it for different values of α_t, is shown in Fig. 3a. Note, that $\alpha = 0$ equals the standard NMF without priors. The SDR results for the combination of priors are shown in Fig. 3b. Our method reached higher SDR values over a broad range of α_t in both cases. An overview over the highest reached values of SDR, SIR



(a) SDR for the temporal continuity prior.

(b) SDR for combined priors.

Fig. 3. Experimental results: Adaptive priors lead to a higher maximum separation quality.

and SAR is given in Table 1. The separation quality is improved for all measures, when using the adaptive cost terms $(\tilde{c}_t, \tilde{c}_s)$ compared to the original ones (c_t, c_s).

	NMF	c_t	$\tilde{c}_t$	$c_t + c_s$	$\tilde{c}_t + \tilde{c}_s$
SDR [dB]	12.69	13.17	**13.36**	13.24	**13.47**
SIR [dB]	18.76	19.10	**19.34**	19.32	**19.56**
SAR [dB]	15.64	16.65	**16.66**	16.87	**17.00**

Table 1. Maximum SDR, SIR and SAR for different priors.

Comparing our method with the method in [3] is difficult because of the different preconditions. Thus, we used a random initialization and only evaluated the mixtures of harmonic and percussive signals in our testset, since these are the limitations of [3]. We evaluated the methods for Virtanens TSD (Eq. (3)) as harmonic prior and an equivalent spectral squared difference as percussive prior. Those are two of the priors that are used in [3], where the harmonic prior is put on one half of the components and the percussive prior on the other half. For our method we used the natural logarithm of the two priors on all components to adapt the priors to the components. We performed the methods with different combinations of the weights, choosing the best combination for evaluation. We also evaluated NMF without priors. The average SDR over all 440 harmonic/percussive mixtures was 14.84 dB for the standard NMF. Method [3] reached a maximum of 15.28 dB, our method reached a maximum of 15.76 dB.

6 Conclusion

In this paper, we introduced a way of adapting a temporal continuity prior to the NMF components so that the prior is used stronger on the components, where it is more helpful. Our method does not need any additional computational steps, but only changes the cost function, leading to less computational complex update rules. We showed, that our method can be generalized for other priors

and showed by experiment, that it leads to better separation results than the original prior. We also evaluated our method for the combination of different priors, verifying, that our method is also benefitial for this scenario.
We conclude, that the adaption to the different components is a helpful extension to existing priors. Our results should motivate future research on this topic. Future work could include methods to decide for the optimal value of α_t and α_s depending on the specific mixture, or combinations of more different priors.

References

1. Becker, J.M., Menzel, M., Rohlfing, C.: Complex SVD initialization for NMF source separation on audio spectrograms. In: DAGA 2015. Nürnberg, Germany (2015)
2. Becker, J.M., Sohn, C., Rohlfing, C.: NMF with spectral and temporal continuity criteria for monaural sound source separation. In: Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 22nd European. pp. 316–320. IEEE (2014)
3. Canadas-Quesada, F.J., Vera-Candeas, P., Ruiz-Reyes, N., Carabias-Orti, J., Cabanas-Molero, P.: Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints. EURASIP Journal on Audio, Speech, and Music Processing 2014(1), 1–17 (2014)
4. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.I.: Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons (2009)
5. Jaiswal, R., FitzGerald, D., Barry, D., Coyle, E., Rickard, S.: Clustering NMF basis functions using shifted NMF for monaural sound source separation. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. pp. 245–248. IEEE (2011)
6. Joder, C., Weninger, F., Virette, D., Schuller, B.: A comparative study on sparsity penalties for nmf-based speech separation: Beyond lp-norms. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. pp. 858–862. IEEE (2013)
7. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems. pp. 556–562 (2001)
8. Marxer, R., Janer, J.: Study of regularizations and constraints in NMF-based drums monaural separation. In: International Conference on Digital Audio Effects Conference (DAFx-13) (2013)
9. Schmidt, M.N., Mørup, M.: Nonnegative matrix factor 2-D deconvolution for blind single channel source separation. In: Independent Component Analysis and Blind Signal Separation. pp. 700–707. Springer (2006)
10. Smaragdis, P.: Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In: Independent Component Analysis and Blind Signal Separation, pp. 494–499. Springer (2004)
11. Spiertz, M., Gnann, V.: Beta divergence for clustering in monaural blind source separation. In: Audio Engineering Society Convention 128. Audio Engineering Society (2010)
12. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. vol. 14, pp. 1462–1469. IEEE (2006)
13. Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. vol. 15, pp. 1066–1074. IEEE (2007)