OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author’s version published in: http://oatao.univ-toulouse.fr/22465

Official URL
DOI: https://doi.org/10.1109/ICASSP.2018.8461803


Any correspondence concerning this service should be sent to the repository administrator: tech-oatao@listes-diff.inp-toulouse.fr
NONNEGATIVE MATRIX FACTORIZATION WITH TRANSFORM LEARNING

Dylan Fagot, Herwig Wendt and Cédric Févotte

IRIT, Université de Toulouse, CNRS, Toulouse, France
firstname.lastname@irit.fr

ABSTRACT
Traditional NMF-based signal decomposition relies on the factorization of spectral data, which is typically computed by means of short-time frequency transform. In this paper we propose to relax the choice of a pre-fixed transform and learn a short-time orthogonal transform together with the factorization. To this end, we formulate a regularized optimization problem reminiscent of conventional NMF, yet with the transform as additional unknown parameters, and design a novel block-descent algorithm enabling to find stationary points of this objective function. The proposed joint transform learning and factorization approach is tested for two audio signal processing experiments, illustrating its conceptual and practical benefits.

Index Terms— Nonnegative matrix factorization (NMF), transform learning, single-channel source separation

1. INTRODUCTION
Nonnegative matrix factorization (NMF) has become a privileged approach to spectral decomposition in several fields such as remote sensing and audio signal processing. In the latter field, it has led to state-of-the-art results in source separation [1] or music transcription [2]. The nonnegative data \( V \in \mathbb{R}^{M \times N} \) is typically the spectrogram \( |X| \) or \(|X|^2 \) of some temporal signal \( y \in \mathbb{R}^T \), where \( X \) is a short-time frequency transform of \( y \), \(| \cdot |\) denotes the entry-wise absolute value and \( \circ \) here denotes entry-wise exponentiation. NMF produces the approximate factorization

\[
V \approx WH,
\]

where \( W \in \mathbb{R}^{M \times K} \) is a nonnegative matrix referred to as dictionary that contains spectral patterns characteristic of the data while \( H \in \mathbb{R}^{K \times N} \) is the nonnegative matrix that contains the activation coefficients that approximate the data samples onto the dictionary. The factorization is usually low-rank \((K < \min(M,N))\) but not necessarily so (in which case regularization constraints should apply on \( W \) and/or \( H \)). The decomposition (1) can then be inverted back to the time domain or post-processed in various ways to solve a large panel of audio signal processing problems. In this traditional setting, a short-time frequency transform such as the short-time Fourier, Cosine or constant-\(Q \) transforms acts as a pre-processing of the raw temporal data \( y \). This is a potential limitation as any ill-chosen specification of the time-frequency transform may harm the quality of the decomposition. As such, we here propose to learn the transform together with the NMF factors. We propose to address this task by solving an optimization problem of the form

\[
\min_{W,H} D(|\phi(y)|^2 |WH)
\]

subject to structure constraints on \( \phi : \mathbb{R}^T \to \mathbb{R}^{M \times N} \) and to non-negativity of \( W \) and \( H \), and where \( D(\cdot | \cdot) \) is a measure of fit. In addition, we study and promote the use of sparsity-inducing penalty on \( H \). We refer to objectives of the form (2) as TL-NMF, which stands for transform-learning NMF.

Connections to other works. TL-NMF is inspired by the work of Ravishankar & Bresler [5] on learning sparsifying transforms, Given a collection of data samples \( Y \) (such as a set of images), their work consists in finding an invertible transform \( \Phi \) such that the output of \( \Phi Y \) is sparse. We are instead looking for a transform \( \phi \) such that \(|\phi(y)|^2 \) can be well approximated by a NMF. TL-NMF can be viewed as finding a one-layer factorizing network, where \( y \) acts as the raw data, \( \phi \) the linear operator, \(| \cdot |^2 \) the nonlinearity and \( WH \) the output of the network. Recent work has proposed combining deep learning and NMF, but in a different way. For instance, [6] considers a discriminative NMF setting and [7] study nonnegative auto-encoders. The TL-NMF framework proposed in this work could in addition be extended to fully bridge deep learning and NMF by looking for a cascade of decompositions \( f_L(\phi_L \ldots f_1(\phi_1(y))) \) such that the output \( y^L \) is a NMF. Yet, this is beyond the scope of this paper and left for future work.

TL-NMF still operates in a transformed domain and is not directly related to synthesis-based NMF models in which the raw data \( y(t) \) is modeled as \( y(t) = \sum_k c_k(t) \) where the spectrogram of \( c_k(t) \) is penalized so as to be closely rank-one [8,9].

Goals and contributions. The goal of this work is to study the TL-NMF problem of form (2). As a first step, we propose in this paper to gently depart from the traditional short-time Fourier or Cosine transform setting by restricting the transform \( \phi(y) \) to be a short-time orthogonal transform (Section 2.1). We consider real-valued transforms for simplicity and use the short-time Cosine transform (STCT) as a baseline. We propose an operational algorithm that returns stationary points of (2) and enables to learn a structured transform \( \phi \) together with the NMF factors \( W \) and \( H \) (Section 2.2). The TL-NMF approach is put to test and compared to conventional STCT-based NMF in two benchmark audio signal processing experiments: music decomposition (Section 3) and speech enhancement (Section 4). The results demonstrate that the proposed approach is operational and yields, in both applications, significant benefits (reduced objective function, data-adapted atoms, improved separation accuracy).

2. NMF MEETS TRANSFORM LEARNING
2.1. Learning a short-time orthogonal transform
Let us denote by \( Y \in \mathbb{R}^{M \times N} \) the matrix that contains adjacent and overlapping short-time frames of size \( M \) of \( y \) and denote by \( \Phi_{\text{DCT}} \in \mathbb{R}^{M \times M} \) the orthogonal real-valued DCT matrix with coefficients \( [\Phi_{\text{DCT}}]_{qm} = (2M)^{-1/2} \cos(\pi(q+1/2)(m+1/2)/M) \). With these notations, the power spectrogram \(|X|^2 \) of \( y \) is simply

This work has received funding from the European Research Council (ERC) under grant agreement No 681839 (project FACTORY). Thanks to Paris Smaragdis for discussions related to this work.
given by $|\Phi_{DCT}Y|^2$. Traditional NMF (with sparsity) may be cast as
\[ \min_{W,H} D(|\Phi_{DCT}Y|^2 WH + \lambda \frac{M}{K} ||H||_1) \]
subject to $H \geq 0, W \geq 0, \forall k, ||w_k||_1 = 1$ \hspace{1cm} (3)

where the notation $A \geq 0$ expresses nonnegativity, $||A||_1 = \sum_{ij} |a_{ij}|$ and $w_k$ denotes the $k^{th}$ column of $W$. Regular unpenalized NMF is simply obtained for $\lambda = 0$. The normalizing ratio $M/K$ measures the measure of fit and the penalty term are of the same order of magnitude. We propose in this work to relax the pre-fixed matrix $\Phi_{DCT}$ and learn it jointly with $W$ and $H$. In other words, we consider the general case where $\Phi \in \mathbb{R}^{M \times M}$ is a parameter of the objective function defined by

\[ C_\lambda(\Phi, W, H) \overset{\text{def}}{=} D(|\Phi Y|^2 WH) + \lambda \frac{M}{K} ||H||_1 \] \hspace{1cm} (4)

and seek a solution to the TL-NMF problem defined by
\[ \min_{\Phi, W, H} C_\lambda(\Phi, W, H) \]
subject to $H \geq 0, W \geq 0, \forall k, ||w_k||_1 = 1, \Phi^T \Phi = I_M$. \hspace{1cm} (5)

We choose at this stage to impose $\Phi$ to be orthogonal though one could consider relaxing this assumption as well. The orthogonality constraint implicitly keeps $\Phi$ nonsingular and excludes trivial solutions such as $(\Phi, W, H) = (0, 0, 0)$ or $(1_{M\times M}, 1_{M\times N}, 1_{1\times M} Y)^2$, where $1_{M\times N}$ denotes the $M \times N$ matrix filled with ones. In this paper, we also choose the measure of fit $D(\cdot)$ to be the Itakura-Saito (IS) divergence $D_{IS}(A|B) = \sum_{ij} (a_{ij}/b_{ij} - \log(a_{ij}/b_{ij}) - 1)$. Used with power spectral data, it is known to underlie a variance-structured Gaussian composite model that is relevant to the representation of audio signals [3] and learn it jointly with $\Phi$. The gradient of the objective function with respect to $(\Phi, W, H)$ can be shown to be given by
\[ \nabla C_\lambda(\Phi, W, H) = 2 (\Delta \odot X) Y^T \] \hspace{1cm} (6)

where $X = \Phi Y, \Delta = \nabla^{\odot -1} - \nabla^{\odot -1}, V = |X|^{\odot 2}, V = WH$. The steepest manifold-dependent descent direction is given by the natural gradient $\Omega = \nabla V^{\odot 2} \nabla - \nabla$. A suitable step-size $\gamma$ is then chosen according to the Armijo rule so that the projection $\pi(\Phi + \gamma \Omega)$ of the updated transform onto the orthogonal constraint induces a significant decrease of the objective function \hspace{1cm} [13]. Our block-coordinate descent algorithm is stopped when the relative variation between iteration $i - 1$ and $i$ falls below a given threshold $\tau$. The resulting TL-NMF algorithm is summarized in Algorithm 1.

3. EXPERIMENT 1: MUSIC DECOMPOSITION

In this section, we report results obtained with the proposed algorithm for decomposing real audio data $y(t)$, consisting of a 23 s excerpt of Mamavatu by Susheela Raman that has been downsampled to $f_s = 16$ kHz. $Y$ is constructed using 40 ms-long, 50%-overlapping temporal segments that are windowed with a sine bell. This construction leads to $M = 640$ and $N = 1191$. The proposed TL-NMF is compared to conventional IS-NMF, which is obtained using Alg. 1 as well, but fixed transform $\Phi = \Phi_{DCT}$. The two algorithms are run with the same stopping threshold $\tau = 10^{-9}$, the same rank $K = 10$, and three values of the regularization parameter $\lambda$.
Fig. 2. The six most significant atoms learnt by TL-NMF from random initializations for \( \lambda \in \{0, 10^{-3}, 10^{0}\} \) (from left to right).

\( \lambda \in \{0, 10^{0}, 10^{3}\} \). The two algorithms are initialized with the same random initializations of \( W \) and \( H \).

Comparison of objective function values. Fig. 1 displays the objective function values w.r.t. iterations for the two approaches for \( \lambda = 10^{0} \) (the initial objective value is cropped for better readability) and permits the following conclusions. IS-NMF reaches the stopping criterion after fewer iterations than TL-NMF, as expected since the transform \( \Phi \) is fixed for the latter, leading to faster convergence for the remaining variables \( W \) and \( H \). Moreover, TL-NMF starts at a larger objective function value due to the random orthogonal transform initialization for \( \Phi \). Yet, the proposed TL-NMF algorithm monotonously decreases the objective function to a value that, at convergence, is of the order of the final value of IS-NMF.

Regularization and learnt transform. We now examine examples of the atoms returned by TL-NMF (rows \( \phi_{nm} \) of \( \Phi \)) for the three values of \( \lambda \). Fig. 2 displays the six atoms which most contribute to the audio signal (i.e., with largest values \( ||\phi_{nm} Y||_2 \)). Clearly, without regularization \( (\lambda = 0) \), the learnt atoms lack apparent structure. Yet, interestingly, as the value for \( \lambda \) is increased, the atoms become oscillatory and smoother. This is a direct consequence of and justifies the use of the sparsity-inducing term in (3), which induces a structure-seeking constraint on the transform. Eventually, for strong regularization \( (\lambda = 10^{0}) \), the learnt atoms resemble packet-like, highly regular oscillations, further analyzed in the next paragraph.

Properties of the learnt atoms. In Fig. 3, pairs of atoms \((1, 2), (3, 6), (4, 5) \) (from left to right) for \( \lambda = 10^{0} \) are plotted, together with the square root of their summed square magnitudes and the sine bell used to window the data. Interestingly, the results demonstrate that the identified pairs of atoms are approximately in quadrature. They hence offer shift invariance properties similar to those of the (complex-valued) short time Fourier transform, which is ubiquitous in audio applications. Yet, the TL-NMF algorithm permits to learn these atoms from random initializations. This provides strong evidence for the relevance of the proposed approach. Further, note that the atoms embrace the sine bell used to taper the observations \( Y \).

4. EXPERIMENT 2: SUPERVISED SOURCE SEPARATION

We now examine whether learning an adaptive transform is actually useful for source separation. To this end, we consider a supervised NMF-based separation setting that follows the approach of [14]. In the following we address the separation of speech from interfering noise, but the method can be applied to any class of sound.

4.1. Principle

We assume that we are given speech and noise reference data \( y_{sp}(t) \) and \( y_{no}(t) \) from which we form short-time matrices \( Y_{sp} \) and \( Y_{no} \) of sizes \( M \times N_{sp} \) and \( M \times N_{no} \), as in Section 2.1. Given a noisy speech recording \( y(t) \) with short-time matrix \( Y \), traditional supervised NMF amounts to estimating activation matrices \( H_{sp} \) and \( H_{no} \) such that

\[
V \approx W_{sp} H_{sp} + W_{no} H_{no},
\]

subject to sparsity of \( H_{sp} \) and \( H_{no} \), where \( V = |\Phi_{DCT} Y|^{2} \), \( W_{sp} = |\Phi_{DCT} Y_{sp}|^{2} \), \( W_{no} = |\Phi_{DCT} Y_{no}|^{2} \) [14]. Temporal source and noise estimates are then reconstructed in a second step by so-called Wiener filtering [3], based on the spectrogram estimates \( \hat{V}_{sp} = W_{sp} H_{sp} \) and \( \hat{V}_{no} = W_{no} H_{no} \).

In this section, we generalize this procedure by again learning an optimal transform within the separation procedure. To this end, we propose to build an approximation like (8) but where the fixed transform \( \Phi = \Phi_{DCT} \) is now relaxed and learnt together with \( H_{sp} \) and \( H_{no} \). This means we propose to minimize

\[
C_A(\Phi, H_{sp}, H_{no}) \overset{\text{def}}{=} \text{DcS} \left( |\Phi Y|^{2} || H_{sp} \right) + \lambda_{sp} M_{N_{sp}} || H_{sp} ||_{1} + \lambda_{no} M_{N_{no}} || H_{no} ||_{1}
\]

s.t. \( \Phi^{T} \Phi = I, H_{sp} \geq 0, H_{no} \geq 0 \),

with \( \Lambda = (\lambda_{sp}, \lambda_{no}) \) defining the possibly different weights in front of the sparsity terms. Note that \( \Phi \) now appears in both sides of the data-fitting term \( \text{DcS}(|\cdot|) \), as the same transform is applied to the mixed data \( Y \) and the reference data \( Y_{sp} \) and \( Y_{no} \). This requires to slightly modify the gradient of \( C_A \) w.r.t. \( \Phi \) as compared to Section 2 and as described in next section. Given a solution to (9) and \( V = |\Phi Y|^{2} \) along with speech and noise spectrogram estimates \( \hat{V}_{sp} = |\Phi_{DCT} Y_{sp}|^{2} H_{sp} \) and \( \hat{V}_{no} = |\Phi_{DCT} Y_{no}|^{2} H_{no} \), temporal estimates may still be produced with Wiener filtering, i.e.,

\[
Y_{s} = \Phi^{T} \left( \frac{\hat{V}_{sp}}{\hat{V}_{sp} + \hat{V}_{no}} \circ |\Phi Y| \right)
\]
followed by standard overlap-adding of the columns of \( \hat{Y}_{sp} \) to return \( y_{sp}(t) \), and likewise for the noise. This is exactly the same procedure as in traditional NMF-based separation except that \( \Phi_{\text{DCT}} \) and \( \Phi_{\text{DCT}}^{\text{re}} \) are replaced by \( \Phi \) and \( \Phi^T \).

### 4.2. Algorithm

Denote \( Y_{\text{ref}} = [Y_{sp}, Y_{no}] \), \( X_{\text{ref}} = \Phi Y_{\text{ref}} \), \( W = |X_{\text{ref}}|^2 \), \( H = [H^T_{\text{ref}}, H^T_{\text{no}}]^T \) and \( V = WH \). Given \( W \), \( H \) can be updated with multiplicative rules derived from majorization-minimization as in [10]. We use again a gradient-descent approach for the update of \( \Phi \). The gradient of the objective function (9) can be expressed as

\[
\nabla \Phi C_\Phi (\Phi, H) = 2 (\Delta \circ X) Y^T + 2 (\Xi \circ X_{\text{ref}}) Y^T_{\text{ref}} \tag{11}
\]

where \( \Delta = \hat{V}^{v-1} - V^{v-1} \) and \( \Xi = \Delta' H^T \) with \( \Delta' = \frac{V - \hat{V}}{\sqrt{\|V\|}} \). Note that the first term of (11) is the gradient in (6). The second term is nothing but the gradient of the data-fitting term \( D_{\text{IS}} \) with its first argument fixed. Based on (11), we again use a line-search step selection in the steepest natural gradient direction followed by a projection, like in Section 2.2 and following [13]. The resulting algorithm is summarized in Alg. 2.

### 4.3. Speech enhancement experiment

We consider clean speech and noise data from the TIMIT corpus [15] and the CHIME challenge, respectively. For speech reference data \( y_{sp}(t) \), we use all utterances but the first one in the train/trl_train/ directory (about 21 s in total). For noise reference data \( y_{no}(t) \), we use 30 s of the file \( \text{audio_icassp2018} / \text{audio} / \text{BIC_150204_01088_0120_0121} \), which contains noise recorded in a bus. A simulated mixed signal \( y(t) \) of duration 3 s is generated by mixing the remaining speech utterance with another segment of the noise file (as such, the test data is not included in the reference data), using signal-to-noise (SNR) ratios of \(-10 \, \text{dB} \) and 0 dB. The audio files’ sampling frequency is \( f_s = 16 \, \text{kHz} \), and short-term matrices \( Y, Y_{sp} \) and \( Y_{no} \) are constructed using 40 ms-long, 50%-overlapping windowed segments like in Section 3, leading to \( M = 640, N = 149, N_{sp} = 1059 \) and \( N_{no} = 1517 \).

Our supervised TL-NMF approach is compared to the traditional supervised NMF procedure (with the IS divergence) described in Section 4.1, based on the same reference data and using the

\[\text{Table 1. Source separation performance.}\]

<table>
<thead>
<tr>
<th>Method</th>
<th>SDR (dB)</th>
<th>SIR (dB)</th>
<th>SAR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR= -10 dB</td>
<td>7.93</td>
<td>13.95</td>
<td>4.12</td>
</tr>
<tr>
<td>Baseline</td>
<td>-9.50</td>
<td>10.00</td>
<td>-9.50</td>
</tr>
<tr>
<td>IS-NMF</td>
<td>-6.75</td>
<td>6.82</td>
<td>-5.00</td>
</tr>
<tr>
<td>TL-NMF</td>
<td>1.73</td>
<td>12.29</td>
<td>12.11</td>
</tr>
<tr>
<td>SNR=0 dB</td>
<td>9.00</td>
<td>2.22</td>
<td>19.20</td>
</tr>
<tr>
<td>Baseline</td>
<td>-9.50</td>
<td>10.00</td>
<td>-9.50</td>
</tr>
<tr>
<td>IS-NMF</td>
<td>1.73</td>
<td>0.69</td>
<td>0.08</td>
</tr>
<tr>
<td>TL-NMF</td>
<td>6.50</td>
<td>5.81</td>
<td>3.06</td>
</tr>
</tbody>
</table>

1\text{http://spandh.dcs.shef.ac.uk/chime_challenge}

2\text{https://www.iirit.fr/~Cedric.Pevotte/extras/audio_icassp2018.zip}
6. REFERENCES


