On the Identifiability of Transform Learning for Non-negative Matrix Factorization

Sixin Zhang, Emmanuel Soubies, Cédric Févotte

On the Identifiability of Transform Learning for Non-negative Matrix Factorization

Sixin Zhang, Emmanuel Soubies, and Cédric Févotte *Senior Member, IEEE*

***Abstract*—Non-negative matrix factorization with transform learning (TL-NMF) aims at estimating a short-time orthogonal transform that projects temporal data into a domain that is more amenable to NMF than off-the-shelf time-frequency transforms. In this work, we study the identifiability of TL-NMF under the Gaussian composite model. We prove that one can uniquely identify row-spaces of the orthogonal transform by optimizing the likelihood function of the model. This result is illustrated on a toy source separation problem which demonstrates the ability of TL-NMF to learn a suitable orthogonal basis.**

***Index Terms*—NMF, transform learning, identifiability, source separation, joint diagonalization.**

I. Introduction

SPECTRAL unmixing by non-negative matrix factorization (NMF) is a standard approach to signal decomposition. It proceeds by transforming the signal into a domain where NMF is applied. For one-dimensional audio signals, it is customary to use a short-time frequency transform, such as the short-time Fourier or discrete cosine transforms. Such transforms first apply a short-time window to divide the signal into shorter segments of equal length and then compute the orthogonal Fourier or cosine transform separately on each segment. However, due to the time-frequency resolution trade-off, there is an increasing interest in replacing the frequency transform by a more discriminative orthogonal transform. Hence, transform-learning NMF (TL-NMF) was proposed in [1] with the goal of learning an adaptive short-time orthogonal transform. The problem is then to estimate the orthogonal transform given a fixed short-time window. TL-NMF also appears to be a special case of independent low-rank tensor analysis (ILRTA) [2] which offers a general framework for short-time modeling of temporal sequences under composite covariance models.

In this work, we study the identifiability of the orthogonal transform in TL-NMF under the Gaussian composite model (GCM) [3]. Under this probabilistic model, we first derive a negative log-likelihood objective for TL-NMF in Section 2.1. It is a variation of the Itakura-Saito divergence that is commonly used in NMF (IS-NMF) [3] and in particular in [1]. In Section 2.2 we establish conditions for identifiability of row-spaces of the orthogonal transform in TL-NMF (*i.e.*, linear subspaces generated by subsets of its rows). The conditions generalize some results about the identifiability of joint-diagonalization proposed in [4]. Finally, Section 3 illustrates our identifiability result on a toy audio decomposition problem.

The authors are with IRIT, Université de Toulouse, CNRS, Toulouse, France. Email: firstname.lastname@irit.fr. This work is supported by the European Research Council (ERC FACTORY-CoG-6681839).

Notations: For a matrix $\mathbf{X} \in \mathbb{R}^{M\times N}$, we denote by $\mathbf{x}_n$, $\underline{\mathbf{x}}_m$ and $[\mathbf{X}]_{mn}$ (or x_{mn}) its n-th column, m-th row, and (m,n)-th element respectively. For a vector signal $\mathbf{y} \in \mathbb{R}^T$, $\mathbf{y}(t)$ denotes its t-th element and $\mathbf{y}^\intercal$ denotes its transpose. We write $\mathrm{Diag}(\mathbf{y})$ for the diagonal matrix formed out of the vector $\mathbf{y}$.

II. Transform Learning NMF

Given a temporal signal $\mathbf{y} \in \mathbb{R}^T$, TL-NMF aims at finding a short-time orthogonal transform $\phi : \mathbb{R}^T \to \mathbb{R}^{M\times N}$ such that the element-wise squared magnitude $|\phi(\mathbf{y})|^{\circ 2}$ of $\phi(\mathbf{y})$ can be well approximated by a low-rank or sparse NMF, *i.e.*,

$$|\phi(\mathbf{y})|^{\circ 2} \approx \mathbf{WH}. \tag{1}$$

Denoting by $\mathbf{Y} \in \mathbb{R}^{M\times N}$ the matrix that contains N adjacent and half-overlapping short-time frames of $\mathbf{y}$ with length M, we may write $\phi(\mathbf{y}) = \mathbf{\Phi Y}$. In this paper, we study how to identify $\mathbf{\Phi}$ given observations of $\mathbf{y}$ which satisfy (1). The problem of how to identify the $\mathbf{W}$ and $\mathbf{H}$ belongs to the identifiability of NMF [5]. In the next section, we derive a new objective (slightly different to [1]) for TL-NMF to learn $\mathbf{\Phi}$, $\mathbf{W}$, and $\mathbf{H}$. Throughout the paper, we consider $\mathbf{\Phi} \in \mathbb{R}^{M\times M}$ to be a real orthogonal matrix. Moreover $\mathbf{W} \in \mathbb{R}_+^{M\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ are non-negative matrices.

A. Gaussian Composite Model

The Gaussian composite model (GCM) is used to characterize sound signals with composite structure [3]. The short-time Fourier coefficients of a signal is modeled as a sum of independent Gaussian random variables with an NMF structure on their variances. We consider this model to characterize the distribution of $\phi(\mathbf{y})$. In short, under the GCM, we have

$$[\mathbf{\Phi Y}]_{mn} \sim \mathcal{N}(0, [\mathbf{WH}]_{mn}), \tag{2}$$

where the variance is $[\mathbf{WH}]_{mn} > 0$. Moreover, conditioned on $\mathbf{\Phi}$, $\mathbf{W}$, and $\mathbf{H}$, $[\mathbf{\Phi Y}]_{mn}$ is assumed to be independent of $[\mathbf{\Phi Y}]_{m'n'}$ for any $(m,n) \neq (m',n')$. As such, the negative log-likelihood function is given, up to a constant, by

$$\begin{aligned} -\log p(\mathbf{Y}|\mathbf{\Phi},\mathbf{W},\mathbf{H}) &= -\sum_{mn} \log \mathcal{N}\big([\mathbf{\Phi Y}]_{mn}^2 \big| 0, [\mathbf{WH}]_{mn}\big) \\ &\overset{c}{=} \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\left(\frac{[\mathbf{\Phi Y}]_{mn}^2}{[\mathbf{WH}]_{mn}} + \log\big([\mathbf{WH}]_{mn}\big)\right). \end{aligned} \tag{3}$$

When $\mathbf{\Phi}$ is fixed (e.g., to the discrete cosine transform, DCT), minimizing (3) with respect to (w.r.t) $\mathbf{W}$ and $\mathbf{H}$ is equivalent

to IS-NMF [3]. To study identifiability, we will consider the expected (w.r.t $\mathbf{Y}$) negative log-likelihood objective:

$$c(\mathbf{\Phi},\mathbf{W},\mathbf{H}) \overset{\text{def}}{=} \mathbb{E}(-\log p(\mathbf{Y}|\mathbf{\Phi},\mathbf{W},\mathbf{H})) \quad (4)$$
$$\overset{\text{c}}{=} \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\left(\frac{\mathbb{E}([\mathbf{\Phi Y}]_{mn}^2)}{[\mathbf{WH}]_{mn}} + \log\left([\mathbf{WH}]_{mn}\right)\right).$$

B. Identifiability of $\mathbf{\Phi}$

Let us assume that there exists an orthogonal transform $\bar{\mathbf{\Phi}}$ such that $\bar{\mathbf{\Phi}}\mathbf{Y}$ follows the GCM with parameters $\bar{\mathbf{W}} \in \mathbb{R}_+^{M\times \bar{K}}$ and $\bar{\mathbf{H}} \in \mathbb{R}_+^{\bar{K}\times N}$. Here $\bar{K}$ is the ground-truth rank which is not known *a priori*. Identifiability is about whether the minimal solution of c as of (4) in $\mathbf{\Phi}$ corresponds to $\bar{\mathbf{\Phi}}$. In the context of TL-NMF, we are interested in identifying row-spaces of $\bar{\mathbf{\Phi}}$ such that signal components of different nature can be transformed into different (orthogonal) row-spaces of $\bar{\mathbf{\Phi}}$. This principle can improve source separation performance as we shall illustrate in Section 3. More formally, we study identifiability in the sense of Definition 1.

Definition 1. *Let $\textbf{\textit{span}}_{\bar{B}}(\bar{\mathbf{\Phi}})$ (resp., $\textbf{\textit{span}}_{B}(\mathbf{\Phi})$) denote the row-space of $\bar{\mathbf{\Phi}}$ (resp., $\mathbf{\Phi}$) spanned by its rows indexed by $\bar{B}$ (resp., B). We say that the rows of $\bar{\mathbf{\Phi}}$ indexed by $\bar{B} \subset \{1,\dots,M\}$ are identifiable if any global minimizer of c is attained for a $\mathbf{\Phi}$ such that there exists $B \subset \{1,\dots,M\}$ for which $\textbf{\textit{span}}_{\bar{B}}(\bar{\mathbf{\Phi}}) = \textbf{\textit{span}}_{B}(\mathbf{\Phi})$.*

We first give sufficient conditions for which $(\bar{\mathbf{\Phi}},\bar{\mathbf{W}},\bar{\mathbf{H}})$ is a global minimum of c.

Lemma 1. *Let $\mathbf{Y}$ be a random matrix such that, for each column $\mathbf{y}_n$ of $\mathbf{Y}$,*

$$\mathbb{E}(\mathbf{y}_n) = \mathbf{0}, \quad (5)$$
$$\bar{\mathbf{\Phi}}\mathbb{E}(\mathbf{y}_n\mathbf{y}_n^\intercal)\bar{\mathbf{\Phi}}^\intercal = \mathrm{Diag}(\bar{\mathbf{v}}_n), \quad (6)$$

for an orthogonal matrix $\bar{\mathbf{\Phi}} \in \mathbb{R}^{M\times M}$ and a vector $\bar{\mathbf{v}}_n = \sum_{k=1}^{\bar{K}} \bar{\mathbf{w}}_k\bar{h}_{kn}$ formed by the non-negative matrices $\bar{\mathbf{W}} \in \mathbb{R}_+^{M\times\bar{K}}$ and $\bar{\mathbf{H}} \in \mathbb{R}_+^{\bar{K}\times N}$. Then, for $K \geq \bar{K}$, the triplet $(\bar{\mathbf{\Phi}},\bar{\mathbf{W}},\bar{\mathbf{H}})$ is a global minimizer of c:

$$c(\mathbf{\Phi},\mathbf{W},\mathbf{H}) \geq c(\bar{\mathbf{\Phi}},\bar{\mathbf{W}},\bar{\mathbf{H}}),\ \forall(\mathbf{\Phi},\mathbf{W},\mathbf{H}). \quad (7)$$

The proof of Lemma 1 is given in Appendix A. In the GCM, each column $\mathbf{y}_n$ of $\mathbf{Y}$ follows $\bar{\mathbf{\Phi}}\mathbf{y}_n \sim \mathcal{N}(\mathbf{0},\mathrm{Diag}(\bar{\mathbf{v}}_n))$. Therefore the conditions (5)–(6) of Lemma 1 are always satisfied. Moreover, it is noteworthy to mention that Lemma 1 applies to a much broader class of signals than those following the GCM. For instance, when $\mathbf{Y}$ is formed by half-overlapping short-time windows, there does not exist a $\bar{\mathbf{\Phi}}$ for which $[\bar{\mathbf{\Phi}}\mathbf{Y}]_{mn}$ is independent of $[\bar{\mathbf{\Phi}}\mathbf{Y}]_{m'n'}$ for adjacent n and n'. Hence the independence assumption from the GCM under which the objective c has been derived is not satisfied. However, the conditions in Lemma 1 can still be satisfied, even if $\mathbf{Y}$ is non-Gaussian. Such examples shall be given in Section 3.

The condition (6) writes as a joint diagonalization of all the covariance matrices $\mathbb{E}(\mathbf{y}_n\mathbf{y}_n^\intercal)$, by a common basis $\bar{\mathbf{\Phi}}$. The identification of the rows of $\bar{\mathbf{\Phi}}$ is a well-studied problem in the literature of joint diagonalization [4]. However, the results are mostly about when $\mathbf{\Phi}$ equals to $\bar{\mathbf{\Phi}}$ up to a signed permutation. Below we go one step further by studying the identification of row-spaces of $\bar{\mathbf{\Phi}}$.

Theorem 1. *Let us partition the rows of $\mathbf{WH}$ into Q sets $\{\bar{B}_q\}_{q=1}^Q$ with the smallest possible Q, such that $\forall(m,m') \in (\bar{B}_q)^2$ and $\forall n \in \{1,\dots,N\}$, $[\bar{\mathbf{W}}\bar{\mathbf{H}}]_{m,n} = [\bar{\mathbf{W}}\bar{\mathbf{H}}]_{m',n}$ (partition into subsets of equal rows, if any). Then, under the conditions of Lemma 1 (i.e. (5), (6), and $K \geq \bar{K}$), $\forall q \in \{1,\dots,Q\}$ the rows of $\bar{\mathbf{\Phi}}$ indexed by $\bar{B}_q$ are identifiable.*

The proof of Theorem 1 is given in Appendix B. Theorem 1 coincides with Theorem 2 in [4] when $Q = M$. Our proof technique is different from [4] as we analyze the global minimizers of c, although they both rely on the orthogonality assumption of $\mathbf{\Phi}$.

Numerical illustration: Let us provide a numerical example to minimize c when $\mathbf{Y}$ follows the GCM with the ground-truth parameters $\bar{\mathbf{\Phi}}$, $\bar{\mathbf{W}}$, and $\bar{\mathbf{H}}$. In such case, the expectation in (4) is given by (see Appendix A)

$$\mathbb{E}(|\mathbf{\Phi Y}|^{\circ 2}) = (\mathbf{\Phi}\bar{\mathbf{\Phi}}^\intercal)^{\circ 2}\bar{\mathbf{W}}\bar{\mathbf{H}}. \quad (8)$$

Let us consider the case where $M = 4$, $N = 3$, $\bar{K} = K = 2$, $\bar{\mathbf{\Phi}}$ is a type-III DCT and

$$\bar{\mathbf{W}}\bar{\mathbf{H}} = \begin{pmatrix} 1 & 0.1 & 1.1 \\ 1 & 0.1 & 1.1 \\ 0.1 & 1 & 1.1 \\ 0.1 & 1 & 1.1 \end{pmatrix}. \quad (9)$$

Given the structure of $\bar{\mathbf{W}}\bar{\mathbf{H}}$ and Theorem 1, there are $Q = 2$ identifiable row-spaces of $\bar{\mathbf{\Phi}}$. The first (resp., second) one is the linear span of the first and second rows (resp., the third and fourth rows) of $\bar{\mathbf{\Phi}}$ (*i.e.*, $\bar{B}_1 = \{1,2\}$ and $\bar{B}_2 = \{3,4\}$). To minimize c, we use a slight modification (that reflects the GCM assumption instead of pure IS divergence) of the block-coordinate descent algorithm described in [6] (projected quasi-Newton step for $\mathbf{\Phi}$, majorization-minimization for $\mathbf{W}$ and $\mathbf{H}$). The algorithm starts from a random initialization and stops when sufficient numerical precision is reached. The estimated basis $\mathbf{\Phi}$ satisfies

$$\mathbf{\Phi}\bar{\mathbf{\Phi}}^\intercal = \begin{pmatrix} 0.703 & 0.711 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.955 & -0.297 \\ 0.711 & -0.703 & 0.000 & 0.000 \\ 0.000 & 0.000 & -0.297 & -0.955 \end{pmatrix}.$$

This shows that the first and second rows (resp., the third and forth row) of $\bar{\mathbf{\Phi}}$ are accurately identified with the first and third row (resp., the second and forth row) of $\mathbf{\Phi}$. We remark that the condition (6) in Lemma 1 implies that existing joint-diagonalization algorithms [7], [8], [9], [10] could also be applicable to identify $\bar{\mathbf{\Phi}}$ directly, prior to NMF.

III. Unsupervised Source Separation

We evaluate the performance of TL-NMF under the GCM in an unsupervised single-channel source separation problem. We study whether TL-NMF can learn a better $\mathbf{\Phi}$ than fixed DCT to improve source separation performance. The code to reproduce the numerical experiments reported in this section is available online.[1]

[1] to be released at publication.

a) Experimental Setting: We consider the signal $\mathbf{y}(t) = \sum_{i=1}^{I} \mathbf{y}^{(i)}(t)$ where $(\mathbf{y}^{(i)})_{i=1}^{I}$ are random sources defined as the sum of cosine functions shifted by a random phase to mimic musical notes with different timbres. More precisely, we set $\mathbf{y}^{(i)}(t) = \sum_{r=1}^{R} \beta_{i,r} \cos(r(\alpha_i t + \theta_i))\mathbf{g}_i(t)$ where $\theta_i \in [0, 2\pi)$ is an uniform random phase, $\alpha_i > 0$ a base frequency, r an integer exponent, $\beta_{i,r} > 0$ a timbre coefficient, and $\mathbf{g}_i \in \mathbb{R}^T$ a positive envelop that varies slowly over $t \in \{1, \dots, T\}$. Moreover, we assume that θ_i is independent of $\theta_{i'}$, for all $i' \neq i$. Our goal is then to assess the ability of TL-NMF to separate those I harmonic notes from the signal $\mathbf{y}$.

Let us analyse to which extent the assumptions of Lemma 1 are satisfied by the considered signal $\mathbf{y}$. Given a short-time window $\mathbf{a} \in \mathbb{R}^M$, the short-time matrix $\mathbf{Y}^{(i)}$ for the i-th source verifies $\mathbf{y}_n^{(i)}(m) = \mathbf{y}^{(i)}(m + \frac{Mn}{2})\mathbf{a}(m)$ (using zero-padding at signal boundaries). From the independence of the uniform random phases $\{\theta_i\}_{i=1}^{I}$, we get that $\mathbb{E}(\mathbf{y}_n) = \sum_i \mathbb{E}(\mathbf{y}_n^{(i)}) = \mathbf{0}$ for all $n \in \{1, \dots, N\}$. Similarly, $\mathbb{E}(\mathbf{y}_n \mathbf{y}_n^\intercal) = \sum_i \mathbb{E}(\mathbf{y}_n^{(i)} \mathbf{y}_n^{(i)\intercal})$ and its (m, m')-th element reads

$$\sum_{i=1}^{I} \frac{s_{m,m',n}^{(i)}}{2} \sum_{r=1}^{R} \beta_{i,r}^2 \cos(r\alpha_i(m - m')), \tag{10}$$

where $s_{m,m',n}^{(i)} = \mathbf{a}(m)\mathbf{a}(m')\mathbf{g}_i(m + \frac{Mn}{2})\mathbf{g}_i(m' + \frac{Mn}{2})$. Hence, the question is whether there exist $\bar{\mathbf{\Phi}}$, $\bar{\mathbf{W}}$, and $\bar{\mathbf{H}}$, such that $\mathbb{E}(\mathbf{y}_n \mathbf{y}_n^T)$ satisfies the joint-diagonalization condition (6). A general analysis from (10) is challenging. Instead, let us provide an intuition through the special case where $s_{m,m',n}^{(i)} = s_n^{(i)}$ (*i.e.*, independent from m and m') and, $\forall i$, $r\alpha_i = r\alpha = \pi p/M$ for $p \in \mathbb{N} \cap (0, M)$. Then, one easily gets that $\mathbb{E}(\mathbf{y}_n \mathbf{y}_n^\intercal) = \bar{\mathbf{\Phi}}^\intercal \mathrm{Diag}(\bar{\mathbf{v}}_n) \bar{\mathbf{\Phi}}$ for

$$\begin{aligned} \bar{\mathbf{v}}_n &= \textstyle\sum_{i=1}^{I} \frac{M s_n^{(i)}}{4} (\beta_{i,1}^2, \beta_{i,1}^2, \cdots, \beta_{i,R}^2, \beta_{i,R}^2, 0 \cdots 0)^\intercal \in \mathbb{R}_+^M, \\ \bar{\mathbf{\Phi}} &= (\mathbf{c}_1, \mathbf{s}_1, \cdots, \mathbf{c}_R, \mathbf{s}_R, * \cdots *)^\intercal \in \mathbb{R}^{M \times M}, \end{aligned} \tag{11}$$

where $\mathbf{c}_r(m) = \sqrt{2/M} \cos(r\alpha m + \varphi_r)$ and $\mathbf{s}_r(m) = \sqrt{2/M} \sin(r\alpha m + \varphi_r)$ for some phase shift $\varphi_r \in [0, 2\pi)$. The last $M - 2R$ atoms are arbitrary as long as $\bar{\mathbf{\Phi}}$ is orthogonal. Hence, the assumptions of Lemma 1 are satisfied with $\bar{K} = I$.[2]

In this work, we consider a scenario with $I = 2$ sources that are generated according to the aforementioned model with $R = 2$, $\beta_{1,r} = \beta_{2,r} = 0.5^r$, as well as the base periods $\frac{2\pi}{\alpha_1} = \frac{40}{\pi}$ms and $\frac{2\pi}{\alpha_2} = \frac{25}{\pi}$ms, respectively. The signal duration is 3s with sampling rate 5000Hz, which is large enough to avoid aliasing. One realization of such a signal is depicted in Figure 1, where one can also see the shape of the used envelopes $\mathbf{g}_i$. To deploy our TL-NMF method, we set $\mathbf{a}$ to be a Tukey window [11] with a cosine fraction parameter set to 0.1. The duration of the window is fixed to 40ms ($M = 200$). As such, each segment covers around 3 or 5 periods of each source. Although this configuration does not follow the (ideal) assumptions we made to derive (11), the covariance matrix $\mathbb{E}(\mathbf{y}_n \mathbf{y}_n^\intercal)$ is still close to be circulant Toeplitz. Indeed, given the very slow variations of the envelopes $\mathbf{g}_i$ with respect to the short duration of the Tukey window $\mathbf{a}(m) \approx 1$, we have $s_{m,m',n}^{(i)} \approx \mathbf{g}_i(\frac{Mn}{2})^2$. Hence, we expect that TL-NMF learns a transform $\mathbf{\Phi}$ containing atoms alike $\cos(r\alpha_i m + \varphi_{r,i})$ and $\sin(r\alpha_i m + \varphi_{r,i})$. As $s_{m,m',n}^{(i)}$ depends on $\mathbf{a}$, the choice of the window will have an impact on the learnt basis $\mathbf{\Phi}$. This phenomenon has been previously observed in [1], [2], [12].

[2]When $\bar{\mathbf{v}}_n$ contains zeros as in (11), we add a small constant ϵ to each $[\mathbf{\Phi Y}]_{mn}$ and $[\mathbf{WH}]_{mn}$ in the objective (4). Here, we set $\epsilon = 5 \times 10^{-7}$.

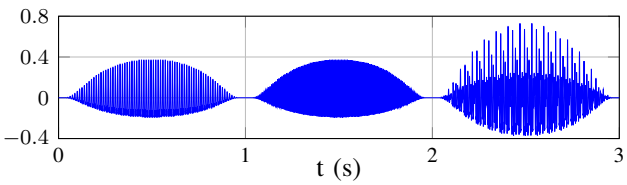

Fig. 1: One random realization $\mathbf{y}$ of a sum of two harmonic processes. Source $\mathbf{y}_1(t)$ (resp., $\mathbf{y}_2(t)$) is nonzero for $t \in [0, 1] \cup [2, 3]$ (resp., $t \in [1, 3]$).

b) Estimation: Unlike the ideal case where c can be computed from (4) and (8), we generate 10 i.i.d realizations of the mixed signal $\mathbf{y}$ to construct an empirical estimate of $\mathbb{E}(|\mathbf{\Phi Y}|^{\circ 2})$, and thus of c. This allows us to apply the numerical method deployed in II.B to optimize $\mathbf{\Phi}$, $\mathbf{W}$, and $\mathbf{H}$. As a baseline, we consider IS-NMF under a fixed type-III DCT basis. It amounts to minimize the estimated c with respect to $\mathbf{W}$ and $\mathbf{H}$ under the fixed $\mathbf{\Phi}$.

c) Results: Figure 2 displays the NMF of $\mathbb{E}(|\mathbf{\Phi Y}|^{\circ 2})$ using $K = 2$. We see that the columns $\mathbf{w}_k$ of $\mathbf{W}$ are better separated under the $\mathbf{\Phi}$ from TL-NMF than under the DCT. Each row $\underline{\mathbf{h}}_k$ of $\mathbf{H}$ varies also smoother over n, following the envelope $\mathbf{g}_i(t)$ of each source. The eight rows of $\mathbf{\Phi}$ with largest energy $\sum_n \mathbb{E}([\mathbf{\Phi Y}]_{mn}^2)$ are depicted in Figure 3. It illustrates that the TL-NMF has found atoms alike $\cos(r\alpha_i t + \varphi_{i,r})$ or $\sin(r\alpha_i t + \varphi_{i,r})$, as expected.

Finally, we apply standard Wiener filtering to separate each realization of $\mathbf{y}$ into $K = 2$ sources. It involves computing two masks from the estimated $\mathbf{W}$ and $\mathbf{H}$ to be applied to $\mathbf{\Phi Y}$ prior to overlap-add reconstruction, see [3]. Quality of separation is evaluated with BSS_Eval 2.1 [13]. We evaluate the separation of the ten realizations of $\mathbf{y}$ and average the BSS_Eval metrics. We then repeat this experiment 5 times and report the mean values with standard deviation of the metrics in Table I. They show that the performance is significantly improved over DCT, thanks to the learning of a more meaningful basis $\mathbf{\Phi}$.

	source	SDR	SIR	SAR
IS-NMF	1	16.00 (0.04)	22.64 (0.11)	17.10 (0.07)
	2	16.13 (0.07)	23.53 (0.17)	17.02 (0.05)
TL-NMF	1	22.53 (0.16)	40.20 (0.99)	22.61 (0.18)
	2	22.52 (0.17)	39.00 (0.23)	22.62 (0.17)

TABLE I: BSS_Eval mean values with standard deviation (in bracket).

IV. Conclusion

In this paper, we have studied the identifiability of transform learning for NMF under the Gaussian composite model (GCM). By minimizing the expected negative log-likelihood, we prove that it is possible to uniquely identify the row-spaces of $\bar{\mathbf{\Phi}}$. This identifiability result is further supported by a source separation example which demonstrates the ability of TL-NMF to separate musical notes with close base frequencies. We may extend our identifiability results to the more general ILRTA [2] in future work.

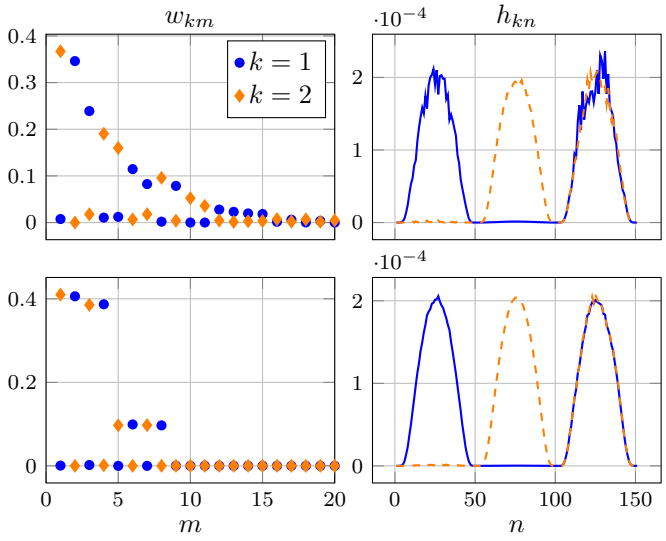

Fig. 2: NMF factors $\mathbf{W}$ and $\mathbf{H}$ of $\mathbb{E}(|\mathbf{\Phi Y}|^{\circ 2})$. Top: with DCT. Bottom: with $\mathbf{\Phi}$ from TL-NMF. The m-axis is reordered with decreasing energy $\sum_n \mathbb{E}([\mathbf{\Phi Y}]^2_{mn})$. We zoom in the m-axis where w_{km} is non-negligible.

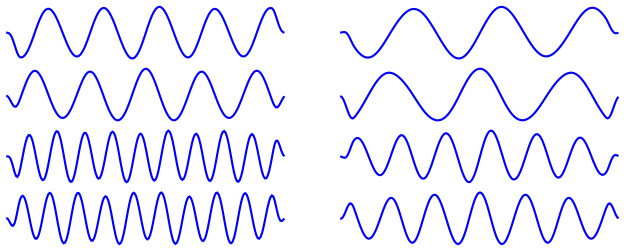

Fig. 3: The eight most significant atoms of $\mathbf{\Phi}$ from TL-NMF.

APPENDIX

A. Proof of Lemma 1

To show that $c(\mathbf{\Phi}, \mathbf{W}, \mathbf{H}) \geq c(\bar{\mathbf{\Phi}}, \bar{\mathbf{W}}, \bar{\mathbf{H}})$, we first minimize c with respect to $\mathbf{W}, \mathbf{H}$. For a fixed orthogonal $\mathbf{\Phi} \in \mathbb{R}^{M \times M}$, let us first show that

$$c^*(\mathbf{\Phi}) = \min_{\mathbf{W},\mathbf{H}} c(\mathbf{\Phi}, \mathbf{W}, \mathbf{H}) = \sum_{mn} (1 + \log \mathbb{E}([\mathbf{\Phi Y}]^2_{mn})), \quad (12)$$

and that this minimal value is attained at any point $(\mathbf{W}, \mathbf{H})$ that verifies $\mathbf{WH} = \mathbb{E}(|\mathbf{\Phi Y}|^{\circ 2})$. To that end, it is sufficient to show that, under conditions (5)–(6), there exists an exact NMF factorization $\mathbf{WH}$ of $\mathbb{E}(|\mathbf{\Phi Y}|^{\circ 2})$ for any $\mathbf{\Phi}$. Let $\mathbf{D} = \mathbf{\Phi}\bar{\mathbf{\Phi}}^\intercal$, so that $\mathbf{\Phi} = \mathbf{D}\bar{\mathbf{\Phi}}$. Denote $\bar{\mathbf{V}} = \bar{\mathbf{W}}\bar{\mathbf{H}}$. From (5) and (6) we have

$$\mathbb{E}([\mathbf{\Phi Y}]^2_{mn}) = \mathbb{E}\big(\big(\sum_{m'} d_{mm'} [\bar{\mathbf{\Phi}}\mathbf{Y}]_{m'n}\big)^2\big) \quad (13)$$

$$= \sum_{m'} d^2_{mm'} \bar{v}_{m'n} \quad (14)$$

Hence, because all the terms in (14) are non-negative, we have derived an exact NNF factorization of $\mathbb{E}(|\mathbf{\Phi Y}|^{\circ 2})$ which proves (12). As $\mathbf{\Phi}$ is orthogonal, one can verify that the objective $c^*(\mathbf{\Phi})$ is equivalent to the one used in [8], [10].

Then, it follows that

$$c^*(\mathbf{\Phi}) = \sum_{mn} \log([(\mathbf{\Phi}\bar{\mathbf{\Phi}}^\intercal)^{\circ 2}\bar{\mathbf{V}}]_{mn}) + NM. \quad (15)$$

It remains to prove that $\min_{\mathbf{\Phi}} c^*(\mathbf{\Phi})$ is attained at $\mathbf{\Phi} = \bar{\mathbf{\Phi}}$. Using the fact that $\mathbf{D} = \mathbf{\Phi}\bar{\mathbf{\Phi}}^\intercal$ is orthogonal and thus that both the columns and the rows of $\mathbf{D}^{\circ 2}$ sum to one, we obtain from Jensen's inequality

$$\log\Big(\sum_{m'} d^2_{mm'} \bar{v}_{m'n}\Big) \geq \sum_{m'} d^2_{mm'} \log(\bar{v}_{m'n}). \quad (16)$$

This implies the following lower bound on $c^*(\mathbf{\Phi})$

$$c^*(\mathbf{\Phi}) \geq \sum_{mn}\sum_{m'} d^2_{mm'} \log(\bar{v}_{m'n}) + NM. \quad (17)$$

$$= \sum_{m'n} \log(\bar{v}_{m'n}) + NM = c(\bar{\mathbf{\Phi}}, \bar{\mathbf{W}}, \bar{\mathbf{H}}), \quad (18)$$

and completes the proof. □

B. Proof of Theorem 1

Let $(\mathbf{\Phi}, \mathbf{W}, \mathbf{H})$ be a global minimizer of c such that $c(\mathbf{\Phi}, \mathbf{W}, \mathbf{H}) = c(\bar{\mathbf{\Phi}}, \bar{\mathbf{W}}, \bar{\mathbf{H}})$. Following the results in Lemma 1, we are going to show that $\forall q \in \{1, \dots, Q\}$, there exists a partition $\{B_q\}_{q=1}^Q$ of $\{1, \dots, M\}$ such that $\mathbf{span}_{B_q}(\mathbf{\Phi}) = \mathbf{span}_{\bar{B}_q}(\bar{\mathbf{\Phi}})$. By definition of $\bar{B}_q$, we denote $\hat{v}_{qn} = \bar{v}_{mn}$ for $m \in \bar{B}_q$. Hence $\forall q' \neq q$, there exists $n \in \{1, \dots, N\}$ such that $\hat{v}_{qn} \neq \hat{v}_{q'n}$. As $c(\mathbf{\Phi}, \mathbf{W}, \mathbf{H}) = c(\bar{\mathbf{\Phi}}, \bar{\mathbf{W}}, \bar{\mathbf{H}})$, the equality holds in (16) and we can write it as

$$\forall n, \quad \log\Big(\sum_{q=1}^{Q} \hat{d}^2_{mq} \hat{v}_{qn}\Big) = \sum_{q=1}^{Q} \hat{d}^2_{mq} \log(\hat{v}_{qn}). \quad (19)$$

where $\hat{d}^2_{mq} = \sum_{m' \in \bar{B}_q} d^2_{mm'}$. We will show at the end of the proof that $\forall m \in \{1, \dots, M\}$ there exists $\tau(m) \in \{1, \dots, Q\}$ such that

$$\hat{d}^2_{mq} = \begin{cases} 1 & \text{if } q = \tau(m), \\ 0 & \text{if } q \neq \tau(m). \end{cases} \quad (20)$$

From (20), we can construct $B_q = \{m : \tau(m) = q, 1 \leq m \leq M\}$ so that

$$\hat{d}^2_{mq} = \sum_{m' \in \bar{B}_q} d^2_{mm'} = 1 \quad \text{if} \quad m \in B_q. \quad (21)$$

By the definition of $d_{mm'}$ and the orthogonality of $\mathbf{\Phi}$ and $\bar{\mathbf{\Phi}}$, (21) implies that each row of $\mathbf{\Phi}$ in the set B_q belong to $\mathbf{span}_{\bar{B}_q}(\bar{\mathbf{\Phi}})$. Therefore $\mathbf{span}_{B_q}(\mathbf{\Phi}) \subset \mathbf{span}_{\bar{B}_q}(\bar{\mathbf{\Phi}})$.

To show that $\mathbf{span}_{B_q}(\mathbf{\Phi}) = \mathbf{span}_{\bar{B}_q}(\bar{\mathbf{\Phi}})$, we first check that

$$\forall q, \quad |B_q| = \sum_{m=1}^{M} \hat{d}^2_{mq} = \sum_{m' \in \bar{B}_q} \sum_{m=1}^{M} d^2_{mm'} = |\bar{B}_q|. \quad (22)$$

From (22), we also obtain a similar formula as (21)

$$\sum_{m \in B_q} d^2_{mm'} = 1 \quad \text{if} \quad m' \in \bar{B}_q. \quad (23)$$

Therefore $\mathbf{span}_{\bar{B}_q}(\bar{\mathbf{\Phi}}) \subset \mathbf{span}_{B_q}(\mathbf{\Phi})$.

Finally we show that (20) is correct by absurd. Assume that (20) is not true. Hence there exist m and $q' \neq q$, such that

$$\hat{d}^2_{mq} \in (0, 1) \text{ and } \hat{d}^2_{mq'} \in (0, 1). \quad (24)$$

Then, the fact that $\sum_q \hat{d}^2_{mq} = \sum_{m'} d^2_{mm'} = 1$ together with the strict concavity of the log function imply that the equality (19) can hold only if $\hat{v}_{qn} = \hat{v}_{q'n}, \forall n$. This contradicts the fact that $\bar{B}_q \cap \bar{B}_{q'} = \emptyset$ and completes the proof. □

REFERENCES

[1] D. Fagot, H. Wendt, and C. Févotte, "Nonnegative matrix factorization with transform learning," in *Proc. ICASSP*, Calgary, Canada, 2018, pp. 2431–2435.

[2] K. Yoshii, K. Kitamura, Y. Bando, E. Nakamura, and T. Kawahara, "Independent Low-Rank Tensor Analysis for Audio Source Separation," in *Proc. EUSIPCO*, Rome, Italy, 2018, pp. 1657–1661.

[3] P. Smaragdis, C. Févotte, G. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 66–75, May 2014.

[4] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. signal Process.*, vol. 45, no. 2, pp. 434–444, Feb. 1997.

[5] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Proc. Neurips*, Vancouver, Canada, 2004, pp. 1141–1148.

[6] P. Ablin, D. Fagot, H. Wendt, A. Gramfort, and C. Févotte, "A quasi-Newton algorithm on the orthogonal manifold for NMF with transform learning," in *Proc. ICASSP*, Brighton, United Kingdom, 2019, pp. 700–704.

[7] J.-F. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization," *SIAM J. Matrix Anal. Appl.*, vol. 17, no. 1, pp. 161–164, Jan. 1996.

[8] D. T. Pham, "Joint approximate diagonalization of positive definite Hermitian matrices," *SIAM J. Matrix Anal. Appl.*, vol. 22, no. 4, pp. 1136–1152, May 2001.

[9] A. Ziehe, P. Laskov, G. Nolte, and K.-R. Müller, "A Fast Algorithm for Joint Diagonalization with Non-Orthogonal Transformations and Its Application to Blind Source Separation," *J. Mach. Learn. Res.*, vol. 5, no. 1, pp. 777–800, Jul. 2004.

[10] P. Ablin, J.-F. Cardoso, and A. Gramfort, "Beyond Pham's algorithm for joint diagonalization," *Proc. ESANN*, pp. 607–612, 2019.

[11] P. Bloomfield, *Fourier analysis of time series: an introduction*. John Wiley & Sons, 2004, p. 69.

[12] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto, "Beyond NMF: Time-Domain Audio Source Separation without Phase Reconstruction." in *Proc. ISMIR*, Curitiba, Brazil, 2013, pp. 369–374.

[13] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.