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 [1], [2]. 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. Using such off-the-shelf transforms can be restrictive and many authors have considered learning adaptive transforms in various settings, e.g., [3], [4], [5]. In particular, transform-learning NMF (TL-NMF) was proposed with the goal of learning an adaptive (short-time) orthogonal transform together with NMF factors [6]. TL-NMF is also a special case of independent low-rank tensor analysis (ILRTA) [7] which offers a general framework for short-time modeling of temporal sequences under composite covariance models. In both TL-NMF and ILRTA, the parameters of the model (e.g. the orthogonal transform and the NMF factors for TL-NMF) are estimated through the minimization of a non-convex objective function [7], [8]. The question of whether the ground-truth parameters of the model correspond to a global minimizer of the objective function (*i.e.*, the identifiability of the parameters) has not been studied so far.

In this work, we study the identifiability of the orthogonal transform in TL-NMF under the Gaussian composite model (GCM) [1]. 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) [1] and in particular in [6]. In Section 2.2, we establish conditions under which the row-spaces of the orthogonal transform in TL-NMF (*i.e.*, the linear subspaces generated by subsets of its rows) are identifiable. These conditions generalize prior results on the identifiability of joint-diagonalization [9]. 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 write $\phi(\mathbf{y}) = \mathbf{\Phi Y}$. In this paper, we study how to identify $\mathbf{\Phi}$ given observations of $\mathbf{y}$ that satisfy (1). The problem of how to identify the $\mathbf{W}$ and $\mathbf{H}$ belongs to the identifiability of NMF [10]. In the next section, we derive a new objective (slightly different to [6]) 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 [1]. 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

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

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

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

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 the identifiability in the sense of Definition 1.

Definition 1. *Let $\boldsymbol{span}_{\bar{B}}(\bar{\mathbf{\Phi}})$ (resp., $\boldsymbol{span}_{B}(\mathbf{\Phi})$) 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 $\boldsymbol{span}_{\bar{B}}(\bar{\mathbf{\Phi}}) = \boldsymbol{span}_{B}(\mathbf{\Phi})$.*

We first give sufficient conditions under which every triplet $(\bar{\mathbf{\Phi}},\mathbf{W}^\star,\mathbf{H}^\star)$ with $\mathbf{W}^\star\mathbf{H}^\star = \bar{\mathbf{W}}\bar{\mathbf{H}}$ is a global minimizer 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 $K \geq \bar{K}$, $\mathbf{W}^\star \in \mathbb{R}_+^{M\times K}$, and $\mathbf{H}^\star \in \mathbb{R}_+^{K\times N}$ such that $\mathbf{W}^\star\mathbf{H}^\star = \bar{\mathbf{W}}\bar{\mathbf{H}}$, the triplet $(\bar{\mathbf{\Phi}},\mathbf{W}^\star,\mathbf{H}^\star)$ is a global minimizer of c:

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

The proof of Lemma 1 is deferred to the supplementary material. 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 [9]. 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 $\bar{\mathbf{W}}\bar{\mathbf{H}}$ 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 deferred to the supplementary material. Theorem 1 coincides with Theorem 2 in [9] when $Q = M$. However, our result derives from the analysis of the global minimizers of c while in [9] the authors study a different objective function.

Numerical illustration: We illustrate the identifiability of $\bar{\mathbf{\Phi}}$ through the minimization of 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 the proof of Lemma 1)

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

Let us consider the case where $M = 4$, $N = 3$, $\bar{K} = K = 2$, $\bar{\mathbf{\Phi}}$ is the DCT (type-II) 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 [8] (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.[1] The estimated basis $\mathbf{\Phi}$ satisfies

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

[1]The code to reproduce the numerical experiments reported in this paper is available online (https://github.com/sixin-zh/tlnmf-gcm).

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, such as [11], [12], [13], [14], 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 an adaptive $\mathbf{\Phi}$ to improve source separation performance.

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 $\mathbf{y}^{(i)}(t) = \sum_{r=1}^{R} \beta_{i,r} \cos(r(2\pi \frac{f_i}{f_0} t + \theta_i))\mathbf{g}_i(t)$ where $\theta_i \in [0, 2\pi)$ is an uniform random phase, $f_i > 0$ a fundamental frequency, $f_0 > 0$ is the sampling frequency, r an integer number, $\beta_{i,r} > 0$ 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.

In this section, 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 fundamental frequency $f_1 = 440$Hz (corresponding to musical note A4) and $f_2 = 466.16$Hz (A4#), respectively. We used $T = 15000$ and $f_0 = 5000$Hz, leading to a signal $\mathbf{y}$ of duration 3s. This sampling rate 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 $\mathbf{g}_i$. Finally, to deploy our TL-NMF method, we consider a Tukey short-time window $\mathbf{a}$ [15] of length $M = 200$ (duration $d_M = 40$ms), with a cosine fraction parameter set to 0.1.

B. Theoretical Analysis

Let us analyse to which extent the assumptions of Lemma 1 are satisfied by the considered signal $\mathbf{y}$. The short-time matrix $\mathbf{Y}^{(i)}$ for the i-th source verifies $\mathbf{y}_n^{(i)}(m) = \mathbf{y}^{(i)}(m + \frac{Mn}{2} - M)\mathbf{a}(m)$ (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

$$\sum_{i=1}^{I} \frac{s_{m,m',n}^{(i)}}{2} \sum_{r=1}^{R} \beta_{i,r}^2 \cos\left(2\pi r \frac{f_i}{f_0}(m - m')\right), \qquad (10)$$

where $s_{m,m',n}^{(i)} = \mathbf{a}(m)\mathbf{a}(m')\mathbf{g}_i(m + \frac{Mn}{2} - M)\mathbf{g}_i(m' + \frac{Mn}{2} - M)$. 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 [6], [7], [16].

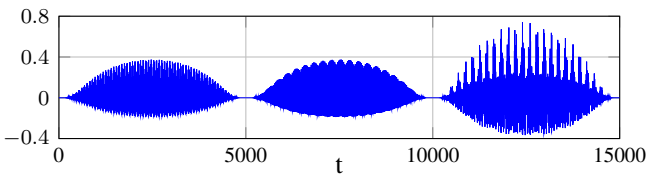

Fig. 1: One random realization $\mathbf{y}$ of a sum of two harmonic processes. Source $\mathbf{y}_1(t)$ (resp., $\mathbf{y}_2(t)$) is nonzero for $t \in \{1, \dots, 5000\} \cup \{10001, \dots, 15000\}$ (resp., $t \in \{5001, \dots, 15000\}$).

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.*, independence from m and m') and, $\forall i$, $2\pi r \frac{f_i}{f_0} = 2\pi r \frac{f}{f_0} = \pi p / M$ 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

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

where $\mathbf{c}_r(m) = \sqrt{2/M} \cos(2\pi r \frac{f}{f_0} m + \varphi_r)$ and $\mathbf{s}_r(m) = \sqrt{2/M} \sin(2\pi r \frac{f}{f_0} m + \varphi_r)$ 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 $\bar{K} = I$.

Although the configuration in Section III-A 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 $\mathbf{g}_i$ and the short duration of the Tukey window, we have $\mathbf{a}(m) \approx 1$ and $\mathbf{g}_i(m + \frac{Mn}{2} - M) \approx \mathbf{g}_i(\frac{Mn}{2} - M)$, which lead to $s_{m,m',n}^{(i)} \approx \mathbf{g}_i(\frac{Mn}{2} - M)^2$. Hence, we expect that TL-NMF learns a transform $\mathbf{\Phi}$ containing signal-adapted atoms alike $\cos(2\pi r \frac{f_i}{f_0} \cdot + \varphi_{r,i})$ and $\sin(2\pi r \frac{f_i}{f_0} \cdot + \varphi_{r,i})$. In contrast, DCT atoms are fixed and defined by the frequencies $\{\frac{r}{2M} f_0 = \frac{r}{2d_M}\text{Hz}\}_{r=0}^{M-1}$ which do not contain the frequencies of the pure harmonics that compose the signal $\mathbf{y}$. It is noteworthy to mention that increasing the window size d_M to refine the frequency grid of the DCT would come at the price of a worse time-localization of the music notes.

C. Numerical 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 Section II.B to optimize $\mathbf{\Phi}$, $\mathbf{W}$, and $\mathbf{H}$. As a baseline, we consider IS-NMF under a fixed DCT basis, which we refer to as DCT-NMF. It amounts to minimize the empirical estimate of c w.r.t $\mathbf{W}$ and $\mathbf{H}$ under the fixed $\mathbf{\Phi}$. The parameters $(\mathbf{\Phi}, \mathbf{W}, \mathbf{H})$ of TL-NMF and $(\mathbf{W}, \mathbf{H})$ of DCT-NMF are initialized randomly. Both methods are used with $K = 2$. To reduce numerical instabilities, we add a small constant ϵ to each $[\mathbf{\Phi Y}]_{mn}$ and $[\mathbf{WH}]_{mn}$ in the objective (4). Here, we set $\epsilon = 5 \times 10^{-7}$.

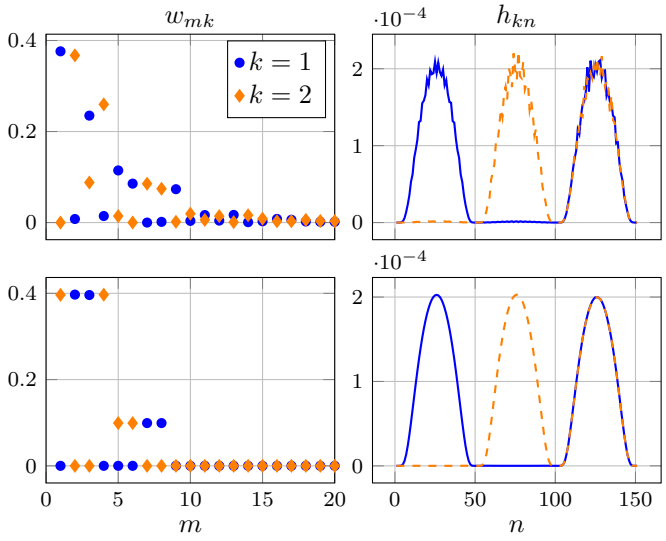

Fig. 2: NMF factors $\mathbf{W}$ and $\mathbf{H}$ of $\mathbb{E}(|\mathbf{\Phi Y}|^{\circ 2})$. Top: DCT-NMF. Bottom: TL-NMF. The m-axis is restricted to where w_{km} is non-negligible ($m = 1, \ldots, 20$).

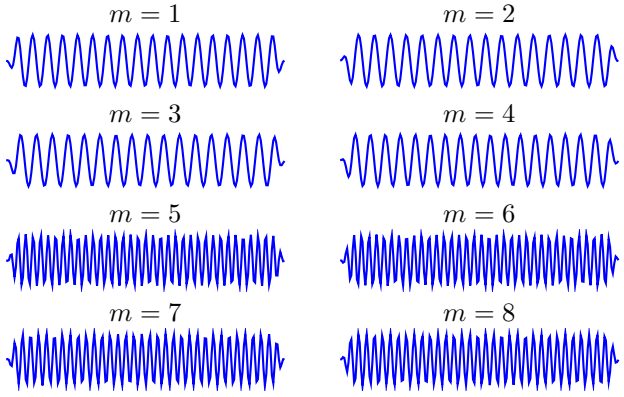

Fig. 3: The eight most significant atoms of $\mathbf{\Phi}$ from TL-NMF.

D. Results and Discussion

a) Adaptivity of TL-NMF: Figure 2 displays the NMF of $\mathbb{E}(|\mathbf{\Phi Y}|^{\circ 2})$ using TL-NMF and DCT-NMF. For TL-NMF, the m-axis is reordered by decreasing influence of the atoms, as measured by the energy $\sum_n \mathbb{E}([\mathbf{\Phi Y}]_{mn}^2)$. 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 $\mathbf{H}$ also varies more smoothly w.r.t n, following the envelope $\mathbf{g}_i(t)$ of each source. The eight rows of $\mathbf{\Phi}$ with largest influence are depicted in Figure 3. It illustrates that TL-NMF has found atoms alike $\cos(2\pi r \frac{f_i}{f_0} \cdot + \varphi_{i,r})$ or $\sin(2\pi r \frac{f_i}{f_0} \cdot + \varphi_{i,r})$, as expected. The frequencies of these atoms are reported in Table I. They have been estimated through a nonlinear least-square regression of the learnt atoms with the harmonic model $(a, f, \theta) \mapsto a\cos(2\pi \frac{f}{f_0} \cdot + \theta)$. To increase the robustness to local minimizers, we used 300 randomized initializations of BFGS and kept the best fit. For comparison, the frequencies of the 16 most significant DCT atoms are also reported in Table I. One can see that these DCT atoms are close to but not concentrated at the fundamental frequencies of the notes A4 (440Hz) and A4# (466.16Hz) that compose the signal $\mathbf{y}$. In contrast, the atoms learnt by TL-NMF are perfectly adapted to the input signal as they recover the frequencies of the two notes. This adaptivity explains that a more discriminative NMF (sparser $\mathbf{W}$ and smoother $\mathbf{H}$) can be obtained by using the learnt $\mathbf{\Phi}$ rather than the DCT (Figure 2).

TABLE I: Frequencies of the DCT and TL-NMF atoms (in the same order as in Figure 2 and Figure 3).

m	DCT	m	DCT	m	TL-NMF
1	437.5	9	887.5	1	**466.37**
2	462.5	10	487.5	2	**440.00**
3	450	11	400	3	**439.85**
4	475	12	500	4	**466.09**
5	425	13	862.5	5	932.39
6	875	14	950	6	932.33
7	937.5	15	912.5	7	879.93
8	925	16	900	8	880.00

TABLE II: BSS_Eval mean values with standard deviation (in bracket). The SDR is a general measure of performance. The SIR and SAR measure the influence of interferences and artifacts, respectively [17]. Decibel values, the higher, the better.

	source	SDR	SIR	SAR
DCT-NMF	1	14.37 (0.02)	19.98 (0.05)	15.80 (0.01)
	2	14.36 (0.00)	19.95 (0.05)	15.81 (0.02)
TL-NMF	1	32.42 (0.09)	40.69 (0.06)	33.12 (0.10)
	2	32.41 (0.10)	40.67 (0.07)	33.12 (0.10)

b) Source Separation Performance: We apply standard Wiener filtering to separate each (of the ten) realizations of $\mathbf{y}$ into $K = 2$ sources. It involves the computation of two masks from the estimated $\mathbf{W}$ and $\mathbf{H}$ to be applied to $\mathbf{\Phi Y}$ prior to overlap-add reconstruction, see [1]. The quality of the separation is evaluated with the standard toolbox BSS_Eval 2.1 [17]. 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 II. They show that the performance is significantly improved over the DCT, thanks to the learning of a more discriminative basis $\mathbf{\Phi}$. We also found that other types of DCT (I, III, IV) lead to similar BSS_Eval metrics than those obtained with the DCT in Table II.

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 the orthogonal transform. This identifiability result is further supported by a source separation example which demonstrates the ability of TL-NMF to separate musical notes with close fundamental frequencies. This is made possible because the atoms learnt by TL-NMF can precisely adapt to the frequencies of the musical notes that compose the mixed signal.

REFERENCES

[1] 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.

[2] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: two converging routes to ILRMA originating from ICA and NMF," *APSIPA Trans. Signal Inf. Process.*, vol. 8, no. 1, p. e12, May 2019.

[3] S. Abdallah and M. Plumbley, "If the independent components of natural images are edges, what are the independent components of natural sounds?" in *Proc. ICA*, San Diego, USA, 2001, pp. 534–539.

[4] S. Ravishankar and Y. Bresler, "Learning Sparsifying Transforms," *IEEE Trans. Signal Process.*, vol. 61, no. 5, pp. 1072–1086, Mar. 2013.

[5] P. Smaragdis and S. Venkataramani, "A neural network alternative to non-negative audio models," in *Proc. ICASSP*, New Orleans, USA, 2017, pp. 86–90.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

[16] 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.

[17] 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.