On Identifiability of Nonnegative Matrix Factorization

Xiao Fu [ID], Kejun Huang [ID], and Nicholas D. Sidiropoulos [ID]

Abstract—In this letter, we propose a new identification criterion that guarantees the recovery of the low-rank latent factors in the nonnegative matrix factorization (NMF) generative model, under mild conditions. Specifically, using the proposed criterion, it suffices to identify the latent factors if the rows of one factor are *sufficiently scattered* over the nonnegative orthant, while no structural assumption is imposed on the other factor except being full-rank. This is by far the mildest condition under which the latent factors are provably identifiable from the NMF model.

Index Terms—Convex analysis, identifiability, nonnegative matrix factorization (NMF), sufficiently scattered.

I. INTRODUCTION

NONNEGATIVE matrix factorization (NMF) [1], [2] aims to decompose a data matrix into low-rank latent factor matrices with nonnegativity constraints on (one or both of) the latent matrices. In other words, given a data matrix $\boldsymbol{X} \in \mathbb{R}^{M \times N}$ and a targeted rank r, NMF tries to find a factorization model $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H}^{\top}$, where $\boldsymbol{W} \in \mathbb{R}^{M \times r}$ and/or $\boldsymbol{H} \in \mathbb{R}^{N \times r}$ take only nonnegative values and $r \leq \min\{M, N\}$.

One notable trait of NMF is model identifiability—the latent factors are uniquely identifiable under some conditions (up to some trivial ambiguities) if the data are truly generated by $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H}^{\top}$. Identifiability is critical in parameter estimation and model recovery. In signal processing, many NMF-based approaches have, therefore, been proposed to handle problems such as blind source separation [3], spectrum sensing [4], and hyperspectral unmixing [5], [6], where model identifiability plays an essential role. In machine learning, identifiability of NMF is also considered essential for applications such as latent mixture model recovery [7], topic mining [8], and social network clustering [9], where model identifiability is entangled with interpretability of the results.

Despite the importance of identifiability in NMF, the analytical understanding of this aspect is still quite limited, and many existing identifiability conditions for NMF are not satisfactory in some sense. Donoho and Stodden [10], Laurberg *et al.* [11], and Huang *et al.* [12] have proven different sufficient conditions for identifiability of NMF, but these conditions all require that both of the generative factors $\boldsymbol{W}$ and $\boldsymbol{H}$ exhibit certain sparsity patterns or properties. The machine learning and remote sensing communities have proposed several factorization criteria and algorithms that have identifiability guarantees, but these methods heavily rely on the so-called *separability condition* [8], [13]–[18]. The separability condition essentially assumes that there is a (scaled) permutation matrix in one of the two latent factors as a submatrix, which is clearly restrictive in practice. Recently, Fu *et al.* [3] and Lin *et al.* [19] proved that the so-called volume minimization (VolMin) criterion can identify $\boldsymbol{W}$ and $\boldsymbol{H}$ without any assumption on one factor (say, $\boldsymbol{W}$) except being full-rank when the other ($\boldsymbol{H}$) satisfies a condition which is much milder than separability. However, the caveat is that VolMin also requires that each row of the nonnegative factor sums up to 1. This assumption implies loss of generality and is not satisfied in many applications.

In this letter, we reveal a new identifiablity result for NMF, which is obtained from a delicate tweak of the VolMin identification criterion. Specifically, we "shift" the sum-to-one constraint on $\boldsymbol{H}$ from its rows to its columns. As a result, we show that this "constraint-altered VolMin criterion" identifies $\boldsymbol{W}$ and $\boldsymbol{H}$ with provable guarantees under conditions that are much more easily satisfied relative to VolMin. This interesting tweak is seemingly slight, yet the result is significant: putting sum-to-one constraints on the columns (instead of rows) of $\boldsymbol{H}$ is without loss of generality, since the bilinear model $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H}^{\top}$ can always be rewritten as $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{D}^{-1}(\boldsymbol{H}\boldsymbol{D})^{\top}$, where $\boldsymbol{D}$ is a full-rank diagonal matrix satisfying $\boldsymbol{D}_{r,r} = 1/\|\boldsymbol{H}_{:,r}\|_1$ and $\boldsymbol{H}_{:,r}$ is the rth column of $\boldsymbol{H}$. Our new result is the only identifiability condition that does not assume *any other* structure beyond the target rank on $\boldsymbol{W}$ (e.g., zero pattern or nonnegativity) and has natural assumptions on $\boldsymbol{H}$ (relative to the restrictive row sum-to-one assumption as in VolMin).

II. BACKGROUND

To facilitate our discussion, let us formally define identifiability of constrained matrix factorization.

Definition 1 (Identifiability): Consider a data matrix that is generated from the model $\boldsymbol{X} = \boldsymbol{W}_{\natural}\boldsymbol{H}_{\natural}^{\top}$, where $\boldsymbol{W}_{\natural}$ and $\boldsymbol{H}_{\natural}$ are the ground-truth factors. Let $(\boldsymbol{W}_{\star}, \boldsymbol{H}_{\star})$ be an optimal solution from an identification criterion

$$(\boldsymbol{W}_{\star}, \boldsymbol{H}_{\star}) = \underset{\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H}^{\top}}{\arg\min} \; g(\boldsymbol{W}, \boldsymbol{H}).$$

If $\boldsymbol{W}_\natural$ and/or $\boldsymbol{H}_\natural$ satisfy some condition such that for all $(\boldsymbol{W}_\star, \boldsymbol{H}_\star)$, we have that $\boldsymbol{W}_\star = \boldsymbol{W}_\natural \boldsymbol{\Pi} \boldsymbol{D}$ and $\boldsymbol{H}_\star = \boldsymbol{H}_\natural \boldsymbol{\Pi} \boldsymbol{D}^{-1}$, where $\boldsymbol{\Pi}$ is a permutation matrix and $\boldsymbol{D}$ is a full-rank diagonal matrix, then we say that the matrix factorization model is identifiable under that condition.[1]

For the "plain NMF" model [1], [10], [12], [20], the identification criterion $g(\boldsymbol{W}, \boldsymbol{H})$ is 1 (or, ∞) if $\boldsymbol{W}$ or $\boldsymbol{H}$ has a negative element, and 0 otherwise. Assuming that $\boldsymbol{X}$ can be perfectly factored under the postulated model, the above is equivalent to the popular least-squares NMF formulation:

$$\underset{\boldsymbol{W} \geq \boldsymbol{0}, \boldsymbol{H} \geq \boldsymbol{0},}{\text{minimize}} \ \left\| \boldsymbol{X} - \boldsymbol{W} \boldsymbol{H}^\top \right\|_F^2. \tag{1}$$

Several sufficient conditions for identifiability of (1) have been proposed. Early results in [10] and [11] require that one factor (say, $\boldsymbol{H}$) satisfies the so-called the separability condition:

Definition 2 (Separability): A nonnegative matrix $\boldsymbol{H} \in \mathbb{R}^{N \times r}$ is separable if for every $k = 1, ..., r$, there exists a row index n_k such that $\boldsymbol{H}_{n_k, :} = \alpha_k \boldsymbol{e}_k^\top$, where $\alpha_k > 0$ is a scalar and $\boldsymbol{e}_k$ is the kth coordinate vector in $\mathbb{R}^r$.

With the separability assumption, the works in [10] and [11] first revealed the reason behind the success of NMF in many applications—NMF is unique under some conditions. The downside is that separability is easily violated in practice—see discussions in [5]. In addition to separability, the conditions in [10] and [11] also need that $\boldsymbol{W}$ to exhibit a certain zero pattern. This is also considered restrictive in practice—e.g., in hyperspectral unmixing, $\boldsymbol{W}_{:,r}$'s spectral signatures, which are always dense. The remote sensing and machine learning communities have come up with many different separability-based identification methods without assuming zero patterns on $\boldsymbol{W}$, e.g., the volume maximization (VolMax) criterion [8], [13] and self-dictionary sparse regression [8], [15], [16], [21], [22], respectively. However, the separability condition was not relaxed in those works.

The stringent separability condition was considerably relaxed by Huang *et al.* [12] based on a so-called *sufficiently scattered* condition from a geometric interpretation of NMF.

Definition 3 (Sufficiently Scattered): A nonnegative matrix $\boldsymbol{H} \in \mathbb{R}^{N \times r}$ is sufficiently scattered if 1) $\text{cone}\{\boldsymbol{H}^\top\} \supseteq \mathcal{C}$, 2) $\text{cone}\{\boldsymbol{H}^\top\}^* \cap \text{bd}\mathcal{C}^* = \{\lambda \boldsymbol{e}_k | \lambda \geq 0, k = 1, ..., r\}$, where $\mathcal{C} = \{\boldsymbol{x} | \boldsymbol{x}^\top \boldsymbol{1} \geq \sqrt{r-1} \|\boldsymbol{x}\|_2\}$, $\mathcal{C}^* = \{\boldsymbol{x} | \boldsymbol{x}^\top \boldsymbol{1} \geq \|\boldsymbol{x}\|_2\}$, $\text{cone}\{\boldsymbol{H}^\top\} = \{\boldsymbol{x} | \boldsymbol{x} = \boldsymbol{H}^\top \boldsymbol{\theta}, \forall \boldsymbol{\theta} \geq \boldsymbol{0}\}$ and $\text{cone}\{\boldsymbol{H}^\top\}^* = \{\boldsymbol{y} | \boldsymbol{x}^\top \boldsymbol{y} \geq 0, \forall \boldsymbol{x} \in \text{cone}\{\boldsymbol{H}^\top\}\}$ are the conic hull of $\boldsymbol{H}^\top$ and its dual cone, respectively, and bd is the boundary of a set.

The main result in [12] is that if both $\boldsymbol{W}$ and $\boldsymbol{H}$ satisfy the sufficiently scattered condition, then the criterion in (1) has identifiability. This is a notable result since it was the first provable result in which separability was relaxed for *both* $\boldsymbol{W}$ and $\boldsymbol{H}$. The sufficiently scattered condition essentially means that $\text{cone}\{\boldsymbol{H}^\top\}$ contains $\mathcal{C}$ as its subset, which is much more relaxed than separability that needs $\text{cone}\{\boldsymbol{H}^\top\}$ to contain the entire nonnegative orthant; see Fig. 1.

On the other hand, the zero-pattern assumption on $\boldsymbol{W}$ and $\boldsymbol{H}$ are still needed in [12]. Another line of work removed the zero pattern assumption from one factor (say, $\boldsymbol{W}$) by using a
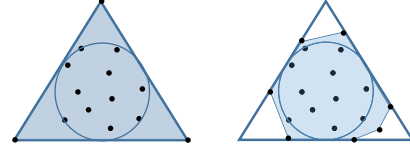

Fig. 1. Illustration of the separability (left) and sufficiently scattered (right) conditions by assuming that the viewer stands in the nonnegative orthant and faces the origin. The dots are rows of $\boldsymbol{H}$, the triangle is the nonnegative orthant, the circle is $\mathcal{C}$, and the shaded region is $\text{cone}\{\boldsymbol{H}^\top\}$. Clearly, separability is special case of the sufficiently scattered condition.
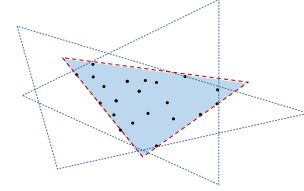

Fig. 2. Intuition of VolMin. The shaded region is $\text{conv}\{\boldsymbol{W}_{:,1}, ..., \boldsymbol{W}_{:,r}\}$, the dots are $\boldsymbol{X}_{:,n}$'s dash lines are enclosing convex hulls, and the bold dashed lines comprise the minimum-volume data-enclosing convex hull.

different identification criterion [3], [19]:

$$\underset{\boldsymbol{W} \in \mathbb{R}^{M \times r}, \boldsymbol{H} \in \mathbb{R}^{N \times r}}{\text{minimize}} \ \det(\boldsymbol{W}^\top \boldsymbol{W}) \tag{2a}$$

$$\text{subject to} \qquad \boldsymbol{X} = \boldsymbol{W} \boldsymbol{H}^\top \tag{2b}$$

$$\boldsymbol{H} \boldsymbol{1} = \boldsymbol{1}, \boldsymbol{H} \geq \boldsymbol{0} \tag{2c}$$

where $\boldsymbol{1}$ is an all-one vector with proper length. Criterion (2) aims at finding the minimum-volume (measured by determinant) data-enclosing convex hull (or simplex) and is also a popular criterion in the NMF literature [23]–[25]. The main result in [3] is that if the ground truth $\boldsymbol{H} \in \{\boldsymbol{Y} \in \mathbb{R}^{N \times r} | \boldsymbol{Y} \boldsymbol{1} = \boldsymbol{1}, \boldsymbol{Y} \geq \boldsymbol{0}\}$ and $\boldsymbol{H}$ is sufficiently scattered, then the VolMin criterion identifies the ground truth $\boldsymbol{W}$ and $\boldsymbol{H}$. This very intuitive result is illustrated in Fig. 2: if $\boldsymbol{H}$ is sufficiently scattered in the nonnegative orthant, $\boldsymbol{X}_{:,n}$'s sufficiently spread in the *convex hull*[2] spanned by the columns of $\boldsymbol{W}$. Then, finding the minimum-volume data-enclosing convex hull recovers the ground truth $\boldsymbol{W}$. This result resolves the long-standing *Craig's conjecture* in remote sensing [26] proposed in the 1990s.

The VolMin identifiability condition is intriguing, since it completely sets $\boldsymbol{W}$ free—there is no assumption on the ground truth $\boldsymbol{W}$ except for being full-column rank, and it has a very mild assumption on $\boldsymbol{H}$. There is a caveat, however: The VolMin criterion needs an extra condition on the ground truth $\boldsymbol{H}$, namely $\boldsymbol{H} \boldsymbol{1} = \boldsymbol{1}$, so that the columns of $\boldsymbol{X}$ all live in the convex hull (not *conic hull* as in the general NMF case) spanned by the columns of $\boldsymbol{W}$—otherwise, the geometric intuition of VolMin in Fig. 2 does not make sense. Many NMF problem instances stemming from applications do not naturally satisfy this assumption. The common trick is to normalize the columns of $\boldsymbol{X}$ using their ℓ_1-norms [14] so that an equivalent model with this sum-to-one assumption holding is enforced—but normalization only works when the ground truth $\boldsymbol{W}$ is also nonnegative. This raises a natural question: Can we essentially keep the advantages of VolMin identifiability (namely, no structural assumption on $\boldsymbol{W}$ (other than low-rank) and no separability requirement on $\boldsymbol{H}$) without assuming sum-to-one on the rows of the ground truth $\boldsymbol{H}$?

---

[1] Whereas identifiability is usually understood as a property of a given model that is independent of identification criteria, NMF can be identifiable under a suitable identification criterion, but not under another, as will be seen.

[2] The convex hull of $\boldsymbol{W}$ is defined as $\text{conv}\{\boldsymbol{W}_{:,1}, ..., \boldsymbol{W}_{:,r}\} = \{\boldsymbol{x} | \boldsymbol{x} = \boldsymbol{W} \boldsymbol{\theta}, \forall \boldsymbol{\theta} \geq \boldsymbol{0}, \boldsymbol{1}^\top \boldsymbol{\theta} = 1\}$.

TABLE I
DIFFERENT ASSUMPTIONS ON W AND H FOR IDENTIFIABILITY OF NMF

	Plain [12]	Self-dict [15], [16], [22]	VolMax [8], [13]	VolMin [3], [19]	Proposed
W	NN, Suff	NN, Full-rank (Full-rank)	NN, Full-rank (Full-rank)	NN, Full-rank (Full-rank)	Full-rank
H	NN, Suff	NN, Sep (NN, Sep row sto)	NN, Sep (NN, Sep row sto)	NN, Suff (NN, Suff row sto)	NN, Suff

Note: "NN" means nonnegativity, "Sep." means separability, "Suff." denotes the sufficiently scattered condition, and "sto" denotes sum-to-one. The conditions in "(·)" give an alternative set of conditions for the corresponding approach.

III. MAIN RESULT

Our main result in this letter fixes the issues with the VolMin identifiability. Specifically, we show that, with a careful and delicate tweak to the VolMin criterion, one can identify the model $X = WH^\top$ without assuming the sum-to-one condition on the rows of H:

Theorem 1: Assume that $X = W_\natural H_\natural^\top$ where $W_\natural \in \mathbb{R}^{M \times r}$ and $H_\natural \in \mathbb{R}^{N \times r}$ and that $\mathrm{rank}(X) = \mathrm{rank}(W_\natural) = r$. Also, assume that $H_\natural$ is sufficiently scattered. Let $(W_\star, H_\star)$ be the optimal solution of the following identification criterion:

$$\underset{W \in \mathbb{R}^{M \times r}, H \in \mathbb{R}^{N \times r}}{\mathrm{minimize}} \quad \det\left(W^\top W\right) \tag{3a}$$

$$\text{subject to} \quad X = WH^\top \tag{3b}$$

$$H^\top \mathbf{1} = \mathbf{1}, H \geq \mathbf{0}. \tag{3c}$$

Then, $W_\star = W_\natural \Pi D$ and $H_\star = H_\natural \Pi D^{-1}$ must hold, where Π and D denotes a permutation matrix and a full-rank diagonal matrix, respectively,

At first glance, the identification criterion in (3) looks similar to VolMin in (2). The difference lies between (2c) and (3c). In (3c), we "shift" the sum-to-one condition to the *columns* of H, rather than enforcing it on the rows of H. This simple modification makes a big difference in terms of generality: Enforcing columns of H to be sum-to-one entails no loss in generality, since in bilinear factorization models like $X = WH^\top$, there is always an intrinsic scaling ambiguity of the columns. In other words, one can always assume the columns of H are scaled by a diagonal matrix and then counter scale the corresponding columns of W, which will not affect the factorization model, i.e., $X = (WD^{-1})(HD)^\top$ still holds. Therefore, there is no need for data normalization to enforce this constraint, as opposed to the VolMin case. In fact, the identifiability of (3) holds for $H^\top \mathbf{1} = \rho \mathbf{1}$ for any $\rho > 0$—we use $\rho = 1$ only for notational simplicity.

We should mention that avoiding normalization may benefit some applications. For example, it was reported in the literature that normalization degrades performance of text mining significantly, since it usually worsens the conditioning of the data matrix [27]. In addition, as mentioned, in applications where W naturally contains negative elements (e.g., channel identification in MIMO communications), even normalization cannot enforce the VolMin model.

It is worth noting that the criterion in Theorem 1 has by far the most relaxed identifiability conditions for NMF. A detailed comparison of different NMF conditions is listed in Table I, where one can see that Criterion (3) works under the mildest conditions on both H and W. Specifically, compared to plain NMF, the

new criterion does not assume any structure on W; compared to VolMin, it does not need the sum-to-one assumption on the rows of H or nonnegativity of W; it also does not need separability, which is inherited from the advantage of VolMin.

In the next section, we will show the proof of Theorem 1. We should remark that although it seems that shifting the sum-to-one constraint to the columns of H is a "small" modification to VolMin, the result in Theorem 1 was not obvious at all before we proved it: by this modification, the clear geometric intuition of VolMin no longer holds—the objective in (3) no longer corresponds to the volume of a data-enclosing convex hull and has no geometric interpretation. Indeed, our proof for the new criterion is purely algebraic rather than geometric.

IV. PROOF OF THEOREM 1

The major insights of the proof are evolved from the VolMin work of the authors and variants [3], [28], [29], with proper modifications to show Theorem 1. To proceed, let us first introduce the following classic lemma in convex analysis.

Lemma 1 (see[30]): If $\mathcal{K}_1$ and $\mathcal{K}_2$ are convex cones and $\mathcal{K}_1 \subseteq \mathcal{K}_2$, then, $\mathcal{K}_2^* \subseteq \mathcal{K}_1^*$, where $\mathcal{K}_1^*$ and $\mathcal{K}_2^*$ denote the dual cones of $\mathcal{K}_1$ and $\mathcal{K}_2$, respectively.

Our purpose is to show that the optimization criterion in (3) outputs $W_\star$ and $H_\star$ that are the column-scaled and permuted versions of the ground truth $W_\natural$ and $H_\natural$. To this end, let us denote $(\widehat{W} \in \mathbb{R}^{M \times r}, \widehat{H} \in \mathbb{R}^{N \times r})$ as a feasible solution of Problem (3) that satisfies the constraints in (3), i.e.,

$$X = \widehat{W}\widehat{H}^\top, \quad \widehat{H}^\top \mathbf{1} = \mathbf{1}, \widehat{H} \geq \mathbf{0}. \tag{4}$$

Note that $X = W_\natural H_\natural^\top$ and that $W_\natural$ has full-column rank. In addition, since $H_\natural$ is sufficiently scattered, $\mathrm{rank}(H_\natural) = r$ also holds [29, Lemma 1]. Consequently, there exists an invertible $A \in \mathbb{R}^{r \times r}$ such that

$$\widehat{H} = H_\natural A, \widehat{W} = W_\natural A^{-\top}. \tag{5}$$

This is because $\widehat{W}$ and $\widehat{H}$ have to admit full-column rank, and thus, $H_\natural$ and $\widehat{H}$ span the same subspace. Otherwise, $\mathrm{rank}(X) = r$ cannot hold. Since (4) holds, one can see that

$$\widehat{H}^\top \mathbf{1} = A^\top H_\natural^\top \mathbf{1} = A^\top \mathbf{1} = \mathbf{1}. \tag{6}$$

Note, here, that we implicitly used $H_\natural^\top \mathbf{1} = \mathbf{1}$, which can be assumed without loss of generality due to the intrinsic scaling ambiguity of any matrix factorization model. By (4), we also have $H_\natural A \geq \mathbf{0}$. By the definition of a dual cone, $H_\natural A \geq \mathbf{0}$ means that $a_i \in \mathrm{cone}\{H_\natural^\top\}^*$, where a_i is the ith column of A, for all $i = 1, \ldots, r$. Because $H_\natural$ is sufficiently scattered, we have that $\mathcal{C} \subseteq \mathrm{cone}\{H_\natural^T\}$ which, together with Lemma 1, leads to $\mathrm{cone}\{H_\natural^\top\}^* \subseteq \mathcal{C}^*$. This further implies that $a_i \in \mathcal{C}^*$, which means $\|a_i\|_2 \leq \mathbf{1}^\top a_i$, by the definition of $\mathcal{C}^*$. Then, we have the following chain:

$$|\det(A)| \leq \prod_{i=1}^r \|a_i\|_2 \tag{7a}$$

$$\leq \prod_{i=1}^r \mathbf{1}^\top a_i \tag{7b}$$

$$= 1 \tag{7c}$$

where (7a) is Hadamard's inequality, and (7b) is due to $a_i \in \mathcal{C}^*$.

Now, suppose that the equality is attained, i.e., $|\det(A)| = 1$; then, all the inequalities in (7) hold as equality, and specifically, (7b) means that the columns of A lie on the boundary of $\mathcal{C}^*$. Recall that $a_i \in \mathrm{cone}\{H_\natural^\top\}^*$, and $H_\natural$ being

sufficiently scattered, according to the second requirement in Definition 3, shows that $\mathrm{cone}\{\boldsymbol{H}_\natural^\top\}^* \cap \mathrm{bd}\mathcal{C}^* = \{\lambda \boldsymbol{e}_k | \lambda \geq 0, k = 1, ..., r\}$; therefore, $\boldsymbol{a}_i$s can only be the $\boldsymbol{e}_k$s. In other words, $\boldsymbol{A}$ can only be a permutation matrix.

Suppose that an optimal solution $\boldsymbol{H}_\star$ of (3) is not a column permutation of $\boldsymbol{H}_\natural$. Since $\boldsymbol{W}_\natural$ and $\boldsymbol{H}_\natural$ are clearly feasible for (3), this means that $\det(\boldsymbol{W}_\star^\top \boldsymbol{W}_\star) \leq \det(\boldsymbol{W}_\natural^\top \boldsymbol{W}_\natural)$. We also know that for every feasible solution, including $\boldsymbol{W}_\star$ and $\boldsymbol{H}_\star$, (5) holds, which means that we have $\boldsymbol{H}_\star = \boldsymbol{H}_\natural \boldsymbol{A}$ and $\boldsymbol{W}_\star = \boldsymbol{W}_\natural \boldsymbol{A}^{-\top}$ hold for a certain invertible $\boldsymbol{A} \in \mathbb{R}^{r \times r}$. Since $\boldsymbol{H}_\natural$ is sufficiently scattered, according to (7b) and our assumption that $\boldsymbol{A}$ is not a permutation matrix, we have $|\det(\boldsymbol{A})| < 1$. However, the optimal objective of (3) is

$$
\begin{aligned}
\det(\boldsymbol{W}_\star^\top \boldsymbol{W}_\star) &= \det(\boldsymbol{A}^{-1} \boldsymbol{W}_\natural^\top \boldsymbol{W}_\natural \boldsymbol{A}^{-\top}) \\
&= \det(\boldsymbol{A}^{-1}) \det(\boldsymbol{W}_\natural^\top \boldsymbol{W}_\natural) \det(\boldsymbol{A}^{-\top}) \\
&= |\det(\boldsymbol{A})|^{-2} \det(\boldsymbol{W}_\natural^\top \boldsymbol{W}_\natural) \\
&> \det(\boldsymbol{W}_\natural^\top \boldsymbol{W}_\natural)
\end{aligned}
$$

which contradicts our first assumption that $(\boldsymbol{W}_\star, \boldsymbol{H}_\star)$ is an optimal solution for (3). Therefore, $\boldsymbol{H}_\star$ must be a column permutation of $\boldsymbol{H}_\natural$.   **Q.E.D.**

As a remark, the proof of Theorem 1 follows the same rationale of that of the VolMin identifiability as in [3]. The critical change is that we have made use of the relationship between sufficiently scattered $\boldsymbol{H}$ and the inequality in (7) here. This inequality appeared in [28] and [29] but was not related to the bilinear matrix factorization criterion in (3)—which might be by far the most important application of this inequality. The interesting and surprising point is that, by this simple yet delicate tweak, the identifiability criterion can cover a substantially wider range of applications, which naturally involve $\boldsymbol{W}$'s that are not nonnegative.

V. VALIDATION AND DISCUSSION

The identification criterion in (3) is a nonconvex optimization problem. In particular, the bilinear constraint $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H}^\top$ is not easy to handle. However, the existing workarounds for handling VolMin can all be employed to deal with Problem (3). One popular method for VolMin is to first take the singular value decomposition (SVD) of the data $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{M \times r}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$, and $\boldsymbol{V} \in \mathbb{R}^{N \times r}$. Then, $\boldsymbol{V}^\top = \widetilde{\boldsymbol{W}}\boldsymbol{H}^\top$ holds, where $\widetilde{\boldsymbol{W}} \in \mathbb{R}^{r \times r}$ is invertible because $\boldsymbol{V}$ and $\boldsymbol{H}$ span the same range space. One can use (3) to identify $\boldsymbol{H}$ from the data model $\widetilde{\boldsymbol{X}} = \boldsymbol{V}^\top = \widetilde{\boldsymbol{W}}\boldsymbol{H}^\top$. Since $\widetilde{\boldsymbol{W}}$ is square and nonsingular, it has an inverse $\boldsymbol{Q} = \widetilde{\boldsymbol{W}}^{-1}$. The identification criterion in (3) can be recast as $\max_{\boldsymbol{Q} \in \mathbb{R}^{r \times r}} |\det(\boldsymbol{Q})|, \text{s.t.} \boldsymbol{Q}\widetilde{\boldsymbol{X}}\mathbf{1} = \mathbf{1}, \boldsymbol{Q}\widetilde{\boldsymbol{X}} \geq \mathbf{0}$. This reformulated problem is much more handy from an optimization point of view. To be specific, one can fix all the columns in $\boldsymbol{Q}$ except one, e.g., $\boldsymbol{q}_i$. Then, the optimization w.r.t. $\boldsymbol{q}_i$ is a linear function, i.e., $\det(\boldsymbol{Q}) = \sum_{i=1}^r (-1)^{i+k} \boldsymbol{Q}_{k,i} \det(\overline{\boldsymbol{Q}}_{k,i}) = \boldsymbol{p}^\top \boldsymbol{q}_i$, where $\boldsymbol{p} = [p_1, \ldots, p_r]^\top$, $p_k = (-1)^{i+k} \det(\overline{\boldsymbol{Q}}_{k,i}), \forall k = 1, ..., r$, and $\overline{\boldsymbol{Q}}_{k,i}$ is a submatrix of $\boldsymbol{Q}$ without the kth row and ith column of $\boldsymbol{Q}$. Maximizing $|\boldsymbol{p}^\top \boldsymbol{q}_i|$ subject to linear constraints can be solved via maximizing both $\boldsymbol{p}^\top \boldsymbol{q}_i$ and $-\boldsymbol{p}^\top \boldsymbol{q}_i$, followed by picking the solution that gives larger absolute objective. Then, cyclically updating the columns of $\boldsymbol{M}$ results in an alternating optimization (AO) algorithm. Similar SVD and AO-based solvers were

TABLE II
MSEs OF THE ESTIMATED $\widehat{\boldsymbol{H}}$

Method	MSE of $\boldsymbol{H}$		
	case 1 (sp. $\boldsymbol{W}$)	case 2 (den. $\boldsymbol{W}$)	case 3 (Gauss. $\boldsymbol{W}$)
Plain ($r = 5$)	5.49E-05	0.0147	0.7468
VolMin ($r = 5$)	1.36E-08	7.31E-10	1.0406
Proposed ($r = 5$)	7.32E-18	7.78E-18	8.44E-18
Plain ($r = 10$)	4.82E-04	0.0403	0.8003
VolMin ($r = 10$)	8.64E-09	8.66E-09	1.2017
Proposed ($r = 10$)	6.54E-18	5.02E-18	6.38E-18

proposed to handle VolMin and its variants in [25], [28], and [29], and empirically good results have been observed. Note that the AO procedure is not the only possible solver here. When the data are very noisy, one can reformulate the problem in (3) as

$\min_{\boldsymbol{W}, \boldsymbol{H}^\top \mathbf{1} = \mathbf{1}, \boldsymbol{H} \geq \mathbf{0}} \left\| \boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}^\top \right\|_F^2 + \lambda \det(\boldsymbol{W}^\top \boldsymbol{W})$, where $\lambda > 0$ balances the determinant term and the data fidelity. Many algorithms for regularized NMF can be employed and modified to handle the above.

An illustrative simulation is shown in Table II to showcase the soundness of the theorem. In this simulation, we generate $\boldsymbol{X} = \boldsymbol{W}_\natural \boldsymbol{H}_\natural^\top$ with $r = 5, 10$ and $M = N = 200$. We tested several cases.
1) $\boldsymbol{W}_\natural \geq \mathbf{0}$, $\boldsymbol{H}_\natural \geq \mathbf{0}$, and both $\boldsymbol{W}_\natural$ and $\boldsymbol{H}_\natural$ are sufficiently scattered.
2) $\boldsymbol{W}_\natural \geq \mathbf{0}$, $\boldsymbol{H}_\natural \geq \mathbf{0}$, and $\boldsymbol{H}_\natural$ is sufficiently scattered but $\boldsymbol{W}_\natural$ is completely dense.
3) $\boldsymbol{W}_\natural$ follows the independent and identically distributed (i.i.d.) normal distribution, and $\boldsymbol{H}_\natural \geq \mathbf{0}$ is sufficiently scattered.

We generate sufficiently scattered factors following [31]—i.e., we generate the elements of a factor following the uniform distribution between zero and one and zero out 35% of its elements, randomly. This way, the obtained factor is empirically sufficiently scattered with an overwhelming probability. We employ the algorithm for fitting-based NMF in [20], the VolMin algorithm in [32], and the described algorithm to handle the new criterion, respectively. We measure the performance of different approaches by measuring the mean squared error (MSE) of the estimated $\widehat{\boldsymbol{H}}$, which is defined as $\mathrm{MSE} = \min_{\boldsymbol{\pi} \in \Pi} \frac{1}{r} \sum_{k=1}^r \| \frac{\boldsymbol{H}_{\natural:,k}}{\|\boldsymbol{H}_{\natural:,k}\|_2} - \frac{\widehat{\boldsymbol{H}}_{:,\pi_k}}{\|\widehat{\boldsymbol{H}}_{:,\pi_k}\|_2} \|_2^2$, where Π is the set of all permutations of $\{1, 2, \ldots, r\}$. The results are obtained by averaging 50 random trials.

Table II matches our theoretical analysis. All the algorithms work very well on case 1, where both $\boldsymbol{W}_\natural$ and $\boldsymbol{H}_\natural$ are sparse (sp.) and sufficiently scattered. In case 2, since $\boldsymbol{W}$ is nonnegative yet dense (den.), plain NMF fails as expected, but VolMin still works, since normalization can help enforce its model when $\boldsymbol{W} \geq \mathbf{0}$. In case 3, when $\boldsymbol{W}$ follows the i.i.d. normal distribution, VolMin fails since normalization does not help, while the proposed method still works perfectly.

To conclude, in this letter, we discussed the identifiability issues with the current NMF approaches. We proposed a new NMF identification criterion that is a simple yet careful tweak of the existing VolMin criterion. We show that, by slightly modifying the constraints of VolMin, the identifiability of the proposed criterion holds under the same sufficiently scattered condition in VolMin, but the modified criterion covers a much wider range of applications including the cases where one factor is not nonnegative. This new criterion offers identifiability to the largest variety of cases among the known results.

REFERENCES

[1] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[2] N. Gillis, "The why and how of nonnegative matrix factorization," in *Regularization, Optimization, Kernels, and Support Vector Machines*, vol. 12. London, U.K.: Chapman & Hall, 2014, p. 257.

[3] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos, "Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2306–2320, May 2015.

[4] X. Fu, W.-K. Ma, and N. Sidiropoulos, "Power spectra separation via structured matrix factorization," *IEEE Trans. Signal Process.*, vol. 64, no. 17, pp. 4592–4605, Sep. 2016.

[5] W.-K. Ma *et al.*, "A signal processing perspective on hyperspectral unmixing," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 67–81, Jan. 2014.

[6] X. Fu, K. Huang, B. Yang, W.-K. Ma, and N. Sidiropoulos, "Robust volume-minimization based matrix factorization for remote sensing and document clustering," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6254–6268, Dec. 2016.

[7] A. Anandkumar, Y.-K. Liu, D. J. Hsu, D. P. Foster, and S. M. Kakade, "A spectral algorithm for latent Dirichlet allocation," in *Proc. Conf. Neural Inf. Process. Syst.*, 2012, pp. 917–925.

[8] S. Arora *et al.*, "A practical algorithm for topic modeling with provable guarantees," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 280–288.

[9] X. Mao, P. Sarkar, and D. Chakrabarti, "On mixed memberships and symmetric nonnegative matrix factorizations," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2324–2333.

[10] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?," in *Proc. Adv. Conf. Neural Inf. Process. Syst.*, 2004. pp. 1141–1148.

[11] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. Jensen, "Theorems on positive data: On the uniqueness of NMF," *Comput. Intell. Neurosci.*, vol. 2008, 2008, Art. no. 764206.

[12] K. Huang, N. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 211–224, Jan. 2014.

[13] T.-H. Chan, W.-K. Ma, A. Ambikapathi, and C.-Y. Chi, "A simplex volume maximization framework for hyperspectral endmember extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4177–4193, Nov. 2011.

[14] N. Gillis and S. Vavasis, "Fast and robust recursive algorithms for separable nonnegative matrix factorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 698–714, Apr. 2014.

[15] X. Fu, W.-K. Ma, T.-H. Chan, and J. M. Bioucas-Dias, "Self-dictionary sparse regression for hyperspectral unmixing: Greedy pursuit and pure pixel search are related," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 6, pp. 1128–1141, Sep. 2015.

[16] B. Recht, C. Re, J. Tropp, and V. Bittorf, "Factoring nonnegative matrices with linear programs," in *Proc. 25th Int. Conf Neural Inf. Process. Syst.*, 2012, pp. 1214–1222.

[17] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[18] E. Esser, M. Moller, S. Osher, G. Sapiro, and J. Xin, "A convex model for nonnegative matrix factorization and dimensionality reduction on physical space," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3239–3252, Jul. 2012.

[19] C.-H. Lin, W.-K. Ma, W.-C. Li, C.-Y. Chi, and A. Ambikapathi, "Identifiability of the simplex volume minimization criterion for blind hyperspectral unmixing: The no-pure-pixel case," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5530–5546, Oct. 2015.

[20] K. Huang and N. Sidiropoulos, "Putting nonnegative matrix factorization to the test: A tutorial derivation of pertinent Cramer-Rao bounds and performance benchmarking," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 76–86, May 2014.

[21] X. Fu and W.-K. Ma, "Robustness analysis of structured matrix factorization via self-dictionary mixed-norm optimization," *IEEE Signal Process. Lett.*, vol. 23, no. 1, pp. 60–64, Jan. 2016.

[22] N. Gillis, "Robustness analysis of hottopixx, a linear programming model for factoring nonnegative matrices," *SIAM J. Matrix Anal. Appl.*, vol. 34, no. 3, pp. 1189–1212, 2013.

[23] G. Zhou, S. Xie, Z. Yang, J.-M. Yang, and Z. He, "Minimum-volume-constrained nonnegative matrix factorization: Enhanced ability of learning parts," *IEEE Trans. Neural Netw.*, vol. 22, no. 10, pp. 1626–1637, Oct. 2011.

[24] R. Schachtner, G. Pöppel, and E. W. Lang, "Towards unique solutions of non-negative matrix factorization problems by a determinant criterion," *Digit. Signal Process.*, vol. 21, no. 4, pp. 528–534, 2011.

[25] T.-H. Chan, C.-Y. Chi, Y.-M. Huang, and W.-K. Ma, "A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4418 –4432, Nov. 2009.

[26] M. D. Craig, "Minimum-volume transforms for remotely sensed data," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 3, pp. 542–552, May 1994.

[27] A. Kumar, V. Sindhwani, and P. Kambadur, "Fast conical hull algorithms for near-separable non-negative matrix factorization," in *Proc. 30th Int. Conf. Mach. Learn. (ICML-13)*, 2013, pp. 231–239.

[28] K. Huang, N. Sidiropoulos, E. Papalexakis, C. Faloutsos, P. Talukdar, and T. Mitchell, "Principled neuro-functional connectivity discovery," in *Proc. SIAM Int. Conf. Data Mining, Society Ind. Appl. Math.*, Jun. 2015, pp. 631–639.

[29] K. Huang, X. Fu, and N. D. Sidiropoulos, "Anchor-free correlated topic modeling: Identifiability and algorithm," in *Proc. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1786–1794.

[30] R. Rockafellar, *Convex Analysis*, vol. 28. Princeton, NJ, USA: Princeton Univ. Press, 1997.

[31] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 2, pp. 713–730, 2008.

[32] J. M. Bioucas-Dias, "A variable splitting augmented Lagrangian approach to linear spectral unmixing," in *Proc. 1st Workshop Hyperspectral Image Signal Process.: Evol. Remote Sens.*, 2009, pp. 1–4.