Learning Nonlinear Mixtures: Identifiability and Algorithm

Bo Yang, Student Member, IEEE, Xiao Fu, Member, IEEE, Nicholas D. Sidiropoulos, Fellow, IEEE, and Kejun Huang, Member, IEEE

Abstract—Linear mixture models have proven very useful in a plethora of applications, e.g., topic modeling, clustering, and source separation. As a critical aspect of the linear mixture models, identifiability of the model parameters is well-studied, under frameworks such as independent component analysis and constrained matrix factorization. Nevertheless, when the linear mixtures are distorted by unknown nonlinear functions—which is well-motivated and more realistic in many cases—the associated identifiability issues are far less studied. This work proposes an identification criterion for a nonlinear mixture model that is well grounded in many real-world applications, and offers identifiability guarantees. A practical implementation based on a judiciously designed neural network is proposed to realize the criterion, and an effective learning algorithm is proposed. Numerical results on synthetic and real application data corroborate the effectiveness of the proposed method.

Index Terms—Nonlinear mixture models, identifiability, unsupervised learning, neural networks, matrix factorization.

I. INTRODUCTION

LINEAR mixture models (LMMs) have found numerous applications in machine learning and signal processing, e.g., topic mining [1], [2], clustering, and source separation. When a LMM is used for parameter estimation, it is critical to ensure that the generative model is identifiable. This is crucial in many data mining problems [3]–[6], as model identifiability is necessary for interpretability. However, a LMM is not identifiable in general—even in the ideal case without noise: a LMM boils down to matrix factorization (MF) that is known to be unidentifiable, unless additional constraints on the factors are imposed.

Identifiability research for LMMs has a long and fruitful history at the confluence of machine learning, statistics, and signal processing, see e.g., [6]–[10]. The arguably most notable line of work is independent component analysis (ICA) ([11], [12]), which was originally motivated by speech source separation. Statistical independence of latent parameters (i.e., signals or data streams corresponding to different sources) is exploited to establish identifiability in ICA. LMM unmixing with correlated latent parameters has also been extensively studied, e.g., in the context of nonnegative matrix factorization (NMF) [6], [8], [13]–[17], bounded component analysis [18], and some other types of constrained MF models [19]–[21].

Despite the relatively good understanding of the identifiability issues of different LMMs, the model is considered oversimplified in many applications. In many cases the observed data cannot be assumed to be linear mixtures of some basis vectors, since nonlinear distortions exist due to various reasons, for example, e.g., sensor saturation/clipping, or nonlinear amplification. A natural question then is: under a reasonable nonlinear mixture model (NMM), is it possible to identify the latent parameters of interest?

This question turns out to be highly nontrivial: most of the analytical tools in the linear mixture case do not apply. One exception is statistical independence of random variables. Based on this observation, many works [22]–[26] tackle nonlinear mixture model identification from a nonlinear ICA viewpoint. This line of work is elegant, but only partially answers our research question when the source signals are statistically independent. In many cases, statistical independence can not be assumed, which is why there has been extensive study on correlated components/sources as mentioned above. The current work targets a scenario in this latter genre.

A. Contributions

In this work, we study the nonlinear mixture model learning problem, under a new setting that is very different from ICA. Specifically, we study a nonlinear mixture model where the observed data vectors are convex combinations of a set of basis vectors followed by nonlinear distortion.

Our specific contributions can be summarized as follows:
1) Flexible generative model: We put forth a novel nonlinear mixture model, which is flexible enough to faithfully capture nonlinear effects naturally arising in many real world applications, e.g., hyperspectral unmixing [7] and MRI [27]. Specifically, the proposed model is related to

Manuscript received May 24, 2019; revised October 25, 2019; accepted April 18, 2020. Date of publication April 22, 2020; date of current version May 15, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Abd-Krim Seghouane. This work was supported by NSF ECCS-1608961, IIS-1447788. (Corresponding author: Bo Yang.)

Bo Yang is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: yang4173@umn.edu).

Xiao Fu is with the School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331 USA (e-mail: xiao.fu@oregonstate.edu).

Nicholas D. Sidiropoulos is with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904 USA (e-mail: nikos@virginia.edu).

Kejun Huang is with the Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: kejun.huang@ufl.edu).

Digital Object Identifier 10.1109/TSP.2020.2989551

1053-587X © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
a classical model in nonlinear ICA (nICA), while the independence assumption in nonlinear ICA is replaced with new assumptions from convex geometry. For source separation, our model can handle the case when sources are dependent, unlike nICA.

2) Identification criterion: We propose a model identification criterion for the considered problem and provide sufficient conditions under which the model is identifiable. The criterion, when its global optimality is achieved, guarantees that an inverse function (possibly scaled and shifted) of the unknown nonlinear distortion is learned, in a completely unsupervised fashion. Thus the nonlinear effects are rectified/equalized, and traditional methods for identifying LMMs can be applied subsequently. Our proof leverages perspectives from functional equations [28], [29], together with novel extensions of existing LMM identifiability results. This fortuitous union yields the right tools for our analysis, as we will see.

3) Neural network-based implementation: We propose a neural network based formulation to implement the proposed criterion. The neural network acts as an adaptive feature extractor, which, after training with the proposed identification criterion, will be the (scaled) inverse function of the unknown nonlinear function in data generation. The employed neural network is judiciously designed, so as to ensure its invertibility, thus the specific constraints imposed by the identification criterion can be satisfied.

4) Numerical validation: We reformulate the criterion into an easy-to-implement form and employ a trust region algorithm for solving the problem efficiently. We also test the algorithm on both synthetic and real data. In particular, the neural network is shown to be able to counter the nonlinear effects after training, and much better parameter estimation performance is observed. Evaluation on a real world hyperspectral image also confirms the effectiveness of the proposed theory and algorithm.

Another salient feature of our method is that it turns the unsupervised parameter estimation problem into a supervised regression problem, which requires little new algorithmic design – see Section IV for more information.

A short version [30] of this work has been submitted to the 2019 Asilomar Conference on Signals, Systems, and Computers for publication. Relative to the conference submission, the current paper studies the solution existence issue (cf. Section III-C) of the proposed optimization criterion, thus addressing an important theoretical question that was left open. Moreover, detailed proofs for all the technical claims are delineated, and several additional numerical experiments concerning different parameter settings are conducted and reported.

B. Related Work

The current work is closely related to the large body of work studying model identifiability. For a parametric statistical model, identifiability is a fundamental question which concerns whether the model parameters can be uniquely pinned down from observed data. In explanatory data analysis, where the goal is to explain/interpret the data, identifiability of the adopted model is critical, as otherwise there exist more than one model settings that fit the data equally well, in which case one does not even know which model to interpret [4], [5], [8].

1) Linear/Nonlinear Mixture Identifiability Through Statistical Independence: ICA [11], [12] is a classical technique in unsupervised learning. Statistical independence and deviation from Gaussianity of components/sources are exploited to establish model identifiability. Several identification criteria have been proposed based on mutual information, entropy, as well as high-order cumulants, see e.g., [12], [31]. Related identifiability issue has also been studied under the nonnegative matrix factorization (NMF) setting, see e.g. [14].

A crucial assumption in ICA is linear mixing: sources are mixed linearly by a so-called mixing matrix. This assumption can be over-simplified in some applications, and considerable effort has been put into developing methods that handle nonlinear effects in the framework of ICA. The work of [22], [23] put forth a post-nonlinear model, where on top of the classical assumption of a linear mixing system, a nonlinear transformation is also assumed. For this model, identifiability is established based on the assumptions of source independence, and some assumptions on the nonlinear function. A recent line of work on nonlinear mixture model includes [24]–[26], which exploit structures in the data, e.g., non-stationarity or temporal dependence, and form identification criterions tailored for these structures such that identifiability of sources is established. Reference [21] studies nonlinear model recovery in the sparse dictionary learning setting. See [32] for additional reference on this topic.

2) Linear Mixture Identifiability Through Convex Geometry: Another line of work establishes identifiability of LMM using geometric properties, instead of statistical properties as in ICA. To the best of our knowledge, this line of work originated in the hyperspectral imaging community, see e.g., [7], [8] and reference therein. In such applications, measurements of a geographical region on the ground are taken in multiple spectral bands, and the measurement for each pixel is assumed to be a convex combination of the spectral signatures of several materials, e.g., water and soil, present on the target ground. In these applications, it is of primary interest to identify the spectral signatures of the materials (known as “endmembers” in the remote sensing community), as well as those convex combination coefficients – in other words, identifiability of LMM is a central topic. Exploiting different model assumptions, considerable work has been done in this area, see [4], [8], [16], [17].

Interestingly, the same LMM also arises in probabilistic topic modeling, see [3], [5], [33]. Again, identifiability of LMM in this application is crucial, as one would like to identify the true topics.

C. Organization

The remainder of the paper is organized as follows. In Section II, we review some preliminary materials that are essential for the presentation of this work, then introduce the considered nonlinear mixture model. In Section III, we detail the main result of this work, i.e., the identifiability guarantees for the proposed
nonlinear mixture model, as well as the existence of solution for
the proposed identification criterion. In Section IV, we develop
a companion learning method for the proposed identification
criterion. Specifically, we put forth a neural network architecture
that guarantees invertibility of the resulting mapping — which
is crucial to implement the requirements set by the proposed
identification criterion. We present numerical experiment results
in Section V. Conclusions are given in Section VI.

D. Notation

Bold capital letters represent matrices, while bold lowercase
letters denote vectors, which are assumed to be column vec-
tors, unless transposed with $(\cdot)^T$. Plain lowercase letters denote
scalars. $X$ and $x$ refer to the observed data, and $A$, $a_i$, $S$, $s_j$
refer to the underlying latent parameters. Symbol $\phi$ denotes the
unknown nonlinear function in data generation, and $f$ denotes
the learning function, which tries to counteract the nonlinear
effects in $\phi$. Symbol $Y$ represents the data transformed by
the learning function $f$, i.e., $Y = f(X)$, and $k$ denotes the
composite function of $f$ and $\phi$. Symbol $[N]$ denotes the set of
integers $\{1, \ldots, N\}$. The vector-valued functions we consider
in this work are all element-wise, and we use the notation $f =
[f_1, \ldots, f_M]^T$ to mean that $[f(x)](i) = f_i(x(i))$ for $x \in \mathbb{R}^M$
and $i \in [M]$. The symbol $\| \cdot \|$ denotes the $\ell_0$ norm, i.e.,
the number of nonzeros, of a vector or matrix. The symbol cone$(X)$
denotes the set formed by conical combinations of columns of
$X$. Finally, $0$ (1) denotes a vector (or matrix) of all 0’s (1’s).

II. PROBLEM STATEMENT

The new model considered in this work is a generalization of
the LMM. In this section, we first briefly review the LMM
and its related identifiability results, then we introduce the new
model, and discuss its motivation. By design, the new model
can be seen as having an extra nonlinear transformation added
to the classical LMM. For this reason, in Section III, we propose
a natural two-stage learning strategy: in the first stage, our new
method “removes” the nonlinear effects. After this step, the data
model is reduced to LMM, and we employ existing techniques
to deal with it in the second stage.

A. Review of Linear Mixture Model

We briefly review existing parameter identification results
concerning the LMM that are related to this work. Relevant
concepts in convex geometry can be found in Appendix A.

To facilitate discussion, we use $\Delta_M := \{x | x \in \mathbb{R}^M, x \geq
0, 1^T x = 1\}$ to denote the $(M - 1)$-dimensional probability
simplex. The LMM is defined as

$$x_j = A s_j \forall j \in [N],$$

where $A \in \mathbb{R}^{M \times r}$ is often a tall matrix, i.e., $M > r$, and $s_j \in
\Delta_r$. Alternatively, we will also write $X = AS$ by collecting all
$x_j$’s into $X$, and $s_j$’s into $S$.

To recover factors $A$ and $S$ from data $X = [x_1, \ldots, x_N]$, the
following so-called Volume Minimization (VolMin, [16], [34],
[35]) criterion is often employed:

$$\min_{B \in \mathbb{R}^{M \times r}, H \in \mathbb{R}^{r \times N}} \det(B^T B)$$

s.t. $X = BH, H \succeq 0, H^T 1 = 1,$

where it is assumed that $r$ is known. The term $\det(B^T B)$ is a
measure of the volume of the simplex formed by using columns
of $B$ as vertices, see e.g. [16]. This criterion means that we want
to find $B$ and $H$ that satisfy the LMM, and we pick the solution
with the minimal volume, hence the name VolMin.

Several algorithms for dealing with (2) have been developed
[17], [35]–[37], and we will use the so-called minimum volume enclosing simplex (MVES): Given data $X$ and the rank
parameter $r$, the MVES algorithm returns a solution $(\hat{B}, \hat{H})$ of
(2). We refer readers to [35] for more on MVES.

B. Proposed Signal Model

We consider the following signal model

$$x_j = \phi(A s_j) \forall j \in [N],$$

where $A \in \mathbb{R}^{M \times r}$ satisfies $A \succeq 0$, and $s_j \in \Delta_r, \forall j \in [N]$. The
function $\phi$ is a nonlinear mapping $\phi : \mathbb{R}^M \to \mathbb{R}^M$, and we consider
element-wise nonlinearity, i.e., $\phi = [\phi_1, \phi_2, \ldots, \phi_M]^T$, so
that

$$\phi(x) = [\phi_1(x(1)), \ldots, \phi_M(x(M))]^T,$$

where $x = [x(1), \ldots, x(M)]^T$. For brevity, we use the short-hand
notation $X = \phi(AS)$ to denote (3), where it should be understood that the $\phi$ is applied to each column of $AS$.

We give a brief discussion on the motivations behind model
(3). Model (3) can be viewed as a generalization of (1), which is
used in various applications. In hyperspectral unmixing (HU),
each $x_j$ is a hyperspectral pixel, each column of $A$ represents the
frequency signature of a certain material (e.g., soil, vegetation,
water), and each $s_j$ denotes the proportion of materials in that
pixel $x_j$, see e.g., [7], [8], [38], [39]. In MRI, LMM is used due
to the so-called “partial volume effect” [27], [40], [41], which
gives rise to the condition $s_j \in \Delta_r$. Both these applications are
of great importance in their respective research fields, where
considerable work has been done based on (1). Yet, it is widely
recognized that in many real world scenarios, the LMM in (1)
is oversimplified. For example, in HU, the measurements $x_j$’s
are obtained by sensors, which have inherent nonlinearity due
to physical limitations of the measuring devices, see [7], [38],
[39]. In MRI, empirical studies show that there exist (currently)
unknown physical/biological processes, which cause nonlinear
relationships between measurements (e.g., changes in cerebral
blood flow) and the underlying signals of interest (brain net-
work connectivity, brain local region energy consumption, etc.),
see [42]–[44]. The existence of such nonlinear distortions in
various problems has motivated us to impose the $\phi$ function in
(3) on top of the classical LMM in (1). This way, many nonlinear
effects happening in practice yet ignored in the LMM literature
may be captured.

However, it is clear that the additional unknown $\phi$ brings con-
siderable complication in recovering $A$ and $S$. Before pursuing
a general result, let us make some simple observations. First, for many nonlinear $\phi$, it is not possible to recover $A$ and $S$, e.g., $\phi(x) = 0 \forall x$. Hence one of the tasks is to impose on $\phi$ reasonable and practical conditions, under which recovery is possible. Second, if $\phi$ is linear, by the element-wise assumption, we have $X = DAS$, where $D$ is a diagonal matrix. From here, we can see that there are no scaling ambiguities on the rows of $A$, even for the simplest $\phi$. In light of this, a crucial question about model (3) is which parts (or aspects) of $A$ and $S$ can be identified, and to what extent?

III. IDENTIFIABILITY ANALYSIS

As mentioned earlier, our overall identification strategy to tackle the new model has two stages. In the first stage, our method “removes” nonlinear effects. In the second stage, since the data model is reduced to the classical LMM, we adapt existing techniques to deal with it. In this section, we first present techniques to remove the nonlinearity; then in the last sub-section, we discuss a technique to identify the model after nonlinearity is dealt with.

A. A Technical Lemma

We aim at understanding the identifiability of (3), as well as designing a method to learn this model. It will be shown (see Theorem 1 and Lemma 2) that $S$ can be identified uniquely, $A$ can be identified up to linear transformation, and the inverse of $\phi$ can be identified up to affine transformation. Towards identifying the model, we will try to learn an adjustable function $f$, and denote

$$y_j = f(\phi(As_j)) \quad \forall j \in [N].$$

The remaining question is how to devise a learning method such that the resulting $f$ will “counteract” the nonlinear effect brought by $\phi$. If this can be done, we can then employ methods designed for LMM (1) to separate the latent factors.

We show that if the unknown functions satisfy a certain functional equation, then we can assert that these functions possess some good properties, which help achieving our end goal of learning an $f$ function to “counteract” the nonlinear effect. To present such a result, let us consider the following functional equation concerning functions $\psi_1, \ldots, \psi_M$ and variables $s \in \text{int } \Delta_r$.

$$\sum_{i=1}^{M} \psi_i(a_i^T s) = 1 \quad \forall s \in \text{int } \Delta_r,$$

where $\text{int } \Delta_r$ denotes the interior of $\Delta_r$. To facilitate presentation, let $A := [a_1, a_2, \ldots, a_M]^T \in \mathbb{R}^{M \times r}$.

**Lemma 1:** Suppose (6) holds, and $M \geq r \geq 3$. Let us further assume that (a) the functions $\psi_1, \ldots, \psi_M$ are twice differentiable, and are all convex (or all concave) in the domain $(0, 1)$; and (b) $A$ is nonnegative and has two positive columns. Then the functions $\psi_1, \ldots, \psi_M$ are all affine.

This lemma asserts that under some technical conditions, only affine functions satisfy (6). We use this result to learn a nonlinear function for each element of the nonlinear mixing function in (3), so that all the composite functions are affine, hence nonlinearity in the signal model is removed. The proof for Lemma 1 can be found in Appendix C.

B. Nonlinear Mixture Model Identification

To proceed, let us suppose that the learning function $f : \mathbb{R}^M \rightarrow \mathbb{R}^M$ in (6) is also element-wise, i.e., $f = [f_1, f_2, \ldots, f_M]^T$, where the $f_i$’s are univariate functions. Denote $k = [k_1, k_2, \ldots, k_M]^T : \mathbb{R}^M \rightarrow \mathbb{R}^M$, where $k_i = f_i \circ \phi_i$, and $\phi$ denotes function composition. To facilitate presentation of the main results, we introduce the following definition.

**Definition 1:** (Incoherence) A tall and full-rank matrix $A \in \mathbb{R}^{m \times r}$ is said to be incoherent if $e_j \notin \text{Range}(A) \quad \forall j \in [m]$.

Note that here incoherence is defined in the same spirit as the incoherence found in well-known compressed sensing literature, see e.g., [45]. The physical meaning of incoherence is that the energy of each column is widespread across the rows.

We make the following assumptions about the generative model (3).

1. **A1)** The functions $\phi_1, \ldots, \phi_M$ are all invertible, and twice differentiable.
2. **A2)** The matrix $A \in \mathbb{R}^{M \times r}$ in (3) satisfies $A \geq 0$, has two strictly positive columns, and is incoherent (see Definition 1). The dimensions satisfy $M \geq r \geq 3$.
3. **A3)** The columns of $S$ satisfy $s_j \in \text{int } \Delta_r \quad \forall j \in [N]$. Let us briefly discuss the roles of the assumptions. For (A1), the invertibility condition is important, as it is in general very hard to recover the unknown parameters if they undergo non-invertible transformations. The twice differentiable condition on $\phi$’s requires the nonlinear functions in data generation to be smooth.

Assumption (A2) is the same as in Lemma 1, except for the additional incoherence assumption. Both conditions in (A2) are important in establishing the identifiability theory. For example, the incoherence assumption ensures that solutions that satisfy (7) exist; see detailed discussion in Section III-C. We should mention that although (A2) seems technical, it is not hard to be satisfied. For instance, if $A$ is generated from an absolutely continuous distribution, and supported on the nonnegative orthant, both conditions are satisfied with probability one.

For brevity, let us define a matrix function that has $k$ acting on the columns of its matrix argument, $Tk(X) = [k(x_1), k(x_2), \ldots, k(x_N)]$ for $X = [x_1, x_2, \ldots, x_N]$. We are ready to state the following results.

**Theorem 1:** (Main results) Under assumptions (A1), (A2), (A3), if $f_1, f_2, \ldots, f_M$ satisfy

$$\sum_{i=1}^{M} k_i(a_i^T s) = 1 \quad \forall s \in \text{int } \Delta_r,$$

where $k_i = f_i \circ \phi_i$, and the composite functions $k_i$ are all convex (or all concave), then the following hold

a) The functions $k_1, k_2, \ldots, k_M$ are affine;

b) The functions $\phi_1^{-1}, \ldots, \phi_M^{-1}$ are identified up to an affine transformation, i.e., $f_i(x) = d_i\phi_i^{-1}(x) + b_i \quad \forall i \in [M]$, where $d_i$’s and $b_i$’s are scalar constants.
The proof can be found in Appendix D. A remark about function $T_k$ is in order.

**Remark 1:** According to (a) in Theorem 1, we can write

$$T_k(X) = DX + b1^T_N,$$

where $D = \text{diag}(d_1, \ldots, d_M)$, and $b = [b_1, \ldots, b_M]^T$, and $d_i$ and $b_i$ are coefficients for the affine function $k_i$. Equation (8) suggests that $T_k$ is an affine function in $X$. However, we would like $T_k$ to be linear in $X$, instead of affine, as later we show that it is possible to identify the LMM parameters under invertible linear transformation (Lemma 2).

Fortunately, for data model (3) satisfying (A1), (A2) and (A3), we can see that $T_k(X)$ is indeed a linear function of $X$ under mild conditions. Let us consider a matrix $X \in \mathbb{R}^{M \times N}$. Due to equation (7), we have $1^T_M T_k(X) = 1^T_M DX + 1^T_M b 1^T_N = 1^T_N$, which means $1^T_N = 1^T_M DX / (1 - 1^T_M b)$. Plugging this into the above equation, we have

$$T_k(X) = DX + b \left( \frac{1}{1 - 1^T_M b} 1^T_M DX \right)$$

$$= (I + \frac{1}{1 - 1^T_M b} b 1^T_M) DX$$

$$= WX$$

where we define $W := (I + \frac{1}{1 - 1^T_M b} b 1^T_M) D$, and $1_M$ is an all-one vector of length $M$. The above equation suggests that $T_k$ is linear in $X$. A subtle point is that the above calculation is invalid when $1 = 1^T_M b$ holds exactly, but this never happens in our extensive simulations.

**Remark 2:** Given the generative model (3), Theorem 1 essentially asserts that if we require $1^T y = 1$ for all input $s$, then the learned functions $f_1, \ldots, f_M$ will remove the nonlinearity in functions $\phi_1, \ldots, \phi_M$. But our main goal is identifying parameters in the latent LMM; $T_k$ being linear is not enough. To see this more clearly, suppose we get a solution for $f_i$’s of this form

$$f_i(x) = 1/M \quad \forall i \in [M].$$

In this case, the $k_i$’s are all constant functions, and hence convex. Moreover, for this solution (10), we have $k_i(As) = DA s + b$, where $D = 0$ and $b = (1/M)1$; meaning that $f$ maps all input $x = \phi_i(As)$ to the single point $y = (1/M)1$, which does satisfy (7). The problem we exposed here is important: we need additional constraints on $y$ beyond $1^T y = 1$, so that $y$ preserves information about the original data $x$. Essentially, we require the $f_i$’s to be invertible, see details in Section IV.

### C. Existence of Solutions

The results in Theorem 1 rely on equation (7). One could be wondering, given the conditions outlined in assumptions (A1), (A2), and (A3), does there exist $f$ such that (7) holds? This amounts to studying feasibility of (7), which is not obvious. For instance, one might guess (incorrectly) that $\{\hat{f}_i = \phi_i^{-1} \quad \forall i\}$ satisfies (7). However, under such $f_i$’s,

$$\sum_{i=1}^M k_i(a_i^T s) = \sum_{i=1}^M a_i^T s \neq 1,$nless we impose more assumptions on $A$ or $S$. This means, for this natural guess, (7) does not hold.

To study this feasibility issue, we first note that if there exists a diagonal matrix $D$, such that $1^T_D A = 1^T$, then letting $\hat{f}_i = \phi_i^{-1}$, we have

$$\sum_{i=1}^M d_i \hat{f}_i(a_i^T s) = \sum_{i=1}^M d_i a_i^T s$$

$$= 1^T D A s$$

$$= 1^T s$$

$$= 1, \quad \forall s \in \text{int} \Delta, \quad (11)$$

where $d_i$ is the $i$-th diagonal element of $D$. Hence, the functions $\{\hat{f}_i(\cdot) = d_i \hat{f}_i(\cdot), i \in [M]\}$ satisfy (7). An additional requirement is that $\{d_i \neq 0 \quad \forall i\}$, otherwise we can get a trivial solution, as explained in the above section.

Building on the above observation, the feasibility problem of (7) boils down to establishing existence of a nonsingular diagonal matrix $D$ (i.e., $d_i \neq 0 \quad \forall i$), such that $1^T_D A = 1^T$, for matrix $A$ that satisfies assumption (A2). We present Proposition 1, which shows that with a mild incoherence condition (see Definition 1) on $A$, such desired $D$ indeed exists.

To simplify notation, here we write $A^T d = 1$ instead of $1^T D A = 1^T$; existence of nonsingular diagonal $D$ is the same as existence of fully dense vector $d$.

**Proposition 1:** For a matrix $A \in \mathbb{R}^{m \times r}$, assume that $m > r$, and $A$ has full column-rank. Also assume that $A$ is incoherent.

There exists a vector $d \in \mathbb{R}^m$, such that

$$A^T d = 1,$$  \hspace{1cm} (12a)

$$\|d\|_0 = m.$$  \hspace{1cm} (12b)

Note that by assumption, $A$ is tall and full rank, so there are infinitely many $d$ vectors that satisfy (12a). However, it is not obvious if there is always a fully dense $d$ (i.e., (12b)) such that (12a) holds for any $A$ that is tall and full rank. For example, consider the matrix (which does not satisfy the incoherence condition)

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}. \quad (13)$$

This matrix is tall and full rank. Let $d = [d_1, d_2, d_3]^T$, and from $A^T d = 1$, we have

$$d_1 + d_3 = 1$$

$$d_1 = 1,$$

which means $d_3 = 0$ when $A^T d = 1$. This suggests there does not exist a fully dense $d$ for the particular $A$ in (13), such that $A^T d = 1$. The proof of Proposition 1 can be found in Appendix E.

**Remark 3:** We established that for an incoherent $A$, there always exist solutions to make (7) hold. Moreover, we point out
that even for some $A$ that is not incoherent, solutions for (7) might also exist. For example, if one or more columns of $A$ are some columns of an identity matrix, then $A$ is not incoherent. However, if we have $1^T A = 1^T$ – which is true when all columns of $A$ are some columns of an identity matrix – then we see that $\{f_i = f_i^{-1} \forall i\}$ is a feasible solution.

D. Parameter Identification After Removing Nonlinearity

To proceed with parameter learning, let us provide the following lemma, concerning parameter identifiability of LMM (1) under a linear transformation. This is needed in the second stage of our learning strategy, where nonlinearity is removed, and only a linear transformation of the LMM is remaining.

**Lemma 2:** Consider the LMM model $X = AS$, where $A \in \mathbb{R}^{M \times r}$ and $S \in \mathbb{R}^{r \times N}$ satisfies the SS condition (cf. Appendix B), and rank$(A) = \text{rank}(S) = r$. Let $Y = WX$, where $W \in \mathbb{R}^{M \times M}$ is nonsingular. Then we can identify $A = WA$ and $S$ up to column permutation by solving

$$
\begin{align*}
\min_{B \in \mathbb{R}^{M \times r}, \Pi \in \mathbb{S}^{r \times N}} & \quad \det(B^T B) \\
\text{s.t.} & \quad Y = BH, H \geq 0, HT^1 = 1.
\end{align*}
$$

That is, if $(B^*, H^*)$ is an optimal solution of the above problem, then $B' = A \Pi$ and $H' = \Pi^T S$, where $\Pi$ is a permutation matrix.

This lemma is a direct consequence of Theorem 2 in (Appendix B, see also [16]). It suggests that when the original model $X = AS$ is identifiable, then after an invertible linear transformation $W$, we can still identify $S$ using VolMin: but it is not possible to identify $A$ due to the linear transformation $W$. This lemma also suggests that we can employ an algorithm designed to tackle LMM to identify $S$, once the nonlinear effects in (3) have been removed, and only an unknown linear transformation is left.

IV. LEARNING ALGORITHM

Theorem 1 suggests the following optimization formulation to learn the desired $f$.

$$
\begin{align*}
\text{find} & \quad f_1, \ldots, f_M \\
\text{s.t.} & \quad f_i \circ \phi_i \text{ is all convex (or all concave)} \quad \forall i \in [M], \\
& \quad \sum_{i=1}^M f_i(x_j(i)) = 1 \quad \forall j \in [N].
\end{align*}
$$

(15)

For this formulation we have the following claim.

**Corollary 1:** For problem (15), suppose the data $X = [x_1, \ldots, x_N] \in \mathbb{R}^{M \times N}$ admit model (3) and assumptions (A1), (A2), (A3) hold. Assume further that $s_j$’s are sampled from a Dirichlet distribution. When $N \rightarrow +\infty$, the optimal solutions to (15) satisfy (7), and the resulting $\{k_i = f_i \circ \phi_i \forall i \in [M]\}$ are all affine.

This corollary follows from the distributional assumption on $s_j$. As $N \rightarrow +\infty$, $s_j$ will cover all the interior of $\Delta_r$ with probability 1, since every point in $\Delta_r$ has nonzero probability. Then the constraints in (15) become the same as the conditions in Theorem 1. Corollary 1 thus guarantees the nonlinear function identification property of formulation (15) in an asymptotic sense. It also suggests that the proposed method should be able to benefit from a large number of samples. In the following, we approximate problem (15) to make it amenable to numerical algorithms.

There may be a number of ways to enforce the constraint that $k_i$’s are all convex (or all concave). For example, we note

$$
\begin{align*}
k_i''(x) &= f_i''((\phi_i(x))[\phi_i'(x)]^2 + f_i'((\phi_i(x))\phi_i''(x). \quad (16)
\end{align*}
$$

To make sure $k_i$ is convex (or concave), we need $k_i''(x) \geq 0$ or $k_i''(x) \leq 0$. This can be done if we somehow know the sign of $\phi_i$. For instance, if we know an \textit{a priori} that the nonlinear distortion is a concave function, i.e., $\phi_i'' \leq 0$, then we can pick a parametric family for $f_i$’s, such that $f_i''(x) \geq 0$ and $f_i''(x) \leq 0$. Then we have $k_i''(x) \geq 0$, i.e., $k_i$ is convex. Similarly, we can constrain $f_i$’s for all $i \in [N]$ to make sure $k_i$’s are all convex (or concave).

To simplify implementation, we adopt an approximation: We only require $f_i$’s to be invertible in this work. This leads to the following optimization problem.

$$
\begin{align*}
\text{find} & \quad f_1, \ldots, f_M \\
\text{s.t.} & \quad f_i \text{ is invertible } \forall i \in [M], \\
& \quad \sum_{i=1}^M f_i(x_j(i)) = 1 \quad \forall j \in [N].
\end{align*}
$$

(17)

In other words, we aim at learning \textit{invertible} functions that add to one. The invertibility condition is crucial, otherwise we can obtain trivial solutions, as explained before.

To parametrize functions $f_j$, we will adopt Neural Networks (NN) with one hidden layer, due to their universal approximation capability [46], [47]. In particular, we employ the following parametric function family

$$
\mathcal{F} = \left\{ f(x) = \sum_{k=1}^K \alpha_k \sigma(\beta_k x + \gamma_k) + \delta_k, \quad \alpha_k > 0, \beta_k > 0 \quad \forall k \in [K] \right\}
$$

(18)

where $K$ is the number of neurons, $\{\alpha_k, \beta_k, \gamma_k, \delta_k\}_{k=1}^K$ are the learnable parameters of this NN, and $\sigma$ denotes the nonlinearity. Importantly, the constraints on $\alpha_k$ and $\beta_k$ are to ensure invertibility, as stated below.

**Lemma 3:** In (18), if $\sigma'(x) > 0$ $\forall x$, the functions in $\mathcal{F}$ are all invertible.

The above lemma can be easily seen to be true. By definition, we have $f'(x) = \sum_{k=1}^K \alpha_k \beta_k \sigma'(\beta_k x + \gamma_k)$. For $\sigma'(x) > 0$, we have $f'(x) > 0$ if $\alpha_k > 0$, $\beta_k > 0$ $\forall k \in [K]$. Note that the requirement for $\sigma'(x) > 0$ is easily satisfied for commonly used neurons, e.g., tanh$(\cdot)$ and the sigmoid function. For this reason, we set $\sigma$ to be tanh$(\cdot)$ in this work.
Algorithm 1: Nonlinear Mixture Learning.

Input: Data $X \in \mathbb{R}^{M \times N}$, number of neurons $K$, latent dimension $r$

Output: Learned functions $\hat{f}_1, \ldots, \hat{f}_M$, estimated $\hat{S}$

1: Learn parameters $\{\hat{\alpha}_k, \hat{\beta}_k, \hat{\gamma}_k, \hat{\delta}_k\}$ by solving (19)
2: Form functions $f_1, \ldots, f_M$ by $f_i(x) = \sum_{k=1}^K \hat{\alpha}_k \sigma(\hat{\beta}_k x + \hat{\gamma}_k) + \hat{\delta}_k$
3: Obtain transformed data by applying the learned functions on input data: $Y = f(X)$
4: Obtain $\hat{S}$ by calling MVES($Y, r$)
5: return $\hat{f}_1, \ldots, \hat{f}_M, \hat{S}$

Utilizing the parametric family $F$ in (18), we arrive at the following optimization problem

$$
\min_{\{\alpha_k, \beta_k, \gamma_k, \delta_k\}} \frac{1}{N} \sum_{j=1}^N \left( 1 - \sum_{i=1}^M \sum_{k=1}^K \alpha_k \sigma(\beta_k x_j(i) + \gamma_k) + \delta_k \right)^2
$$

s.t. $\alpha_k > 0, \beta_k > 0 \ \forall k \in [K], i \in [M].$  \hspace{1cm} (19)

This is a nonlinear least-squares regression problem, with bound constraints. We employ a trust-region algorithm [48] for optimization.

After obtaining parameters $\{\hat{\alpha}_k, \hat{\beta}_k, \hat{\gamma}_k, \hat{\delta}_k\}$ via (19), we obtain $\hat{f}_i(x) = \sum_{k=1}^K \hat{\alpha}_k \sigma(\hat{\beta}_k x + \hat{\gamma}_k) + \hat{\delta}_k$, and form the transformed data $Y = f(X)$. Theorem 1 predicts that $Y \approx W \mathbf{AS}$ for some nonsingular matrix $W$. From Lemma 2, we see that we can employ an algorithm for LMM to identify $S$. For this purpose, we employ the classical MVES algorithm [35] for LMM, and obtain an estimate $\hat{S}$.

The overall procedure is summarized in Algorithm 1. We emphasize again that the method is unsupervised: the only training data is $X$, not $\{x_j, y_j\}_{j=1}^N$ (feature-label pairs) as in, e.g., the generalized additive models [31, Ch. 9] setting, or recent works on nonlinear estimation [49], [50].

V. NUMERICAL EXPERIMENTS

We evaluate the presented theoretical results, as well as the learning algorithm in this section. To probe different aspects of the proposed method, we design and execute several numerical experiments, leveraging controlled synthetic data sets, as well as a real-world data set.

A. Synthetic Data Study

We start by providing a qualitative assessment of the proposed theory and algorithm. For this purpose, we will visualize the learned functions to see if nonlinearity in data generation is indeed resolved. We randomly generate $S$ according to a Dirichlet distribution – such that the generated $s_j$’s are non-negative and sum to one. The dimensions are $M = r = 4$ and $N = 1000$. The parameter of this Dirichlet distribution is set to $\mu = [0.1, 0.1, 0.1, 0.1]$, so that the generated $s_j$’s are well spread on the probability simplex, hence the sufficiently scattered condition (cf. Appendix B) is likely to be satisfied. For this experiment, we take $\mathbf{A}$ to be $2I_4$. The four nonlinear functions in data generation are $\phi_1(x) = x, \phi_2(x) = \sqrt{x}, \phi_3(x) = \sqrt{x}$, and $\phi_4(x) = \log(x + 1)$. Note that these functions are not revealed to the learning algorithm, and are only used to visualize the results after learning is completed. For learning, each function $f_i$ is parametrized by a constrained one-hidden-layer NN defined in (18), with $K = 20$ neurons. The learned functions $f_1, \ldots, f_4$ and the composite functions $f_1 \circ \phi_1 \cdots f_4 \circ \phi_4$ are shown in Fig. 1.

One can immediately see that the learned functions indeed resolve nonlinearity in data generating nonlinear functions: the learned $f_1$ is a linear function since $\phi_1$ is a linear function; the other learned functions are all visually similar to the corresponding inverse functions of $\phi_i$’s. Moreover, one can clearly see that the composite functions all look affine.

Next, we test the parameter estimation performance. For this experiment, we generate data with five different nonlinear functions: (a) $e^x$, (b) $x + x^2$, (c) $\log(e^x + 1)$, (d) $\log(x + 1)$, (e) $x + \tanh(x)$. For each case, one of the five functions is used for all coordinates (features), i.e. $\phi_1 = \cdots = \phi_M$. The parameter settings are $M = 10, N = 1000$, and $r = 4$. We generate $\mathbf{A} \in \mathbb{R}^{10 \times 4}$ by sampling an i.i.d. normal distribution, and then take the absolute values, followed by a column normalization step. $S$ is similarly generated as in the first experiment. For this experiment, the $f_i$ functions are constrained to be the same: a constrained one-hidden-layer NN defined in (18), with $K = 40$ for all cases, to avoid unrealistic parameter tuning. In other words, all the NN share the same parameters. Since problem
(19) is nonconvex, different initialization could lead to different results. For this reason, the formulation (19) is optimized five times with different random initialization, and the result of smallest cost function value is used for subsequent steps of Algorithm 1. The performance metric we employ is mean squared error (MSE): $\text{MSE} = \| \hat{S} - S \|^2_F$.

Since our method is the first work dealing with this nonlinear model, the only baseline we can employ is MVES without considering nonlinear effects. For each setting, 100 trials with different randomly generated data are performed, and the empirical cumulative distribution function (CDF) of the resulting MSEs is reported in Fig. 2.

From Fig. 2, one can see that the proposed method yields significant improvements over applying MVES directly, in all the cases. Note that the x-axis in Fig. 2 is $\log_{10}(\text{MSE})$, hence our method yields several orders of magnitude improvement in accuracy over the baseline.

In the third experiment, we examine the estimation accuracy of matrix $S$ with different number of data samples – as our theory offers asymptotic guarantees, we would like to examine how the learning method performs in the case of finite data. The setting is similar to above: we generate $A$ randomly, with $M = 10$ and $r = 4$. Then different number of data samples (columns of $S$) are generated following the Dirichlet distribution. For each setting, we generate 50 data sets, and measure the resulting MSE of estimating $S$. Two nonlinear functions, $\phi(x) = e^x$ and $\phi(x) = \log(x + 1)$ are employed in data generation. For all these experiments, the number of neurons in $f$ is fixed to $K = 40$. The results are presented in Fig. 3. As can be seen, when the number of data samples is small, larger MSEs of estimating $S$ are observed. However, much better estimation performance is achieved with more data. Note also that the results presented in Fig. 3 are relatively worse than those in Fig. 2, which is mainly due to the small amount of available data.

In the next experiment, we vary the number of neurons in the neural network. The setting is similar to above: we generate $A$ randomly, with $M = 10$ and $r = 4$. We draw columns of $S$ by following a Dirichlet distribution, where the number of columns is $N = 1000$. The nonlinear function in data generation is $\phi(x) = e^x$. In total 50 data sets are generated. We run Algorithm 1 for different number of neurons $K$. The results are presented in Fig. 4. As can be seen from the figure, very good estimation results can be achieved even when there are only 4 neurons, and better results are obtained when we employ more neurons in the network. However, using only 1 neuron does not work: in many cases we manually examined, the learned $\hat{f}(X)$ is almost constant – the entries are all the same, and no meaningful $\hat{S}$ can be estimated. This can be explained by the fact that a single neuron is unlikely to be able to approximate the (scaled) inverse of $\phi$.

Next, we examine the estimation performance for different (but known) rank in the latent space. As before, we set $M = 10$, and $N = 1000$. The number of neurons are set to $K = 15$. We
use $\phi(x) = e^x$ in data generation, and 50 data sets are generated for each rank setting. The results are presented in Fig. 5. As can be seen from this figure, the estimation performance is insensitive to the different rank settings.

In the last synthetic data experiment, we test the proposed method when additive noise is present. Notice that the identifiability theory in this work is applicable under noiseless condition. It is therefore of great interest to assess how the proposed method performs when noise is present. For this task, we generate data via the following natural noise model

$$X = \phi(AS) + W,$$

where $W$ denotes a matrix with i.i.d. zero mean Gaussian noise. The variance of the entries is determined by setting the signal-to-noise-ratio (SNR), which is defined as

$$\text{SNR} = 10 \log_{10} \frac{\|\phi(AS)\|_F^2}{MN\sigma^2}. \quad (21)$$

The symbol $\sigma^2$ denotes the variance of the entries of $W$. We set $M = r = 4$, and $N = 1000$ for this simulation, and use $\phi(x) = e^x$ as the unknown nonlinear function for generating data. We vary SNR and compute the corresponding $\sigma^2$, and we repeat the experiment 10 times for each SNR setting. The averaged MSEs are reported in Table I. As can be seen from the table, the estimation results are reasonably good even when noise is present. In addition, the estimation quality improves when noise becomes smaller.

### B. Case Study With a Hyperspectral Image

We next perform an experiment on hyperspectral unmixing (HU). Unlike normal RGB images, a pixel in a hyperspectral image contains information on hundreds of spectral bands. With the more detailed spectral information, it is often assumed that different materials have their distinct spectral signatures (see e.g., [7] for details). Physically, each pixel is represented by a convex combination of materials that are present in that geographical patch. However, it is known that the collected measurement may encounter nonlinear distortion. The HU task involves identifying the spectra of the materials in a given region and the material concentrations in the different pixels (patches of the scene).

The image employed in this experiment is the Moffett Field captured in California, USA. The region has three main materials: water, soil, and vegetation. This scene is known for the existence of nonlinear mixture pixels, which usually poses a challenge to LMM-based HU algorithms such as MVES. The size of the image is $50 \times 50$, hence we have 2500 pixels. Each pixel is measured on 224 spectral bands, and we remove the water-absorbing bands following standard preprocessing [37], and end up with 200 spectral bands. In this case, each pixel $\{x_j, j \in [2500]\}$ is measured as a vector of length 200, i.e. $x_j \in \mathbb{R}^{200}$. A single band of the image is shown in Fig. 7: the top part is water, while the bottom part is land (sand and vegetation). To apply our method, we use the same $f_1$ on each of the 200 spectral bands, and fix $K = 40$. The latent dimension $r = 3$ is used since we know there are 3 materials in this scene. For comparison, in addition to MVES, we employ another three baselines that tackle the nonlinearity problem from a robust LMM identification viewpoint: SISAL [36], robust volume-minimization (RVoMin) [37], and robust nonnegative matrix factorization (rNMF) [51].

We plot the estimated $S$ in the known water region (top $15 \times 50$ part of Fig. 7). We take this part to compare the performance of the algorithms as it is clear that there is only one material (water) in this region, so the ground truth for each $s_j$ is a permutation of $[1, 0, 0]^T$, whereas the ground truth $s_j$ for other regions is unavailable and hard to obtain. Since the resulting columns of
\( \bar{S} \) live in a dimension-2 simplex, we project all the points into a 2D space, with the vertices of the triangle corresponding to the original vertices in the 3D space, as shown in Fig. 6.

From Fig. 6, we see that results of the proposed method coalesce around a coordinate vector \([0, 0, 1]^T\), which is almost identical to the ground truth. However results of the other methods are far away from any coordinate vector. This shows that by explicitly incorporating nonlinearity, the proposed method yields much improved results over prior art.

VI. CONCLUSION

This work proposed a novel and well-grounded nonlinear extension of the LMM data model, to account for nonlinear effects that naturally arise in real-world applications. The proposed model augments a widely used model, by allowing additional nonlinear distortions. It is an important problem to consider in practice, but a concrete study was sorely missing prior to this work. Much to one’s surprise, the seemingly impossible mission of figuring out unknown nonlinearities in the new data model can actually be accomplished up to affine transformations, as we show in this paper. A learning algorithm leveraging the power of neural networks was proposed to equalize the unknown nonlinear functions. Numerical experiments show clear advantages of the proposed method over classic LMM learning algorithms when nonlinearity is present.

APPENDIX A

SOME DEFINITIONS IN CONVEX GEOMETRY

Definition 2: (Convex cone) The convex cone of \( \{x_1, \ldots, x_N\} \) is defined as

\[
\text{cone}\{x_1, \ldots, x_N\} = \left\{ x \mid x = \sum_{j=1}^{N} x_j \theta_j, \theta_j \geq 0 \quad \forall j \in [N] \right\}.
\]

(22)

Definition 3: (Convex hull) The convex hull of \( \{x_1, \ldots, x_N\} \) is defined as

\[
\text{conv}\{x_1, \ldots, x_N\} = \left\{ x \mid x = \sum_{j=1}^{N} x_j \theta_j, \sum_{j=1}^{N} \theta_j = 1, \theta_j \geq 0 \quad \forall j \in [N] \right\}.
\]

(23)

Definition 4: (Simplex) A convex hull \( \text{conv}\{x_1, \ldots, x_N\} \) is called a simplex if \( x_1, \ldots, x_N \) are affinely independent, i.e., \( x_1 - x_N, \ldots, x_{N-1} - x_N \) are linearly independent.

A probability simplex is a special simplex, with all vertex vectors being the coordinate vectors, i.e. \( \forall i \in [N], x_i = e_i \) for some \( j \), where \( e_j \) has 1 at its \( j \)-th coordinate, and 0 for all other coordinates.

APPENDIX B

IDENTIFIABILITY OF LMM

We include some identifiability results for LMM, as they are important prior art that we build upon. In order to characterize identifiability of (1), let us introduce the following definition.

Definition 5: (Sufficiently scattered, [5], [16]) Let matrix \( S \in \mathbb{R}^{r \times N} \), where \( \mathbb{R}_{+}^{r \times N} \) is the nonnegative subset of \( \mathbb{R}^{r \times N} \). Matrix \( S \) is said to be sufficiently scattered (SS) if cone(\( S \)) satisfies: (a) \( \mathcal{C} \subseteq \text{cone}(S) \), where \( \mathcal{C} \) is a second order cone: \( \mathcal{C} = \{ x \in \mathbb{R}^r \mid x \geq \sqrt{r-1}\|x\|_2 \} \), (b) \( \text{cone}(S) \subseteq \text{cone}(Q) \), for any unitary matrix \( Q \in \mathbb{R}^{r \times r} \) that is not a permutation matrix.

An illustration for \( r = 3 \) of this definition is given in Fig. 8. We are viewing from the nonnegative orthant towards the origin. The 3 vertices of the triangle correspond to coordinates \([1, 0, 0], [0, 1, 0], [0, 0, 1]\), thus the triangle is the simplex where all columns of \( S \) reside. The inner circle corresponds to \( \mathcal{C} \). The sufficient scattered condition, shown in Fig. 8b, requires columns of \( S \) to be well-scattered on the simplex, so that cone(\( S \)) encloses \( \mathcal{C} \). This condition is in fact fairly relaxed, as discussed in [52].

We should also mention that a stronger condition on \( S \), termed “separability,” which is shown in Fig. 8a, can also help establish identifiability of LMM. We refer the interested readers to [4],
Here we present SS, as it is more relaxed than the separability condition.

Based on this VolMin criterion, the following theorem established identifiability of model (1).

Theorem 2: [16] Let the matrices $A$ and $S$ satisfy $\text{rank}(A) = \text{rank}(S) = r$. Suppose $S$ satisfies the SS condition. Under the generative model (1), the VolMin criterion (2) uniquely identifies both $A$ and $S$ up to a permutation. Specifically, any optimal solution to (2) takes the form

\[ B = AP, H = \Pi^TS, \]

where $\Pi$ is a permutation matrix.

A proof of this result can be found in [16]. We mention that by Theorem 2, given that $S$ satisfies SS, the only remaining indeterminacy is a permutation of the columns (rows) of $A$ (resp. $S$), which is unavoidable – but also inconsequential in most applications.

### Appendix C

**Proof of Lemma 1**

**Proof:** Assume without loss of generality that the two nonzero columns are the first and second column. Let us denote

\[ \zeta(s_1, s_2, \ldots, s_{r-1}) := \sum_{i=1}^{M} \psi_i(a_i^T s) = 1, \quad s \in \text{int } \Delta_r. \quad (24) \]

Note that $\zeta$ is a function of $(r-1)$ variables $s_1, s_2, \ldots, s_{r-1}$, since $1^T s = 1$. Equation (24) suggests that $\zeta$ is a constant function on $\Delta_r$. Taking derivative with respect to (w.r.t.) $s_1$ and $s_2$, we get

\[ \frac{\partial \zeta}{\partial s_1} = \sum_{i=1}^{M} a_i(1) \psi'_i(a_i^T s), \quad (25) \]

and

\[ \frac{\partial^2 \zeta}{\partial s_1 \partial s_2} = \sum_{i=1}^{M} a_i(1) a_i(2) \psi''_i(a_i^T s) = 0. \quad (26) \]

By the assumption on $A$, we have $a_i(1) a_i(2) > 0 \ \forall i$. The assumption that $\psi_i$’s are all convex (or concave) translates to $\psi''_i \geq 0$ (or $\psi''_i \leq 0$), for all $i \in [M]$. From (26), we conclude that $\psi''_i = 0 \ \forall i$, which suggests that all the $\psi_i$’s are affine.

While our result is novel and we conceived it for use in our work, other interesting results concerning functional equations can be found in [28], [29].

### Appendix D

**Proof of Theorem 1**

**Proof:** Given assumptions (A2) and equation (7), (a) is a direct consequence of Lemma 2.

For (b), we note that from (a), $k_i(t) = d_i t + b_i$ for some constants $d_i$ and $b_i$. Let $x = \phi_i(t)$, then $t = \phi^{-1}_i(x)$. Plugging into $f_i(\phi_i(t)) = d_i t + b_i$, we obtain $f_i(x) = d_i \phi^{-1}_i(x) + b_i$. 

### Appendix E

**Proof of Proposition 1**

To prove Proposition 1, we need Lemma 4 and Lemma 5, which are presented here followed by their respective proof.

**Lemma 4:** Suppose $A \in \mathbb{R}^{m \times r}$ is full rank and incoherent, i.e. $e_i \notin \text{Range}(A) \ \forall i \in [m]$. Then $A = [\mathbf{1}^n_1]$ is incoherent.

This lemma asserts that if a matrix $A$ is incoherent, then appending a row of all 1’s preserves incoherence.

**Proof:** The incoherence condition means that there is no such $y \in \mathbb{R}_r^n$, such that $Ay = e_j$ for any $i \in [m]$. Suppose there is a $\tilde{y} \in \mathbb{R}_r^n$, such that $A\tilde{y} = e_j$ for some $j \in [m+1]$. There are two cases

1) \[ 1 \leq j \leq m: \] This means we have $\tilde{y}$ such that $A\tilde{y} = e_j$ for some $j \in [m]$ -- a contradiction to the assumption that $A$ is incoherent.

2) \[ j = m + 1: \] This means that $A\tilde{y} = 0_m$ for $\tilde{y} \neq 0$ -- a contradiction to the assumption that $A$ is full rank.

Hence $A$ is incoherent if $A$ is full rank and incoherent.

**Lemma 5:** For a tall and full rank matrix $A \in \mathbb{R}^{m \times r}$, where $A$ is incoherent, there exists a $d \in \mathbb{R}^m$, such that

\[ A^Td = 0, \quad (27a) \]

\[ \|d\|_0 = m. \quad (27b) \]

**Proof:** Let $U \in \mathbb{R}^{m \times (m-r)}$ be a set of bases of the null space of $A$, i.e.

\[ \text{Range}(U) = \text{Null}(A). \quad (28) \]

By assumption, $A$ is incoherent, hence $e_j \notin \text{Range}(A) \ \forall j \in [m]$. For any $j$, we have the decomposition

\[ e_j = \hat{e}_j + \xi_j, \quad (29) \]

where $\hat{e}_j \in \text{Range}(A)$ and $\xi_j \in \text{Range}(U)$. Since $e_j \notin \text{Range}(A)$, we have $e_j^T U = \xi_j^T U \neq 0_{m-r} \forall j \in [m]$, which means $U$ does not have a row that is all-zero.

Let $I_1, \ldots, I_{m-r}$ be the index sets of nonzero entries in each column of $U$, then we have $\bigcup_{j=1}^{I_{m-r}} I_j = [m]$ since $U$ does not have an all-zero row. Let us present the following useful fact.

**Fact 1:** Let $x, y \in \mathbb{R}^m$, with sets $I_x$ and $I_y$ being the sets of indices of nonzero entries, then we can find a vector $z \in \text{Span}(x, y)$, such that $I_z = I_x \cup I_y$.

**Proof:** Let $a = \frac{1}{\max_{j \notin I_y} |x_j|}$ and $b = \frac{2}{\min_{j \in I_y, y_j = 0} |y_j|}$. The denominator of $b$ is the minimum of absolute value of the nonzero entries of $y$. Consider the vector

\[ z = ax + by. \quad (30) \]

By the choice of $a$ and $b$, we have $\max_j |ax_j| = 1$ and $\min_{j: y_j \neq 0} |by_j| = 2$. Hence for any $j$ where $x_j \neq 0$ and $y_j \neq 0$,
we have $a x + b y \neq 0$. This shows that there exists a $z \in \text{Span}\{x, y\}$, such that $T_z = T_x \cup T_y$.

We can now use Fact 1 to show that there exists a fully dense $d \in \text{Range}(U)$. Consider the first two columns of $U$: $U_1$ and $U_2$. From Fact 1, we can find a vector $u \in \text{Span}\{U_1, U_2\}$, such that $T_u = T_1 \cup T_2$. Now consider $u$ and $U_3$, invoking Fact 1 again, we can find a vector $\nu \in \text{Span}\{u, U_3\}$, such that $T_\nu = T_1 \cup T_3$. Continuing this process, we can find a vector $d \in \text{Span}\{U_1, \ldots, U_{m-r}\} = \text{Range}(U)$, such that $T_d = \{ r + m-r \} = \{ m \}$; meaning that $d \in \text{Range}(U)$ and is fully dense. Since $d \in \text{Range}(U)$, we have $A^T d = 0$.

**Proof of Proposition 1:** Consider a matrix $A \in \mathbb{R}^{m \times r}$ that is tall, full rank, and incoherent, we can rewrite (12a) as

\[
\begin{bmatrix} A^T & 1 \end{bmatrix} \begin{bmatrix} d \\ -1 \end{bmatrix} = 0
\]  

(31)

Let us denote $\hat{A}^T = [A^T 1]$. Then we can see that 1) $\hat{A} \in \mathbb{R}^{(m+1)\times r}$ is tall and full rank, 2) $\hat{A}$ is incoherent by Lemma 4. We see that $A$ satisfies all the conditions in Lemma 5, hence there exists a $d \in \mathbb{R}^{m+1}$ such that $\hat{A}^T d = 0$, and $\|d\|_0 = m + 1$. Since $d$ is fully dense, we construct $\tilde{d} \in \mathbb{R}^{m+1}$ as

\[
\tilde{d} := -d/(m + 1).
\]  

(32)

By this construction, we have $\tilde{d}(m + 1) = -1$. In addition, $A^T \tilde{d} = 0$ as it is merely a scaled version of $d$. Let $\tilde{A} = \tilde{d}^T : m \in \mathbb{R}^m$, then we have

\[
A^T \tilde{A} \tilde{d} = 1, \quad \|\tilde{d}\|_0 = m.
\]  

(33)

Hence we managed to show the existence of a $d$ that satisfies both (12a) and (12b) for any $A$ that satisfies the conditions in Proposition 1.

**ACKNOWLEDGMENT**

The authors would like to thank NVIDIA for a GPU donation, which was used to conduct the experiments in this paper.

**REFERENCES**

[1] T. Hofmann, “Probabilistic latent semantic analysis,” in Proc. 15th Conf. Uncertainty Artif. Intell., 1999, pp. 289–296.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, Jan. 2003.

[3] S. Arora, R. Ge, and A. Moitra, “Learning topic models—going beyond ICA,” in Proc. 20th Int. Conf. Artificial Intell. Statist., 2013, pp. 280–288.

[4] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, vol. 1. Springer Series in Statistics, 2009. Berlin, Germany: Springer.

[5] Y.-W. Yang, C.-Y. Chi, T.-H. Chan, and Y. Wang, “Blind separation of positive dependent sources by non-negative least-correlated component analysis,” in Proc. 16th IEEE Signal Process. Soc. Workshop Mach. Learn. Signal Process., 2006, pp. 73–78.

[6] C. Efthimiou, Introduction to Functional Equations, 2010. [Online]. Available: https://www.msri.org/people/staff/levy/files/MCL/Efthimiou/100914book.pdf, Accessed: May 9, 2020.

[7] M. Kuczma, An Introduction to the Theory of Functional Equations and Inequalities: Cauchy’s Equation and Jensen’s Inequality. Berlin, Germany: Springer, 2009.

[8] Y. Deville and L. T. Duarte, “An overview of blind source separation methods for linear-quadratic and post-nonlinear mixtures,” in Proc. Int. Conf. Latent Variable Anal. Signal Separation, 2015, pp. 155–167.

[9] J. M. Bioucas-Dias et al., “Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 5, no. 2, pp. 354–379, Apr. 2012.

[10] X. Miao, P. Sarkar, and D. Chakrabarti, “On mixed memberships and symmetric nonnegative matrix factorizations,” in Proc. 34th Int. Conf. Mach. Learn., 2017, pp. 2324–2333.

[11] J. M. Bioucas-Dias et al., “Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 5, no. 2, pp. 354–379, Apr. 2012.

[12] F.-Y. Wang, C.-Y. Chi, T.-H. Chan, and Y. Wang, “Blind separation of positive dependent sources by non-negative least-correlated component analysis,” in Proc. 16th IEEE Signal Process. Soc. Workshop Mach. Learn. Signal Process., 2006, pp. 73–78.
[35] 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.

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

[37] X. Fu, K. Huang, B. Yang, W.-K. Ma, and N. D. Sidirooulos, “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.

[38] R. Heylen, M. Parente, and P. Gader, “A review of nonlinear hyperspectral unmixing methods,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 7, no. 6, pp. 1844–1868, Jun. 2014.

[39] N. Dobiogeon, J.-Y. Tournier, C. Richard, J. C. M. Bermudez, S. McLaughlin, and A. O. Hero, “Nonlinear unmixing of hyperspectral images: Models and algorithms,” IEEE Signal Process. Mag., vol. 31, no. 1, pp. 82–94, Jan. 2014.

[40] T.-H. Chan, W.-K. Ma, C.-Y. Chi, and Y. Wang, “A convex analysis framework for blind separation of non-negative sources,” IEEE Trans. Signal Process., vol. 56, no. 10, pp. 5120–5134, Oct. 2008.

[41] P. Santigo and H. D. Gage, “Statistical models of partial volume effect,” IEEE Trans. Image Process., vol. 4, no. 11, pp. 1531–1540, Nov. 1995.

[42] M. S. Vafaee and A. Gjedde, “Model of blood–brain transfer of oxygen explains nonlinear flow-metabolism coupling during stimulation of visual cortex,” J. Cerebral Blood Flow Metabolism, vol. 20, no. 4, pp. 747–754, 2000.

[43] A.-L. Lin, P. T. Fox, J. Hardies, T. Q. Duong, and J.-H. Gao, “Nonlinear coupling between cerebral blood flow, oxygen consumption, and ATP production in human visual cortex,” Proc. Nat. Acad. Sci., vol. 107, no. 18, pp. 8446–8451, 2010.

[44] X. Liang, A. Connelly, and F. Calamante, “Graph analysis of resting-state ASL perfusion MRI data: Nonlinear correlations among CBF and network metrics,” Neuroimage, vol. 87, pp. 265–275, 2014.

[45] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” Found. Comput. Math., vol. 9, 2009.

[46] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” Neural Netw., vol. 2, no. 5, pp. 359–366, 1989.

[47] A. R. Barron, “Universal approximation bounds for superpositions of a sigmoidal function,” IEEE Trans. Inf. Theory, vol. 39, no. 3, pp. 930–945, May 1993.

[48] T. F. Coleman and Y. Li, “An interior trust region approach for nonlinear minimization subject to bounds,” SIAM J. Optim., vol. 6, no. 2, pp. 418–445, 1996.

[49] X. Yi, Z. Wang, C. Caramanis, and H. Liu, “Optimal linear estimation under unknown nonlinear transform,” in Proc. Advances Neural Inf. Process. Syst., 2015, pp. 1549–1557.

[50] S. Chen and A. Banerjee, “Sparse linear isotonic models,” in Proc. 21st Int. Conf. Artif. Intell. Statist., 2018, pp. 1270–1279.

[51] C. Févotte and N. Dobiogeon, “Nonlinear hyperspectral unmixing with robust nonnegative matrix factorization,” IEEE Trans. Image Process., vol. 24, no. 12, pp. 4810–4819, Dec. 2015.

[52] K. Huang, X. Fu, and N. D. Sidirooulos, “Learning hidden Markov models from pairwise co-occurrences with applications to topic modeling,” in Proc. 35th Int. Conf. Mach. Learn., 2018, pp. 2068–2077.

[53] 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.

[54] S. Arora, R. Ge, F. Koehler, T. Ma, and A. Moitra, “Provable algorithms for inference in topic models,” in Proc. 33rd Int. Conf. Mach. Learn., 2016, pp. 2859–2867.

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

[56] X. Fu, K. Huang, B. Yang, W.-K. Ma, and N. D. Sidirooulos, “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.

[57] R. Heylen, M. Parente, and P. Gader, “A review of nonlinear hyperspectral unmixing methods,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 7, no. 6, pp. 1844–1868, Jun. 2014.

[58] N. Dobiogeon, J.-Y. Tournier, C. Richard, J. C. M. Bermudez, S. McLaughlin, and A. O. Hero, “Nonlinear unmixing of hyperspectral images: Models and algorithms,” IEEE Signal Process. Mag., vol. 31, no. 1, pp. 82–94, Jan. 2014.

[59] T.-H. Chan, W.-K. Ma, C.-Y. Chi, and Y. Wang, “A convex analysis framework for blind separation of non-negative sources,” IEEE Trans. Signal Process., vol. 56, no. 10, pp. 5120–5134, Oct. 2008.

[60] P. Santigo and H. D. Gage, “Statistical models of partial volume effect,” IEEE Trans. Image Process., vol. 4, no. 11, pp. 1531–1540, Nov. 1995.

[61] M. S. Vafaee and A. Gjedde, “Model of blood–brain transfer of oxygen explains nonlinear flow-metabolism coupling during stimulation of visual cortex,” J. Cerebral Blood Flow Metabolism, vol. 20, no. 4, pp. 747–754, 2000.

[62] A.-L. Lin, P. T. Fox, J. Hardies, T. Q. Duong, and J.-H. Gao, “Nonlinear coupling between cerebral blood flow, oxygen consumption, and ATP production in human visual cortex,” Proc. Nat. Acad. Sci., vol. 107, no. 18, pp. 8446–8451, 2010.

[63] X. Liang, A. Connelly, and F. Calamante, “Graph analysis of resting-state ASL perfusion MRI data: Nonlinear correlations among CBF and network metrics,” Neuroimage, vol. 87, pp. 265–275, 2014.

[64] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” Found. Comput. Math., vol. 9, 2009.

[65] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” Neural Netw., vol. 2, no. 5, pp. 359–366, 1989.

[66] A. R. Barron, “Universal approximation bounds for superpositions of a sigmoidal function,” IEEE Trans. Inf. Theory, vol. 39, no. 3, pp. 930–945, May 1993.

[67] T. F. Coleman and Y. Li, “An interior trust region approach for nonlinear minimization subject to bounds,” SIAM J. Optim., vol. 6, no. 2, pp. 418–445, 1996.

[68] X. Yi, Z. Wang, C. Caramanis, and H. Liu, “Optimal linear estimation under unknown nonlinear transform,” in Proc. Advances Neural Inf. Process. Syst., 2015, pp. 1549–1557.

[69] S. Chen and A. Banerjee, “Sparse linear isotonic models,” in Proc. 21st Int. Conf. Artif. Intell. Statist., 2018, pp. 1270–1279.

[70] C. Févotte and N. Dobiogeon, “Nonlinear hyperspectral unmixing with robust nonnegative matrix factorization,” IEEE Trans. Image Process., vol. 24, no. 12, pp. 4810–4819, Dec. 2015.

[71] K. Huang, X. Fu, and N. D. Sidirooulos, “Learning hidden Markov models from pairwise co-occurrences with applications to topic modeling,” in Proc. 35th Int. Conf. Mach. Learn., 2018, pp. 2068–2077.

[72] 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.

[73] S. Arora, R. Ge, F. Koehler, T. Ma, and A. Moitra, “Provable algorithms for inference in topic models,” in Proc. 33rd Int. Conf. Mach. Learn., 2016, pp. 2859–2867.