Nonlinear Dependent Component Analysis: Identifiability and Algorithm

Qi Lyu
School of EECS, Oregon State University
Corvallis, OR 97331, USA
lyuqi@oregonstate.edu

Xiao Fu
School of EECS, Oregon State University
Corvallis, OR 97331, USA
xiao.fu@oregonstate.edu

***Abstract*—This work studies the model identifiability of a class of post-nonlinear mixture models with dependent latent components, in particular, components that form vectors residing in the probability simplex. This problem is motivated by applications such as hyperspectral unmixing under nonlinear distortion effects. Prior works tackle nonlinear component analysis using statistical independence, which is no longer applicable in our case. A recent work by Yang *et al.* offered a solution leveraging functional equations, but the model identifiability conditions are somewhat restrictive. The implementation there also has difficulties; e.g., the function approximator used in their work may not be able to approximate general nonlinear distortions and the formulated optimization problem is hard to handle. In this work, we substantially improve both the theoretical and practical aspects. To be specific, we offer a much tighter identifiability result and an easy-to-implement neural network-based algorithm—without sacrificing function approximation capabilities. Numerical experiments are employed to support our design.**

***Index Terms*—post-nonlinear mixture, dependent source, probability simplex, identifiability, neural networks**

I. Introduction

Latent component analysis has been an essential tool for a large variety of applications in signal processing (SP) and machine learning (ML). Many component analysis tools have been proposed, e.g., principal component analysis (PCA) [1], independent component analysis (ICA) [2], [3], nonnegative matrix factorization (NMF) [4], [5], dictionary learning/sparse coding [6], and tensor decomposition models [7], just to name a few.

One of the most important theoretical aspects pertaining to component analysis is *model identifiability*—since these tools are oftentimes associated with unsupervised learning and blind signal processing tasks, e.g., topic model learning [8], community detection [9], and blind source separation [2]. With model identifiabilty, the latent components of interest can be identified (often up to trivial ambiguities such as scaling and permutation) through learning the model parameters of the employed component analysis models from the observed data.

Establishing model identifiability is a nontrivial task. In a nutshell, many component analysis models can be understood as matrix factorization models—which is in general non-unique, thereby lacking identifiability. A common practice to circumvent this issue is introducing structural information as constraints, e.g., statistical independence in ICA, nonnegativity in NMF, and sparsity in dictionary learning. The identifiability analyses for these classic component analysis models are elegant, and the model uniqueness results have improved many core tasks in SP and ML; see [2], [4], [6], [7], [10].

On the other hand, the classic component analysis models are mostly variants of matrix and tensor factorization models. These models essentially assume that all the data vectors are generated from a *linear* subspace (a Khatri-Rao subspace for tensors). This is of course over-simplified for reality—since nonlinear distortions happen ubiquitously. Starting from the 1980s, efforts have been made towards incorporating nonlinear distortions into component analysis [11]. One notable line of work is the so-called *nonlinear independent component analysis* (nICA) [12]–[19]. The nICA framework considers *unknown* nonlinear distortions on top of the ICA model. One take-home point learned is that general nonlinear distortion is not identifiable under the framework of ICA [12]. To circumvent this, one may exploit certain structures of nonlinear distortions—e.g., under the so-called post-nonlinear mixture model [15]–[18] that is considered realistic for many sensing problems. One may also utilize more prior information from the data to remove nonlinear distortions, e.g., temporal correlations; see [13], [14], [19].

The model identifiability results under the nonlinear ICA frameworks are encouraging—showing that nonlinear distortions may be provably removable, under some conditions. However, assuming statistical independence among the latent components may be stringent. To relax this assumption, the recent work in [20] addresses the problem of *dependent component* identification under the post-nonlinear mixture model. To be specific, the authors of [20] considered a model where the latent components are nonnegative and sum-to-one—which is often considered in weighted mixture models such as soft clustering [21] and hyperspectral imaging [22]. Working from there, and combining insights from NMF identifiability, latent component identifiability was shown.

However, some challenges remain. First, the work in [20] assumes that the model parameters are all nonnegative, which may restrict the applicability in some cases. Second, the

This work is supported in part by the National Science Foundation (NSF) under projects ECCS 1808159 and NSF III-1910118, and the Army Research Office (ARO) under project ARO W911NF-19-1-0407.

nonlinear model identifiability hinges on a special assumption that the composition of the learned nonlinearity-compensating function and the nonlinear distortion is convex or concave—which is hard to verify or control, thereby being restrictive. Third, the work in [20] utilizes a neural network (NN) to approximate the "inverse" of the unknown nonlinear distortion, but the NN used there has positive weights for regularity purposes. This may cause performance losses—although NNs are universal function approximators, the function-approximation capacity of positive NNs is unknown.

In this work, we offer a new solution for the nonlinear model identification problem in [20] and its extensions. Our contribution is twofold: First, we offer a new identifiability result that does not rely on the restrictive assumptions used in [20]. This may substantially enlarge the range of applicable cases for this nonlinear component analysis model. Second, we propose a general-purpose neural network (other than a special positive neural network) based implementation for the formulated problem. This way, the risk of not being able to approximate certain nonlinear functions is removed. The associated optimization problem is also much easier to handle, leveraging existing optimizers for NNs, e.g., `Adam` [23]—which makes our implementation easily scalable. Extensive experiments are employed to showcase the performance of the proposed approach.

II. Background

A. Linear and Nonlinear Independent Component Analysis

Many classic latent component analysis models start with the following linear mixture model (LMM):

$$\boldsymbol{x}_\ell = \boldsymbol{A}\boldsymbol{s}_\ell,\ \ell = 1, \cdots, N \tag{1}$$

where $\boldsymbol{x}_\ell \in \mathbb{R}^M$ denotes the ℓth observed data sample, $\boldsymbol{A} \in \mathbb{R}^{M\times K}$ the "mixing system" (or the basis of the subspace where $\boldsymbol{x}_\ell$'s reside), and $\boldsymbol{s}_\ell \in \mathbb{R}^K$ the vector that holds the K latent components.

Many latent component analysis models are concerned with identifying $\boldsymbol{A}$ and $\{\boldsymbol{s}_\ell\}$ from $\{\boldsymbol{x}_\ell\}$. Model identifiability has been established under various conditions. For example, ICA assumes $s_{k,\ell}$'s are statistically independent [24], and NMF models assume $\boldsymbol{A}$ and $\boldsymbol{s}_\ell$ are element-wise nonnegative [4], [25], [26]. Beyond the classic linear mixture models, nonlinear mixtures have also been considered, mostly under the umbrella of nonlinear ICA. The notable line of work in [12]–[14], [19] considers the model

$$\boldsymbol{x}_\ell = \boldsymbol{g}(\boldsymbol{s}_\ell),\ \ell = 1, 2, \ldots, N$$

where $\boldsymbol{g}(\cdot) : \mathbb{R}^K \to \mathbb{R}^K$ is an invertible continuous nonlinear distortion applied on to the latent components. This model is in general not identifiable, even if one assumes that $s_{k,\ell}$'s are statistically independent [12]. A series of additional structural information on $\boldsymbol{s}_\ell$ (e.g., temporal correlations) has been exploited to establish identifiability. Another way for establishing model identifiability is to exploit structural information of the nonlinear distortions, other than that of the latent components. The post-nonlinear mixture (PNM) model is often considered [15]–[18], [20], where we have

$$\boldsymbol{x}_\ell = \boldsymbol{g}(\boldsymbol{A}\boldsymbol{s}_\ell),\ \ell = 1, 2, \cdots, N \tag{2}$$

in which $\boldsymbol{g} = [g_1(\cdot), \cdots, g_M(\cdot)]^\top$ and $g_m(\cdot) : \mathbb{R} \to \mathbb{R}$ is a scalar-to-scalar nonlinear continuous invertible function. The PNM model has many applications in sensing-related problems, e.g., radar and biomedical sensing; see discussions in [17]. Under the PNM model, the identifiability of $\boldsymbol{s}_\ell$ has also been established—mainly using the statistical independence of the latent components [17], [18].

B. Nonlinear Dependent Component Analysis: Prior Art

Very recently, Yang et al. [20] considered an interesting problem under the PNM model. Instead of having $s_{k,\ell}$'s be statistically independent, the model assumption is that

$$\boldsymbol{s}_\ell \in \Delta, \quad \Delta = \{\boldsymbol{x} \in \mathbb{R}^K | \boldsymbol{1}^\top \boldsymbol{x} = 1, \boldsymbol{x} \geq \boldsymbol{0}\}. \tag{3}$$

Under this model, the observations $\boldsymbol{x}_\ell$'s are generated as weighted combinations of $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N$ and then distorted by $g_1(\cdot), \ldots, g_M(\cdot)$. Note that weighted combination is a particularly important model that finds applications in topic modeling [8], soft clustering [21], and hyperspectral unmixing [22]. The nonlinear distortion part is also important in modeling additional distortions happening in practice—e.g., nonlinearity is often observed in hyperspectral imaging [27]. Note that since $\boldsymbol{1}^\top \boldsymbol{s}_\ell = 1$, the latent components are dependent—which means that the classic results from nICA do not apply to this case.

The work in [20] utilizes the sum-to-one structure to construct a functional equation, and shows that under some conditions a carefully constructed model identification criterion can "remove" $\boldsymbol{g}(\cdot)$ through learning a nonlinear function $\boldsymbol{f}$. Then, the problem for identifying $\boldsymbol{s}_\ell$ becomes a classic NMF problem. There are a number of caveats. First, the assumption for $\boldsymbol{g}(\cdot)$ removal might be too restrictive. The assumptions include that $\boldsymbol{A}$ being nonnegative and incoherent, and the learned function satisfies that $\boldsymbol{f} \circ \boldsymbol{g}$ is a convex or concave function. This condition is particularly hard since one has no control for it—or a way of checking it. Second, when implementing the learning criterion, the authors in [20] uses a neural network to represent $\boldsymbol{f}$. However, the NN employed is with positive network weights for enforcing function invertibility. This construction makes optimization easier, but may have hindered the function approximation capability of NNs.

III. Proposed Approach

In this work, we offer a new solution under the PNM model and (3) that effectively circumvent the challenges in [20].

A. Proposed Formulation

Ideally, we expect to learn element-wise invertible nonlinear function $\boldsymbol{f}$ such that the following holds:

$$\boldsymbol{1}^\top \boldsymbol{f}(\boldsymbol{x}_\ell) = \boldsymbol{1}^\top \boldsymbol{f} \circ \boldsymbol{g}(\boldsymbol{A}\boldsymbol{s}_\ell) = \boldsymbol{1}^\top \boldsymbol{h}(\boldsymbol{A}\boldsymbol{s}_\ell) = 1, \tag{4}$$

for all ℓ. Here $\boldsymbol{h}$ is also an element-wise function with $h_i = f_i \circ g_i$. In other words, our learning objective is to find an invertible function to reverse the distortions introduced by $\boldsymbol{g}$ so that the sum-to-one condition can be satisfied.

Formally, we wish to have the following criterion satisfied in terms of $\boldsymbol{f}$-searching:

$$\begin{aligned} &\text{find } \boldsymbol{f} && \text{(5a)}\\ &\text{s.t. } \mathbf{1}^\top \boldsymbol{f}(\boldsymbol{x}) = 1,\ \forall \boldsymbol{x} \in \mathcal{X} && \text{(5b)}\\ &\quad f_i \text{ is invertible over } \mathcal{X},\ \forall i, && \text{(5c)} \end{aligned}$$

where

$$\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^M |\ \boldsymbol{x} = \boldsymbol{g}(\boldsymbol{A}\boldsymbol{s}),\ \forall \boldsymbol{s} \in \text{int}\Delta,\ \boldsymbol{s} \in \mathbb{R}^K\}$$

in which $\text{int}\Delta$ means the interior of Δ. Note that the criterion is identical to what was proposed in [20]. The difference lies in model assumptions. In particular, Yang *et al.* assumed that $\boldsymbol{A}$ is generic, nonnegative and incoherent in [20]. In this work, we only require that $\boldsymbol{A}$ is generic (i.e., the entries are drawn from any jointly continuous distribution). Yang *et al.* showed the following

Theorem 1 *[20] Consider the post-nonlinear mixture model $\boldsymbol{x}_\ell = \boldsymbol{g}(\boldsymbol{A}\boldsymbol{s}_\ell)$ with the constraint $\boldsymbol{s}_\ell \in \text{int}\Delta$, where $g_m(\cdot)$ for all m are continuous and invertible. Assume that we have infinite samples such that $\mathcal{X}$ is available. Assume that $\boldsymbol{A} \in \mathbb{R}^{M\times K}$ is drawn from any joint continuous distribution. In addition, assume that $M \geq K$ and that*

1) *$\boldsymbol{A}$ is nonnegative and is incoherent (see definition in [20]); and that*
2) *by solving problem (5), the resulting h_i's are convex or concave.*

Then h_i has to be an affine functions almost surely; i.e., any f_i satisfying (5) makes the following holds:

$$h_i(x) = f_i \circ g_i(x) = c_i x + d_i,\ i = 1, \ldots, M,$$

where c_i, d_i are constants. In addition, if $\sum_{i=1}^M d_i \neq 0$, we have

$$h_i(x) = f_i \circ g_i(x) = \alpha_i x,\ i = 1, \ldots, M,\ \alpha_i \neq 0,\ \forall i.$$

The theorem is of interest, since it for the first time showed that the PNM model is identifiable even under dependent latent components. The challenge is that both conditions 1) and 2) may be restrictive—especially condition 2). In this work, we show that the key conditions in 1) and 2) are in fact not needed. To proceed, we show the following:

Lemma 1 *Consider $\boldsymbol{s} = [s_1, \ldots, s_K]^\top \in \text{int}\Delta_K$. The following always holds true:*

$$\frac{\partial s_i}{\partial s_j} = 0,$$

for $i \neq j$ where $i, j = 1, \cdots, K-1$.

Proof: Lemma 1 can be shown as follows. First, for $\boldsymbol{s}_\ell \in \text{int } \Delta_K$, we only have $K-1$ free variables, i.e., without loss of generality, s_i for $i = 1, \ldots, K-1$. For any fixed $\bar{s}_i$, s_j can be any possible values in a nonempty continuous domain (e.g., if $s_i = 0.5$ then the domain of s_j is $(0, 0.5)$ regardless of other components). Hence, if one treats s_i as a function of s_j, then the sensitivity of s_i w.r.t. s_j is defined as

$$\begin{aligned} \frac{\partial s_i}{\partial s_j} &= \lim_{\Delta s_j \to 0} \frac{s_i(s_j + \Delta s_j) - s_i(s_j)}{\Delta s_j}\\ &= \lim_{\Delta s_j \to 0} \frac{\bar{s}_i - \bar{s}_i}{\Delta s_j}\\ &= 0. \end{aligned}$$

This completes the proof. ■

This lemma is important for deriving our main theorem:

Theorem 2 (Nonlinearity Removal) *Consider the post-nonlinear mixture model $\boldsymbol{x}_\ell = \boldsymbol{g}(\boldsymbol{A}\boldsymbol{s}_\ell)$ with the constraint $\boldsymbol{s}_\ell \in \text{int}\Delta$, where $g_m(\cdot)$ for all m are continuous and invertible. Assume that we have infinite samples such that $\mathcal{X}$ is available. Assume that*

$$M \geq K \geq 3$$

and that $\boldsymbol{A} \in \mathbb{R}^{M\times K}$ is drawn from any joint continuous distribution. Then, by solving problem (5), the resulting h_i's are affine functions almost surely; i.e., any f_i satisfying (5) makes the following holds:

$$h_i(x) = f_i \circ g_i(x) = c_i x + d_i,\ i = 1, \ldots, M,$$

where c_i, d_i are constants. In addition, if $\sum_{i=1}^M d_i \neq 0$, we have

$$h_i(x) = f_i \circ g_i(x) = \alpha_i x,\ i = 1, \ldots, M,\ \alpha_i \neq 0,\ \forall i. \quad (6)$$

The proof sketch is as follows. According to Lemma 1 we have $\frac{\partial s_i}{\partial s_j} = 0$ for $\boldsymbol{s} \in \text{int } \Delta_K$. Then, by taking second order derivatives of the equality constraint in (4) w.r.t. s_i and s_j, it ends up with a system of linear equations that involves the vector $\boldsymbol{h}'' = [h_1'', \cdots, h_M'']^\top$, i.e., $\boldsymbol{H}\boldsymbol{h}'' = \mathbf{0}$. By utilizing the assumptions, one can show that $\boldsymbol{H}$ has full column rank thus it immediately implies $\boldsymbol{h}'' = 0$, which further leads to that all h_i's are affine.

B. Latent Component Identification

Note that under (6), the following holds:

$$\boldsymbol{f}(\boldsymbol{x}_\ell) = \boldsymbol{C}\boldsymbol{A}\boldsymbol{s}_\ell = \boldsymbol{B}\boldsymbol{s}_\ell,\ \ell = 1, 2, \ldots, N$$

where $\boldsymbol{C} = \text{Diag}(\alpha_1, \ldots, \alpha_M)$. This model is identical to the structural matrix factorization model in [25], [28], [29], which is identifiable if $\{\boldsymbol{s}_\ell\}$ satisfies certain conditions, e.g., the separability condition or the sufficiently scattered condition; see details in [4], [25], [29], [30]. Hence, a simple strategy is to first implementing the criterion in (4) for nonlinearity removal. Then, any structural matrix factorization algorithm proposed in the literature, e.g., those in [29], [30], can be employed for identifying $\boldsymbol{s}_\ell$ from $\boldsymbol{f}(\boldsymbol{x}_\ell)$. In this work, we utilize the minimum-volume enclosing simplex (MVES) algorithm from [30] for $\boldsymbol{s}_\ell$-identification after nonlinearity removal.

C. Neural Network-based Implementation

We have shown that solving Problem (5) removes the nonlinear distortions. However, Problem (5) is not really "workable" since it involves continuous functional searching. To approach this formulation, we parameterize the function $\boldsymbol{f}$ with neural networks due to their universal approximation ability. Each f_i is approximated by an individual neural network. Hence, the practical formulation is as follows:

$$\min_{\boldsymbol{\theta_f},\boldsymbol{\theta_g}} \sum_{\ell=1}^{N} \left(1 - \mathbf{1}^\top \boldsymbol{f}_{\rm NN}(\boldsymbol{x}_\ell)\right)^2 + \lambda \sum_{\ell=1}^{N} \|\boldsymbol{x}_\ell - \boldsymbol{g}_{\rm NN}(\boldsymbol{f}_{\rm NN}(\boldsymbol{x}_\ell))\|_2^2 \tag{7}$$

where two neural networks $\boldsymbol{f}_{\rm NN} = [f_{\rm NN}^{(1)}, \cdots, f_{\rm NN}^{(M)}]$ and $\boldsymbol{g}_{\rm NN} = [g_{\rm NN}^{(1)}, \cdots, g_{\rm NN}^{(M)}]$ are parameterized by $\boldsymbol{\theta_f}$ and $\boldsymbol{\theta_g}$, respectively. Note that $\boldsymbol{g}_{\rm NN}$ can be regarded as an estimate of the ground-truth $\boldsymbol{g}$.

To explain the above formulation, first note that the first fitting term is for approximating the equality constraint in (5b). The second term articulates the difference between our implementation and that of [20]. The latter does not have the second term in (7). Instead, a constraint

$$\boldsymbol{\theta}_f > \mathbf{0}$$

is employed. The reason is that under this positivity constraint, the function $\boldsymbol{f}_{\rm NN}$ is always invertible, which satisfies the problem specification in (5c). However, this may be problematic since positive NNs may not retain the universal approximation property for nonlinear functions—which is the reason why one uses NNs in the first place.

In our implementation, we use the second term in (7) to promote invertibility of the learned $\boldsymbol{f}_{\rm NN}$. It is straightforward to see the following:

Lemma 2 *Assume that there exists a function $\boldsymbol{g}_{\rm NN}$ such that for all $\boldsymbol{x} \in \mathcal{X}$, the following holds:*

$$\boldsymbol{x} = \boldsymbol{g}_{\rm NN}(\boldsymbol{f}_{\rm NN}(\boldsymbol{x})),$$

then $\boldsymbol{f}_{\rm NN}$ is invertible over $\mathcal{X}$.

Hence, when N is large, the regularization approximately enforces invertibility over $\mathcal{X}$.

Another benefit of employing our formulation other than the positivity constraint formulation as in [20] is that unconstrained optimization for NNs is much easier. We defer detailed discussion on this aspect in a pertinent journal version.

IV. Numerical Results

We use two baselines in our simulations, i.e., nonlinear matrix factor recovery (NMFR) [20] that was developed under the same model and MVES [30] that does not consider nonlinear distortion.

Our formulation is tackled by `PyTorch`-based `Adam` algorithm [23] with the initial step size of $1e^{-3}$. `Adam` is a stochastic gradient algorithm that works under mini-batch settings. The batch size is $5,000$ in our simulations. The algorithm stops after running $5,000$ epochs. The parameter λ is set to be $1e^{-5}$. For $\boldsymbol{f}_{\rm NN}$ and $\boldsymbol{g}_{\rm NN}$, each channel is modeled with a single hidden layer neural network with 256 neurons.

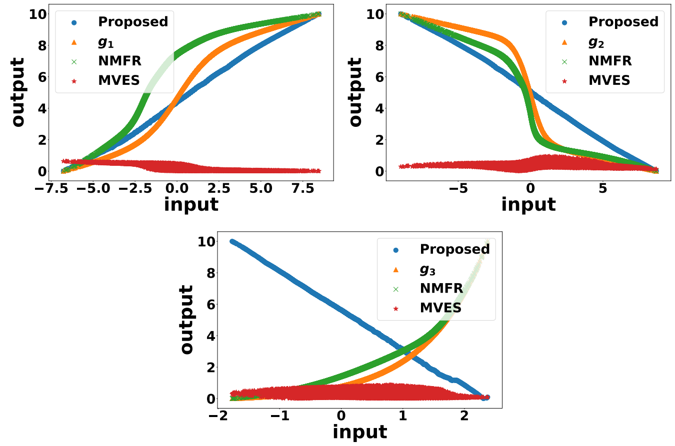

Fig. 1. Nonlinearity removal effects compared with baselines.

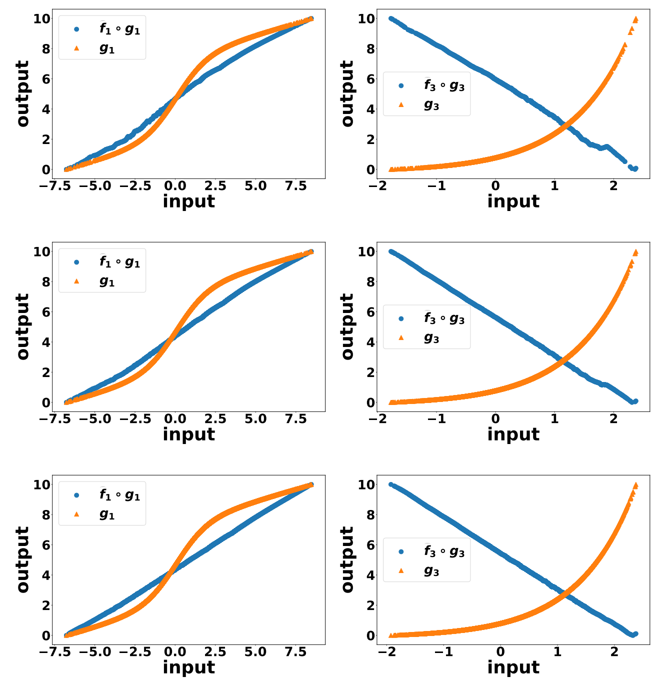

Fig. 2. Impact of N. From top to bottom, N is 5000, 10000 and 20000, respectively.

For the first simulation, we set $M = K = 3$. The three nonlinear functions are: $g_1(x) = 5\text{sigmoid}(x) + 0.3x$, $g_2(x) = -3\tanh(x) - 0.2x$ and $g_3 = 0.4\exp(x)$. The number of samples is $10,000$. The matrix $\boldsymbol{A}$ is drawn from standard Gaussian distribution. The learned $f_{\rm NN}^{(i)} \circ g_i$'s are shown in Fig. 1. One can see that the proposed method works remarkably better than the NMFR from [20]. Although the two methods start with the same conceptual formulation in (5), the performance difference may come from implementation strategies: since Yang *et al.*'s implementation uses positive

TABLE I
MSE BETWEEN $\boldsymbol{S}$ AND ESTIMATED $\widehat{\boldsymbol{S}}$ AFTER OPTIMAL MATCH.

N	Proposed	NMFR	MVES
5000	$3.67e^{-3} \pm 1.53e^{-3}$	$1.30e^{-1} \pm 4.48e^{-2}$	$5.49e^{-2} \pm 4.50e^{-3}$
10000	$1.11e^{-3} \pm 5.54e^{-4}$	$9.68e^{-2} \pm 2.84e^{-2}$	$5.84e^{-2} \pm 8.60e^{-3}$
20000	$1.58e^{-4} \pm 1.65e^{-4}$	$8.01e^{-2} \pm 1.78e^{-2}$	$4.96e^{-2} \pm 3.11e^{-3}$

NNs, it may not be able to approximate the true solution $\boldsymbol{f}$; in addition, constrained optimization may be much harder than dealing with our unconstrained formulation.

For the next simulation, we qualitatively show the influence of the sample size N. With the same setting as in Fig. 1, we randomly select two channels of the observations and show the learned composite functions in Fig. 2. The figure clearly illustrates that as more samples are available, the nonlinearity removal performance improves substantially.

In the last simulation, we combine nonlinearity removal and $\boldsymbol{s}_\ell$-identification, where the second phase is conducted by applying MVES to $\boldsymbol{f}_{\rm NN}(\boldsymbol{x}_\ell)$ for $\ell = 1, \ldots, N$. The performance measure here is the mean squared error (MSE) of the estimated $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N]$. The results are shown in Tab. I, which is averaged over 10 random trials. It can be seen from the table that the performance of the proposed approach shows a notable margin over the baselines. In particular, the MSE performance is one or two orders of magnitude lower compared to NMFR and MVES.

V. CONCLUSION

In this work, we address the nonlinearity removal and latent component identification problems under the post-nonlinear model with nonnegative and sum-to-one dependent components. Our contribution is two fold: First, we have tightened the sufficient conditions under which the nonlinearity is removable—which offers substantially more relaxed conditions relative to a recently derived result. Second, we offer a new NN-based formulation that has better function approximation ability and is easier to optimize. As a result, the numerical performance is largely improved compared to prior work.

REFERENCES

[1] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[2] P. Comon and C. Jutten, *Handbook of Blind Source Separation*. Elsevier, 2010.

[3] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[4] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications," *IEEE Signal Process. Mag.*, vol. 36, no. 2, pp. 59–80, March 2019.

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

[6] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, 2011.

[7] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3551–3582, 2017.

[8] K. Huang, X. Fu, and N. D. Sidiropoulos, "Anchor-free correlated topic modeling: Identifiability and algorithm," in *Proceedings of NIPS 2016*, 2016, pp. 1786–1794.

[9] K. Huang and X. Fu, "Detecting overlapping and correlated communities without pure nodes: Identifiability and algorithm," in *Proceedings of ICML 2019*, vol. 97, 09–15 Jun 2019, pp. 2859–2868.

[10] N. Gillis, "The why and how of nonnegative matrix factorization," *Regularization, Optimization, Kernels, and Support Vector Machines*, vol. 12, p. 257, 2014.

[11] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE journal*, vol. 37, no. 2, pp. 233–243, 1991.

[12] A. Hyvärinen and P. Pajunen, "Nonlinear independent component analysis: Existence and uniqueness results," *Neural Networks*, vol. 12, no. 3, pp. 429–439, 1999.

[13] A. Hyvarinen and H. Morioka, "Unsupervised feature extraction by time-contrastive learning and nonlinear ICA," in *Proceedings of NIPS 2016*, 2016, pp. 3765–3773.

[14] ——, "Nonlinear ICA of temporally dependent stationary sources," in *Proceedings of AISTATS 2017*, vol. 54, 20–22 Apr 2017, pp. 460–469.

[15] E. Oja, "The nonlinear PCA learning rule in independent component analysis," *Neurocomputing*, vol. 17, no. 1, pp. 25–45, 1997.

[16] A. Ziehe, M. Kawanabe, S. Harmeling, and K.-R. Müller, "Blind separation of post-nonlinear mixtures using linearizing transformations and temporal decorrelation," *Journal of Machine Learning Research*, vol. 4, no. Dec, pp. 1319–1338, 2003.

[17] A. Taleb and C. Jutten, "Source separation in post-nonlinear mixtures," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2807–2820, 1999.

[18] S. Achard and C. Jutten, "Identifiability of post-nonlinear mixtures," *IEEE Signal Process. Lett.*, vol. 12, no. 5, pp. 423–426, 2005.

[19] A. Hyvärinen, H. Sasaki, and R. E. Turner, "Nonlinear ica using auxiliary variables and generalized contrastive learning," in *The 22nd International Conference on Artificial Intelligence and Statistics*. Journal of Machine Learning Research, 2019, pp. 859–868.

[20] B. Yang, X. Fu, N. D. Sidiropoulos, and K. Huang, "Learning nonlinear mixtures: Identifiability and algorithm," *arXiv preprint arXiv:1901.01568*, 2019.

[21] G. Hamerly and C. Elkan, "Alternatives to the k-means algorithm that find better clusterings," in *Proceedings of the eleventh international conference on Information and knowledge management*, 2002, pp. 600–607.

[22] W.-K. Ma, J. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A. Plaza, A. Ambikapathi, and C.-Y. Chi, "A signal processing perspective on hyperspectral unmixing," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 67–81, Jan 2014.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287 – 314, 1994.

[25] X. Fu, K. Huang, and N. D. Sidiropoulos, "On identifiability of nonnegative matrix factorization," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 328–332, 2018.

[26] 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, 2014.

[27] N. Dobigeon, J.-Y. Tourneret, 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, 2014.

[28] 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.

[29] X. Fu, K. Huang, B. Yang, W. Ma, and N. D. 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.

[30] 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.