Nonlinear Multiview Analysis: Identifiability and Neural Network-based Implementation

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*—Multiview analysis aims to extract common information from data entities across different domains (e.g., acoustic, visual, text). Canonical correlation analysis (CCA) is one of the classic tools for this problem, which estimates the shared latent information via linear transforming the different views of data. CCA has also been generalized to the nonlinear regime, where kernel methods and neural networks are introduced to replace the linear transforms. While the theoretical aspects of linear CCA are relatively well understood, nonlinear multiview analysis is still largely intuition-driven. In this work, our interest lies in the identifiability of shared latent information under a nonlinear multiview analysis framework. We propose a model identification criterion for learning latent information from multiview data, under a reasonable data generating model. We show that minimizing this criterion leads to identification of the latent shared information up to certain indeterminacy. We also propose a neural network based implementation and an efficient algorithm to realize the criterion. Our analysis is backed by experiments on both synthetic and real data.**

***Index Terms*—Unsupervised learning, mixture separation, multiview analysis, neural networks, identifiability**

I. Introduction

Data is oftentimes acquired in different feature domains (e.g., audio, video, and image). The different domains for representing data give rise to diverse "views" of the same data—which may be able to characterize distinct aspects of the entities of interest. The purpose of multiview analysis is to extract the "essence" of the data that is in common across different views. Compared to traditional single-view analytical tools like principal component analysis (PCA), independent component analysis (ICA) [1] and nonnegative matrix factorization (NMF) [2], multiview approaches like canonical correlation analysis (CCA) [3] have some unique traits. For instance, it is shown that CCA is robust under covariance-unknown colored noise [4] and strong interference [5], whose presences are considered very challenging for single-view methods such as PCA.

Classic CCA finds common information across views through seeking for linear mappings such that the mapped data have maximum correlation in a latent space [3], [6]. Long after the advent of CCA, interesting interpretations for its effectiveness from a factor analysis viewpoint appeared in a number of papers [4], [5]. In particular, Bach *et al.* [4] explained CCA's robustness to colored noise from a maximum likelihood estimation perspective. In [5], the authors analyzed CCA by modeling each view as a mixture of shared and view-specific interference components and showed that the classic CCA extracts the shared components up to certain ambiguities, no matter how strong are the interference terms. Recently, nonlinear versions of CCA have been actively studied, since linear transformations are inadequate to capture the reality in many applications. Kernel CCA [7], [8] employs various kernel functions to transform the data. Furthermore, deep CCA [9], [10] has been proposed to incorporate neural networks to realize nonlinear transformations. These nonlinear CCA approaches have brought up boosted performance in applications like image embedding [10] and speech processing [11].

While the classic CCA has been extensively studied in both computational and theoretical aspects, the understanding to nonlinear CCA methods is quite limited. It still remains unclear *in theory* why the nonlinear methods have improved performance. To answer this intriguing question, we take a model-based perspective to analyze the nonlinear multiview approaches. Our contribution is twofold. First, we propose a generative model for nonlinear multiview analysis that is a natural extension of those for classic CCA in [4], [5]. In particular, we model the acquired data as nonlinearly distorted multiview linear mixtures. Based on this model, we propose a model identification criterion to extract the shared latent components across views, and show that our criterion leads to the removal of the unknown nonlinear distortions. Notably, our nonlinearity removal method does not rely on strong assumptions on the latent components in the model, e.g., statistical independence that is often utilized in nonlinear ICA [12], [13]. Second, we implement our proposed continuous function learning formulation leveraging neural networks, and propose a simple block coordinate descent (BCD) based algorithm to handle the neural network searching problem. Using this implementation, our theoretical claims are backed by a series of experiment results.

II. Background

A. Multiview Data and Analysis

Data entities often have different appearances in different feature domains; e.g., a car can be represented by its audio,

This work is supported in part by the National Science Foundation under NSF ECCS 1808159 and ECCS 1608961, and the Army Research Office under ARO W911NF-19-1-0247 and ARO W911NF-19-1-0407.

video, and text description, respectively. These different representations are called "views" of the data entities. Intuitively, integrating multiple views of data would have benefits for extracting essential information that can better represent the data samples—which eventually can improve performance of downstream tasks such as clustering and classification.

The classic method in statistical machine learning for this purpose is CCA, whose mathematical programming form can be expressed as follows [3], [14]:

$$\min_{\boldsymbol{B}^{(1)},\boldsymbol{B}^{(2)}} \sum_{\ell=1}^{N} \left\| \boldsymbol{B}^{(1)}\boldsymbol{y}_\ell^{(1)} - \boldsymbol{B}^{(2)}\boldsymbol{y}_\ell^{(2)} \right\|_2^2 \quad (1)$$
$$\text{s.t. } \frac{1}{N}\sum_{\ell=1}^{N} \boldsymbol{B}^{(q)}\boldsymbol{y}_\ell^{(q)}(\boldsymbol{y}_\ell^{(q)})^\top(\boldsymbol{B}^{(q)})^\top = \boldsymbol{I},\ q=1,2.$$

where $\boldsymbol{y}_\ell^{(q)} \in \mathbb{R}^{M_q}$ is the ℓth observation of the qth view, $\boldsymbol{B}^{(q)} \in \mathbb{R}^{R\times M_q}$ with $1 \le R \le K$ and the constraints avoid degenerate solutions. Simply speaking, CCA aims to find two linear mappings to transform the views—such that the transformed views are closely matched with each other in a Euclidean distance sense. The cost function of the above is equivalent to $\max_{\boldsymbol{B}^{(q)}} \ \mathrm{Tr}(\frac{1}{N}\sum_{\ell=1}^{N}\boldsymbol{B}^{(1)}\boldsymbol{y}_\ell^{(1)}(\boldsymbol{B}^{(2)}\boldsymbol{y}_\ell^{(2)})^\top)$, which, under the normalization constraints, is exactly maximizing the (sample average form of) cross-correlation between $\boldsymbol{B}^{(1)}\boldsymbol{y}_\ell^{(1)}$ and $\boldsymbol{B}^{(2)}\boldsymbol{y}_\ell^{(2)}$.

CCA oftentimes exhibits better performance relative to its single-view counterparts, e.g., PCA. There are also theoretical analyses that support this observations [4], [5]. In particular, the recent work [5] assumes the following generative model for multiview analysis problems:

$$\boldsymbol{y}_\ell^{(q)} = \boldsymbol{A}^{(q)}\boldsymbol{s}_\ell^{(q)}, \quad \boldsymbol{s}_\ell^{(q)} = [\boldsymbol{s}_\ell^\top, (\boldsymbol{c}_\ell^{(q)})^\top]^\top, \quad (2)$$

where $q=1,2$, $\ell=1,\dots,N$, $\boldsymbol{A}^{(q)} \in \mathbb{R}^{M_q\times(K+R_q)}$, $\boldsymbol{s}_\ell \in \mathbb{R}^K$ denotes the shared components across views (i.e., the essential information), and $\boldsymbol{c}_\ell^{(q)} \in \mathbb{R}^{R_q}$ is the view-specific component (or, an interference term). The model is a reasonable extension of linear mixture model (i.e., $\boldsymbol{y}_\ell = \boldsymbol{A}\boldsymbol{s}_\ell$) to multiview scenarios. The model in (2) assumes that the differences across views are due to both $\boldsymbol{A}^{(q)}$ and $\boldsymbol{c}_\ell^{(q)}$.

It is obvious that with $R=K$ and $\boldsymbol{B}^{(q)} = [\boldsymbol{\Theta},\boldsymbol{0}](\boldsymbol{A}^{(q)})^\dagger$, where $\boldsymbol{\Theta} \in \mathbb{R}^{K\times K}$ is an arbitrary non-singular matrix, the objective value of Eq. (1) becomes zero, under the model in Eq. (2). The work in [5] shows that this is indeed the only solution under the CCA formulation. Hence, $\boldsymbol{B}^{(q)}\boldsymbol{y}_\ell^{(q)} = \boldsymbol{\Theta}\boldsymbol{s}_\ell$, $\ell=1,\dots,L$ always holds; i.e., the subspace $\mathrm{range}(\boldsymbol{S}^\top)$, where $\boldsymbol{S}=[\boldsymbol{s}_1,\dots,\boldsymbol{s}_L]$, can be identified via CCA. Note that the view-specific $\boldsymbol{c}_\ell^{(q)}$ component is always removed no matter how strong its energy is. This is quite different compared to PCA, which always first returns the latent components with strongest energy (variation). This simple model and its elegant proof support the effectiveness of CCA for extracting essential cross-view information in the presence of strong interference.

III. Nonlinear Multiview Analysis

The model $\boldsymbol{y}_\ell^{(q)} = \boldsymbol{A}^{(q)}s_\ell^{(q)}$ is a typical linear mixture model (LMM), since the elements of $\boldsymbol{y}_\ell^{(q)}$ are linear combinations of the latent components $[\boldsymbol{s}_\ell^{(q)}]_k$ for $k=1,\dots,K$. The LMM is simple and useful, but might not be able to capture the essence of realistic data generating processes.

To handle more challenging scenarios where nonlinearity is involved in data generation, we propose a nonlinear extension of the linear CCA model in [5]. Our model is as follows:

$$\boldsymbol{y}_\ell^{(q)} = \boldsymbol{g}^{(q)}(\boldsymbol{A}^{(q)}\boldsymbol{s}_\ell^{(q)}),\ \boldsymbol{s}_\ell^{(q)} \quad = [\boldsymbol{s}_\ell^\top, (\boldsymbol{c}_\ell^{(q)})^\top]^\top, \quad (3)$$

where the nonlinear function $\boldsymbol{g}^{(q)}(\cdot) = [g_1^{(q)}(\cdot),\dots,g_{M_q}^{(q)}(\cdot)]^\top$ and $g_m^{(q)}(\cdot): \mathbb{R}\to\mathbb{R}$ represents the nonlinear distortion at channel m of view q. Note that the unknown nonlinear functions can be different for each view and each dimension. We also assume that the latent components are random variables defined on continuous open sets, i.e., $\boldsymbol{s}_\ell \in \mathcal{S}$, $\boldsymbol{c}_\ell^{(q)} \in \mathcal{C}_q$, where $\mathcal{S}$ and $\mathcal{C}_q$ are both open sets.

We should remark that the nonlinear model in (3) for an individual view is called the post-nonlinear (PNL) model in the context of nonlinear independent component analysis (nICA) [12], [13], which is a nonlinear version of ICA-based single view analysis. The PNL model is not the most general nonlinear model [15], but is effective in modeling nonlinear distortions happening in the receiver/sensor end—which fits applications like bio-signal processing and brain-computer interface (BCI) [16]–[18]. Classic methods in nICA hinge on statistical independence among the latent components (i.e., $[\boldsymbol{s}_\ell^{(q)}]_k$) to remove the nonlinearity. In this work, we will show that nonlinearity can be removed even if the components are dependent, if two views of the data entities are available.

A. A Function Learning Based Formulation

Like in the linear CCA case, our goal is to identify the $\boldsymbol{\Theta}\boldsymbol{s}_\ell$ for $\ell=1,\dots,L$ with a nonsingular $\boldsymbol{\Theta} \in \mathbb{R}^{K\times K}$ (representing a rotation ambiguity that can be removed by any blind source separation techniques as post-processing) under the model in (3). Note that we do not assume statistical independence among the latent components, which means that existing nonlinear ICA techniques are not applicable.

We seek an element-wise nonlinear mapping $\boldsymbol{f}^{(q)}(\cdot): \mathbb{R}^{M_q}\to\mathbb{R}^{M_q}$ where $\boldsymbol{f}^{(q)}(\cdot) = [f_1^{(q)}(\cdot),\dots,f_{M_q}(\cdot)]^\top$ and $f_m(\cdot):\mathbb{R}\to\mathbb{R}$, and a linear transformation $\boldsymbol{B}^{(q)} \in \mathbb{R}^{K\times M_q}$ so that the following criterion is minimized:

$$\min_{\boldsymbol{B}^{(q)},\boldsymbol{f}^{(q)}} \sum_{\ell=1}^{N} \left\| \boldsymbol{B}^{(1)}\boldsymbol{f}^{(1)}\left(\boldsymbol{y}_\ell^{(1)}\right) - \boldsymbol{B}^{(2)}\boldsymbol{f}^{(2)}\left(\boldsymbol{y}_\ell^{(2)}\right) \right\|_2^2.$$
$$\text{s.t.} \quad \boldsymbol{f}^{(q)} \text{ is invertible},\ q=1,2, \quad (4)$$
$$\frac{1}{N}\sum_{\ell=1}^{N}\left(\boldsymbol{B}^{(q)}\boldsymbol{f}^{(q)}\left(\boldsymbol{y}_\ell^{(q)}\right)\boldsymbol{f}^{(q)}\left(\boldsymbol{y}_\ell^{(q)}\right)^\top(\boldsymbol{B}^{(q)})^\top\right) = \boldsymbol{I}.$$

Ideally, we wish to obtain $\boldsymbol{B}^{(q)} = [\boldsymbol{\Theta},\boldsymbol{0}](\boldsymbol{A}^{(q)})^\dagger \in \mathbb{R}^{K\times M_q}$, $\boldsymbol{f}^{(q)}(\cdot) = [(g_1^{(q)}(\cdot))^{-1},\cdots,(g_{M_q}^{(q)}(\cdot))^{-1}]^\top$. The solution above will extract the shared row subspace of $\boldsymbol{S}$—i.e.,

the essential shared information across views. Again, the key question lies in the uniqueness of this solution. In other words, does the formulation in (4) have identifiability for the shared subspace spanned by the rows of $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N]$?

B. Nonlinearity Removal

To see how we approach the identifiability problem, we rewrite problem (4) in its population form:

$$\begin{aligned}
&\text{find } \boldsymbol{B}^{(1)}, \boldsymbol{B}^{(2)}, \boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} && \text{(5a)}\\
&\text{s.t. } \boldsymbol{B}^{(1)}\boldsymbol{f}^{(1)}\left(\boldsymbol{y}_\ell^{(1)}\right) = \boldsymbol{B}^{(2)}\boldsymbol{f}^{(2)}\left(\boldsymbol{y}_\ell^{(2)}\right), && \text{(5b)}\\
&\qquad \forall \boldsymbol{y}_\ell^{(q)} = \boldsymbol{g}^{(q)}(\boldsymbol{A}^{(q)}\boldsymbol{s}_\ell^{(q)}),\ \boldsymbol{s}_\ell \in \mathcal{S},\ \boldsymbol{c}_\ell^{(q)} \in \mathcal{C}_q,\\
&\quad \boldsymbol{f}^{(q)} \text{ is invertible, } q = 1, 2, && \text{(5c)}\\
&\quad \mathbb{E}\left[\boldsymbol{B}^{(q)}\boldsymbol{f}^{(q)}\left(\boldsymbol{y}_\ell^{(q)}\right)\boldsymbol{f}^{(q)}\left(\boldsymbol{y}_\ell^{(q)}\right)^\top(\boldsymbol{B}^{(q)})^\top\right] = \boldsymbol{I}.
\end{aligned}$$

Note that the above is derived from (4) assuming that one has uncountably infinite $\boldsymbol{y}_\ell^{(q)}$'s such that all possible values of $\boldsymbol{y}_\ell^{(q)}$ are exhausted under the assumed generative model. We use an equality constraint in (5b), since when there is no noise, the optimal value in (4) should be zero under the model in Eq. (3).

To proceed, we will be using the following condition:

Definition 1 (Ubiquitously Unanchored) *Consider a collection of d real-valued random components $\boldsymbol{v} = [v_1, \ldots, v_d]^\top \in \mathbb{R}^d$, where v_i resides in a continuous and open set $\mathcal{V}_i \subseteq \mathbb{R}$ such that $\text{volume}(\mathcal{V}_i) > 0$. Denote $\bar{v}_j$ as any fixed value from $\mathcal{V}_j$. Assume that for any $i \in \{1, \ldots, d\}$ and any $\bar{v}_j$ where $j \neq i$, the vectors $[\bar{v}_1, \ldots, \bar{v}_{i-1}, v, \bar{v}_{i+1}, \ldots, \bar{v}_d]^\top,\ \forall v \in \mathcal{V}_i$ are contained in the domain of $\boldsymbol{v}$. Then, the components in $\boldsymbol{v}$ are called ubiquitously unanchored.*

With the above definition, we show the following:

Theorem 1 (Nonlinearity Removal) *Consider the nonlinear model in Eq. (3). Assume that $M_q \geq K + R_q,\ q = 1, 2$, and that the mixing matrices $\boldsymbol{A}^{(q)}$ for $q = 1, 2$ are drawn from any absolutely continuous distributions. Assume that the components in $[s_{k,\ell}, (\boldsymbol{c}_\ell^{(1)})^\top, (\boldsymbol{c}_\ell^{(2)})^\top]^\top$ are ubiquitously unanchored for any k. Further assume that the dimensions of the components satisfy $\frac{R_q(R_q+1)}{2} \geq M_q$. Suppose that $(\boldsymbol{B}^{(q)}, \boldsymbol{f}^{(q)})$ for $q = 1, 2$ are solutions of Eq. (5) with $\|\boldsymbol{B}^{(q)}\|_0 = KM_q$. Then, the composition $f_i^{(q)} \circ g_i^{(q)}(x)$ for all i, q are affine functions with probability one.*

There are a couple of notable points: First, removing nonlinearity does not rely on strong assumptions such as statistical independence between the latent components. In fact, the ubiquitously unanchored condition is very mild—variables that are strongly dependent can satisfy this condition; see [19]. In addition, the theorem has no restriction on the relationship between the shared components. Hence, $s_{k,\ell}$ and $s_{j,\ell}$ can be completely dependent without affecting the removal of nonlinear distortions, which would have been impossible if one resorts to nonlinear ICA, e.g., those in [12], [13]—this shows the power of multiview analysis. Additionally, even if the energy of $\boldsymbol{s}_\ell$ is significantly smaller compared to that of $\boldsymbol{c}_\ell^{(q)}$, the proposed criterion can still recover the shared subspace. This property is inherited from linear CCA [5].

Note that after the nonlinearity is removed, the remaining problem boils down to a linear CCA problem under the model in (2). Hence, the shared components $\boldsymbol{\Theta}\boldsymbol{s}_\ell$ for $\ell = 1, 2, \ldots$ can be then identified up to the rotation ambiguity $\boldsymbol{\Theta}$.

IV. Implementation and Algorithm

In this section, we propose a practical implementation and propose an algorithm to tackle it. We parameterize $f_m^{(q)}(\cdot)$ for $q = 1, 2$ and all m using neural networks (NNs), since the NNs are known as "universal function approximators". With the NN-based parametrization, we consider the following optimization problem:

$$\begin{aligned}
\min_{\boldsymbol{U}, \boldsymbol{\theta}_f^{(q)}, \boldsymbol{\theta}_g^{(q)}, \boldsymbol{B}^{(q)}} & \sum_{q=1}^{2}\sum_{\ell=1}^{N}\left\|\boldsymbol{u}_\ell - \boldsymbol{B}^{(q)}\boldsymbol{f}_{\mathsf{NN}}^{(q)}\left(\boldsymbol{y}_\ell^{(q)}\right)\right\|_2^2 \qquad (6)\\
&+ \lambda\sum_{q=1}^{2}\sum_{\ell=1}^{N}\left\|\boldsymbol{y}_\ell^{(q)} - \boldsymbol{g}_{\mathsf{NN}}^{(q)}\left(\boldsymbol{f}_{\mathsf{NN}}^{(q)}\left(\boldsymbol{y}_\ell^{(q)}\right)\right)\right\|_2^2\\
\text{s.t. } & \frac{1}{N}\left[\sum_{\ell=1}^{N}\boldsymbol{u}_\ell\boldsymbol{u}_\ell^\top\right] = \boldsymbol{I}, \quad \frac{1}{N}\sum_{\ell=1}^{N}\boldsymbol{u}_\ell = \boldsymbol{0}
\end{aligned}$$

where a slack variable $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N] \in \mathbb{R}^{K\times N}$ that represents the extracted shared components is introduced, $\boldsymbol{f}_{\mathsf{NN}}^{(q)}(\cdot) = [f_{\mathsf{NN},1}^{(q)}(\cdot), \ldots, f_{\mathsf{NN},M_q}^{(q)}(\cdot)]^\top$ is a collection of neural network (NN)-parametrized element-wise non-linear mappings that we attempt to learn for nonlinearity removal, and $\boldsymbol{\theta}_f^{(q)}$ denotes all the NN parameters of $\boldsymbol{f}_{\mathsf{NN}}^{(q)}$. Similarly, $\boldsymbol{g}_{\mathsf{NN}}^{(q)}(\cdot) = [g_{\mathsf{NN},1}^{(q)}(\cdot), \ldots, g_{\mathsf{NN},M_q}^{(q)}(\cdot)]^\top$ is another set of neural networks for learning the generative function in (3) which approximately ensures the learned $\boldsymbol{f}^{(q)}$ is invertible (i.e., the reflect the invertibility constraints in (4)), at least for the available data samples $\boldsymbol{y}_\ell^{(q)}$ for $\ell = 1, \ldots, N$. Note that if $\boldsymbol{f}_{\mathsf{NN}}^{(q)}$ is invertible, then there exists a $\boldsymbol{g}_{\mathsf{NN}}^{(q)}$ such that the second term is zero—but the converse is not necessarily true. This reconstruction idea is known as the autoencoder in the context of deep learning [20]. We should remark that the constraint $\frac{1}{N}\sum_{\ell=1}^{N}\boldsymbol{u}_\ell = \boldsymbol{0}$ is vital for avoiding numerical problems. The reason is that nonlinearity removal is up to affine transformations (cf. Theorems 1). Hence, if not handled carefully, the constants in the affine transformations (i.e., the d_i's) may dominate, but these constants are not physically meaningful. Hence, adding a zero-mean constraint can effectively remove these constants and retain the variations of the latent components of interest.

To handle (6), we propose a block coordinate descent (BCD) based algorithm. The variables $\boldsymbol{\theta}_f^{(q)}, \boldsymbol{\theta}_g^{(q)}$ and $\boldsymbol{B}^{(q)}$ variables are treated as one block, and $\boldsymbol{U}$ as another. It is readily seen that the subproblem w.r.t. the first block is unconstrained thus can be handled with stochastic gradient descent—which is easy to implement leveraging back-propagation. While the $\boldsymbol{U}$-subproblem is a nonconvex, we show that this subproblem can

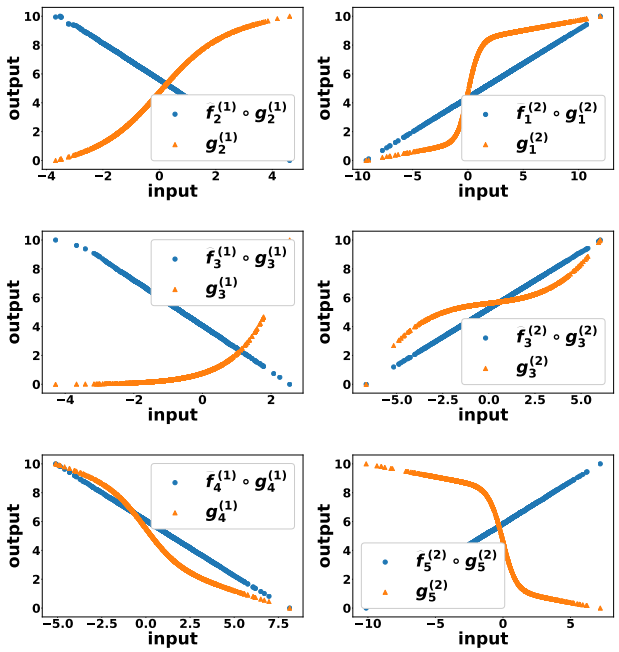

Fig. 1. Nonlinear distortions and the learned composite functions. Left to right: view 1-2. Top to bottom: 3 randomly selected dimensions.

TABLE I
THE SUBSPACE DISTANCE OF DIFFERENT APPROACHES.

	Proposed,	w/o zero mean,	DCCA	DCCAE	KCCA	CCA	PCA
dist	0.035	0.998	0.523	0.530	0.998	0.993	0.999

be solved efficiently through centering and truncated SVD. Details can be found in the pre-print in [19].

V. Numerical Experiments

The baselines include PCA, CCA [3], kernel CCA (KCCA) [21], Deep CCA (DCCA) [9] and DCCAE [10] (which is a performance-enhanced version of DCCA). To evaluate the results, we measure the distance between ground-truth subspace and the learned $\text{range}(\widehat{\boldsymbol{S}}^\top)$, which is between 0 and 1 with 0 being the best [22].

A. Synthetic Data Simulations

For the first experiment, we construct the shared component $\boldsymbol{s}_\ell \in \mathbb{R}^2$ which is uniformly sampled from a parabola ($[\boldsymbol{s}_\ell]_1 = [\boldsymbol{s}_\ell]_2^2, [\boldsymbol{s}_\ell]_2 \in [-1, 1]$). This way, the components in $\boldsymbol{s}_\ell$ are completely dependent. View-specific components $\boldsymbol{c}_\ell^{(q)} \in \mathbb{R}^3$ for $q = 1, 2$ are set to be i.i.d. Gaussian components with different means and variances. Note that the condition of Theorem 1 is satisfied in this case. The elements of $\boldsymbol{A}^{(q)} \in \mathbb{R}^{5\times 5}$ follow zero-mean unit-variance i.i.d. Gaussian distribution. The sample size for each view is $N = 1,000$. The nonlinear functions are set as follows: $g_1^{(1)}(x) = 3\text{sigmoid}(x) + 0.1x$, $g_2^{(1)}(x) = 5\text{sigmoid}(x) + 0.2x$, $g_3^{(1)}(x) = 0.2\exp(x)$, $g_4^{(1)}(x) = -4\text{sigmoid}(x) - 0.3x$, $g_5^{(1)}(x) = -3\text{sigmoid}(x) - 0.2x$; $g_1^{(2)}(x) = 5\tanh(x) + 0.2x$, $g_2^{(2)}(x) = 2\tanh(x) + 0.1x$, $g_3^{(2)}(x) = 0.1x^3 + x$, $g_4^{(2)}(x) = -5\tanh(x) - 0.4x$, $g_5^{(2)}(x) = -6\tanh(x) - 0.3x$, some of which are plotted in orange in Fig. 1.

Fig. 1 shows the learned $\widehat{f}_m^{(q)} \circ g_m^{(q)}$ for $q = 1, 2$ for 3 randomly selected dimensions. One can see that all the function compositions are visually affine. Table I shows the averaged subspace distance of 10 random trials under the same settings. One can see that the proposed approach admits a subspace distance that is almost zero. It is much lower than those of the baselines, perhaps because our method exploits the model information. In addition, as expected, without the zero-mean constraint, the result is much worse—which echos our remark on the importance of having the zero-mean constraint.

To observe the impact of the view-specific interference, we define the *Shared Component to Interference Ratio* (SCIR) $\text{SCIR} = 10\log_{10}\left(\frac{\|\boldsymbol{S}\|_F^2/K}{\frac{1}{Q}\sum_{q=1}^Q \|\boldsymbol{C}^{(q)}\|_F^2/R_q}\right)$ dB. Table II shows the result under different SCIRs. One can see that even if the ratio is -20 dB, the performance of the proposed approach is still very good—the average subspace distance metric is 0.064 which is much better than other baselines. This observation is consistent with our analysis—the proposed method is robust to strong view-specific interference.

TABLE II
THE SUBSPACE DISTANCE UNDER DIFFERENT SCIR.

SCIR	Proposed	DCCA	DCCAE	KCCA	CCA	PCA
-10 dB	0.035	0.523	0.530	0.998	0.993	0.999
-20 dB	0.064	0.999	0.998	0.999	0.998	0.997

TABLE III
SPECTRAL CLUSTERING ACC (%) OF DIFFERENT ALGORITHMS.

	raw	PCA	CCA	KCCA	DCCA	DCCAE	Proposed
view1	64.15	65.75	68.60	63.65	60.55	62.15	70.35
view2	50.05	51.05	64.15	59.45	62.15	59.20	66.15

B. Real Dataset Results

To show the usefulness of the model, we test it on a multiview handwritten digits clustering [23]. There are 200 samples per class (digit) and they are represented in different views, among which 64 Karhunen-Love coefficients and 47 Zernike moments are used as two views in this experiments. The training set includes 1,200 samples, the validation and testing sets both have 400 samples. Spectral clustering [24] is performed on the learned representations on the testing set. The clustering accuracy on the testing set is shown in Table III. The proposed approach shows promising results. This also suggests that combining linear and nonlinear mixture models may improve performance in data representation learning.

VI. Conclusion

To conclude, we analyzed the nonlinear multiview approach from an identifiability-driven and model-based perspective. A criterion was proposed to recover the shared components with identifiability guarantees under reasonable conditions. Our claims were backed by a series of numerical results.

REFERENCES

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

[2] 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.

[3] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[4] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," 2005.

[5] M. S. Ibrahim and N. D. Sidiropoulos, "Cell-edge interferometry: Reliable detection of unknown cell-edge users via canonical correlation analysis," in *Proc. IEEE SPAWC 2019*, 2019, pp. 1097–1105.

[6] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[7] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, no. Jul, pp. 1–48, 2002.

[8] K. Fukumizu, F. R. Bach, and A. Gretton, "Statistical consistency of kernel canonical correlation analysis," *Journal of Machine Learning Research*, vol. 8, no. Feb, pp. 361–383, 2007.

[9] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proceedings of ICML 2013*, vol. 28, no. 3, 17–19 Jun 2013, pp. 1247–1255.

[10] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proceedings of ICML 2015*, 2015, pp. 1083–1092.

[11] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "Unsupervised learning of acoustic features via deep canonical correlation analysis," in *Proc. IEEE ICASSP 2015*, 2015, pp. 4590–4594.

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

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

[14] X. Fu, K. Huang, M. Hong, N. D. Sidiropoulos, and A. M.-C. So, "Scalable and flexible multiview MAX-VAR canonical correlation analysis," *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4150–4165, 2017.

[15] C. Jutten and J. Karhunen, "Advances in blind source separation (bss) and independent component analysis (ica) for nonlinear mixtures," *International journal of neural systems*, vol. 14, no. 05, pp. 267–292, 2004.

[16] I. Rustandi, M. Just, and T. Mitchell, "Integrating multiple-study multiple-subject fMRI datasets using canonical correlation analysis," in *Proceedings of the MICCAI 2009 Workshop: Statistical modeling and detection issues in intra-and inter-subject functional MRI data analysis*, 2009.

[17] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

[18] X. Fu, K. Huang, O. Stretcu, H. A. Song, E. Papalexakis, P. Talukdar, T. Mitchell, N. Sidiropoulo, C. Faloutsos, and B. Poczos, "Brainzoom: High resolution reconstruction from multi-modal brain signals," in *Proc. SIAM ICDM 2017*. SIAM, 2017, pp. 216–227.

[19] Q. Lyu and X. Fu, "Nonlinear multiview analysis: Identifiability and neural network-assisted implementation," *IEEE Tran. Signal Process.,* to appear, 2020.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[21] D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, and B. Schölkopf, "Randomized nonlinear component analysis," in *Proceedings of ICML 2014*, 2014, pp. 1359–1367.

[22] G. H. Golub and C. F. V. Loan, *Matrix Computations*. The Johns Hopkins University Press, 1996.

[23] M. van Breukelen, R. P. Duin, D. M. Tax, and J. Den Hartog, "Handwritten digit recognition by combined classifiers," *Kybernetika*, vol. 34, no. 4, pp. 381–386, 1998.

[24] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proceedings of NIPS 2002*, 2002, pp. 849–856.