This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3333212

Revisiting Deep Generalized Canonical Correlation Analysis

Paris A. Karakasis, *Graduate Student Member, IEEE* and Nicholas D. Sidiropoulos, *Fellow, IEEE*

***Abstract*—Canonical correlation analysis (CCA) is a classic statistical method for discovering latent co-variation that underpins two or more observed random vectors. Several extensions and variations of CCA have been proposed that have strengthened our capabilities in terms of revealing common random factors from multiview datasets. In this work, we first revisit the most recent deterministic extensions of deep CCA and highlight the strengths and limitations of these state-of-the-art methods. Some methods allow trivial solutions, while others can miss weak common factors. Others overload the problem by also seeking to reveal what is *not common* among the views – i.e., the private components that are needed to fully reconstruct each view. The latter tends to overload the problem and its computational and sample complexities. Aiming to improve upon these limitations, we design a novel and efficient formulation that alleviates some of the current restrictions. The main idea is to model the private components as *conditionally* independent given the common ones, which enables the proposed compact formulation. Judicious experiments with synthetic and real datasets showcase the validity of our claims and the effectiveness of the proposed approach.**

***Index Terms*—Generalized Canonical Correlation Analysis, Deep Learning, Conditional Independence**

I. Introduction

SEEKING out and reasoning about similarity enables us to abstract common patterns, learn new patterns, and ultimately reason about our world. In many applications, we nowadays have rich multimodal information about an entity, topic, or concept of interest, e.g., audio and video, images and text, or different medical sensing and imaging modalities, such as EEG, MEG, and fMRI, which can be combined to provide more complete information in support of critical decision making. The different sources of information can be considered as different "views" of an underlying phenomenon that we are interested in analyzing to accomplish several downstream tasks, such as predicting missing pieces of information or improving the quality of the observed signals. Multiview/Multimodal learning lies under the umbrella of unsupervised learning and studies how the information of multiple jointly observed views can be fused for the aforementioned problems. Under certain assumptions, the advantages of multiview techniques for several downstream tasks have also been theoretically corroborated [1], [2], [3].

Canonical Correlation Analysis (CCA) is a well-known statistical tool [4] that can be used to extract shared information from multiple views. CCA can be considered as a generalization of Principal Component Analysis (PCA). In a scenario with two views, CCA treats the views as different random vectors and it poses the problem of finding individual linear combinations of the views that generate a common latent random vector. CCA and Generalized CCA (GCCA – the extension to more than two views) have found many successful applications in speech processing [5], [6], [7], communications [8], [9], biomedical signal processing [10], [11], [12], [13], and many other areas.

Various attempts to extend CCA to deal with nonlinear transformations have been made over the years. Kernel CCA in [14] was one of the first attempts, while many more have followed after adopting a probabilistic point of view, as in Deep Variational CCA [15] and Nonparametric CCA [16], or a deterministic one as in Deep CCA (DCCA) [17] and other works that built upon it, as in [18], [19], [20], [21]. Nonlinear CCA is an active research area, but one clear lesson that has emerged is that DCCA and its variants can significantly outperform the classical linear CCA in many downstream tasks.

In this work, we first highlight what are the advantages and limitations of the previous nonlinear GCCA approaches under a deterministic setting, and then propose a novel formulation of DCCA that bypasses most of the current limitations. We provide careful motivation and experiments on synthetic and real datasets that compare the proposed approach to the prior art and showcase its promising performance.

Reproducible Research: The codes for reproducing all the presented experimental results have been submitted, for reviewing purposes, as Supplementary material in a .rar file. As for the considered datasets, they can be automatically downloaded/generated using the provided codes.

II. Background and Related Prior Work

Consider a collection of K random vectors $\mathbf{x}^{(k)} \in \mathbb{R}^{D_k} \sim \mathcal{D}_{\mathbf{x}^{(k)}}$, for $k \in [K] := \{1, \ldots, K\}$. When these K random vectors provide different indirect "views" of the same latent random vector $\mathbf{g} \in \mathbb{R}^F \sim \mathcal{D}_{\mathbf{g}}$, estimating $\mathbf{g}$ from realizations of the views is often the main problem of interest. This problem lies at the heart of (G)CCA and many prior works have considered nonlinear extensions of (G)CCA for more or less general problem instances [20], [22], [23], [24], [12]. In this section, we consider the state-of-art formulations of this problem, which can be viewed as nonlinear extensions of the most popular formulation of GCCA: the Maximum Variance (MAX-VAR) formulation.

Paris A. Karakasis and Nicholas D. Sidiropoulos are with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904 USA (e-mail: {karakasis, nikos}@virginia.edu).

P. A. Karakasis and N. D. Sidiropoulos were partially supported by NSF IIS-1908070 and ECCS-2118002.

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3333212

A. (Generalized) Canonical Correlation Analysis

Assume that we observe jointly drawn realizations of a pair of random vectors $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$, i.e., $K = 2$. CCA [4] poses the problem of estimating two maximally correlated random vectors of the form $\mathbf{z}^{(k)} = \mathbf{Q}_k^T \mathbf{x}^{(k)}$. To avoid trivial solutions, CCA additionally requires that the extracted random vectors are uncorrelated. Without loss of generality, we can assume that all random vectors $\mathbf{x}^{(k)}$ are zero-mean, since the empirical mean can be subtracted as a pre-processing step. Then, the two-view CCA problem can be mathematically expressed as

$$\begin{aligned} \max_{\{\mathbf{Q}_k \in \mathbb{R}^{D_k \times F}\}_{k=1}^2} \quad & \mathbb{E}\left[\text{Trace}\left(\mathbf{Q}_1^T \mathbf{x}^{(1)} \mathbf{x}^{(2)^T} \mathbf{Q}_2\right)\right] \\ \text{s.t.} \quad & \mathbb{E}\left[\mathbf{Q}_k^T \mathbf{x}^{(k)} \mathbf{x}^{(k)^T} \mathbf{Q}_k\right] = \mathbf{I}_F, \text{ for } k = 1, 2, \end{aligned} \tag{1}$$

where matrices $\mathbf{Q}_k \in \mathbb{R}^{D_k \times F}$ denote the reducing operators that will transform the observed random vectors $\mathbf{x}^{(k)}$ into random vectors, of uncorrelated components, that are maximally correlated. As a result, random vectors $\mathbf{z}^{(k)}$ can be considered as different estimates of a common latent random vector, which coincide when the maximum value of the objective in (1) is attained.

The relation between solving the CCA problem and estimating the latent common random vector becomes more clear after considering the following, equivalent when $K = 2$, formulation of CCA given by

$$\begin{aligned} \min_{\mathbf{g}, \{\mathbf{Q}_k \in \mathbb{R}^{D_k \times F}\}_{k=1}^K} \quad & \sum_{k=1}^K \mathbb{E}\left[\left\|\mathbf{Q}_k^T \mathbf{x}^{(k)} - \mathbf{g}\right\|^2\right] \\ \text{s.t.} \quad & \mathbb{E}\left[\mathbf{g}\mathbf{g}^T\right] = \mathbf{I}_F, \end{aligned} \tag{2}$$

where $\mathbf{g}$ denotes the common latent random vector. As one can see, this formulation of CCA naturally generalizes to multiple views and is known as the MAX-VAR formulation [25]. After considering the subspace that is spanned from the realizations of $\mathbf{g}$, CCA can be also considered as a method for the estimation of a linear subspace which is "common" to the K sets of random variables [8].

The generalization of CCA to multiple views is known as Generalized CCA (GCCA), and several distinct formulations have been proposed over the years. More specifically, five different formulations are presented in [26] with all of them boiling down to the classical CCA formulation in (1) when $K = 2$ [27]. Nowadays, MAX-VAR has prevailed as one of the most popular formulations of GCCA, in good part because it admits a closed form solution via eigendecomposition. Moreover, GCCA identifiability has been recently considered from the common subspace point of view in [28]. In this manuscript we focus on the MAX-VAR formulation of GCCA.

B. Deep Canonical Correlation Analysis (DCCA)

The CCA formulation is restricted to consider only linear transformations of the observed random vectors, which cannot compensate for potentially more appropriate nonlinear transformations. Andrew et al. [17] proposed using deep neural networks (DNNs) in order to approximate nonlinear transformations and hence overcome this limitation. The formulation proposed in [17] is equivalent to

$$\begin{aligned} \max_{\{\boldsymbol{f}_k \in \mathcal{C}\}_{k=1}^2} \quad & \mathcal{R}_2\left(\boldsymbol{f}_1, \boldsymbol{f}_2\right) := \mathbb{E}\left[\text{Trace}\left(\boldsymbol{f}_1\left(\mathbf{x}^{(1)}\right) \boldsymbol{f}_2\left(\mathbf{x}^{(2)}\right)^T\right)\right] \\ \text{s.t.} \quad & \mathbb{E}\left[\boldsymbol{f}_k\left(\mathbf{x}^{(k)}\right) \boldsymbol{f}_k\left(\mathbf{x}^{(k)}\right)^T\right] = \mathbf{I}_F, \\ & \mathbb{E}\left[\boldsymbol{f}_k\left(\mathbf{x}^{(k)}\right)\right] = \mathbf{0}_F, \text{ for } k = 1, 2, \end{aligned} \tag{3}$$

where $\boldsymbol{f}_k : \mathbb{R}^{D_k} \to \mathbb{R}^F$ and $\mathcal{C}$ is the class of all functions that can be generated through a given DNN parametrization. Note that a zero-mean constraint has been added, because $\boldsymbol{f}_k\left(\mathbf{x}^{(k)}\right)$ is not automatically zero-mean even if $\mathbf{x}^{(k)}$ is. Without this constraint, the orthogonality constraints by themselves are not enough to guarantee uncorrelateness.

C. Deep Canonically Correlated Autoencoders (DCCAE)

The above formulation in (3) has the disadvantage that for rich enough classes $\mathcal{C}$ of functions $\boldsymbol{f}_k$, trivial solutions could satisfy the considered criterion, as we discuss in the next subsection. Later, the formulation of [18] remedied this issue, although its original motivation was different. For functions $\boldsymbol{w}_k : \mathbb{R}^F \to \mathbb{R}^{D_k}$, let

$$\mathcal{L}^{(K)}(\boldsymbol{f}_1, \boldsymbol{w}_1, ..., \boldsymbol{f}_K, \boldsymbol{w}_K) = \sum_{k=1}^K \mathbb{E}\left[\left\|\mathbf{x}^{(k)} - \boldsymbol{w}_k\left(\boldsymbol{f}_k\left(\mathbf{x}^{(k)}\right)\right)\right\|^2\right].$$

Instead of (3), Wang et al. [18] proposed using

$$\begin{aligned} \max_{\{\boldsymbol{f}_k, \boldsymbol{w}_k \in \mathcal{C}\}_{k=1}^2} \quad & (1-\lambda)\,\mathcal{R}_2\left(\boldsymbol{f}_1, \boldsymbol{f}_2\right) - \lambda \mathcal{L}^{(2)}\left(\boldsymbol{f}_1, \boldsymbol{w}_1, \boldsymbol{f}_2, \boldsymbol{w}_2\right) \\ \text{s.t.} \quad & \text{the constraints in (3).} \end{aligned} \tag{4}$$

For certain distributions, it can be shown that CCA, and hence the first term of the objective in (4), maximizes the mutual information between the projected views [29]. On the other hand, minimizing over the reconstruction error of the each view can be seen as maximizing a lower bound on the mutual information between the corresponding view and its learned latent representation [30]. As a result, the DCCAE objective offers a trade-off between maximizing the mutual information between each view and its encoding, on the one hand, and maximizing the mutual information across the encodings of the two views, on the other [18]. Since the second term in the objective tends to capture the strongest, but potentially non common, latent principal components of each view, non common information could leak in the learned embeddings. This possibility could deteriorate the quality of the learned embeddings, especially in the case of weak latent common components, as we will see in our experiments.

D. Deep Generalized Canonical Correlation Analysis

The extension of the MAX-VAR formulation to the nonlinear regime, using DNNs, was proposed in [19] under the name Deep Generalized Canonical Correlation Analysis (DGCCA).

$$\begin{aligned} \min_{\mathbf{g}, \{\boldsymbol{f}_k \in \mathcal{C}\}_{k=1}^K} \quad & \mathcal{R}^{(K)}\left(\boldsymbol{f}_1, ..., \boldsymbol{f}_K, \mathbf{g}\right) := \sum_{k=1}^K \mathbb{E}\left[\left\|\boldsymbol{f}_k\left(\mathbf{x}^{(k)}\right) - \mathbf{g}\right\|^2\right] \\ \text{s.t.} \quad & \mathbb{E}\left[\mathbf{g}\mathbf{g}^T\right] = \mathbf{I}_F, \quad \mathbb{E}\left[\mathbf{g}\right] = \mathbf{0}_F. \end{aligned} \tag{5}$$

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3333212

DGCCA can be seen as a direct multiview extension of DCCA and hence it inherits the disadvantages of DCCA. Since in practice we deal with finite number of samples, let us assume M, the constraints of (5) translate to $\mathbf{G}^T\mathbf{G} = M \cdot \mathbf{I}_F$ and $\mathbf{G}^T\mathbf{1}_M = \mathbf{0}_F$, where the rows of $\mathbf{G} \in \mathbb{R}^{M\times F}$ correspond to the realizations of $\mathbf{g}$. One can see that the constraints can be satisfied even if $M - F - 1$ samples are mapped to $\mathbf{0}_F$, while the remaining $F + 1$ to the rows of any matrix $\mathbf{U} \in \mathbb{R}^{F+1\times F}$ satisfying $\mathbf{U}^T\mathbf{U} = M\cdot\mathbf{I}_F$ and $\mathbf{U}^T\mathbf{1}_{F+1} = \mathbf{0}_F$. More generally, in the case of continuous distributions $\mathcal{D}_{\mathbf{x}^{(k)}}$, for any $\mathbf{G}$ satisfying the constraints above, we can find functions $\boldsymbol{f}_k$ that map the samples of all $\mathbf{x}^{(k)}$ to the corresponding rows of $\mathbf{G}$. As a result, trivial solutions may arise leading to non informative embeddings.

E. Correlation-based learning of common and private latent random vectors

In [21], Lyu et al. consider the problem of estimating disentangled representations of common and uncommon (private per view) latent random vectors, from a pair of views, by building upon DCCA. Specifically, they consider the following multiview generative model

$$\mathbf{x}^{(k)} = \boldsymbol{v}_k\left(\begin{bmatrix}\mathbf{g}\\ \mathbf{c}^{(k)}\end{bmatrix}\right),\ \text{for } k = 1, 2, \tag{6}$$

where $\mathbf{g} \in \mathbb{R}^F$ denotes the common (shared) latent random vector and $\mathbf{c}^{(k)} \in \mathbb{R}^{L_k}$ denote the non-common (not shared, or *private*) latent random vectors. Furthermore, they assume that all the random vectors together are jointly group independent, i.e.

$$p\left(\mathbf{g}, \mathbf{c}^{(1)}, \mathbf{c}^{(2)}\right) = p\left(\mathbf{g}\right) p\left(\mathbf{c}^{(1)}\right) p\left(\mathbf{c}^{(2)}\right). \tag{7}$$

Finally, functions $\boldsymbol{v}_k : \mathbb{R}^{F+L_k} \to \mathbb{R}^{D_k}$ are assumed to be smooth and invertible functions.

Based on the above generative model, Lyu et al. [21] proposed estimating common and uncommon latent factors jointly by solving the problem

$$\begin{aligned}\min_{\substack{\boldsymbol{f}_k\in\mathcal{C}_f,\\ \boldsymbol{v}_k\in\mathcal{C}_v,\\ \mathbf{g}}}\ \max_{\substack{\phi_k\in\mathcal{C}_\phi,\\ \tau_1\in\mathcal{C}_\tau}}\ & (1-\lambda)\,\mathcal{R}^{(2)}\left(\boldsymbol{f}_{1_S}, \boldsymbol{f}_{2_S}, \mathbf{g}\right) + \lambda\mathcal{L}^{(2)}\left(\boldsymbol{f}_1, \boldsymbol{v}_1, \boldsymbol{f}_2, \boldsymbol{v}_2\right)\\ & + \beta\mathcal{V}^{(2)}\left(\boldsymbol{f}_1, \phi_1, \tau_k, \boldsymbol{f}_2, \phi_2, \tau_2\right)\\ \text{s.t.}\quad & \mathbb{E}\left[\mathbf{g}\mathbf{g}^T\right] = \mathbf{I}_F,\ \ \mathbb{E}\left[\mathbf{g}\right] = \mathbf{0}_F,\end{aligned} \tag{8}$$

where functions $\boldsymbol{f}_k\left(\mathbf{x}^{(k)}\right)$ admit the following block decomposition $\boldsymbol{f}_k\left(\mathbf{x}^{(k)}\right) = \left[\boldsymbol{f}_{k_S}\left(\mathbf{x}^{(k)}\right)^T \boldsymbol{f}_{k_P}\left(\mathbf{x}^{(k)}\right)^T\right]^T$, with subscripts S and P denoting the shared and private components, respectively. The second term of the objective enforces the invertibility of functions $\boldsymbol{f}_k$. At last, for functions $\phi_k : \mathbb{R}^F \to \mathbb{R}$ and $\tau_k : \mathbb{R}^{L_k} \to \mathbb{R}$, each of the terms in the sum below

$$\mathcal{V}^{(K)}\left(\boldsymbol{f}_1, \phi_1, \tau_1, ..., \boldsymbol{f}_K, \phi_K, \tau_K\right) := \sum_{k=1}^{K} \frac{\left|\mathbb{C}\text{ov}\left[\phi_k\left(\boldsymbol{f}_{k_S}\left(\mathbf{x}^{(k)}\right)\right), \tau_k\left(\boldsymbol{f}_{k_P}\left(\mathbf{x}^{(k)}\right)\right)\right]\right|}{\sqrt{\mathbb{V}\left[\phi_k\left(\boldsymbol{f}_{k_S}\left(\mathbf{x}^{(k)}\right)\right)\right]}\sqrt{\mathbb{V}\left[\tau_k\left(\boldsymbol{f}_{k_P}\left(\mathbf{x}^{(k)}\right)\right)\right]}}. \tag{9}$$

is used to promote group independence between the pairs of random vectors $\mathbf{g}$ and $\mathbf{c}^{(k)}$. This can be verified after noticing that the maximum correlation between several mappings of the two blocks is minimized, for each view, and recalling that when all functions of two random variables are uncorrelated, the two random variables are independent.

When the dimensions of the latent components are known, under the generative model in (6) and the assumption in (7), the authors have shown that the criterion in (8) is capable of recovering the common and the private components up to nonlinear invertible transformations. On the other hand, when the dimensions are unknown, the possibility of modelling common components as uncommon appears, as the independence promoting term does not enforce independence between random vectors $\mathbf{c}^{(k)}$. Moreover, as we show in the sequel, the assumption in (7) can be relaxed when we are interested only in estimating $\mathbf{g}$.

III. Proposed Framework

In this section, we propose a multiview generative model that links the latent and the observed random vectors, $\mathbf{g}$ and $\mathbf{x}^{(k)}$. Based on this model, we consider the problem of estimating random vector $\mathbf{g}$ from realizations of $\mathbf{x}^{(k)}$ and propose a problem formulation for learning such estimators.

A. Proposed Multiview Generative Model

For a collection of K random vectors $\mathbf{x}^{(k)} \in \mathbb{R}^{D_k} \sim \mathcal{D}_{\mathbf{x}^{(k)}}$, for $k \in [K]$, we propose the following generative model

$$\mathbf{x}^{(k)} = \boldsymbol{v}_k\left(\begin{bmatrix}\mathbf{g}\\ \mathbf{c}^{(k)}\end{bmatrix}\right), \tag{10}$$

where $\boldsymbol{v}_k : \mathbb{R}^{F+L_k} \to \mathbb{R}^{D_k}$ are measurable functions, $\mathbf{g} \in \mathbb{R}^F \sim \mathcal{D}_{\mathbf{g}}$ is an F-dimensional random vector, and $\mathbf{c}^{(k)} \in \mathbb{R}^{L_k} \sim \mathcal{D}_{\mathbf{c}^{(k)}}$ are *conditionally* independent random vectors given $\mathbf{g}$, i.e.

$$p\left(\mathbf{g}, \mathbf{c}^{(1)}, \ldots, \mathbf{c}^{(K)}\right) = p\left(\mathbf{g}\right)\prod_{k=1}^{K} p\left(\mathbf{c}^{(k)}|\mathbf{g}\right). \tag{11}$$

The generative model above implies that

$$p\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)}\right) = \int p\left(\mathbf{g}\right)\prod_{k=1}^{K} p\left(\mathbf{x}^{(k)}|\mathbf{g}\right) d\mathbf{g}. \tag{12}$$

Notice that the proposed generative model is more general that the one of Lyu et al. [21], since it allows dependencies among random vectors $\mathbf{c}^{(k)}$, but also between random vectors $\mathbf{c}^{(k)}$ and $\mathbf{g}$. Next, we consider the problem of estimating $\mathbf{g}$ from complete realizations of the views.

On estimating random vector $\mathbf{g}$*:* Given joint realizations of random vectors $\mathbf{x}^{(k)}$, we consider the problem of estimating $\mathbf{g}$. Several known estimators can be used for such a task that are optimal under different criteria. For example, under the Mean Squared Error (MSE) criterion, it is known that the optimal estimator is given by the conditional expectation

$$\mathbb{E}\left[\mathbf{g}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)}\right]. \tag{13}$$

However, two practical concerns appear with this option. One is that approximating this function of all $\mathbf{x}^{(k)}$ can be challenging in terms of complexity and scalability. Another one is that we often wish to learn $\mathbf{g}$ from only partial realizations of the views. The conditional expectation in (13) can in principle still be used in this case, but such an

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3333212

approach is impractical (see discussion in the Appendix). In that direction, DGCCA explores the other extreme and poses the problem of finding optimal single-view-based estimators of $\mathbf{g}$ that achieve as strong agreement as possible. Although the resulting estimators may be suboptimal, compared to the one in (13), this is a practical alternative with respect to the aforementioned concerns.

Another advantage of the DGCCA approach is that the obtained solution directly provides estimators of $\mathbf{g}$, without having to estimate the remainder of the assumed generative model. This characteristic is consistent with the main principle in statistical learning theory for solving problems using a restricted amount of information: "*When solving a given problem, try to avoid solving a more general problem as an intermediate step*" [31]. Estimating the full generative model is a far more general problem which may entail much higher sample complexity, and therefore worse performance when working with limited training data, as we will verify in our experiments.

The fact we do not have access to any realizations of $\mathbf{g}$ poses some additional challenges. For example, as we mentioned earlier, there exist trivial solutions that satisfy the criterion of DGCCA in (5). Trivial solutions often appear in unsupervised learning problem formulations [32], [33], [21]. Several approaches have been considered for addressing this issue, such as adding view reconstruction regularization terms as in (4), or considering flow-based [34] and entropy-based [33] regularizations. Moreover, there also exist certain inherently unresolvable ambiguities. Specifically, it can be shown that there exist several invertible functions $\boldsymbol{\gamma} : \mathbb{R}^F \to \mathbb{R}^F$, such that random vector $\boldsymbol{\gamma}(\mathbf{g})$ is a valid alternative representation of the common latent random vector.

B. Proposed Formulation

Our goal is to exclude trivial solutions that could emerge with DCCA and DGCCA, but also to avoid: (i) having information leakage from the uncommon latent components to the estimates of $\mathbf{g}$ that could emerge with DCCAE in (4), (ii) the need to estimate the private components of the different views, which are often not needed for the downstream tasks after (G)CCA – especially when these represent strong noise in the individual views.

After adopting our proposed generative model, we have that all the observed random vectors $\mathbf{x}^{(k)}$ are conditionally independent given $\mathbf{g}$. As a result, given joint realizations of $\mathbf{g}$ and $\mathbf{x}^{(l)}$, for $l \neq k$, the optimal Mean Squared Error (MSE) estimator of view $\mathbf{x}^{(k)}$ can be expressed as a function of $\mathbf{g}$. This observation motivates us to consider the following formulation of multiview DGCCA

$$\begin{aligned} \min_{\substack{\{\boldsymbol{f}_k \in \mathcal{C}_f\}_{k=1}^K, \\ \hat{\mathbf{g}}, \\ \{\boldsymbol{w}_k \in \mathcal{C}_w\}_{k=1}^K}} & \mathcal{B}^{(K)}\left(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K, \hat{\mathbf{g}}\right) := \sum_{k=1}^{K} \mathbb{E}\left[\left\|\boldsymbol{w}_k\left(\hat{\mathbf{g}}\right) - \mathbf{x}^{(k)}\right\|^2\right] \\ \text{s.t.} \quad & \mathbb{E}\left[\hat{\mathbf{g}}\hat{\mathbf{g}}^T\right] = \mathbf{I}_F, \quad \mathbb{E}\left[\hat{\mathbf{g}}\right] = \mathbf{0}_F, \\ & \mathcal{R}^{(K)}\left(\boldsymbol{f}_1, ..., \boldsymbol{f}_K, \hat{\mathbf{g}}\right) = 0. \end{aligned} \tag{14}$$

The formulation in (14) is designed to guard against leakage of individual components into the estimate of $\mathbf{g}$, and we rigorously prove this property in the following Theorem. Note that invertible nonlinear transformations of (sharing the same MSE prediction capabilities as) random vector $\mathbf{g}$ can be obtained even from the problem formulation in (14) – this is inherently unresolvable, as discussed earlier. We therefore limit ourselves to identifying an invertible nonlinear transformation of $\mathbf{g}$ and we provide a sufficient condition for this to happen in the following Theorem, whose proof can be found in the Appendix.

Theorem 1. *Under the proposed generative model in (10), consider a solution of the optimization problem in (14). If the following additional assumptions,*

(i) *functions $\boldsymbol{v}_k$ are also partially invertible w.r.t. $\mathbf{g}$, i.e., there exist functions $\boldsymbol{u}_k : \mathbb{R}^{D_k} \to \mathbb{R}^F$, such that $\boldsymbol{u}_k\left(\mathbf{x}^{(k)}\right) = \mathbf{g}$, for all $\mathbf{x}^{(k)}$ and $k \in [K]$,*

(ii) *there exists mean dependence between at least one pair of random vectors $\mathbf{g}$ and $\mathbf{x}^{(k')}$, i.e.*

$$\mathbb{E}\left[\mathbf{x}^{(k')}|\mathbf{g}\right] \neq \mathbb{E}\left[\mathbf{x}^{(k')}\right] \text{ for a } k' \in [K] \text{ and all } \mathbf{g} \in \mathbb{R}^F,$$

hold, then the learned encodings $\boldsymbol{f}_k\left(\mathbf{x}^{(k)}\right)$ have to be non trivial functions only of $\mathbf{g}$, i.e. $\boldsymbol{f}_k\left(\boldsymbol{v}_k\left(\begin{bmatrix}\mathbf{g} \\ \mathbf{c}^{(k)}\end{bmatrix}\right)\right) = \boldsymbol{\gamma}(\mathbf{g})$, where $\boldsymbol{\gamma} : \mathbb{R}^F \to \mathbb{R}^F$. Moreover, if

(iii) *the conditional expectation, $\mathbb{E}\left[\left[\mathbf{x}^{(1)^T}, \ldots, \mathbf{x}^{(K)^T}\right]^T | \mathbf{g}\right]$, is an invertible function of $\mathbf{g}$,*

then function $\boldsymbol{\gamma}$ is also invertible and the latent common random vector $\mathbf{g}$ is identifiable up to invertible nonlinearities.

Remark III.1. *$E[\mathbf{x}^{(k)}|\mathbf{g}] = \phi(\mathbf{g})$ is the MMSE estimate of $\mathbf{x}^{(k)}$ given random vector $\mathbf{g}$, which is a function of $\mathbf{g}$. Our assumption that $E[\mathbf{x}^{(k)}|\mathbf{g}] \neq E[\mathbf{x}^{(k)}]$ simply means that this function is non-trivial, in the sense that it is not a constant (note that the right hand side $E[\mathbf{x}^{(k)}]$ is a constant, not a random vector). This is a very mild assumption, especially considering that we only assume it for* one *view.*

Remark III.2. *Assumption (ii) would not be satisfied only if every pair of random vectors $\mathbf{g}$ and $\mathbf{x}^{(k)}$ were mean-independent. However, this would require a quite careful setting of circumstances in order to happen. To see that, consider the following example of two mean-independent random variables that are actually dependent. Let $\mathbf{z}$ be the discrete two-dimensional random vector that equiprobably has the following 4 outcomes $\{(0,-1),(0,+1),(-1,0),(0,+1)\}$, as appears in Fig. 1. It can be easily seen that it consists of quite dependent elements as the value of the one coordinate may reveal the value of the other. For example, $\mathbf{z}[2] = 0$ whenever $\mathbf{z}[1] = \pm 1$. However, we have that*

$$\mathbb{E}\left[\mathbf{z}\left[2\right]|\mathbf{z}\left[1\right] = 0\right] = \mathbb{E}\left[\mathbf{z}\left[2\right]|\mathbf{z}\left[1\right] = \pm 1\right] = \mathbb{E}\left[\mathbf{z}\left[2\right]\right] = 0.$$

As a result, the elements of $\mathbf{z}$, although dependent, they are mean-independent. Nevertheless, this example empirically shows that mean-independence requires a very careful design of outcomes and assigned probabilities to them, in order to take place.

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3333212

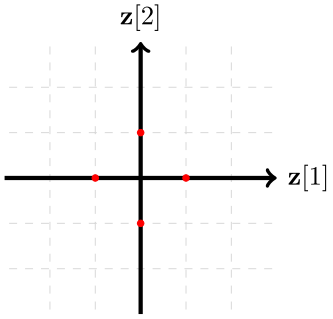

Fig. 1. Example of two discrete mean-independent random variables that are actually dependent.

The appearance of random vector $\hat{\mathbf{g}}$ in the objective of (14) but also in the constraints of (14) complicates the design of optimization algorithms for solving it. For example, even if we did not have to deal with the nonlinear constraints, $\mathcal{R}^{(K)}\left(\boldsymbol{f}_1,...,\boldsymbol{f}_K,\hat{\mathbf{g}}\right)=0$, the direct approach of updating the realizations of $\hat{\mathbf{g}}$ would require methods that restrain the realizations on the Stiefel manifold, which are computationally expensive and sensitive to local minima. Next we consider two modifications that lead into a tractable approximation of (14). First, under the first additional assumption of Theorem 1, there exist feasible solutions, with respect to random vector $\hat{\mathbf{g}}$ and functions $\{\boldsymbol{f}_k\}_{k=1}^K$, such that the encoding of each view, $\boldsymbol{f}_k\left(\mathbf{x}^{(k)}\right)$, will be equal to $\hat{\mathbf{g}}$ in the mean square sense. As a result, an encoding of a view can be used in place of $\hat{\mathbf{g}}$ in the objective of (14). Furthermore, we can get that

$$\begin{aligned}\mathcal{B}^{(K)}\left(\boldsymbol{w}_1,\ldots,\boldsymbol{w}_K,\hat{\mathbf{g}}\right)&=\mathcal{Q}^{(K)}(\boldsymbol{f}_1,\boldsymbol{w}_1,..,\boldsymbol{f}_K,\boldsymbol{w}_K)\\&:=\frac{\sum_{k=1}^{K}\sum_{\substack{j=1\\j\neq k}}^{K}\mathbb{E}\left[\left\|\boldsymbol{w}_k\left(\boldsymbol{f}_j\left(\mathbf{x}^{(j)}\right)\right)-\mathbf{x}^{(k)}\right\|^2\right]}{(K-1)},\end{aligned}$$

which is not a function of $\hat{\mathbf{g}}$. As for the nonlinear constraints $\mathcal{R}^{(K)}\left(\boldsymbol{f}_1,...,\boldsymbol{f}_K,\hat{\mathbf{g}}\right)=0$, we consider using a Lagrangian based approach as in the formulations of DCCAE in (4) and of Lyu et al. in (8). The aforementioned modifications lead to the following tractable reformulation of (14)

$$\begin{aligned}\min_{\substack{\{\boldsymbol{f}_k\in\mathcal{C}_f\}_{k=1}^K,\\ \hat{\mathbf{g}},\\ \{\boldsymbol{w}_k\in\mathcal{C}_w\}_{k=1}^K}}&(1-\lambda)\mathcal{R}^{(K)}(\boldsymbol{f}_1,...,\boldsymbol{f}_K,\hat{\mathbf{g}})+\lambda\mathcal{Q}^{(K)}(\boldsymbol{f}_1,\boldsymbol{w}_1,...,\boldsymbol{f}_K,\boldsymbol{w}_K)\\ \text{s.t.}\quad&\mathbb{E}\left[\hat{\mathbf{g}}\hat{\mathbf{g}}^T\right]=\mathbf{I}_F,\quad\mathbb{E}\left[\hat{\mathbf{g}}\right]=\mathbf{0}_F.\end{aligned}\tag{15}$$

The intuition behind this reformulation should be clear: even if the mappings of the realizations of all the views are not equal, as long as they are close enough to $\hat{\mathbf{g}}$ they can still be used in lieu of $\hat{\mathbf{g}}$ for the reconstruction of any view. By excluding $\mathbf{x}_k$ from participating in the reconstruction of itself, we can guarantee that there will be no information leakage from the corresponding private random vector $\mathbf{c}_k$, regardless of how strong it is. In other words, by considering cross-reconstructions among all the views, we enforce capturing only the common latent factor, while the private latent components, that are potentially strong, will be ignored. As a result, the proposed formulation does not suffer from the downsides of the DCCAE method. Also, as we show next, this reformulation is convenient from an algorithmic point of view, as it enables the development of alternating optimization algorithms.

C. Proposed Algorithm

In this section, we propose an alternating optimization algorithm for updating the estimates of the realizations of the common latent random vector $\mathbf{g}$, as well as the nonlinear mappings $\boldsymbol{f}_k$ and $\boldsymbol{w}_k$. Next, we describe the corresponding optimization subproblems and the proposed algorithm.

1) Updating functions $\boldsymbol{f}_k$ and $\boldsymbol{w}_k$: For fixed estimates (realizations) of random vector $\mathbf{g}$, the problem of updating functions $\boldsymbol{f}_k$ and $\boldsymbol{w}_k$ is given by

$$\begin{aligned}\min_{\substack{\{\boldsymbol{f}_k\in\mathcal{C}_f\}_{k=1}^K,\\ \{\boldsymbol{w}_k\in\mathcal{C}_w\}_{k=1}^K}}&(1-\lambda)\,\mathcal{R}^{(K)}\left(\boldsymbol{f}_1,\ldots,\boldsymbol{f}_K,\mathbf{g}\right)\\&+\lambda\mathcal{Q}^{(K)}\left(\boldsymbol{f}_1,\boldsymbol{w}_1,\ldots,\boldsymbol{f}_K,\boldsymbol{w}_K,\mathbf{g}\right).\end{aligned}\tag{16}$$

This is an unconstrained problem that can be tackled using stochastic optimization methods.

2) Updating the realizations of $\hat{\mathbf{g}}$: For fixed functions $\boldsymbol{f}_k$ and $\boldsymbol{w}_k$, the problem of updating the realizations of $\hat{\mathbf{g}}$ is given by

$$\begin{aligned}\min_{\hat{\mathbf{g}}}&\sum_{k=1}^{K}\mathbb{E}\left[\left\|\boldsymbol{f}_k\left(\mathbf{x}^{(k)}\right)-\hat{\mathbf{g}}\right\|^2\right]\\ \text{s.t.}\quad&\mathbb{E}\left[\hat{\mathbf{g}}\hat{\mathbf{g}}^T\right]=\mathbf{I}_F,\quad\mathbb{E}\left[\hat{\mathbf{g}}\right]=\mathbf{0}_F.\end{aligned}\tag{17}$$

Given jointly drawn realizations of random vectors $\mathbf{x}^{(k)}$, $\mathbf{x}_m^{(k)}$, the goal is to estimate the corresponding realization $\mathbf{g}_m$, for $m\in[M]$. Consider the matrices $\mathbf{G}$ and $\mathbf{Y}\in\mathbb{R}^{M\times F}$, where $\mathbf{G}(m,:)=\hat{\mathbf{g}}_m$ and $\mathbf{Y}(m,:)=\sum_{k=1}^K\boldsymbol{f}_k\left(\mathbf{x}_m^k\right)$. Then, updating the realizations of $\mathbf{g}$ boils down to the Orthogonal Procrustes problem [35]

$$\max_{\mathbf{G}\in\mathbb{R}^{M\times F}}\text{Trace}\left(\mathbf{G}^T\bar{\mathbf{Y}}\right),\ \text{s.t.}\ \mathbf{G}^T\mathbf{G}=\mathbf{I}_F,\tag{18}$$

where $\bar{\mathbf{Y}}$ denotes the columnwise centered matrix $\mathbf{Y}$. Given the Singular Value Decomposition (SVD) of $\bar{\mathbf{Y}}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, an optimal solution can be computed as $\mathbf{G}^o=\mathbf{U}\mathbf{V}^T$.

3) Algorithm and Complexity: The proposed algorithm is outlined in Algorithm 1. The update of $\mathbf{G}$ requires the computation of SVD with complexity $\mathcal{O}\left(MF^2\right)$. Performing a stochastic update of $\boldsymbol{f}_k$ and $\boldsymbol{w}_k$, based on a $|\mathcal{B}|$-sized batch, has complexity $\mathcal{O}\left(|\mathcal{B}|\sum_{k=1}^K d_k\right)$, where d_k denotes the number of parameters used for approximating $\boldsymbol{f}_k$ and $\boldsymbol{w}_k$. Hence, the overall complexity of the proposed algorithm is $\mathcal{O}\left(MF^2+|\mathcal{B}|\sum_{k=1}^K d_k\right)$ per iteration. As for the stopping criteria, we use completing a predetermined number of iterations for the outer loop and a full epoch for the inner one.

IV. Experimental Results

In this section, we present the experimental results that emerged after comparing the proposed framework to the state-of-the-art methods on three datasets; one synthetic and two real world. For our experiments, we have extended all the methods that are restricted to only two views by substituting the terms that coincide with the objective of DCCA to the

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3333212

Algorithm 1 Proposed Algorithm

Input: $\left\{\mathbf{x}_m^{(k)}\right\}_{m=1}^{M}, \boldsymbol{f}_k, \boldsymbol{w}_k, \forall\, k \in [K]$

Output: $\boldsymbol{f}_k^*, \boldsymbol{w}_k^*\ \forall k \in [K]$

while *stopping criterion is not met* **do**

- Update matrix $\mathbf{G}$ by solving (17)

while *stopping criterion is not met* **do**

- Draw at random $\mathcal{B} = \left\{\mathbf{x}_m^{(k)}, \forall\ k \in [K]\right\}_{m \in \mathcal{M}' \subset [M]}$

- Perform a stochastic gradient update of $\boldsymbol{f}_k$ and $\boldsymbol{w}_k, \forall\ k \in [K]$, by restricting (16) on $\mathcal{B}$

end while

end while

one of DGCCA. Regarding the method proposed in [21], such an extension can be achieved in a straightforward way by considering an extra set of terms in (8) for the third view. In all the experiments and for all the methods, all the MSE terms in the corresponding objectives have been normalized to reflect the relative MSE. As a result, a fair comparison of all the methods over the same value of λ is enabled.

A. Synthetic Dataset

We begin the comparison of all the methods on a two-view synthetic dataset. For its generation, we consider the following setting. First, a sample is drawn from a categorical random variable Z, with alphabet $\{0, 1, 2, 3\}$ and corresponding probabilities $0.1, 0.2, 0.3$, and 0.4. As common random vector $\mathbf{g}$, we consider the one-hot encoding of Z. Regarding the private random vectors, we assume that each $\mathbf{c}^{(k)} \in \mathbb{R}^4$ follows a mixture of four zero mean Gaussian distributions with random covariance matrices. For each view, the Gaussian distribution specified by the drawn value of Z is sampled to obtain the realization of the corresponding random vector $\mathbf{c}^{(k)}$. The power ratio between common and uncommon random vectors is set to $-18\,\text{dB}$.

Based on the above procedure, we form the latent training, validation, and testing datasets, consisting of $3,000$, $1,500$, and $1,500$ samples, respectively. These samples are mapped through functions $\boldsymbol{v}_k(\cdot) : \mathbf{R}^8 \to \mathbf{R}^{64}$, for $k = 1, 2$, so as to create samples of the two views. For the construction of each function $\boldsymbol{v}_k$, we use a three-hidden-layer neural network with 32 neurons per layer. The first two layers are followed by a ReLU activation function, while all the network weights are drawn i.i.d. from the standard normal distribution.

In order to approximate functions $\boldsymbol{f}_k$ and $\boldsymbol{w}_k$, we use the same architecture as above. To train them, we use the AdamW optimizer [36] with initial learning rate equal to 0.001 and a regularization parameter equal to 0.001. We set the batch size equal to 100 samples and we run the algorithm for 40 epochs. For the method of Lyu et al. [21], we set the batch size for the maximization problem equal to 300 and the learning rate equal to 0.01. Moreover, we approximate τ_k and ϕ_k using the same architecture as described above. For the regularization parameter of the independence-promoting term, we consider two different settings. Note that random vectors $\mathbf{c}^{(k)}$ are not independent under the considered generative model. They are *conditionally* independent given $\mathbf{g}$, but this allows for (unconditional) dependence. This is not in agreement with the generative model of Lyu et al. [21]. For this reason, we consider two different values of the regularization parameter for the independence-promoting term: 0 and 0.001 (the independence-promoting term is ignored/considered, respectively). We employ an early stopping policy, by keeping the model that achieved the lowest objective value on the validation dataset throughout the considered number of iterations.

The embedded views ideally should convey only information related to the value of Z, which can be considered as label. Hence, all the pairs sharing the same value/label should have similar embeddings in the learned latent space. We compare the proposed method to all the state-of-the-art methods in terms of how well it encodes the label information. Towards this end, we follow [18] and [21] and consider a supervised and an unsupervised test. In more detail, we train the models of all methods for λ taking all the values in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. For each value of λ and method, we consider 10 different initializations. For each learned model, we compute the averages of the embeddings of the two views for the three datasets (training/validation/testing).

Regarding the unsupervised test, we use K-means in order to split the learned embeddings into 4 clusters and evaluate how well the clusters agree with ground-truth labels. In the previous step, K-means is run 10 times with random initialization and the run with the best K-means objective is finally considered. We measure the clustering performance with three criteria, clustering accuracy (ACC), normalized mutual information (NMI), and adjusted Rand index (ARI) [37]. For the supervised test, we train a linear support vector machine (SVM) using the averaged embeddings of the training data. The performance of the trained SVM is evaluated on the averaged embeddings of the testing dataset and measured by the achieved classification accuracy (CLA-ACC). Finally, the final performance metrics of each method, for each value of λ, are determined by computing the averages over all 10 corresponding learned models.

In Tables I, II, III, and IV, we present the performance metrics of all methods considered versus the various values of λ. Since CCA and DGCCA do not depend on the λ parameter, we present the corresponding results only in the forth column ($\lambda = 0.5$). We can observe that the proposed method achieves the highest scores across all the metrics, and is also the least sensitive with respect to the choice of λ. In Fig. 2, we present the average correlation coefficient that each method achieved, over the 10 learned models, for each value of λ. Regarding CCA and DGCCA, for the reason explained above, their results are depicted here as horizontal lines. We observe that the proposed method also achieves the highest correlation coefficient.

Finally, in Fig. (3) we depict the average walltimes of all the considered methods for the synthetic dataset. The linear CCA method (denoted as LCCA in the two figures) is much faster than all the other methods, but, as we discussed above, its performance is restricted by its limited capabilities in modeling nonlinear transformations. We can also observe that DGCCA

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3333212

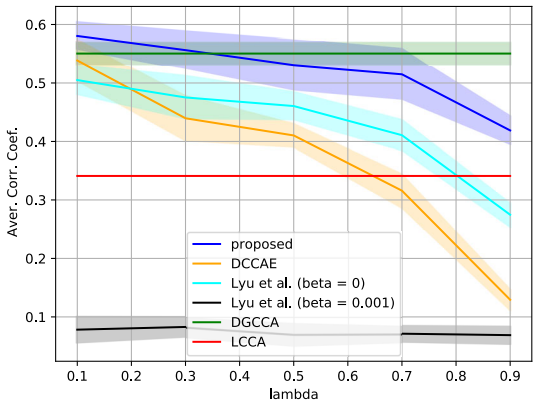

Fig. 2. Average Correlation Coefficient for the synthetic dataset across different values of the λ parameter.

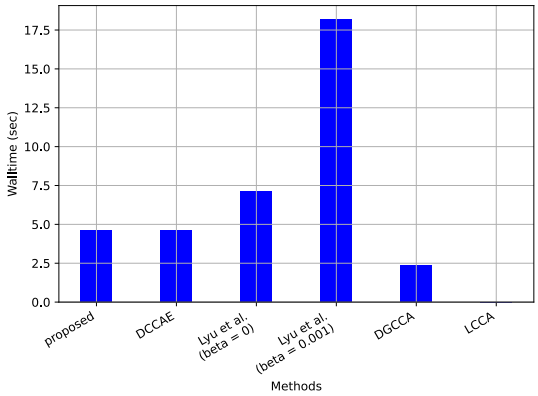

Fig. 3. Average walltimes of all the considered methods for the synthetic dataset.

is the next faster, which is expected since DGCCA does not learn functions $\boldsymbol{w}_k$ that allow the reconstruction of each view and as we discussed earlier, allows the possibility of trivial solutions. The proposed algorithm has comparable speed with the DCCAE method, while the method of Lyu et al. [21] is slower than all the other methods, especially when the independence promoting term is activated where it is significantly much slower. This is expected since the algorithmic complexity analysis that we provide in the main paper holds for all the deep methods and shows that the complexity of each method is proportional to the total number of parameters that each method uses to model the corresponding encoding and decoding functions. Since the method of [21] needs significantly more parameters in order to approximate the full autoencoders of the views, but also the functions for promoting independence, the resulting computational complexity is much higher.

B. Acoustic-articulatory data for speech recognition

Next, we consider a dataset from the University of Wisconsin X-Ray Micro-Beam Database (XRMB) [38]. XRMB consists of simultaneously recorded speech and articulatory measurements from 47 American English speakers. From those measurements, Wang et al. [6], [7] created a multiview

TABLE I
CLUSTERING ACCURACY (ACC) FOR THE SYNTHETIC DATASET ACROSS DIFFERENT VALUES OF THE λ PARAMETER.

	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$
CCA	-	-	0.92±0	-	-
DGCCA	-	-	0.97±0.01	-	-
DCCAE	0.96±0.05	0.86±0.05	0.84±0.05	0.74±0.03	0.48±0.06
Lyu et al. (2021) $\beta = 0$	0.97±0.01	0.95±0.03	0.95±0.02	0.88±0.05	0.75±0.04
Lyu et al. (2021) $\beta = 0.001$	0.40±0.06	0.39±0.04	0.40±0.03	0.36±0.03	0.38±0.03
Proposed	**0.98±0.01**	0.97±0.01	0.95±0.04	0.95±0.04	0.90±0.04

TABLE II
NORMALIZED MUTUAL INFORMATION (NMI) FOR THE SYNTHETIC DATASET ACROSS DIFFERENT VALUES OF THE λ PARAMETER.

	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$
CCA	-	-	0.77±0	-	-
DGCCA	-	-	0.90±0.02	-	-
DCCAE	0.89±0.04	0.80±0.03	0.77±0.03	0.63±0.05	0.26±0.05
Lyu et al. (2021) $\beta = 0$	0.90±0.02	0.87±0.05	0.87±0.03	0.80±0.04	0.61±0.05
Lyu et al. (2021) $\beta = 0.001$	0.10±0.06	0.08±0.04	0.08±0.02	0.08±0.04	0.08±0.04
Proposed	**0.92±0.03**	0.90±0.03	0.88±0.04	0.87±0.04	0.78±0.02

TABLE III
ADJUSTED RAND INDEX (ARI) FOR THE SYNTHETIC DATASET ACROSS DIFFERENT VALUES OF THE λ PARAMETER.

	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$
CCA	-	-	0.84±0	-	-
DGCCA	-	-	0.94±0.01	-	-
DCCAE	0.93±0.05	0.81±0.06	0.78±0.03	0.63±0.06	0.24±0.08
Lyu et al. (2021) $\beta = 0$	0.94±0.02	0.91±0.05	0.91±0.03	0.82±0.05	0.66±0.03
Lyu et al. (2021) $\beta = 0.001$	0.07±0.06	0.05±0.03	0.06±0.02	0.04±0.02	0.05±0.03
Proposed	**0.95±0.02**	0.94±0.02	0.91±0.05	0.91±0.05	0.84±0.04

TABLE IV
CLASSIFICATION ACCURACY (CLA-ACC) FOR THE SYNTHETIC DATASET ACROSS DIFFERENT VALUES OF THE λ PARAMETER.

	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$
MAX-VAR	-	-	0.94±0	-	-
DGCCA	-	-	0.98±0.01	-	-
DCCAE	0.98±0.01	0.93±0.03	0.91±0.02	0.87±0.01	0.67±0.04
Lyu et al. (2021) $\beta = 0$	0.98±0.01	0.97±0.01	0.97±0.01	0.94±0.02	0.85±0.03
Lyu et al. (2021) $\beta = 0.001$	0.54±0.05	0.54±0.06	0.53±0.05	0.50±0.07	0.53±0.04
Proposed	**0.99±0.01**	0.98±0.01	0.97±0.03	0.97±0.02	0.93±0.02

dataset[1], which has been used for multiview based phonetic recognition [5], [7], [18], [19]. It consists of two views, where acoustic and articulatory features have been concatenated over a 7-frame window around each frame, giving 273-dimensional acoustic and 112-dimensional articulatory features, respectively. Since each pair of samples here is tied together by two factors, the label of the spoken phoneme and the identity of the speaker, the authors of [39] and [19] considered using one-hot vector encodings of the phoneme labels $(0 - 39)$ as a third view, in order to find latent embeddings that convey information only related to the phoneme label. Finally, three datasets are provided for training, validation, and testing pur-

[1]The dataset can be found here: https://home.ttic.edu/~klivescu/XRMB_data/full/README

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3333212

poses, respectively, where each one of them contains samples of non-overlapping speakers.

In our experiments, we adopt the same framework and consider the three-view setup for phonetic recognition. To limit the experiment runtime, we follow [19] and we use a subset of speakers for our experiments. Specifically, we use as training dataset the samples that correspond to speakers 1 to 5, while we use the remaining two datasets (validation and testing) intact. The underlying task of our experiments is to exploit the latent embeddings for supervised phonetic recognition, as in [5], [7], [18], [19]. Following [18] and [19], we consider a dimension equal to 30 for the common factor $\mathbf{g}$. For the approximation of functions $\boldsymbol{w}_k$ and $\boldsymbol{f}_k$, we use three-hidden layer neural networks with 512 neurons per layer, where the first two layers are followed by a ReLU activation function. For the training of all the models, we use the AdamW optimizer [36] with an initial learning rate of 0.001 and a regularization parameter equal to 0.001. We set the batch size equal to 500 samples and we run the algorithm for 200 epochs. For the method proposed of Lyu et al. [21], we set the batch size for the maximization problem equal to 1000 and the learning rate equal to 10^{-6}. The dimensions of the individual components, for the first two views, are set to be equal to 60 and 20, after verifying that the auto-reconstruction errors are satisfactory low ($< 1\%$). For the approximation of functions τ_k and ϕ_k, we use three-hidden layer neural networks with 64 neurons per layer, where the first two layers are followed by a ReLU activation function. As the third view conveys no private information, neither private latent factors nor independence-promoting terms are considered for the third view. Regarding the regularization parameter for the independence-promoting term, we set it equal to 10^{-6}. Finally, we adopt an early stopping policy as we did in the previous subsection.

We train the models for all methods for regularization parameter $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. For each value and method, we consider 5 different initializations. We test each learned model on the classification task, by using the K-nearest neighbors algorithm (with K=5) on the averages of the embeddings of all three views and evaluating them on the averages of the embeddings of the first two views since the third view relies on the phoneme labels. Finally, we consider as performance metric the classification accuracy score after averaging all 5 learned models for each method. Since the concept of correlation coefficient does not naturally extend to more than two random vectors, without having access to their joint distribution, we consider the following metric for measuring the attained correlation among the three views. For each pair of views, we compute the total correlation coefficient defined as the average of the cosines of the canonical angles between the subspaces spanned by the two views. Then, we compute the average of the three total correlation coefficients and use that as total correlation coefficient of a triplet. We report as final correlation coefficient the averages of the total correlation coefficient of each triplet over the 5 learned models for each method.

In Table V, we present the classification accuracy score that each method attained. Notice that MAX-VAR and DGCCA have no tuning parameter, hence we report their attained score in the fourth column ($\lambda = 0.5$). In Fig. 4 we depict the total correlation coefficient as described above. The correlation coefficients for MAX-VAR and DGCCA are presented using straight lines. We observe that although our method does not attain the highest correlation coefficient, it achieves the highest classification accuracy. This observation comes to validate our statement that the quality of the embeddings of all the other methods can be deteriorated because of potential information leakage from the corresponding private components, but also because they open to capturing trivial solutions. As we see here, fictitious high correlation coefficients can be found by deep CCA-based methods, without however capturing useful information. At last, in Fig. 5, we depict the average walltimes that emerged after training all the methods. We can see that the observations we noted in our experiments with the synthetic data carry on here, too, with the difference that the DCCAE method and the proposed method are more than three times faster than the method of Lyu et al. [21].

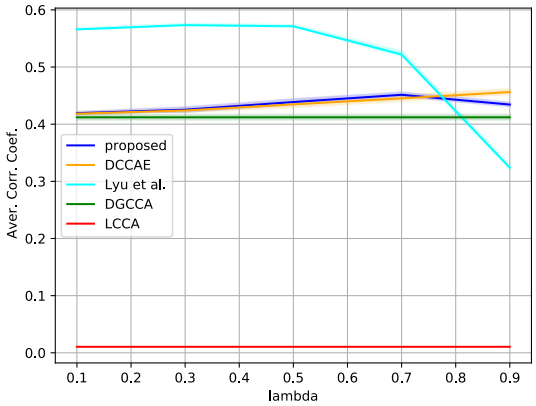

Fig. 4. Average Correlation Coefficient for the XRMB dataset across different values of the λ parameter.

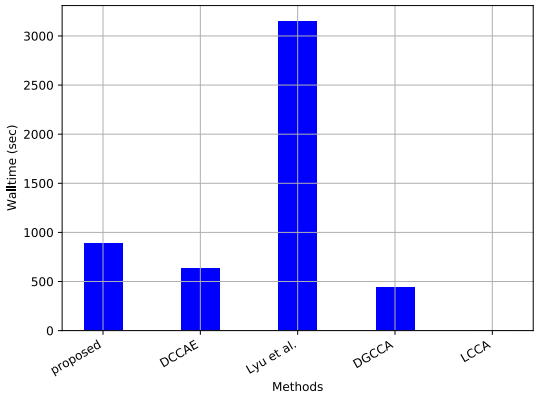

Fig. 5. Average walltimes of all the considered methods for the XRMB dataset.

C. Multiview Noisy MNIST digits

The Multiview Noisy MNIST dataset was introduced in [18] and it is a popular dataset in multiview learning [18], [21]. It is based on the well-known MNIST dataset [40], which consists of 28×28 grayscale digit images, with 60,000 and 10,000 images for training and testing, respectively. Wang et al. [18]

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3333212

TABLE V
AVERAGE CLASSIFICATION SCORES FOR THE PHONEME CLASSIFICATION TASK ACROSS DIFFERENT VALUES OF THE λ PARAMETER.

	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$
MAX-VAR	-	-	0.35±0	-	-
DGCCA	-	-	0.46±0	-	-
DCCAE	0.47±0.01	0.50±0.01	0.52±0.01	0.54±0.01	0.57±0
Lyu et al. (2021)	0.60±0.01	0.60±0.01	0.60±0	0.60±0	0.60±0
Proposed	0.51±0.01	0.54±0.01	0.57±0.01	0.58±0.01	**0.63±0.01**

TABLE VI
NETWORK ARCHITECTURES FOR THE MULTIVIEW MNIST DATASET.

Encoders ($\boldsymbol{f}_k$)	Decoders ($\boldsymbol{w}_k$)
input: $28 \times 28 \times 1$	input: latent_dimension
4×4 Conv, 64, stride 2, ReLU	FC 256, ReLU
4×4 Conv, 32, stride 2, ReLU	FC $7 \times 7 \times 32$, ReLU
FC 256, ReLU	4×4 Conv_Trans, 64 ReLU, stride 2
FC latent_dimension	4×4 Conv_Trans, 1, stride 2

considered generating a two-view variation of the MNIST dataset, where the first view consists of randomly rotated digit images, while the second view consists of noisy digit images. More specifically, the dataset is generated as follows. The pixel values of all images of the initial dataset are first rescaled to the $[0, 1]$ interval. Then, all the images are rotated by random angles uniformly sampled from $[-\pi/4, \pi/4]$ and the resulting images are used to construct the first view. For each image of the first view, an image of the same digit (0-9) is randomly selected from the original dataset. Then, a noisy version of that image is obtained by adding noise to each pixel independently and uniformly sampled from $[0, 1]$. The pixel values of the resulting image are truncated to the $[0, 1]$ interval and the final image is used as the corresponding sample of the second view. At last, the original training set is further split into training and tuning sets of size 50,000 and 10,000, respectively.

What is common between the realizations of the two views is the identity of the depicted digits. As a result, the encodings of the two views ideally should convey only information related to the identity of the pair. Hence, all the pairs sharing the same identity should have similar embeddings in the learned latent space. In this section, we compare the proposed method to all the state-of-the-art methods in the task of encoding the class label information. To achieve that, we follow [18], [21] and we consider a supervised setup and an unsupervised setup, as we did in our experiments with the synthetic dataset. For the supervised part, we test the performance of all the methods using a classifier, while for the unsupervised one, we measure the class separation in the learned latent space.

Following [18], [21], we consider a dimension equal to 10 for the common factor $\mathbf{g}$. In Table VI we present the architectures of the encoders and decoders that we used for our experiments with the Multiview MNIST dataset. The structures are common for the two views. In order to approximate the functions τ_k and ϕ_k of the method proposed in [21], we use three hidden-layer neural networks with 64 neurons per layer, where the first two layers are followed by a ReLU activation function. For the training of all the models, we use the AdamW optimizer [36] with initial learning rate of 0.001 and a regularization parameter of 0.001. We set the batch size equal to 500 samples and we run the algorithm for 40 epochs. For the method of [21], we set the batch size for the maximization problem equal to 1000 and the learning rate equal to 10^{-3}. The dimensions of the individual components are set equal to 20 and 50, respectively, as it is mentioned in their experiments. Based on how this dataset is constructed, one can detect similarities with the synthetic dataset we considered above. In order to examine whether the proposed generative model or the generative model proposed by Lyu et al. [21] fits better on this dataset, we consider two different values for the regularization parameter for the independence promoting term, 0 (the independence promoting term is ignored) and 0.001 (the independence promoting term is considered), as we did in our experiments with the synthetic dataset. At last, we adopt the same early stopping policy as we did with the previous two datasets.

We train the models of all methods for $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. For each value of λ and method, we consider 5 different initializations resulting in 5 different learned models. For each learned model, we consider two tests corresponding to the supervised and the unsupervised evaluation setups, respectively, as we did with the synthetic dataset. To make the tests more challenging, we create three datasets in the latent space (training/validation/testing), by considering the embeddings of the samples of the noisy view for the three initial datasets (training/validation/testing), instead of the averages of the embeddings of the samples of both views. For the supervised test, we train a linear support vector machine (SVM) using the latent training data. The performance of the trained SVM is evaluated on the latent testing dataset and is measured by classification accuracy (CLA-ACC). Regarding the unsupervised test, we consider using K-means in order to split the latent training dataset into 10 clusters and evaluate how well the clusters agree with ground-truth labels. In the previous step, K-means is run 10 times with random initialization and the run with the best K-means objective is finally considered. We measure the clustering performance with three criteria, clustering accuracy (ACC), normalized mutual information (NMI), and adjusted Rand index (ARI) [37]. Finally, the final performance metrics of each method, for each value of λ, is determined by computing the average over the performance metrics obtained from all the 5 corresponding learned models.

In Tables VII, VIII, IX, and X, we present the final performance matrices of all the considered methods versus the considered values of the λ parameter. We can observe that the proposed method achieves the best scores across all the considered metrics, consistently, and that also is the most robust to the variation of the λ parameter. In Fig. 6, we present the average correlation coefficient that each method achieved over the 5 learned models for each value of the λ parameter. We observe that the method proposed by [21] achieves a slightly higher correlation coefficient for small values of λ and $\beta = 0$, but it has significantly higher variance over the different initializations. On the other hand, the proposed method achieves comparably high correlation coefficients, with significantly less variance over the different initializations, and is less sensitive to the variation of the λ parameter. Moreover, we can observe that all the perfor-

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3333212

mance scores and the correlation coefficient drop significantly when we consider the independence-promoting term for [21] method, which is an indication that the proposed generative model is more appropriate for this dataset compared to the one in [21]. In Fig. 7, we depict the walltime of all the methods for the Multiview Mnist dataset. We can see that the observations we noted for the previous two datasets carry on for the Multiview Mnist dataset, too.

Next, we consider the t-SNE visualizations [41] of the learned latent embeddings for the Multiview MNIST dataset. More specifically, we consider the samples appearing in the noisy view from the testing dataset and the learned models that achieved the lowest objective values, over the validation dataset, for all the methods. The resulting embeddings are fed to the t-SNE algorithm. In Fig. 8, we present the obtained visualizations for the CCA and the DGCCA methods. In Fig. 9, we present the obtained visualizations for the DCCAE method over different values of λ. In Fig. 10 and Fig. 11, we present the obtained visualizations for the method proposed in [21] over different values of λ, for $\beta = 0$ and $\beta = 0.001$, respectively. At last, in Fig. 12, we present the obtained visualizations for the proposed method over different values of λ. We observe that the 10 clusters are well separated for all the deep methods for small values of λ, except for the method of Lyu et al. [21] and $\beta = 0.001$. In addition, we can observe that the proposed method is the only one that produces consistently clustered embeddings for all values of λ.

V. CONCLUSIONS

In this work, we revisited the problem of DGCCA and we highlighted the advantages and limitations that the previous works have. We proposed a novel formulation of the problem that overcomes most of those limitations and an algorithm for estimating the common latent random vector given jointly observed views. We tested the proposed concepts in artificial and real-life datasets. The obtained results corroborate our claims that the proposed framework is more appropriate and performs better in a series of tasks.

APPENDIX

ESTIMATING RANDOM VECTOR $\mathbf{g}$ - FURTHER DISCUSSION

We estimate random vector $\mathbf{g}$ from realizations of random vectors $\mathbf{x}^{(k)}$ via the conditional expectation

$$\mathbb{E}\left[\mathbf{g}\middle|\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(K)}\right]. \tag{19}$$

Although this estimator is optimal under the Mean Squared Error criterion, it comes with two important practical caveats. First, it is a very high-dimensional and generally unstructured function of all random vectors $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(K)}$, the approximation of which can be challenging in terms of complexity and scalability. The second caveat comes into play when we would also like to be able to estimate $\mathbf{g}$ from partial realizations of $\left\{\mathbf{x}^{(k)}\right\}_{k\in\mathbb{S}}$, with $\mathbb{S} \subset [K]$ – which is often important in practice. Since the optimal estimators, in this case, are functions of the one in (19) via

$$\mathbb{E}\left[\mathbf{g}\middle|\left\{\mathbf{x}^{(k)}\right\}_{k\in\mathbb{S}}\right] = \mathbb{E}_{\left\{\mathbf{x}^{(k)}\right\}_{k\notin\mathbb{S}}}\left[\mathbb{E}\left[\mathbf{g}\middle|\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(K)}\right]\right], \tag{20}$$

we can see that given the estimator in (19), the problem of computing the optimal estimators with respect to the subset $\mathbb{S}$ boils down to finding ways to efficiently compute different expectations of (19) over the views that belong to the complement of $\mathbb{S}$.

Considering the extreme case where each $\mathbb{S}$ is a singleton, we may consider combining multiple single-view estimators of $\mathbf{g}$, each of them still given by (20). Although each of these estimators may have significantly poorer prediction capabilities compared to the one in (19), they have the advantage that they may allow designing efficient and scalable frameworks for estimating $\mathbf{g}$ by combining the single view based estimates from each random vector $\mathbf{x}^{(k)}$. The problem that arises here is how may one combine such estimators in a way that is

TABLE VII
CLUSTERING ACCURACY (ACC) FOR THE MULTIVIEW MNIST ACROSS DIFFERENT VALUES OF THE λ PARAMETER.

	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$
CCA	-	-	0.89±0	-	-
DGCCA	-	-	0.97±0	-	-
DCCAE	0.97±0	0.97±0	0.97±0	0.94±0.01	0.49±0.15
Lyu et al. (2021) $\beta = 0$	0.97±0	0.97 ±0	0.94±0.05	0.94±0.01	0.62±0.11
Lyu et al. (2021) $\beta = 0.001$	0.21±0.02	0.22±0.02	0.22±0.02	0.21±0.01	0.21±0.02
Proposed	**0.98±0**	**0.98±0**	0.97±0	0.97±0	0.87±0.21

TABLE VIII
NORMALIZED MUTUAL INFORMATION (NMI) FOR THE MULTIVIEW MNIST ACROSS DIFFERENT VALUES OF THE λ PARAMETER.

	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$
CCA	-	-	0.79±0	-	-
DGCCA	-	-	0.93±0	-	-
DCCAE	**0.93±0**	**0.93±0**	0.92±0	0.87±0.01	0.42±0.17
Lyu et al. (2021) $\beta = 0$	**0.93±0**	0.92±0	0.90±0.03	0.87±0.01	0.56±0.11
Lyu et al. (2021) $\beta = 0.001$	0.09±0.01	0.11±0.03	0.11±0.02	0.10±0.01	0.10±0.02
Proposed	**0.93±0**	**0.93±0**	**0.93±0.01**	**0.93±0**	0.82±0.21

TABLE IX
ADJUSTED RAND INDEX (ARI) FOR THE MULTIVIEW MNIST DATASET ACROSS DIFFERENT VALUES OF THE λ PARAMETER.

	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$
CCA	-	-	0.78±0	-	-
DGCCA	-	-	0.94±0	-	-
DCCAE	**0.95±0**	0.94±0	0.93±0	0.88±0.01	0.33±0.16
Lyu et al. (2021) $\beta = 0$	0.94±0	0.93±0	0.90±0.05	0.87±0.02	0.48±0.12
Lyu et al. (2021) $\beta = 0.001$	0.04±0.01	0.06±0.02	0.06±0.01	0.05±0.01	0.05±0.02
Proposed	**0.95±0**	**0.95±0**	**0.95±0.01**	0.94±0	0.81±0.26

TABLE X
CLASSIFICATION ACCURACY (CLA-ACC) FOR THE MULTIVIEW MNIST ACROSS DIFFERENT VALUES OF THE λ PARAMETER.

	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$
CCA	-	-	0.91±0	-	-
DGCCA	-	-	0.97±0	-	-
DCCAE	**0.98±0**	0.97±0	0.97±0	0.95±0.01	0.66±0.12
Lyu et al. (2021) $\beta = 0$	0.97±0	0.97±0	0.96±0.01	0.95±0.01	0.82±0.05
Lyu et al. (2021) $\beta = 0.001$	0.38±0.05	0.39±0.05	0.44±0.03	0.41±0.03	0.39±0.02
Proposed	**0.98±0**	**0.98±0**	0.97±0	0.97±0	0.94±0.07

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3333212

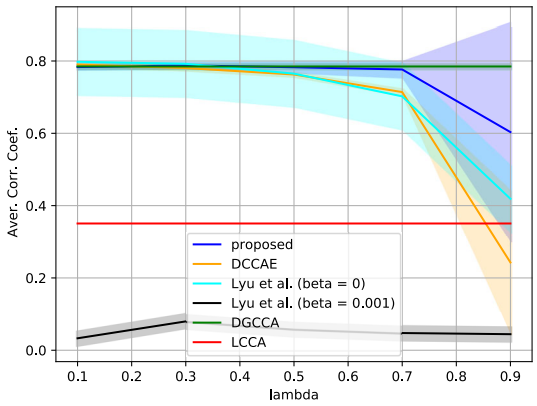

Fig. 6. Average Correlation Coefficient for the Multiview dataset across different values of the λ parameter.

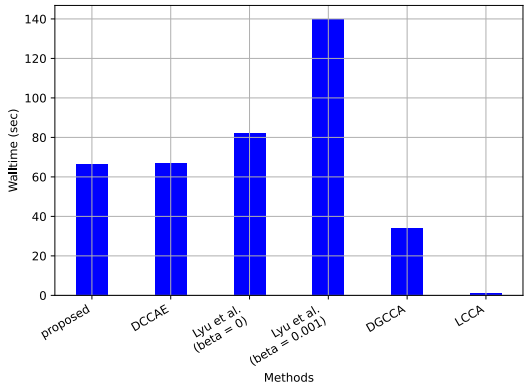

Fig. 7. Average walltimes of all the considered methods for the Multiview MNIST dataset.

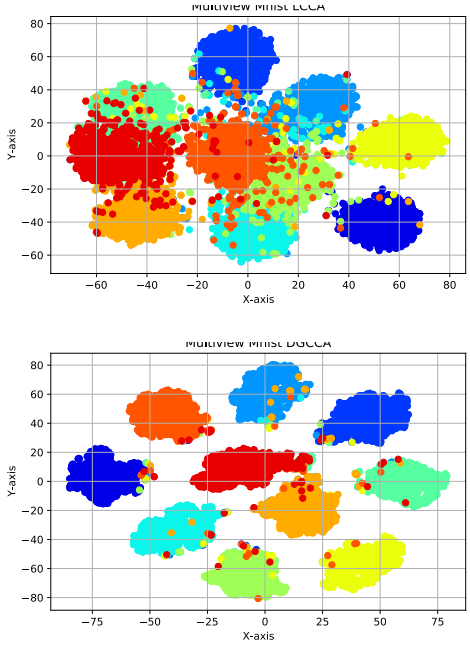

Fig. 8. t-SNE visualizations of the embeddings of the Multiview MNIST dataset for the CCA (top) and the DGCCA (bottom) methods.

efficient, consistent, and also enables the estimation of $\mathbf{g}$ even from a single view. The first formulation of DGCCA falls along those lines, since it aims at finding single view-based estimators of $\mathbf{g}$ that are in agreement, and hence potentially suboptimal but consistent with one another.

Proof of Theorem 1

Theorem 1. *Under the proposed generative model in (10), consider a solution of the optimization problem in (14). If the following additional assumptions,*

(i) *functions $\boldsymbol{v}_k$ are also partially invertible w.r.t. $\mathbf{g}$, i.e., there exist functions $\boldsymbol{u}_k : \mathbb{R}^{D_k} \to \mathbb{R}^F$, such that $\boldsymbol{u}_k\left(\mathbf{x}^{(k)}\right) = \mathbf{g}$, for all $\mathbf{x}^{(k)}$ and $k \in [K]$,*

(ii) *there exists mean dependence between at least one pair of random vectors $\mathbf{g}$ and $\mathbf{x}^{(k')}$, i.e.*

$$\mathbb{E}\left[\mathbf{x}^{(k')}|\mathbf{g}\right] \neq \mathbb{E}\left[\mathbf{x}^{(k')}\right] \text{ for a } k' \in [K] \text{ and all } \mathbf{g} \in \mathbb{R}^F,$$

hold, then the learned encodings $\boldsymbol{f}_k\left(\mathbf{x}^{(k)}\right)$ have to be non trivial functions only of $\mathbf{g}$, i.e. $\boldsymbol{f}_k\left(\boldsymbol{v}_k\left(\begin{bmatrix}\mathbf{g}\\ \mathbf{c}^{(k)}\end{bmatrix}\right)\right) = \boldsymbol{\gamma}(\mathbf{g})$, where $\boldsymbol{\gamma} : \mathbb{R}^F \to \mathbb{R}^F$. Moreover, if

(iii) *the conditional expectation, $\mathbb{E}\left[\left[\mathbf{x}^{(1)^T}, \ldots, \mathbf{x}^{(K)^T}\right]^T |\mathbf{g}\right]$, is an invertible function of $\mathbf{g}$,*

then function $\boldsymbol{\gamma}$ is also invertible and the latent common random vector $\mathbf{g}$ is identifiable up to invertible nonlinearities.

Proof. We begin by considering the function

$$\mathcal{R}^{(K)}\left(\boldsymbol{f}_1, ..., \boldsymbol{f}_K, \hat{\mathbf{g}}\right) := \sum_{k=1}^{K} \mathbb{E}\left[\left\|\boldsymbol{f}_k\left(\mathbf{x}^{(k)}\right) - \hat{\mathbf{g}}\right\|^2\right].$$

Because of the generative model in (10) and additional assumption (i), we have that the level set

$$\mathcal{S}_0 := \left\{\left(\{\boldsymbol{f}_k\}_{k=1}^K, \hat{\mathbf{g}}\right) : \mathcal{R}^{(K)}\left(\boldsymbol{f}_1, ..., \boldsymbol{f}_K, \hat{\mathbf{g}}\right) = 0\right\} \quad (21)$$

is non empty, since $\left(\{\boldsymbol{u}_k\}_{k=1}^K, \mathbf{g}\right) \in \mathcal{S}_0$. Moreover, notice that for any function $\boldsymbol{\gamma} : \mathbb{R}^F \to \mathbb{R}^F$, also $\left(\{\boldsymbol{\gamma} \circ \boldsymbol{u}_k\}_{k=1}^K, \boldsymbol{\gamma}(\mathbf{g})\right) \in \mathcal{S}_0$. Now, let $\left(\{\bar{\boldsymbol{f}}_k\}_{k=1}^K, \bar{\mathbf{g}}\right)$ be an element of $\mathcal{S}_0$ and functions $\bar{\boldsymbol{h}}_k : \mathbb{R}^{F+L_k} \to \mathbb{R}^F$ be defined according to

$$\bar{\boldsymbol{h}}_k\left(\begin{bmatrix}\mathbf{g}\\ \mathbf{c}^{(k)}\end{bmatrix}\right) := \bar{\boldsymbol{f}}_k\left(\boldsymbol{v}_k\left(\begin{bmatrix}\mathbf{g}\\ \mathbf{c}^{(k)}\end{bmatrix}\right)\right).$$

For any $k_1, k_2 \in [K]$, with $k_1 \neq k_2$, we have that for any realization of $\mathbf{g}$, $\mathbf{g}' \in \mathbb{R}^F$, it holds that

$$\bar{\boldsymbol{h}}_{k_1}\left(\begin{bmatrix}\mathbf{g}'\\ \mathbf{c}^{(k_1)}\end{bmatrix}\right) = \bar{\boldsymbol{h}}_{k_2}\left(\begin{bmatrix}\mathbf{g}'\\ \mathbf{c}^{(k_2)}\end{bmatrix}\right),$$

for all $\mathbf{c}^{(k_1)} \in \mathbb{R}^{L_{k_1}}$ and $\mathbf{c}^{(k_2)} \in \mathbb{R}^{L_{k_2}}$. Hence functions $\{\bar{\boldsymbol{h}}_k\}_{k=1}^K$ have to be functions only of $\mathbf{g}$, since, for any fixed $\mathbf{g}'$, the equalities hold no matter what the values of $\mathbf{c}^{(k_1)}$ and $\mathbf{c}^{(k_2)}$ are. In other words, there exists a function $\bar{\boldsymbol{\gamma}} : \mathbf{R}^F \to \mathbf{R}^F$, such that

$$\bar{\boldsymbol{h}}_k\left(\begin{bmatrix}\mathbf{g}'\\ \mathbf{c}^{(k)}\end{bmatrix}\right) = \bar{\boldsymbol{\gamma}}(\mathbf{g}'), \text{ for all } \mathbf{g}' \in \mathbb{R}^F \text{ and } k \in [K].$$

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3333212

Therefore, $\bar{\mathbf{g}}$ must be equal to $\bar{\boldsymbol{\gamma}}(\mathbf{g})$. Since the above arguments hold for any feasible $\bar{\mathbf{g}}$, it follows that the elements of $\mathcal{S}_0$ are given by $\left(\{\boldsymbol{\gamma}\circ\boldsymbol{u}_k\}_{k=1}^K, \boldsymbol{\gamma}(\mathbf{g})\right)$, where $\boldsymbol{\gamma}$ can be any function $\mathbb{R}^F \to \mathbb{R}^F$.

Now, let us focus on the objective of (14),

$$\mathcal{B}^{(K)}(\boldsymbol{w}_1,\ldots,\boldsymbol{w}_K,\hat{\mathbf{g}}) = \sum_{k=1}^K \mathbb{E}\left[\left\|\boldsymbol{w}_k(\hat{\mathbf{g}}) - \mathbf{x}^{(k)}\right\|^2\right].$$

The only dependency between the objective and the constraints of (14) is through random variable $\hat{\mathbf{g}}$. On the other hand, as we showed above, a feasible random vector $\hat{\mathbf{g}}$ has to be a function of the common latent random vector $\mathbf{g}$. Given that, the optimization problem of (14) can be written as

$$\begin{aligned}\min_{\boldsymbol{\gamma},\{\boldsymbol{w}_k\in\mathcal{C}_w\}_{k=1}^K} & \sum_{k=1}^K \mathbb{E}\left[\left\|\boldsymbol{w}_k(\boldsymbol{\gamma}(\mathbf{g})) - \mathbf{x}^{(k)}\right\|^2\right] \\ \text{s.t.} \quad & \mathbb{E}\left[\boldsymbol{\gamma}(\mathbf{g})\boldsymbol{\gamma}(\mathbf{g})^T\right] = \mathbf{I}_F, \quad \mathbb{E}\left[\boldsymbol{\gamma}(\mathbf{g})\right] = \mathbf{0}_F\end{aligned} \tag{22}$$

As long as the distribution of $\mathbf{g}$, $\mathcal{D}_{\mathbf{g}}$, is non-singular, the existence of an invertible function $\boldsymbol{\gamma}$, under which $\mathbb{E}\left[\boldsymbol{\gamma}(\mathbf{g})\boldsymbol{\gamma}(\mathbf{g})^T\right] = \mathbf{I}_F$ and $\mathbb{E}\left[\boldsymbol{\gamma}(\mathbf{g})\right] = \mathbf{0}_F$, emerges naturally after considering whitening transformations. Therefore, the feasible sets of problems (14) and (22) are nonempty.

Given a function $\boldsymbol{\gamma}$, the conditionally optimal functions $\{\tilde{\boldsymbol{w}}_k\}_{k=1}^K$ can be obtained by $\tilde{\boldsymbol{w}}_k(\boldsymbol{\gamma}(\mathbf{g})) = \mathbb{E}\left[\mathbf{x}^{(k)}|\boldsymbol{\gamma}(\mathbf{g})\right]$. Because of our second additional assumption, optimal functions $\{\boldsymbol{w}_k\circ\boldsymbol{\gamma}\}_{k=1}^K$ have to be non trivial functions in the sense that they have to be nonconstant. To see that, notice that letting $\boldsymbol{w}_{k'}(\boldsymbol{\gamma}(\mathbf{g})) = \mathbb{E}\left[\mathbf{x}^{(k')}\right]$ would induce a strictly higher MSE than $\boldsymbol{w}_{k'}(\boldsymbol{\gamma}^*(\mathbf{g})) = \mathbb{E}\left[\mathbf{x}^{(k')}|\boldsymbol{\gamma}^*(\mathbf{g})\right] = \mathbb{E}\left[\mathbf{x}^{(k')}|\mathbf{g}\right]$, which is the best MSE predictor of $\mathbf{x}^{(k')}$ over all possible functions $\boldsymbol{\gamma}$. Non triviality of all optimal functions $\boldsymbol{w}_{k'}\circ\boldsymbol{\gamma}$ implies non triviality of all optimal functions $\boldsymbol{\gamma}:\mathbb{R}^F\to\mathbb{R}^F$, which have to satisfy the following three properties

(a) $\mathbb{E}\left[\mathbf{x}^{(k)}|\boldsymbol{\gamma}(\mathbf{g})\right] = \mathbb{E}\left[\mathbf{x}^{(k)}|\mathbf{g}\right],\ \forall k\in[K]$,
(b) $\mathbb{E}\left[\boldsymbol{\gamma}(\mathbf{g})\boldsymbol{\gamma}(\mathbf{g})^T\right] = \mathbf{I}_F$,
(c) $\mathbb{E}\left[\boldsymbol{\gamma}(\mathbf{g})\right] = \mathbf{0}_F$.

Let $\boldsymbol{\phi}:\mathbb{R}^F\to\mathbb{R}^{\sum_{k=1}^K D_k}$ be given by the conditional expectation $\boldsymbol{\phi}(\mathbf{g}) := \mathbb{E}\left[\left[\mathbf{x}^{(1)^T},\ldots,\mathbf{x}^{(K)^T}\right]^T\middle|\mathbf{g}\right]$. Under the third additional assumption, there exists a function $\boldsymbol{\psi}:\mathbb{R}^{\sum_{k=1}^K D_k}\to\mathbb{R}^F$, such that $\boldsymbol{\psi}(\boldsymbol{\phi}(\mathbf{g})) = \mathbf{g}$, for all $\mathbf{g}\in\mathbb{R}^F$. For an optimal collection of functions $\{\boldsymbol{w}_k^*\}_{k=1}^K$ and $\boldsymbol{\gamma}^*$ c.f. (14), let $\boldsymbol{\phi}^*(\boldsymbol{\gamma}^*(\mathbf{g})) := \left[\boldsymbol{w}_1^*(\boldsymbol{\gamma}^*(\mathbf{g}))^T,\ldots,\boldsymbol{w}_K^*(\boldsymbol{\gamma}^*(\mathbf{g}))^T\right]^T$. Because of property (a), we have that

$$\boldsymbol{\phi}(\mathbf{g}) = \boldsymbol{\phi}^*(\boldsymbol{\gamma}^*(\mathbf{g})), \text{ for all } \mathbf{g}\in\mathbb{R}^F, \tag{23}$$

which implies that an optimal function $\boldsymbol{\gamma}^*$ has an inverse function, given by $\boldsymbol{\gamma}^{*^{-1}} = \boldsymbol{\psi}\circ\boldsymbol{\phi}^*$, since

$$\mathbf{g} = \boldsymbol{\psi}(\boldsymbol{\phi}^*(\boldsymbol{\gamma}^*(\mathbf{g}))), \text{ for all } \mathbf{g}\in\mathbb{R}^F. \tag{24}$$

As a result, under the additional assumption (iii), the latent common random vector is identifiable up to invertible nonlinearities expressed through $\boldsymbol{\gamma}^*$.

□

REFERENCES

[1] S. M. Kakade and D. P. Foster, "Multi-view regression via canonical correlation analysis," in *International Conference on Computational Learning Theory*. Springer, 2007, pp. 82–96.

[2] D. P. Foster, S. M. Kakade, and T. Zhang, "Multi-view dimensionality reduction via canonical correlation analysis," 2008.

[3] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 129–136.

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

[5] R. Arora and K. Livescu, "Kernel cca for multi-view learning of acoustic features using articulatory measurements," in *Symposium on machine learning in speech and language processing*, 2012.

[6] W. Wang, R. Arora, and K. Livescu, "Reconstruction of articulatory measurements with smoothed low-rank matrix completion," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 54–59.

[7] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "Unsupervised learning of acoustic features via deep canonical correlation analysis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4590–4594.

[8] M. S. Ibrahim and N. D. Sidiropoulos, "Cell-edge interferometry: Reliable detection of unknown cell-edge users via canonical correlation analysis," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2019, pp. 1–5.

[9] M. S. Ibrahim, P. A. Karakasis, and N. D. Sidiropoulos, "A simple and practical underlay scheme for short-range secondary communication," *IEEE Transactions on Wireless Communications*, 2022.

[10] Y.-O. Li, T. Adali, W. Wang, and V. D. Calhoun, "Joint blind source separation by multiset canonical correlation analysis," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3918–3929, 2009.

[11] N. M. Correa, T. Adali, Y.-O. Li, and V. D. Calhoun, "Canonical correlation analysis for data fusion and group inferences," *IEEE signal processing magazine*, vol. 27, no. 4, pp. 39–50, 2010.

[12] J. R. Katthi and S. Ganapathy, "Deep correlation analysis for audio-eeg decoding," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2742–2753, 2021.

[13] P. A. Karakasis, A. P. Liavas, N. D. Sidiropoulos, P. G. Simos, and E. Papadaki, "Multi-subject task-related fmri data processing via a two-stage generalized canonical correlation analysis," *IEEE Transactions on Image Processing*, 2022.

[14] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *International Journal of Neural Systems*, vol. 10, no. 05, pp. 365–377, 2000.

[15] W. Wang, X. Yan, H. Lee, and K. Livescu, "Deep variational canonical correlation analysis," *arXiv preprint arXiv:1610.03454*, 2016.

[16] T. Michaeli, W. Wang, and K. Livescu, "Nonparametric canonical correlation analysis," in *International conference on machine learning*. PMLR, 2016, pp. 1967–1976.

[17] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*. PMLR, 2013, pp. 1247–1255.

[18] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International conference on machine learning*. PMLR, 2015, pp. 1083–1092.

[19] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, and R. Arora, "Deep generalized canonical correlation analysis," *arXiv preprint arXiv:1702.02519*, 2017.

[20] Q. Lyu and X. Fu, "Nonlinear multiview analysis: Identifiability and neural network-assisted implementation," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2697–2712, 2020.

[21] Q. Lyu, X. Fu, W. Wang, and S. Lu, "Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective," in *International Conference on Learning Representations*, 2021.

[22] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[23] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 310–12 320.

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3333212

[24] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 750–15 758.

[25] P. Horst, "Generalized canonical correlations and their applications to experimental data," *Journal of Clinical Psychology*, vol. 17, no. 4, pp. 331–347, 1961.

[26] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[27] N. A. Asendorf, "Informative data fusion: Beyond canonical correlation analysis," Ph.D. dissertation, 2015.

[28] M. Sørensen, C. I. Kanatsoulis, and N. D. Sidiropoulos, "Generalized canonical correlation analysis: A subspace intersection approach," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2452–2467, 2021.

[29] M. Borga, "Canonical correlation: a tutorial," *On line tutorial http://people. imt. liu. se/magnus/cca*, vol. 4, no. 5, 2001.

[30] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." *Journal of machine learning research*, vol. 11, no. 12, 2010.

[31] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 1999.

[32] A. Hyvarinen, H. Sasaki, and R. Turner, "Nonlinear ica using auxiliary variables and generalized contrastive learning," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 859–868.

[33] J. Von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello, "Self-supervised learning with data augmentations provably isolates content from style," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[34] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *Advances in neural information processing systems*, vol. 31, 2018.

[35] P. H. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[37] K. Y. Yeung and W. L. Ruzzo, "Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

[38] J. R. Westbury, G. Turner, and J. Dembowski, "X-ray microbeam speech production database user's handbook," *University of Wisconsin*, 1994.

[39] R. Arora and K. Livescu, "Multi-view learning with supervision for transformed bottleneck features," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2499–2503.

[40] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[41] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3333212

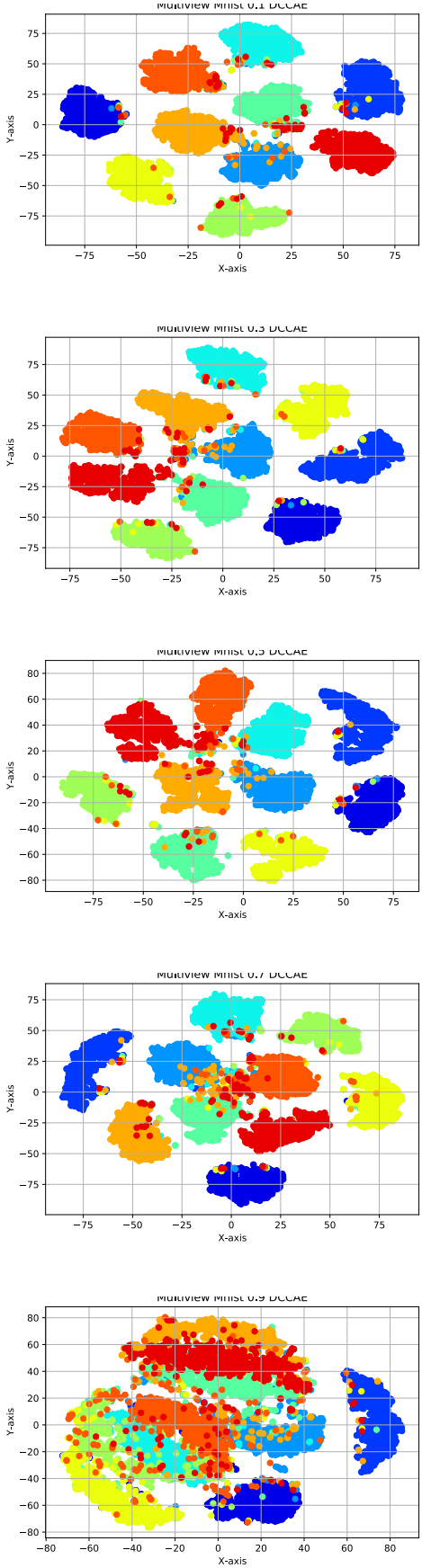

Fig. 9. t-SNE visualizations of the embeddings of the Multiview MNIST dataset for the DCCAE method for different values of λ ($\lambda = 0.1$-top to $\lambda = 0.9$-bottom).

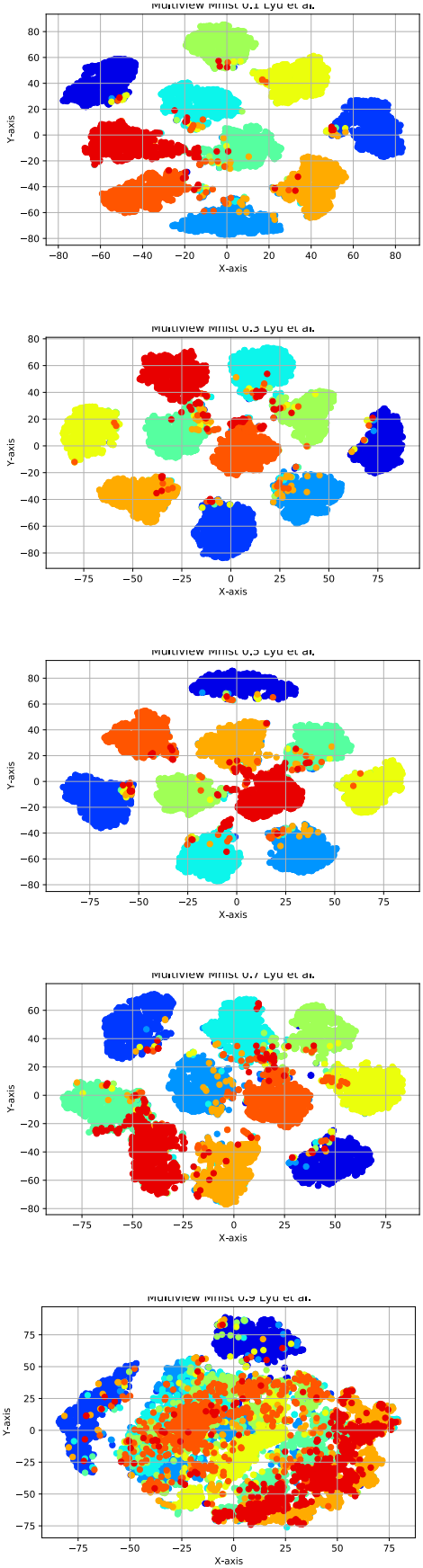

Fig. 10. t-SNE visualizations of the embeddings of the Multiview MNIST dataset for the method of [21] for $\beta = 0$ and different values of λ ($\lambda = 0.1$-top to $\lambda = 0.9$-bottom).

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3333212

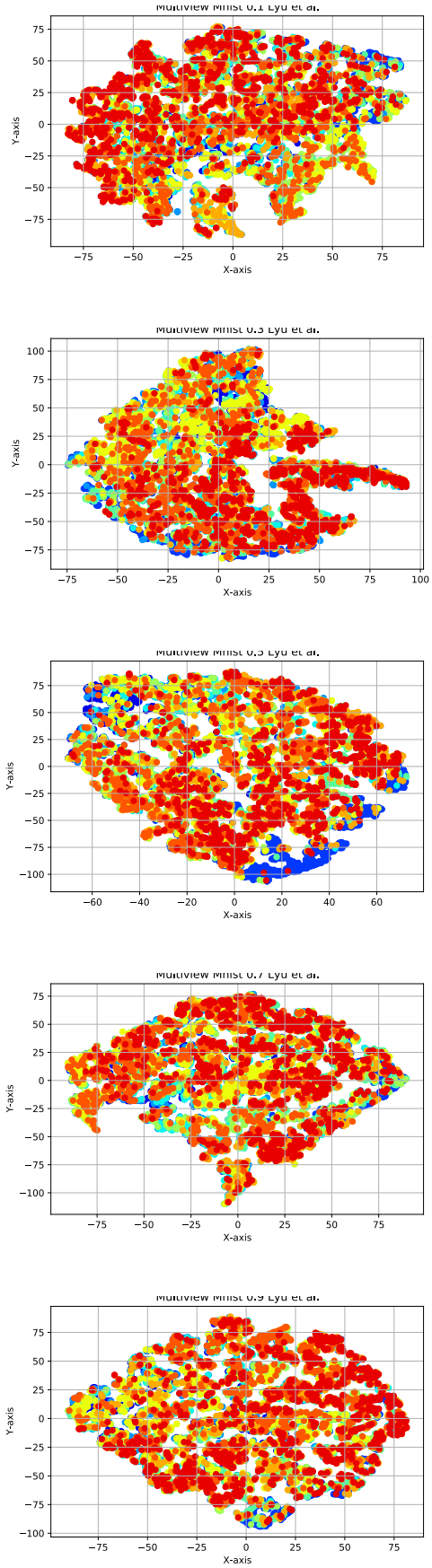

Fig. 11. t-SNE visualizations of the embeddings of the Multiview MNIST dataset for the method of [21] for $\beta = 10^{-3}$ and different values of λ ($\lambda = 0.1$-top to $\lambda = 0.9$-bottom).

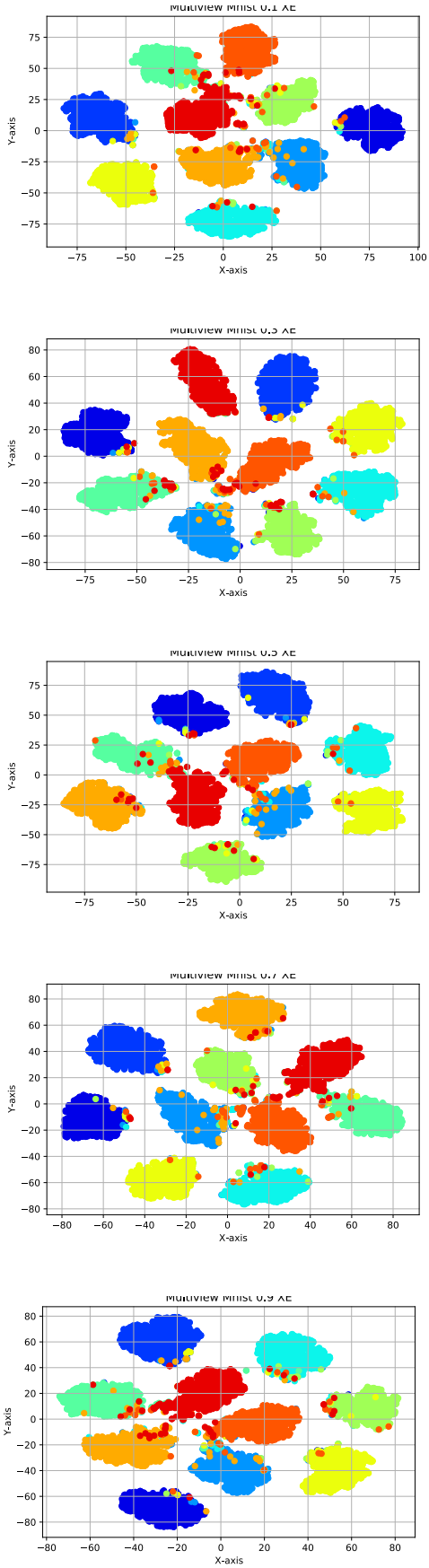

Fig. 12. t-SNE visualizations of the embeddings of the Multiview MNIST dataset for the proposed method for different values of λ ($\lambda = 0.1$-top to $\lambda = 0.9$-bottom).

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.