2023 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 17–20, 2023, ROME, ITALY

GEODESIC-BASED RELAXATION FOR DEEP CANONICAL CORRELATION ANALYSIS

Maurice Kuschel[⋆] *Timothy Marrinan*[†] *Tanuj Hasija*[⋆]

[⋆]Signal and System Theory Group, Paderborn University
[†] Pacific Northwest National Lab, Seattle

ABSTRACT

Deep canonical correlation analysis (DCCA) is often applied to paired data samples from diverse sources to extract meaningful common information. However, when the data sources are heterogeneous, some of the useful information may be complementary but not exactly common. In spite of this fact, existing techniques learn maximally correlated representations from multiple views and are formulated so that they aim to yield identical latent subspaces for each view. This approach is sub-optimal in estimating the true signal subspaces for heterogeneous data sources. We propose a residual relaxation for deep canonical correlation analysis (RDCCA) based on a subspace distance metric, which generalizes the existing problem formulation and extracts representations that are better estimates of the actual, non-identical subspaces. We demonstrate that when using such a relaxation, the learned representations are closer to the true ones and that RDCCA outperforms CCA and DCCA in scenarios with heterogeneous data.

Index Terms— Deep canonical correlation analysis, Residual relaxation, Grassmann manifolds, Unsupervised learning, Multiview representation learning

1. INTRODUCTION

Multiview representation learning (MRL) is concerned with learning meaningful, compact representations from high-dimensional multiview data. The idea is that multiple views observing the same object or phenomenon contain shared information. We can then extract this information by analyzing those views jointly [1]. In the example of a video, we can understand the observed scene better when interpreting audio and visual data at the same time compared to considering either modality alone. With more and more multiview data being available through smart sensors and IoT devices, MRL strategies attract increasing attention.

Canonical correlation analysis (CCA) and its extensions [2, 3] are commonly used for unsupervised MRL problems and have been shown to perform well in many applications [4, 5]. One reason for the popularity of CCA is the lightweight and, at the same time, capable algorithm. It can extract the set of correlated representations from two views. At the same time, it neglects private information, which is present in only one of the views. However, the method has one significant limitation: it can only learn linear mappings and is, therefore, ineffective in settings where nonlinear relationships exist between the views [6].

Using neural networks is a widespread practice to generalize CCA for nonlinear relationships, yielding deep CCA (DCCA) [7]. This data-driven approach has the advantage of being scalable and not restricted to a particular class of nonlinearities [8]. Depending on the architecture and size of the neural networks, we can use them to learn highly complex nonlinearities.

The typical assumption in CCA (and subsequently DCCA) is that the correlation between the underlying signals is unknown and that identifying representations that are maximally correlated will reveal the true latent signals [3]. This assumption is reasonable when the data generation processes are homogeneous. For example, if samples from the two views are photographs taken by identical cameras with slightly different vantage points of the same scene, we would expect their images to be highly correlated. However, the assumption may not hold, for example, if the view of one camera is partly obstructed or if the views represent different modalities [9]. In the presence of heterogeneous data or structured interference, we expect the correlation between latent representations of these two 'views' to be less than one. Due to the expressive power of neural networks as universal function approximators [10], it is possible for DCCA to learn latent representations that are more highly correlated than the true signals. This type of overfitting leads to poor estimation of the signals and can be a barrier to any downstream tasks for which the latent representations are used.

Many previous attempts to improve the estimation of signal subspaces for heterogeneous data have tried to indirectly prevent over-correlated representations by encouraging neural networks to learn simple functions. This has been done by including regularization terms that limit network capacity [11], by using autoencoders to enforce invertibility [8], or by including independence constraints on the latent representations [12]. Wang *et al.* tried to address the problem of over-correlation with a residual relaxation technique for data

979-8-3503-2411-2/23/$31.00

augmentation [13]. However, this technique focuses on individual samples and essentially learns a fuzzy mapping to within a neighborhood of the latent representation of the original unaugmented training sample.

In this work, we directly address the problem of learning over-correlated signal subspaces at its core. We modify the optimization problem, such that the partly-correlated subspaces are the target of the optimization, instead of just restricting the method implicitly and hindering the optimizer to reach perfectly-correlated subspaces. We do this by proposing a generalization of the DCCA objective function that allows for a nonzero residual between the learned subspaces. The residual relaxation accounts for the possibility of intrinsic differences between the signal subspaces associated with different views. This difference is measured using a subspace distance known as the chordal distance, and tools from Riemannian geometry allow us to constrain the learned representations to be correlated but not identical. We show that the residual relaxation leads to better subspace estimation and improved downstream clustering accuracy than CCA or DCCA.

2. BACKGROUND AND RELATED WORK

Let $\mathbf{x}_m^{(q)} \in \mathbb{R}^N$ be an observation from View $q = 1, 2$, where N denotes the dimensions of the observations. The M paired observations from each view form the columns of data matrices, $\mathbf{X}^{(q)} = [\mathbf{x}_1^{(q)}, \ldots, \mathbf{x}_M^{(q)}]$.

2.1. Canonical correlation analysis (CCA)

Given data matrices $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, the goal of CCA is to learn linear mappings, $\mathbf{A}^{(1)}, \mathbf{A}^{(2)} \in \mathbb{R}^{N \times N}$, such that the mapped representations, $\mathbf{Z}^{(1)} = \mathbf{A}^{(1)}\mathbf{X}^{(1)}$ and $\mathbf{Z}^{(2)} = \mathbf{A}^{(2)}\mathbf{X}^{(2)}$ are maximally correlated. The whole process happens in a completely unsupervised manner. The CCA solution is obtained by solving the following optimization problem,

$$\underset{\mathbf{A}^{(1)},\mathbf{A}^{(2)}}{\arg\max} \quad \text{trace}\left(\frac{1}{M}\mathbf{A}^{(1)}\mathbf{X}^{(1)}\mathbf{X}^{(2)\top}\mathbf{A}^{(2)\top}\right) \tag{1}$$

$$\text{such that} \quad \frac{1}{M}\mathbf{A}^{(q)}\mathbf{X}^{(q)}\mathbf{X}^{(q)\top}\mathbf{A}^{(q)\top} = \mathbf{I}, \tag{2}$$

$$\frac{1}{M}\mathbf{A}^{(1)}\mathbf{X}^{(1)}\mathbf{X}^{(2)\top}\mathbf{A}^{(2)\top} \text{ is diagonal} \tag{3}$$

for $q = 1, 2$. It is well-known that we can solve this problem algebraically via the singular value decomposition (SVD) [2]. Consider the matrix

$$\mathbf{C}^{(1,2)} = \left(\mathbf{X}^{(1)}\mathbf{X}^{(1)\top}\right)^{-\frac{1}{2}}\mathbf{X}^{(1)}\mathbf{X}^{(2)\top}\left(\mathbf{X}^{(2)}\mathbf{X}^{(2)\top}\right)^{-\frac{1}{2}},$$

and let $\mathbf{PDQ}^\top = \mathbf{C}^{(1,2)}$ be the associated SVD. The solution to (1) can then be written as

$$\mathbf{A}^{(1)} = \mathbf{P}^\top\left(\frac{1}{M}\mathbf{X}^{(1)}\mathbf{X}^{(1)\top}\right)^{-\frac{1}{2}},$$

$$\mathbf{A}^{(2)} = \mathbf{Q}^\top\left(\frac{1}{M}\mathbf{X}^{(2)}\mathbf{X}^{(2)\top}\right)^{-\frac{1}{2}}.$$

2.2. Deep CCA (DCCA)

In order to extract nonlinear relationships between $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, DCCA extends the correlation maximization framework from CCA by additionally learning two nonlinear mappings, $\mathbf{f}^{(1)} : \mathbb{R}^N \to \mathbb{R}^K$ and $\mathbf{f}^{(2)} : \mathbb{R}^N \to \mathbb{R}^K$, such that $\mathbf{Z}^{(1)} = \mathbf{A}^{(1)}\mathbf{f}^{(1)}(\mathbf{X}^{(1)})$, $\mathbf{Z}^{(2)} = \mathbf{A}^{(2)}\mathbf{f}^{(2)}(\mathbf{X}^{(2)})$ are maximally correlated, now with $\mathbf{A}^{(1)}, \mathbf{A}^{(2)} \in \mathbb{R}^{K \times K}$. DCCA then solves the problem

$$\underset{\mathbf{A}^{(q)},\mathbf{f}^{(q)}}{\arg\max} \quad \text{trace}\left(\frac{1}{M}\mathbf{A}^{(1)}\mathbf{f}^{(1)}(\mathbf{X}^{(1)})\mathbf{f}^{(2)}(\mathbf{X}^{(2)})^\top\mathbf{A}^{(2)\top}\right) \tag{4}$$

$$\text{such that} \quad \frac{1}{M}\mathbf{A}^{(q)}\mathbf{f}^{(q)}(\mathbf{X}^{(q)})\mathbf{f}^{(q)}(\mathbf{X}^{(q)})^\top\mathbf{A}^{(q)\top} = \mathbf{I}, \tag{5}$$

$$\frac{1}{M}\mathbf{A}^{(1)}\mathbf{f}^{(1)}(\mathbf{X}^{(1)})\mathbf{f}^{(2)}(\mathbf{X}^{(2)})^\top\mathbf{A}^{(2)\top} \text{ is diagonal} \tag{6}$$

for $q = 1, 2$. In DCCA, we use two separate neural networks to learn $\mathbf{f}^{(1)}$ and $\mathbf{f}^{(2)}$ [7]. However, due to the non-trivial learning of the parameters of $\mathbf{f}^{(1)}$ and $\mathbf{f}^{(2)}$, the solution to (4) cannot be obtained in a closed form like that of CCA. Instead Problem (4) is typically solved iteratively using either full-batch optimization algorithms or stochastic gradient descent with mini-batches.

2.3. Reformulation of DCCA optimization

At optimality, the representations which maximize correlation are the same as those which minimize the Euclidean distance between the projections $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$. Therefore, Problem (4) can be reformulated according to [14] as

$$\underset{\mathbf{A}^{(q)},\mathbf{f}^{(q)}}{\arg\min} \quad \left\|\mathbf{A}^{(1)}\mathbf{f}^{(1)}(\mathbf{X}^{(1)}) - \mathbf{A}^{(2)}\mathbf{f}^{(2)}(\mathbf{X}^{(2)})\right\|_2^2 \tag{7}$$

$$\text{such that} \quad (5) \text{ and } (6),$$

which can be equivalently written by introducing a slack variable, $\mathbf{U}$, as:

$$\underset{\mathbf{A}^{(q)},\mathbf{f}^{(q)},\mathbf{U}}{\arg\min} \quad \sum_{q=1}^{2}\left\|\mathbf{U} - \mathbf{A}^{(q)}\mathbf{f}^{(q)}(\mathbf{X}^{(q)})\right\|_2^2 \tag{8}$$

$$\text{such that} \quad \frac{1}{M}\mathbf{U}\mathbf{U}^\top = \mathbf{I}, \tag{9}$$

$$\frac{1}{M}\mathbf{U}\mathbf{1} = \mathbf{0}, \tag{10}$$

where $\mathbf{1}$ denotes an all-one vector. The slack variable, $\mathbf{U}$, represents the extracted shared components and, ideally, is equal to the truly shared components. This distance-based reformulation of the correlation maximization problem has the advantage of being solvable by a lightweight algorithm [12]. In this reformulation, (9) is analogous to (5), and (10) ensures that the extracted latent components are zero-mean.

3. DCCA OPTIMIZATION PROBLEM WITH RESIDUAL

The optimization problems presented above are designed so that the projections $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ are maximally correlated. However, for heterogeneous data which is not perfectly correlated, the solution to Problem (8) may not approximate the true signal subspaces, as the maximally correlated subspaces are not good estimates of the true, partly-correlated subspaces. To address this misalignment, we generalize the DCCA optimization problem so that the correlation between the optimal subspaces is controlled by a parameter, which is equivalent to learning subspaces that are a fixed distance apart. Let $\mathbf{G}$ and $\mathbf{H}$ be $K \times M$ matrices that contain samples from two sets of zero-mean, unit variance random vectors. Let $\boldsymbol{c} = [c_1, c_2, \ldots, c_K]^\top$ be the vector of canonical correlations between $\mathbf{G}$ and $\mathbf{H}$. We denote by $d_{\text{chord}}(\mathbf{G}, \mathbf{H}) = \sqrt{\frac{1}{K}\sum_{k=1}^{K} 1 - c_k^2}$, the normalized chordal distance between the rowspaces of $\mathbf{G}$ and $\mathbf{H}$.

To prevent the learned subspaces from being overly correlated, the single slack variable in Equation (8) is replaced by view-specific slack variables, $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$, which are constrained to be separated by a given chordal distance. Then minimizing the subspace distance between $\mathbf{A}^{(q)}\mathbf{f}^{(q)}(\mathbf{X}^{(q)})$ and $\mathbf{U}^{(q)}$ for each q, while maintaining a fixed distances between $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$, avoids over-correlation between $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$. Consider the following residual relaxation of Problem (8) with two slack variables:

$$\underset{\mathbf{A}^{(q)},\mathbf{f}^{(q)},\mathbf{U}^{(q)}}{\arg\min} \quad \sum_{q=1}^{2} d_{\text{chord}}(\mathbf{U}^{(q)}, \mathbf{A}^{(q)}\mathbf{f}^{(q)}(\mathbf{X}^{(q)})) \tag{11}$$

$$\text{such that} \quad d_{\text{chord}}(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = r \tag{12}$$

$$d_{\text{chord}}(\mathbf{U}, \mathbf{U}^{(1)}) = d_{\text{chord}}(\mathbf{U}, \mathbf{U}^{(2)}) \tag{13}$$

$$\mathbf{U} = \underset{\mathbf{U}}{\arg\min} \sum_{q=1}^{2} d_{\text{chord}}(\mathbf{U}, \mathbf{A}^{(q)}\mathbf{f}^{(q)}(\mathbf{X}^{(q)})) \tag{14}$$

$$\frac{1}{M}\mathbf{U}\mathbf{U}^\top = \mathbf{I}, \quad \frac{1}{M}\mathbf{U}^{(q)}\mathbf{U}^{(q)\top} = \mathbf{I}, \tag{15}$$

$$\frac{1}{M}\mathbf{U}\mathbf{1} = \mathbf{0}, \quad \frac{1}{M}\mathbf{U}^{(q)}\mathbf{1} = \mathbf{0} \tag{16}$$

for $q = 1, 2$.

Finding transformations that minimize the chordal distance between two sets of data is equivalent to finding transformations that maximize the correlation between the resulting representations [15, 16]. Thus, the major difference between Problem (8) and Problem (11) is the inclusion of view-specific slack variables and the constraint in Equation (12), which requires these subspaces to remain a fixed distance apart. Constraint (13) implies that $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ must also be equidistant from $\mathbf{U}$. The question then becomes, how can we enforce this constraint? The answer requires tools from the Grassmann manifold, whose points can be parameterized subspaces of a fixed dimension. Since the chordal distance is invariant to the choice of basis for the row spaces of its arguments, we can identify $\mathbf{U}, \mathbf{U}^{(q)}$, and $\mathbf{A}^{(q)}\mathbf{f}^{(q)}(\mathbf{X}^{(q)})$ with the linear subspaces spanned by their rows.

4. SUBSPACE DISTANCE REGULARIZATION

For fixed values of $\mathbf{A}^{(q)}$ and $\mathbf{f}^{(q)}$ for $q = 1, 2$, we can compute the slack variable, $\mathbf{U}$, that solves (14). In order to define each view-specific slack variable, $\mathbf{U}^{(q)}$, we need to find a matrix whose row space lies on the shortest path between $\mathbf{U}$ and $\mathbf{A}^{(q)}\mathbf{f}^{(q)}(\mathbf{X}^{(q)})$, with the chordal distance of $\mathbf{U}^{(1)}$ from $\mathbf{U}^{(2)}$ being r and both $\mathbf{U}^{(q)}$ having the same chordal distance from $\mathbf{U}$. This path of shortest distance is referred to as a geodesic, and points along a geodesic can be computed on a Grassmann manifold using the exponential and logarithmic maps.

4.1. Geodesic on Grassmann manifolds

Since we are interested in the rank-K subspaces spanned by the signals, it is natural to represent these subspaces as points on a Grassmann manifold, that is, the set of all K-dimensional subspaces of $\mathbb{R}^M$, denoted $\text{Gr}(K, M)$. Let $\mathbf{G} \in \text{Gr}(K, M)$ be represented by a $K \times M$ slice of an orthonormal matrix whose row space spans $\mathbf{G}$. Given a tangent vector, $\boldsymbol{Y} \in \text{T}_{\mathbf{G}} \subseteq \mathbb{R}^{K \times M}$, and a scale parameter $t \geq 0$, we can compute a geodesic retraction in the direction of $\boldsymbol{Y}$ using the exponential map, $\underset{\mathbf{G}}{\text{Exp}} : \text{T}_{\mathbf{G}} \to \text{Gr}(K, M)$ defined as

$$\underset{\mathbf{G}}{\text{Exp}}(\boldsymbol{Y}, t) = \mathbf{P}\cos(t\mathbf{D})\mathbf{P}^\top\mathbf{G} + \mathbf{P}\sin(t\mathbf{D})\mathbf{Q}^\top, \tag{17}$$

where $\mathbf{P}\mathbf{D}\mathbf{Q}^\top$ is the 'econ' SVD of $\boldsymbol{Y}$. Furthermore, for $\mathbf{H} \in \text{Gr}(K, M)$ within a neighborhood of $\mathbf{G}$, we can compute a vector in the tangent space of $\mathbf{G}$ in the direction of $\mathbf{H}$ using the logarithmic map, $\text{Log}_{\mathbf{G}} : \text{Gr}(K, M) \to \text{T}_{\mathbf{G}}$, which is defined by

$$\underset{\mathbf{G}}{\text{Log}}(\mathbf{H}) = \mathbf{P}\mathbf{E}\mathbf{Q}^\top \tag{18}$$

where $\mathbf{P}\mathbf{D}\mathbf{Q}^\top$ is the 'econ' SVD of $(\mathbf{H}\mathbf{G}^\top)^{-1}\mathbf{H}(\mathbf{I} - \mathbf{G}^\top\mathbf{G})$ and $\mathbf{E} = \arctan(\mathbf{D})$. If $t \in [0, 1]$ and $\boldsymbol{Y} = \text{Log}_{\mathbf{G}}(\mathbf{H})$, the point $\text{Exp}_{\mathbf{G}}(\boldsymbol{Y}, t)$ will be a point on the geodesic between $\mathbf{G}$ and $\mathbf{H}$. Further details on Grassmannian geometry and optimization can be found in the following references [17, 18][1].

4.2. Proposed RDCCA Algorithm

We propose a block coordinate descent (BCD)-based algorithm to solve the optimization problem in (11). Algorithm 1 consists of two loops, iteratively solving two objective functions sequentially. In the outer loop, the view-specific slack variables $\mathbf{U}_\tau^{(q)}$ are computed for timestep τ. This

[1]The notation in this paper reflects the fact that our signal subspaces are represented by the row spaces of the matrices, which is nonstandard in Riemannian geometry literature.

is done by computing the current learned representations $\mathbf{Z}_\tau^{(q)} = \mathbf{A}_\tau^{(q)}\mathbf{f}_\tau^{(q)}(\mathbf{X}^{(q)})$, orthonormalizing them, yielding $\mathbf{Z}_{\tau,\text{orth}}^{(q)}$, and computing $\mathbf{U}_\tau$ by solving problem (14) via the SVD in (line 6). We then compute $\mathbf{U}_\tau^{(q)}$ by moving $\mathbf{U}_\tau$ along the geodesic to $\mathbf{Z}_{\tau,\text{orth}}^{(q)}$ with the exponential and logarithmic maps, using (17) and (18). The scale parameter, t, is computed via line search such that the resulting chordal distance between $\mathbf{U}_\tau^{(1)}$ and $\mathbf{U}_\tau^{(2)}$ equals the desired residual (line 7).

In the inner loop, we compute the chordal distance between the orthonormal projections and the slack variables, $d_{\text{chord}}(\mathbf{U}_\tau^{(q)}, \mathbf{Z}_{\tau,s,\text{orth}}^{(q)})$, to update the linear and nonlinear mappings $\mathbf{f}_{\tau,s+1}^{(q)}$ and $\mathbf{A}_{\tau,s+1}^{(q)}$ via gradient backpropagation (line 14 & 15). The representations have to be orthonormal to be on the same manifold as $\mathbf{U}_\tau^{(q)}$. Therefore, we always do an orthonormalization of the projections via the Procrustes method (line 12).

Algorithm 1 RDCCA

Require: Data $\mathbf{X}^{(q)}$, K and target residual r.

1: $\tau \leftarrow 1$;
2: Initialize $\mathbf{f}_\tau^{(q)}$ randomly and $\mathbf{A}_\tau^{(q)} \leftarrow \mathbf{I}$;
3: **while** stopping criterion is not reached **do**
4: $\quad\mathbf{Z}_\tau^{(q)} \leftarrow \mathbf{A}_\tau^{(q)}\mathbf{f}_\tau^{(q)}(\mathbf{X}^{(q)})$;
5: $\quad\mathbf{Z}_{\tau,\text{orth}}^{(q)} \leftarrow \mathbf{P}\mathbf{Q}^\top, \mathbf{P}\mathbf{D}\mathbf{Q}^\top = \mathbf{Z}_\tau^{(q)}$;
6: $\quad\mathbf{U}_\tau \leftarrow$ first K rows of $\mathbf{Q}^\top$, where $\mathbf{P}\mathbf{D}\mathbf{Q}^\top = [\mathbf{Z}^{(1)\top}_{\tau,\text{orth}}, \mathbf{Z}^{(2)\top}_{\tau,\text{orth}}]^\top$;
7: $\quad\mathbf{U}_\tau^{(q)} \leftarrow \text{Exp}_{\mathbf{U}_\tau}(\text{Log}_{\mathbf{U}_\tau}(\mathbf{Z}_{\tau,\text{orth}}^{(q)}), t)$, with t such that $d_{\text{chord}}(\mathbf{U}_\tau^{(1)}, \mathbf{U}_\tau^{(2)}) = r$;
8: $\quad s \leftarrow 1$;
9: $\quad\mathbf{f}_{\tau,s}^{(q)} \leftarrow \mathbf{f}_\tau^{(q)}$, $\mathbf{A}_{\tau,s}^{(q)} \leftarrow \mathbf{A}_\tau^{(q)}$;
10: $\quad$**while** stopping criterion is not reached **do**
11: $\quad\quad\mathbf{Z}_{\tau,s}^{(q)} \leftarrow \mathbf{A}_{\tau,s}^{(q)}\mathbf{f}_{\tau,s}^{(q)}(\mathbf{X}^{(q)})$;
12: $\quad\quad\mathbf{Z}_{\tau,s,\text{orth}}^{(q)} \leftarrow \mathbf{P}\mathbf{Q}^\top, \mathbf{P}\mathbf{D}\mathbf{Q}^\top = \mathbf{Z}_{\tau,s}^{(q)}$;
13: $\quad\quad\text{loss}_{\tau,s} \leftarrow \sum_{q=1}^{2} d_{\text{chord}}(\mathbf{U}_\tau^{(q)}, \mathbf{Z}_{\tau,s,\text{orth}}^{(q)})$;
14: $\quad\quad\mathbf{A}_{\tau,s+1}^{(q)} \leftarrow \text{backprop}(\mathbf{A}_{\tau,s}^{(q)}, \nabla_{\mathbf{A}_{\tau,s}^{(q)}}\text{loss}_{\tau,s})$;
15: $\quad\quad\mathbf{f}_{\tau,s+1}^{(q)} \leftarrow \text{backprop}(\mathbf{f}_{\tau,s}^{(q)}, \nabla_{\mathbf{f}_{\tau,s}^{(q)}}\text{loss}_{\tau,s})$;
16: $\quad\quad s \leftarrow s + 1$;
17: $\quad$**end while**
18: $\quad\mathbf{A}_{\tau+1}^{(q)} = \mathbf{A}_{\tau,s}^{(q)}$;
19: $\quad\mathbf{f}_{\tau+1}^{(q)} = \mathbf{f}_{\tau,s}^{(q)}$;
20: $\quad\tau \leftarrow \tau + 1$;
21: **end while**

5. EXPERIMENTS

In this section, we show that including residual relaxation in the correlation maximization problem results in extracting more meaningful representations for heterogeneous data using both synthetic and real datasets. When referring to the proposed, residually relaxed method with a certain residual, we include the residual in parentheses.

5.1. Synthetic Data

First, we compare the different methods on synthetic data generated with a post-nonlinear (PNL) mixture model [19], given as

$$\mathbf{X}^{(q)} = \mathbf{l}^{(q)}\Big(\mathbf{\Theta}^{(q)}\mathbf{S}^{(q)}\Big), \tag{19}$$

where $\mathbf{S}^{(q)} \in \mathbb{R}^{K\times M}$ denotes the matrix containing the K shared components from view q and $\mathbf{\Theta}^{(q)} \in \mathbb{R}^{N\times K}$ is the linear mixing matrix, assumed to be full rank and contain all non-zero elements. The non-linear mixing is denoted as $\mathbf{l}^{(q)} = \begin{bmatrix} l^{(q)(1)}(\cdot) & \cdots & l^{(q)(N)}(\cdot)\end{bmatrix}^\top$, encapsulating the N scalar non-linear mixing functions $l^{(q)(n)}(\cdot)$, each applied on the n-th dimension of view q. For $\mathbf{S}^{(1)} = \mathbf{S}^{(2)} = \mathbf{S}$, [12] showed that (19) is identifiable, i.e., the signal subspaces can be recovered by learning the nonlinear functions $\mathbf{l}^{(q)}$. However, as we want to investigate data with shared subspaces that are correlated but not identical, we consider correlated $\mathbf{S}^{(q)}$.

Let $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ each contain $K = 2$ random vectors, Gaussian distributed with a zero-mean and unit variance. The correlation coefficient between the components of $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ is 0.6. The matrices $\mathbf{\Theta}^{(q)}$ are randomly generated and the non-linear mixings are as follows:

$$l^{(1)(n)}(\cdot) = 0.2\ \text{sigmoid}(\cdot) + (\cdot)^3,\ \text{for } n = 1, 2 \tag{20}$$

$$l^{(2)(n)}(\cdot) = \tanh(\cdot) + 0.2\ \exp(\cdot),\ \text{for } n = 1, 2. \tag{21}$$

We generated[2] $M = 200$ independent and identically distributed (i.i.d.) observations to obtain $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. As the original subspaces $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ are known, we can evaluate the extracted subspaces from the different methods by measuring the distance between the ground truth subspace and the extracted one for every view q, using

$$\text{dist}(\mathbf{Z}^{(q)}, \mathbf{S}^{(q)}) = 1 - \frac{c_1 + c_2}{2}, \tag{22}$$

where c_1 and c_2 denote canonical correlation values between the rowspaces of the matrices $\mathbf{Z}^{(q)}$ and $\mathbf{S}^{(q)}$. We report the distance measure averaged over both views,

$$\text{Avg. distance} = \frac{1}{2}(\text{dist}(\mathbf{Z}^{(1)}, \mathbf{S}^{(1)}) + \text{dist}(\mathbf{Z}^{(2)}, \mathbf{S}^{(2)})). \tag{23}$$

Both DCCA and RDCCA use fully connected networks with two hidden layers with 256 neurons each and sigmoid activation. The final layer consists of two neurons with linear activation. Weights of the fully connected layers are regularized via L_2-regularization with a regularization parameter of 10^{-4} for DCCA and 10^{-6} for RDCCA. Table 1 compares CCA,

[2]Code available at: github.com/SSTGroup/GeodesicRelaxationDCCA

	CCA	DCCA	RDCCA(.0)	RDCCA(.5)	RDCCA(.7)	RDCCA(.8)	RDCCA(.9)
Avg. distance	0.23	0.79	0.79	0.39	0.15	**0.08**	0.13
Avg. correlation	0.33	0.99	0.99	0.86	0.70	**0.59**	0.42

Table 1: Comparison of the average subspace distance between $\mathbf{Z}^{(q)}$ and $\mathbf{S}^{(q)}$ and the learned correlation between $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ for CCA, DCCA and RRCCA for synthetic data. All results are averaged over five runs.

DCCA, and RDCCA with different residuals with respect to the average subspace distance between $\mathbf{Z}^{(q)}$ and $\mathbf{S}^{(q)}$ and the learned correlation between $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$. Results are averaged over five runs, and deep techniques are trained for 2000 epochs.

We can see that CCA is not able to well-extract the underlying subspaces as the correlation coefficients of the extracted subspaces are smaller than the true value of 0.6. DCCA, on the other hand, is able to find over complex representations, as the extracted subspaces are nearly perfectly correlated. The performance of RDCCA depends on the input residual value. For a residual of 0, it performs similarly to DCCA. However, for a residual of 0.8, which is the ground truth normalized chordal distance for a correlation coefficient of 0.6, RDCCA outperforms all other configurations by far with an average correlation coefficient of 0.59 and a low average distance. Even for non-optimal residual values like 0.5, 0.7, and 0.9, RDCCA outperforms DCCA. We chose those values to show that subspace extraction can be improved even without hitting the optimal residual value. They all allow for more accurate extraction of subspaces compared to the original DCCA.

5.2. Occluded Multiview MNIST

For the next investigation, we use the well-known MNIST dataset [20], however, modified to have multiview data [8]. The original dataset consists of 28×28 pixel gray-scale images of handwritten digits. To every image in the set, we assign another image from the same class. The matched image acts as a second view, such that we have pairs of images of the same digit. With no noise or interference, the only common information in the images is the digit itself.

To make the two views heterogeneous, we add noise and structural interference to the images. Thereby, we also ensure that the shared subspaces are not identical. We use the available MNIST version with spatter noise [21] and add two white boxes on top. Those boxes are placed randomly over the images of both views and the height and width of the boxes are sampled uniformly and independently from the interval $[0, 10]$. An exemplary image pair can be seen in Figure 1. Pixel values are normalized to $[0, 1]$ and we split the data into training, evaluation, and test sets with 50k, 10k, and 10k samples, respectively.

To evaluate the meaningfulness of the learned representations, we cluster them with the K-means algorithm [22] for every view. We compute the clustering accuracy via majority voting and average over both views. For RDCCA, we use the evaluation data to determine the residual that works best. For both DCCA and RDCCA, we use neural networks with five fully connected layers, each consisting of 1536 neurons with sigmoid activation. The final layer consists of $K = 15$ neurons with linear activation. Both methods are regularized via L_2-regularization with a regularization parameter of 10^{-5}. We report the clustering accuracy on the test data, averaged over five runs each.

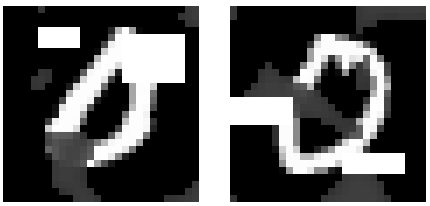

Fig. 1: An example image pair (class 0) from the occluded multiview MNIST dataset with two boxes added.

	CCA	DCCA	RDCCA(.6)
Avg. test accuracy	51.0%	73.1%	$\mathbf{79.4}\%$
Avg. training correlation	0.34	0.93	0.69

Table 2: Comparison of the K-means clustering accuracy on the test data and the learned correlation between the extracted representations, $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$, on the training data for CCA, DCCA and RRCCA for the occluded multiview MNIST dataset. All results are averaged over five runs.

Table 2 shows the comparison of K-means clustering accuracy of CCA, DCCA, and RDCCA with a residual of 0.6, chosen accordingly to accuracy on the evaluation data. We can see that RDCCA clearly outperforms CCA and DCCA, with more than 6% improvement over DCCA. When looking at the averaged correlation coefficients between the two $K = 15$ dimensional representations, $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$, it stands out that DCCA learned nearly perfectly correlated representations on the training data. This negatively affects the clustering accuracy on the test data. Similarly to the synthetic experiments, CCA just found a small correlation between the representations and performed worst. RDCCA learned representations that are correlated with an average correlation coefficient of around 0.7. As the dataset is designed with noise and structural interference, representations learned by RDCCA are more meaningful and lead to better clustering accuracy.

6. CONCLUSION

Current multiview learning methods of CCA and DCCA aim to find maximally correlated representations. We propose a novel, residual relaxation-based optimization problem, RDCCA, that generalizes the DCCA problem to aim for partly-correlated, non-identical representations. Further, we propose an iterative algorithm that learns those representations by computing view-specific slack variables on the geodesic between a central slack variable and the learned representations. Experiments with synthetic data, where we know the ground truth subspace, show that RDCCA learns better estimates of the underlying subspaces. Even for suboptimal residual values, RDCCA outperforms DCCA. On the occluded multiview MNIST dataset, we showed that RDCCA learns representations that lead to higher clustering accuracies compared to CCA and DCCA.

7. REFERENCES

[1] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.

[2] H. Hotelling, "Relations between two sets of variates," *Biometrika*, pp. 321–377, 1936.

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

[4] Wei Shao, Shunian Xiang, Zuoyi Zhang, Kun Huang, and Jie Zhang, "Hyper-graph based sparse canonical correlation analysis for the diagnosis of alzheimer's disease from multi-dimensional genomic data," *Methods*, vol. 189, pp. 86–94, 2021.

[5] Xiaowei Zhuang, Zhengshi Yang, and Dietmar Cordes, "A technical review of canonical correlation analysis for neuroscience applications," *Human Brain Mapping*, vol. 41, no. 13, pp. 3807–3833, 2020.

[6] Karl Friston, Jacquie Phillips, Dave Chawla, and Christian Buchel, "Nonlinear PCA: characterizing interactions between modes of brain activity," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 355, no. 1393, pp. 135–146, 2000.

[7] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *Proc. of ICML*. PMLR, 2013, pp. 1247–1255.

[8] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes, "On deep multi-view representation learning," in *Proc. of ICML*. PMLR, 2015, pp. 1083–1092.

[9] Tülay Adali, Yuri Levin-Schwartz, and Vince D Calhoun, "Multimodal data fusion using source separation: Application to medical imaging," *Proc. of the IEEE*, vol. 103, no. 9, pp. 1494–1506, 2015.

[10] Kurt Hornik, Maxwell Stinchcombe, and Halbert White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[11] Yingzhen Yang, Jiahui Yu, Xingjian Li, Jun Huan, and Thomas S Huang, "An empirical study on regularization of deep neural networks by local rademacher complexity," *arXiv preprint arXiv:1902.00873*, 2019.

[12] Qi Lyu and Xiao Fu, "Nonlinear multiview analysis: Identifiability and neural network-assisted implementation," *IEEE Trans. Signal Process.*, vol. 68, pp. 2697–2712, 2020.

[13] Yifei Wang, Zhengyang Geng, Feng Jiang, Chuming Li, Yisen Wang, Jiansheng Yang, and Zhouchen Lin, "Residual relaxation for multi-view representation learning," *NeurIPS*, vol. 34, pp. 12104–12115, 2021.

[14] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[15] Tim Marrinan, J Ross Beveridge, Bruce Draper, Michael Kirby, and Chris Peterson, "Finding the subspace mean or median to fit your need," in *Proc. IEEE CVPR*, 2014, pp. 1082–1089.

[16] Bruce Draper, Michael Kirby, Justin Marks, Tim Marrinan, and Chris Peterson, "A flag representation for finite collections of subspaces of mixed dimensions," *Linear Algebra and its Applications*, vol. 451, pp. 15–32, 2014.

[17] P-A Absil, Robert Mahony, and Rodolphe Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, 2008.

[18] Nicolas Boumal, *An introduction to optimization on smooth manifolds*, Cambridge University Press, 2023.

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

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[21] Norman Mu and Justin Gilmer, "MNIST-C: A robustness benchmark for computer vision," *arXiv preprint arXiv:1906.02337*, 2019.

[22] J MacQueen, "Classification and analysis of multivariate observations," in *Berkeley Symp. Math. Statist. Probability*. University of California Los Angeles LA USA, 1967, pp. 281–297.