Canonical Correlation Analysis of Datasets with a Common Source Graph

Jia Chen, Gang Wang, *Student Member, IEEE*,
Yanning Shen, *Student Member, IEEE*, and Georgios B. Giannakis, *Fellow, IEEE*

***Abstract*—Canonical correlation analysis (CCA) is a powerful technique for discovering whether or not hidden sources are commonly present in two (or more) datasets. Its well-appreciated merits include dimensionality reduction, clustering, classification, feature selection, and data fusion. The standard CCA however, does not exploit the geometry of the common sources, which may be available from the given data or can be deduced from (cross-) correlations. In this paper, this extra information provided by the common sources generating the data is encoded in a graph, and is invoked as a graph regularizer. This leads to a novel graph-regularized CCA approach, that is termed graph (g) CCA. The novel gCCA accounts for the graph-induced knowledge of common sources, while minimizing the distance between the wanted canonical variables. Tailored for diverse practical settings where the number of data is smaller than the data vector dimensions, the dual formulation of gCCA is also developed. One such setting includes kernels that are incorporated to account for nonlinear data dependencies. The resultant graph-kernel (gk) CCA is also obtained in closed form. Finally, corroborating image classification tests over several real datasets are presented to showcase the merits of the novel linear, dual, and kernel approaches relative to competing alternatives.**

***Index Terms*—Dimensionality reduction, correlation analysis, signal processing over graphs, Laplacian regularization, generalized eigen-decomposition**

I. Introduction

In many fields, exploratory data analysis depends critically on dimensionality reduction, a process to discover compact representations of large volumes of high-dimensional data [1]. Dimensionality reduction has been a crucial first step to obtain tractable learning tasks, such as classification, clustering, and regression [2], [1]. Principal component analysis (PCA) is arguably the most widely used dimensionality reduction method, finding low-dimensional representations from high-dimensional data while preserving most of the data variance [3]. Yet, ordinary PCA presumes that data vectors lie close to a hyperplane - a gross geometrical approximation for several datasets. Local linear embedding on the other hand, preserves linear relationships between neighboring data [1], while Laplacian eigenmaps ensure that data close in the original manifold are mapped to close by locations in the low-dimensional space, thus aiming to preserve local distances [4].

Nonetheless, such dimensionality reduction methods deal with one dataset at a time. They are challenged when it comes to analyzing two (or more) datasets jointly. Moreover, they require all data vectors to have the same dimension. Canonical correlation analysis (CCA) is a well-known method for extracting low-dimensional representations from two datasets that can have different dimensions, while maximizing their correlations [5]. Although recent PCA variants such as discriminative PCA can deal with two datasets at a time, their goal is to extract the most discriminative features from the data of interest relative to the other [6]. Formally, CCA aims at finding latent low-dimensional common structure from a paired dataset collected from different views of the same entities, also known as common sources. Each view contains high-dimensional representations of the sources in a certain feature space. For example, images of an individual captured by two cameras can be interpreted as two different views of this individual (here playing the role of a source). The ability of CCA to handle multiple datasets of different dimensions is a key enabler in tasks such as multi-mode data fusion, where the need arises to fuse information from different domains [7]. Ever since its proposition [5], CCA benefits have been documented in diverse applications, such as blind source separation, brain imaging, clustering and classification, word embedding, and natural language processing, to name a few [7], [8], [9].

To account for nonlinearities present in the data, kernel and deep CCA generalizations have also been developed based on kernels or deep neural networks [7], [10]. Sparse CCA looking for sparse canonical vectors was investigated by [11]. Multi-view CCA on the other hand, generalizes ordinary CCA to handle data from more than two modalities. Even though CCA solutions can be found via generalized eigen-decomposition, the resultant computational complexity may not scale well with the problem dimensionality. This motivated decentralized CCA alternatives [12].

However, all aforementioned PCA and CCA tools do not exploit structural graph-induced information on the sources that may be available. Such information may be inferred from alternative views of the data, or it can be provided by the physics that dictates the underlying graph. Indeed, graph-aware dimensionality reduction methods have lately demonstrated promising performance [13], [14], [15], [16], [17].

Building on recent advances in graph-aware dimensionality reduction [13], [14], the present paper introduces a neat link between graph embedding and canonical correlations, by putting forward a novel graph (g) CCA approach. Our gCCA pursues maximally correlated linear projections, while also leveraging statistical dependencies due to the common sources hidden in the paired dataset. The underlying source

This work was supported in part by NSF grants 1500713 and 1514056. The authors are with the Digital Technology Center and the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA. Emails: {chen5625, gangwang, shenx513, georgios}@umn.edu.

graph encoding these dependencies can be either given, or be constructed based on prior knowledge. When the number of data samples is smaller than the data vector dimensions, we advocate the graph dual (gd) CCA. Relative to gCCA, our gdCCA not only bypasses the inversion of ill-conditioned data covariance matrices, but also incurs lower complexity in high-dimensional setups. To further account for nonlinearities, we also develop what we term graph kernel (gk) CCA. Interestingly, solutions to all three gCCA variants can be found analytically through generalized eigenvalue decompositions.

Different from [18], [19], where CCA was regularized by two graph Laplacians separately per view, gCCA here jointly leverages a single graph induced by the common sources. This is of major practical importance, e.g., in brain mapping, where besides functional magnetic resonance imaging (MRI) and diffusion-weighted MRI data collected at different brain regions [20], one has also access to the connectivity patterns among these regions. Finally, numerical tests on several real-world datasets are presented to corroborate the merits of our proposed approaches for classification tasks over their competing alternatives.

The rest of this paper is structured as follows. Upon introducing the standard CCA in Section II, our gCCA is motivated, and derived in Section III. Its dual counterpart is developed in Section IV. Generalizing linear gCCA variants, the kernel version of gCCA is devised in Section V. Numerical tests on several real-world datasets are presented in Section VI, and the paper is concluded in Section VII.

Notation: Bold uppercase (lowercase) letters denote matrices (column vectors). Operators $\mathrm{Tr}(\cdot)$, $(\cdot)^{-1}$ and $(\cdot)^\top$ are matrix trace, inverse and transpose, respectively; $\|\cdot\|_2$ stands for the ℓ_2-norm of vectors; $\mathbf{0}$ is an all-zero vector whose dimension is clear from the context; $\langle \mathbf{a},\, \mathbf{b}\rangle$ denotes the inner product of vectors $\mathbf{a}$ and $\mathbf{b}$; and $\mathbf{I}$ represents the identity matrix of suitable size.

II. Preliminaries

Consider two datasets $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{y}_i\}_{i=1}^N$ with corresponding dimensionality D_x and D_y, collected from two different views of the same sources $\mathbf{s}_i \in \mathbb{R}^\rho$ with possibly $\rho \ll \min\{D_x,\, D_y\}$. CCA amounts to finding low-dimensional subspaces $\mathbf{U} \in \mathbb{R}^{D_x \times d}$ and $\mathbf{V} \in \mathbb{R}^{D_y \times d}$ with $d \le \rho$, such that the Euclidean distance between the low-dimensional representations $\{\mathbf{U}^\top \mathbf{x}_i\}$ and $\{\mathbf{V}^\top \mathbf{y}_i\}$ is minimized. Assume without loss of generality that both datasets are centered, meaning their corresponding sample means have been removed from the datasets. For ease of exposition, this section focuses on $d = 1$ first, while generalization to $d \ge 2$ will be discussed later. CCA solves the following problem

$$(\mathbf{u}^*,\, \mathbf{v}^*) := \arg\min_{\mathbf{u},\, \mathbf{v}} \quad \frac{1}{N}\sum_{i=1}^N \left(\mathbf{u}^\top \mathbf{x}_i - \mathbf{v}^\top \mathbf{y}_i\right)^2 \tag{1a}$$

where $\mathbf{u} \in \mathbb{R}^{D_x}$ and $\mathbf{v} \in \mathbb{R}^{D_y}$ are also termed a canonical pair. To ensure unique nonzero solutions however, the ensuing standard constraints are imposed

$$\mathbf{u}^\top \mathbf{\Sigma}_x \mathbf{u} = 1, \quad \text{and} \quad \mathbf{v}^\top \mathbf{\Sigma}_y \mathbf{v} = 1 \tag{1b}$$

where $\mathbf{\Sigma}_x := (1/N)\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top$ and $\mathbf{\Sigma}_y := (1/N)\sum_{i=1}^N \mathbf{y}_i \mathbf{y}_i^\top$ denote the sample covariance matrices of $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$, respectively. Projections $\{\mathbf{x}_i^\top \mathbf{u}^*\}_{i=1}^N$ and $\{\mathbf{y}_i^\top \mathbf{v}^*\}_{i=1}^N$ form a pair of canonical variables, which can be interpreted as low-dimensional approximations of the common sources $\{\mathbf{s}_i\}_{i=1}^N$.

After simple manipulations, (1) leads to the following popular formulation of CCA [7]

$$(\mathbf{u}^*, \mathbf{v}^*) := \arg\max_{\mathbf{u},\, \mathbf{v}} \quad \mathbf{u}^\top \mathbf{\Sigma}_{xy} \mathbf{v} \tag{2a}$$

$$\text{s. to} \quad \mathbf{u}^\top \mathbf{\Sigma}_x \mathbf{u} = 1, \;\text{and}\; \mathbf{v}^\top \mathbf{\Sigma}_y \mathbf{v} = 1 \tag{2b}$$

where $\mathbf{\Sigma}_{xy} := (1/N)\sum_{i=1}^N \mathbf{x}_i \mathbf{y}_i^\top$ is the sample cross-covariance matrix of $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$.

Using Lagrange duality theory, the solution of (2) will be given next in analytical form. To this end, letting $\lambda,\, \mu \in \mathbb{R}$ be the dual variables associated with the two constraints in (2b), one can write the Lagrangian as

$$\mathcal{L}(\mathbf{u}, \mathbf{v}; \lambda, \mu) = \mathbf{u}^\top \mathbf{\Sigma}_{xy} \mathbf{v} - \lambda(\mathbf{u}^\top \mathbf{\Sigma}_x \mathbf{u} - 1) - \mu(\mathbf{v}^\top \mathbf{\Sigma}_y \mathbf{v} - 1).$$

At the optimum $(\mathbf{u}^*,\, \mathbf{v}^*)$, the KKT conditions assert that

$$\mathbf{\Sigma}_{xy}\mathbf{v}^* = 2\lambda^* \mathbf{\Sigma}_x \boldsymbol{u}^*, \qquad (\mathbf{u}^*)^\top \mathbf{\Sigma}_x \mathbf{u}^* = 1 \tag{3a}$$

$$\mathbf{\Sigma}_{xy}^\top \mathbf{u}^* = 2\mu^* \mathbf{\Sigma}_y \mathbf{v}^*, \qquad (\mathbf{v}^*)^\top \mathbf{\Sigma}_y \mathbf{v}^* = 1. \tag{3b}$$

Left-multiplying the first equations in (3a) and (3b) by $(\mathbf{u}^*)^\top$ and $(\mathbf{v}^*)^\top$, respectively, lead to $(\mathbf{u}^*)^\top \mathbf{\Sigma}_{xy} \mathbf{v}^* = 2\lambda^* = 2\mu^*$. Hence, solving (2) reduces to solving the generalized eigenvalue problem, see e.g., [7]

$$\begin{bmatrix} \mathbf{\Sigma}_{xy}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{xy} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = 2\lambda \begin{bmatrix} \mathbf{0} & \mathbf{\Sigma}_y \\ \mathbf{\Sigma}_x & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}. \tag{4}$$

Maximizing the objective function (2a) is tantamount to finding the largest generalized eigenvalue $\lambda^* := \lambda_1$ in (4), and the optimal canonical vectors $[(\mathbf{u}^*)^\top \; (\mathbf{v}^*)^\top]^\top$ to (2) are obtained from the corresponding generalized eigenvector.

In order to find $d \le \min(D_x,\, D_y)$ pairs of canonical vectors, say $\{(\mathbf{u}_i, \mathbf{v}_i)\}_{i=1}^d$, one can basically repeat the steps leading to (5) with extra constraints. Specifically, if the first $(k-1)$ pairs $\{(\mathbf{u}_i^*,\, \mathbf{v}_i^*)\}_{i=1}^{k-1}$ have been found, the k-th pair can be obtained by solving (2) with the orthogonality constraints $(\mathbf{u}_k^*)^\top \mathbf{\Sigma}_x \mathbf{u}_i^* = 0$ and $(\mathbf{v}_k^*)^\top \mathbf{\Sigma}_y \mathbf{v}_i^* = 0$ for $i = 1,\, 2,\, \ldots,\, k-1$; that is,

$$\max_{\mathbf{u}_k,\, \mathbf{v}_k} \quad \mathbf{u}_k^\top \mathbf{\Sigma}_{xy} \mathbf{v}_k \tag{5a}$$

$$\text{s. to} \quad \mathbf{u}_k^\top \mathbf{\Sigma}_x \mathbf{u}_k = 1, \quad \mathbf{v}_k^\top \mathbf{\Sigma}_y \mathbf{v}_k = 1 \tag{5b}$$

$$\mathbf{u}_k^\top \mathbf{\Sigma}_x \mathbf{u}_i^* = 0, \quad \mathbf{v}_k^\top \mathbf{\Sigma}_y \mathbf{v}_i^* = 0 \tag{5c}$$

$$\forall i = 1,\, 2,\, \ldots,\, k-1 \tag{5d}$$

and the same steps can be repeated until d canonical pairs are found. For brevity, let us concatenate the d canonical vectors $\{\mathbf{u}_i\}$ and $\{\mathbf{v}_i\}$ to form matrices $\mathbf{U}$ and $\mathbf{V}$ accordingly, and rewrite (5) in the following compact form

$$\max_{\mathbf{U},\, \mathbf{V}} \quad \mathrm{Tr}(\mathbf{U}^\top \mathbf{\Sigma}_{xy} \mathbf{V}) \tag{6a}$$

$$\text{s. to} \quad \mathbf{U}^\top \mathbf{\Sigma}_x \mathbf{U} = \mathbf{I}, \quad \text{and} \quad \mathbf{V}^\top \mathbf{\Sigma}_y \mathbf{V} = \mathbf{I} \tag{6b}$$

which yields simultaneously multiple canonical vectors. As deduced earlier, the m-th columns of minimizers $\mathbf{U}^* \in \mathbb{R}^{D_x \times d}$

and $\mathbf{V}^* \in \mathbb{R}^{D_y \times d}$ of (6) correspond to the left and right generalized eigenvectors of (4) associated with the m-th largest generalized eigenvalue, respectively.

III. CCA over Graphs

In diverse applications, the common sources $\{\mathbf{s}_i\}_{i=1}^N$ may be viewed as nodal vectors of a graph having N nodes. This structural prior information can be leveraged when finding the canonical vectors. In this paper, this extra knowledge of common sources is encoded in a graph, and will be embodied in the canonical variables through graph regularization.

We outline some basics of the graph theory first. A graph is represented by a tuple $\mathcal{G} = \{\mathcal{N}, \mathcal{W}\}$, where $\mathcal{N} := \{1, 2, \ldots, N\}$ is the vertex set, and $\mathcal{W} := \{w_{ij}\}_{(i,j)\in\mathcal{N}\times\mathcal{N}}$ stacks up edge weights w_{ij} over all vertex pairs (i, j). For ease of exposition, this paper focuses on undirected graphs, for which $w_{ij} = w_{ji}$ for all $i, j \in \mathcal{N}$. Moreover, a graph is said to be unweighted if all w_{ij}'s take binary values 0 or 1. Upon forming the so-called weighted adjacency matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$ with its (i, j)-th entry being w_{ij}, and defining $d_i := \sum_{j=1}^N w_{ij}$, the Laplacian matrix of graph $\mathcal{G}$ is given by

$$\mathbf{L}_{\mathcal{G}} := \mathbf{D} - \mathbf{W} \in \mathbb{R}^{N\times N} \tag{7}$$

where the diagonal matrix $\mathbf{D} \in \mathbb{R}^{N\times N}$ holds ordered entries $\{d_i\}_{i=1}^N$ on its diagonal.

Having introduced basic graph notation, we present a neat link between canonical correlations and graph embedding next. Consider for instance a graph $\mathcal{G}$ with adjacency matrix $\mathbf{W}$, over which the underlying sources $\{\mathbf{s}_i\}_{i=1}^N$ are assumed to be smooth. In other words, vectors $(\mathbf{s}_i, \mathbf{s}_j)$ residing on two connected nodes $i, j \in \mathcal{G}$ are deemed close to each other in Euclidean distance. As remarked earlier, canonical variables $\mathbf{u}^\top\mathbf{x}_i$ and $\mathbf{v}^\top\mathbf{y}_j$ are accordingly one-dimensional approximates of $\mathbf{s}_i$ and $\mathbf{s}_j$. Building on this fact, let us now focus on the weighted sum of distances between any two pairs of canonical variables from $\{\mathbf{u}^\top\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{v}^\top\mathbf{y}_i\}_{i=1}^N$ over $\mathcal{G}$, namely the quadratic term

$$\sum_{i=1}^N\sum_{j=1}^N w_{ij}\big(\mathbf{u}^\top\mathbf{x}_i - \mathbf{v}^\top\mathbf{y}_j\big)^2. \tag{8}$$

It is clear that by minimizing (8) over $\mathbf{u}$ and $\mathbf{v}$, canonical variables $\mathbf{u}^\top\mathbf{x}_i$ and $\mathbf{v}^\top\mathbf{y}_j$ corresponding to adjacent nodes $i, j \in \mathcal{G}$ with large edge weights w_{ij} will be promoted to stay close to each other. As such, invoking this term as a regularizer accounts for the additional graph knowledge of the common sources, while maximizing the linear correlation coefficient between the canonical variables, yielding

$$\begin{aligned}\min_{\mathbf{u},\,\mathbf{v}}\;\; & \frac{1}{2N}\sum_{i=1}^N\big(\mathbf{u}^\top\mathbf{x}_i-\mathbf{v}^\top\mathbf{y}_i\big)^2+\frac{\gamma}{2}\sum_{i=1}^N\sum_{j=1}^N w_{ij}\big(\mathbf{u}^\top\mathbf{x}_i-\mathbf{v}^\top\mathbf{y}_j\big)^2\\ \text{s. to}\;\; & \mathbf{u}^\top\boldsymbol{\Sigma}_x\mathbf{u}=1, \quad \text{and} \quad \mathbf{v}^\top\boldsymbol{\Sigma}_y\mathbf{v}=1\end{aligned}$$

in which $\gamma \geq 0$ is a hyper-parameter that balances the distance between canonical variable estimates with their smoothness over $\mathcal{G}$. After expanding the squares and removing the constant terms, the problem at hand can be equivalently rewritten as

$$\max_{\mathbf{u},\,\mathbf{v}}\;\; \mathbf{u}^\top\boldsymbol{\Sigma}_{xy}\mathbf{v}-\gamma\mathbf{u}^\top\mathbf{X}\mathbf{L}_{\mathcal{G}}\mathbf{Y}^\top\mathbf{v}-\frac{\gamma}{2}\sum_{i=1}^N d_i\big(\mathbf{u}^\top\mathbf{x}_i-\mathbf{v}^\top\mathbf{y}_i\big)^2 \tag{9a}$$

$$\text{s. to}\;\; \mathbf{u}^\top\boldsymbol{\Sigma}_x\mathbf{u}=1, \quad \text{and} \quad \mathbf{v}^\top\boldsymbol{\Sigma}_y\mathbf{v}=1. \tag{9b}$$

Evidently, problem (9) is non-convex and is not amenable to efficient solvers due to the bilinear terms as well as the quadratic equality constraints. Even though block coordinate descent-type solvers can be employed, only convergence to a stationary point can be guaranteed in general [12]. Instead of coping with the objective function (9a) directly, we shall pursue a lower bound of it, which will turn out to afford an analytical solution.

Toward that end, it is easy to verify that with all $\{d_i \geq 0\}_{i=1}^N$, the following holds for all $\mathbf{u} \in \mathbb{R}^{D_x}$ and $\mathbf{v} \in \mathbb{R}^{D_y}$:

$$\sum_{i=1}^N d_i\big(\mathbf{u}^\top\mathbf{x}_i - \mathbf{v}^\top\mathbf{y}_i\big)^2 \leq 2d_{\max}N\left(\mathbf{u}^\top\boldsymbol{\Sigma}_x\mathbf{u} + \mathbf{v}^\top\boldsymbol{\Sigma}_y\mathbf{v}\right) \tag{10}$$

where $d_{\max} := \max_{1\leq i\leq N} d_i$, and the equality is achieved when $d_i = d_{\max}$ and $\mathbf{u}^\top\mathbf{x}_i = -\mathbf{v}^\top\mathbf{y}_i$ for all $i = 1, 2, \ldots, N$. Subsequently, we replace the last term in (9a) with the right-hand-side term, which contributes to a valid lower bound of (9a). Formally stated, we have the following reformulation.

Proposition 1. *Replacing the sum in* (9a) *with its upper bound in* (10) *leads to an objective that lower bounds* (9a)*. Merging and ignoring the constant terms due to the equality constraints* (9b) *leads to our novel gCCA formulation*

$$\max_{\mathbf{u},\,\mathbf{v}}\;\; \mathbf{u}^\top\boldsymbol{\Sigma}_{xy}\mathbf{v} - \gamma\mathbf{u}^\top\mathbf{X}\mathbf{L}_{\mathcal{G}}\mathbf{Y}^\top\mathbf{v} \tag{11a}$$

$$\text{s. to}\;\; \mathbf{u}^\top\boldsymbol{\Sigma}_x\mathbf{u}=1, \quad \text{and} \quad \mathbf{v}^\top\boldsymbol{\Sigma}_y\mathbf{v}=1. \tag{11b}$$

Clearly, when $\gamma = 0$, our gCCA finds $(\mathbf{u}, \mathbf{v})$ that only maximizes the linear correlation between the pair of canonical variables. In this case, no graph knowledge is exploited, and our gCCA reduces to the standard CCA. With γ increasing gradually, gCCA accounts progressively for extra graph information of the common sources when finding the canonical variables.

Next, let us consider multiple canonical pairs $\{(\mathbf{u}_i, \mathbf{v}_i)\}_{i=1}^d$, and collect them to form matrices $\mathbf{U} := [\mathbf{u}_1 \;\cdots\; \mathbf{u}_d]$ and $\mathbf{V} := [\mathbf{v}_1 \;\cdots\; \mathbf{v}_d]$. We can then generalize gCCA in (11) to $d \geq 2$ as

$$\max_{\mathbf{U},\mathbf{V}}\;\; \operatorname{Tr}\big(\mathbf{U}^\top\boldsymbol{\Sigma}_{xy}\mathbf{V} - \gamma\mathbf{U}^\top\mathbf{X}\mathbf{L}_{\mathcal{G}}\mathbf{Y}^\top\mathbf{V}\big) \tag{12a}$$

$$\text{s. to}\;\; \mathbf{U}^\top\boldsymbol{\Sigma}_x\mathbf{U}=\mathbf{I}, \quad \text{and} \quad \mathbf{V}^\top\boldsymbol{\Sigma}_y\mathbf{V}=\mathbf{I}. \tag{12b}$$

Interestingly, even with the extra graph-inducing regularization term, our gCCA in (12) still admits an analytical solution, under the standard assumption that data covariance matrices $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$ are both nonsingular. For concreteness, the solution is summarized in the following result, and for self-contained presentation, its proof is provided in Appendix A.

Algorithm 1 CCA with a common source graph

```
Input: $\{\mathbf{x}_i\}_{i=1}^N$, $\{\mathbf{y}_i\}_{i=1}^N$, $d$, $\mathbf{W}$, and $\gamma$.
Form (cross-)covariance matrices, $\mathbf{\Sigma}_x$, $\mathbf{\Sigma}_y$ and $\mathbf{\Sigma}_{xy}$.
Build $\mathbf{L}_{\mathcal{G}}$ using (7).
Perform SVD on $\mathbf{\Sigma}_x^{-1/2}\left(\mathbf{\Sigma}_{xy} - \gamma\mathbf{X}\mathbf{L}_{\mathcal{G}}\mathbf{Y}^\top\right)\mathbf{\Sigma}_y^{-1/2}$
Extract the first $d$ leading eigenvectors to obtain $\bar{\mathbf{U}}^*$ and $\bar{\mathbf{V}}^*$.
Compute $\mathbf{U}^* = \mathbf{\Sigma}_x^{-1/2}\bar{\mathbf{U}}^*$ and $\mathbf{V}^* = \mathbf{\Sigma}_y^{-1/2}\bar{\mathbf{V}}^*$.
Output: $\mathbf{U}^*$ and $\mathbf{V}^*$.
```

Theorem 1. *Given zero-mean data $\{\mathbf{x}_i \in \mathbb{R}^{D_x}\}_{i=1}^N$ and $\{\mathbf{y}_i \in \mathbb{R}^{D_y}\}_{i=1}^N$, suppose that $\mathbf{\Sigma}_x = (1/N)\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^\top$ and $\mathbf{\Sigma}_y = (1/N)\sum_{i=1}^N \mathbf{y}_i\mathbf{y}_i^\top$ are nonsingular. Then the optimal solution $(\mathbf{U}^* \in \mathbb{R}^{D_x\times d}, \mathbf{V}^* \in \mathbb{R}^{D_y\times d})$ to the gCCA problem* (12) *with $d \le \min(D_x, D_y)$, is given by*

$$\mathbf{U}^* := \mathbf{\Sigma}_x^{-1/2}\bar{\mathbf{U}}^*, \quad \text{and} \quad \mathbf{V}^* := \mathbf{\Sigma}_y^{-1/2}\bar{\mathbf{V}}^* \tag{13}$$

where the columns of $\bar{\mathbf{U}}^* \in \mathbb{R}^{D_x\times d}$ and $\bar{\mathbf{V}}^* \in \mathbb{R}^{D_y\times d}$ are the d left and right singular vectors of $\mathbf{\Sigma}_x^{-1/2}\left(\mathbf{\Sigma}_{xy} - \gamma\mathbf{X}\mathbf{L}_{\mathcal{G}}\mathbf{Y}^\top\right)\mathbf{\Sigma}_y^{-1/2}$ associated with its d largest singular values. Moreover, the maximum objective value of (12a) *is the sum of the d largest singular values.*

Our proposed gCCA scheme is summarized in Alg. 1. Two remarks are now in order.

Remark 1. Different from our single regularizer in (12), the approaches in [18], [19] rely on two regularizers or two constraints involving graph priors $\mathbf{U}^\top\mathbf{X}\mathbf{L}_{\mathcal{G}_x}\mathbf{X}^\top\mathbf{U}$ and $\mathbf{V}^\top\mathbf{Y}\mathbf{L}_{\mathcal{G}_y}\mathbf{Y}^\top\mathbf{V}$ for the two-view data $\mathbf{X}$ and $\mathbf{Y}$, respectively. However, the problem formulation in [19] does not admit an analytical solution. Although iterative algorithms can be used to solve the involved nonconvex optimization problem, only convergence to a stationary point can be ensured in general [21]. When the two datasets lie in two distinct graphs $\mathcal{G}_x$ and $\mathcal{G}_y$, using the graph-Laplacian regularized constraints can improve standard CCA performance [22]. This approach is mainly suggested for semi-supervised learning, where $\mathbf{\Sigma}_{xy}$ is fully available. In contrast, (12) leverages the graph induced by the common sources, and our source graph regularizer $\mathbf{U}^\top\mathbf{X}\mathbf{L}_{\mathcal{G}}\mathbf{Y}^\top\mathbf{V}$ directly exploits correlations between the low-dimensional approximations of common sources over $\mathcal{G}$. This is critical in certain practical setups, in which one has prior knowledge about the common sources besides the given datasets. In brain imaging for instance, in addition to the functional MRI and diffusion-weighted MRI data collected at different brain regions [20], one has also access to the connectivity patterns among these regions. Furthermore, our proposed gCCA framework admits an analytical solution.

Remark 2. To induce different graph properties, rather than relying on $\mathbf{L}_{\mathcal{G}}$, a family of graph regularizations of the form $r(\mathbf{L}_{\mathcal{G}}) := \sum_{i=1}^N r(\lambda_i^w)\mathbf{u}_i^w(\mathbf{u}_i^w)^\top$ can be also employed [23], where $r(\cdot) : \mathbb{R} \to \mathbb{R}^+$ is a scalar function, and appropriate choices of $r(\lambda_i^w)$ are helpful for inducing diverse graph properties; while $\mathbf{u}_i^w \in \mathbb{R}^N$ is the eigenvector of $\mathbf{L}_{\mathcal{G}}$ associated with its i-th largest eigenvalue λ_i^w.

IV. Dual CCA over Graphs

Similar to dual PCA [14], various practical scenarios involving high-dimensional data vectors, have $N \ll \min\{D_x, D_y\}$, in which case $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_y$ become singular, and the results in Theorem 1 do not apply. Even though this rank deficiency can be remedied with appropriate Tikhonov regularization [7], the resultant computational complexity can be considerably higher than the alternative of investigating gCCA in the dual domain. In this direction, consider first expressing $\mathbf{u} \in \mathbb{R}^{D_x}$ and $\mathbf{v} \in \mathbb{R}^{D_y}$ in terms of their corresponding parts of the data matrices $\mathbf{X}$ and $\mathbf{Y}$ as

$$\mathbf{u} := \mathbf{X}\boldsymbol{\alpha}, \quad \text{and} \quad \mathbf{v} := \mathbf{Y}\boldsymbol{\beta} \tag{14}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^N$ and $\boldsymbol{\beta} \in \mathbb{R}^N$ are the so-termed dual vectors. Substituting (14) into (11) gives rise to our graph dual (gd) CCA formulation for one pair of canonical vectors

$$\max_{\boldsymbol{\alpha},\,\boldsymbol{\beta}} \quad \boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\beta} - \gamma\boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\mathbf{L}_{\mathcal{G}}\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\beta} \tag{15a}$$

$$\text{s. to} \quad \boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha} = 1 \tag{15b}$$

$$\boldsymbol{\beta}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\beta} = 1. \tag{15c}$$

Similar to Section III, introducing variables $\lambda_x \in \mathbb{R}$ and $\lambda_y \in \mathbb{R}$ to be the Lagrange multipliers corresponding to constraints (15b) and (15c), respectively, one can write the Lagrangian for (15) as

$$\mathcal{L}(\boldsymbol{\alpha},\,\boldsymbol{\beta};\,\lambda_x,\,\lambda_y) := -\boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}(\mathbf{I} - \gamma\mathbf{L}_{\mathcal{G}})\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\beta} + \frac{\lambda_x}{2}(\boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha} - 1) + \frac{\lambda_y}{2}(\boldsymbol{\beta}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\beta} - 1).$$

Setting derivatives of the Lagrangian with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to zero further leads to

$$-\mathbf{X}^\top\mathbf{X}(\mathbf{I} - \gamma\mathbf{L}_{\mathcal{G}})\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\beta} + \lambda_x\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha} = \mathbf{0} \tag{16a}$$

$$-\mathbf{Y}^\top\mathbf{Y}(\mathbf{I} - \gamma\mathbf{L}_{\mathcal{G}})\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha} + \lambda_y\mathbf{Y}^\top\mathbf{Y}\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\beta} = \mathbf{0}. \tag{16b}$$

Left-multiplying (16a) and (16b) by $\boldsymbol{\alpha}^\top$ and $\boldsymbol{\beta}^\top$, respectively, and subsequently subtracting the latter from the former, we arrive at

$$\lambda_x\boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha} - \lambda_y\boldsymbol{\beta}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\beta} = 0. \tag{17}$$

Taking into account (17), (15b), and (15c), it follows that at the optimal solution, we have $\lambda^* := \lambda_x^* = \lambda_y^*$. Supposing for now that $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{Y}^\top\mathbf{Y}$ are nonsingular, we find

$$\boldsymbol{\alpha}^* := \frac{1}{\lambda^*}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{Y}^\top\mathbf{Y} - \gamma\mathbf{L}_{\mathcal{G}}\mathbf{Y}^\top\mathbf{Y}\right)\boldsymbol{\beta}^*. \tag{18}$$

Plugging (18) into (16b) yields

$$\left(\mathbf{Y}^\top\mathbf{Y}\right)^{-1}(\mathbf{I} - \gamma\mathbf{L}_{\mathcal{G}})^2\,\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\beta}^* = (\lambda^*)^2\boldsymbol{\beta}^* \tag{19}$$

and similarly, one obtains that

$$\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}(\mathbf{I} - \gamma\mathbf{L}_{\mathcal{G}})^2\,\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha}^* = (\lambda^*)^2\boldsymbol{\alpha}^*. \tag{20}$$

The last two equalities show that $\boldsymbol{\alpha}^*$ depends solely on $\mathbf{X}$, and $\boldsymbol{\beta}^*$ solely on $\mathbf{Y}$. This holds without any assumption about the paired dataset $\mathbf{X}$ and $\mathbf{Y}$ whatsoever. Furthermore, when $\gamma = 0$, both (19) and (20) lead to trivial solutions. However, recall that our goal is to extract relations between data $\mathbf{X}$ and $\mathbf{Y}$. As with the dual CCA [7], in order to avoid such trivial solutions, we invoke a Tikhonov regularization term that leads us to our graph dual (gd) CCA formulation

$$\max_{\boldsymbol{\alpha},\,\boldsymbol{\beta}} \quad \boldsymbol{\alpha}^\top\big(\mathbf{X}^\top\mathbf{X}\mathbf{Y}^\top\mathbf{Y} - \gamma\boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\mathbf{L}_{\mathcal{G}}\mathbf{Y}^\top\mathbf{Y}\big)\boldsymbol{\beta} \tag{21a}$$

$$\text{s.to} \quad \boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha} + \epsilon\boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha} = 1 \tag{21b}$$

$$\boldsymbol{\beta}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\beta} + \epsilon\boldsymbol{\beta}^\top\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\beta} = 1. \tag{21c}$$

Here, the coefficient $\epsilon > 0$ is a pre-selected penalty parameter. Appealing to Lagrange duality theory again, the minimizers $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ are the eigenvectors of (22a) and (22b) associated with the largest eigenvalue, namely $(\lambda_1^*)^2$; that is,

$$(\mathbf{I} - \gamma\mathbf{L}_{\mathcal{G}})\mathbf{Y}^\top\mathbf{Y}(\mathbf{Y}^\top\mathbf{Y} + \epsilon\mathbf{I})^{-1}(\mathbf{I} - \gamma\mathbf{L}_{\mathcal{G}})\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha}^* \\ = (\lambda^*)^2(\mathbf{X}^\top\mathbf{X} + \epsilon\mathbf{I})\boldsymbol{\alpha}^* \tag{22a}$$

$$(\mathbf{I} - \gamma\mathbf{L}_{\mathcal{G}})\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \epsilon\mathbf{I})^{-1}(\mathbf{I} - \gamma\mathbf{L}_{\mathcal{G}})\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\beta}^* \\ = (\lambda^*)^2(\mathbf{Y}^\top\mathbf{Y} + \epsilon\mathbf{I})\boldsymbol{\beta}^*. \tag{22b}$$

Moreover, the optimal objective function value coincides with λ_1^*.

When looking for d pairs of dual vectors $\{(\boldsymbol{\alpha}_i,\,\boldsymbol{\beta}_i)\}_{i=1}^d$, which are collected to form matrices $\mathbf{A} := [\boldsymbol{\alpha}_1 \,\cdots\, \boldsymbol{\alpha}_d]$ and $\mathbf{B} := [\boldsymbol{\beta}_1 \,\cdots\, \boldsymbol{\beta}_d]$, our gdCCA becomes

$$\max_{\mathbf{A},\,\mathbf{B}} \quad \operatorname{Tr}\big(\mathbf{A}^\top\mathbf{X}^\top\mathbf{X}\mathbf{Y}^\top\mathbf{Y}\mathbf{B} - \gamma\mathbf{A}^\top\mathbf{X}^\top\mathbf{X}\mathbf{L}_{\mathcal{G}}\mathbf{Y}^\top\mathbf{Y}\mathbf{B}\big) \tag{23a}$$

$$\text{s.to} \quad \mathbf{A}^\top\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top\mathbf{X}\mathbf{A} + \epsilon\mathbf{A}^\top\mathbf{X}^\top\mathbf{X}\mathbf{A} = \mathbf{I} \tag{23b}$$

$$\mathbf{B}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{Y}^\top\mathbf{Y}\mathbf{B} + \epsilon\mathbf{B}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{B} = \mathbf{I} \tag{23c}$$

for which the i-th column of its optimal solution $\mathbf{A}^*$ ($\mathbf{B}^*$) is provided by the generalized eigenvector in (22a) [(22b)] associated with the i-th largest generalized eigenvalue. Once $\mathbf{A}^*$, $\mathbf{B}^*$ are found, the optimal canonical vectors sought can be obtained via (14) as $\mathbf{U}^* = \mathbf{X}\mathbf{A}^*$ and $\mathbf{V}^* = \mathbf{Y}\mathbf{B}^*$.

V. KCCA over Graphs

Although linear models are attractive due to their simplicity, they cannot capture complex nonlinear data dependencies that are common in real-world applications, including genomics [24], functional MRI [18], and acoustic feature learning [10].

Going beyond linearity, we generalize our linear models of CCA over graphs in Sections III and IV to take into account nonlinear relationships between data $\mathbf{X}$ and $\mathbf{Y}$ using kernel methods. In this context, a graph (g) KCCA framework is developed. We begin with transforming the two datasets using two nonlinear functions to higher (possibly infinite) dimensional feature spaces, and subsequently find low-dimensional canonical variables. Specifically, let $\boldsymbol{\phi}_x$ be a mapping from space $\mathbb{R}^{D_x}$ to space $\mathbb{R}^{D_h}$ (possibly with $D_h = \infty$). It is clear from (23) that both the objective and the constraints depend on the data $\mathbf{X}$ only through the similarities $\{\langle\mathbf{x}_i,\,\mathbf{x}_j\rangle\}_{i,j=1}^N$. Therefore, upon 'lifting' all data vectors $\{\mathbf{x}_i\}_{i=1}^N$ to obtain $\{\boldsymbol{\phi}(\mathbf{x}_i)\}_{i=1}^N$, all similarities $\{\langle\mathbf{x}_i,\,\mathbf{x}_j\rangle\}_{i,j=1}^N$ can be readily replaced with $\{\langle\boldsymbol{\phi}(\mathbf{x}_i),\,\boldsymbol{\phi}(\mathbf{x}_j)\rangle\}_{i,j=1}^N$. Nonetheless, evaluating $\{\langle\boldsymbol{\phi}(\mathbf{x}_i),\,\boldsymbol{\phi}(\mathbf{x}_j)\rangle\}_{i,j=1}^N$ can be computationally intractable due to the high-dimensionality.

To circumvent the cost of explicitly working in the high-dimensional space, the so-called 'kernel trick' is employed [25]. To this end, we select some kernel function κ_x, such that $\kappa_x(\mathbf{x}_i,\,\mathbf{x}_j) := \langle\boldsymbol{\phi}_x(\mathbf{x}_i),\,\boldsymbol{\phi}_x(\mathbf{x}_j)\rangle$ for all $i,\,j = 1,\,2,\,\ldots,\,N$, which form the $(i,\,j)$-th entries of the so-termed kernel matrix $\bar{\mathbf{K}}_x \in \mathbb{R}^{N\times N}$. Similarly, we can build the kernel matrix $\bar{\mathbf{K}}_y \in \mathbb{R}^{N\times N}$ for data $\mathbf{Y}$ using a different kernel function κ_y. As in linear gCCA and gdCCA discussed is Sections III and IV, we require that the data in the mapped feature spaces $\{\boldsymbol{\phi}_x(\mathbf{x}_i)\}_{i=1}^N$ and $\{\boldsymbol{\phi}_y(\mathbf{y}_i)\}_{i=1}^N$ be centered, where $\boldsymbol{\phi}_y(\mathbf{y}_i)$ is the nonlinear mapping for 'lifting' data $\mathbf{y}_i$ to render kernel matrix $\mathbf{K}_y$. Using the kernel trick again, the required centering in the high-dimensional space can be realized by centering the kernel matrix for data $\mathbf{X}$ as

$$\mathbf{K}_x(i,j) := \bar{\mathbf{K}}_x(i,j) - \frac{1}{N}\sum_{\ell=1}^N \bar{\mathbf{K}}_x(\ell,j) - \frac{1}{N}\sum_{\ell=1}^N \bar{\mathbf{K}}_x(i,\ell) \\ + \frac{1}{N^2}\sum_{m=1}^N\sum_{n=1}^N \bar{\mathbf{K}}_x(m,n) \tag{24}$$

and likewise for centering $\mathbf{K}_y$.

Upon replacing $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{Y}^\top\mathbf{Y}$ in (23) with $\mathbf{K}_x$ and $\mathbf{K}_y$, we arrive at our gKCCA

$$\max_{\mathbf{A},\,\mathbf{B}} \quad \operatorname{Tr}(\mathbf{A}^\top\mathbf{K}_x\mathbf{K}_y\mathbf{B} - \gamma\mathbf{A}^\top\mathbf{K}_x\mathbf{L}_{\mathcal{G}}\mathbf{K}_y\mathbf{B}) \tag{25a}$$

$$\text{s.to} \quad \mathbf{A}^\top\mathbf{K}_x^2\mathbf{A} + \epsilon\mathbf{A}^\top\mathbf{K}_x\mathbf{A} = \mathbf{I} \tag{25b}$$

$$\mathbf{B}^\top\mathbf{K}_y^2\mathbf{B} + \epsilon\mathbf{A}^\top\mathbf{K}_y\mathbf{B} = \mathbf{I}. \tag{25c}$$

It is clear that with properly selected kernel matrices $\mathbf{K}_x$ and $\mathbf{K}_y$, gKCCA is able to capture nonlinear correlations between $\mathbf{X}$ and $\mathbf{Y}$, while also leveraging the graph prior information of the common sources. Following the steps used to solve the gCCA problem (12), the solution to (25) is summarized in Theorem 2, with its proof deferred to Appendix B. The main steps of the gKCCA are listed in Alg. 2.

Theorem 2. *If $\mathbf{K}_x$ and $\mathbf{K}_y$ are nonsingular, the optimal solutions $\mathbf{A}^*$ and $\mathbf{B}^*$ to* (25) *are given by*

$$\mathbf{A}^* := \mathbf{K}_x^{-1/2}(\mathbf{K}_x + \epsilon\mathbf{I})^{-1/2}\bar{\mathbf{A}}^* \tag{26a}$$

$$\mathbf{B}^* := \mathbf{K}_y^{-1/2}(\mathbf{K}_y + \epsilon\mathbf{I})^{-1/2}\bar{\mathbf{B}}^* \tag{26b}$$

where matrices $\bar{\mathbf{A}}^* \in \mathbb{R}^{N\times d}$ and $\bar{\mathbf{B}}^* \in \mathbb{R}^{N\times d}$ collect as columns the top d left and right singular vectors of

$$\mathbf{C} := (\mathbf{K}_x + \epsilon\mathbf{I})^{-1/2}\mathbf{K}_x^{1/2}(\mathbf{I} - \gamma\mathbf{L}_{\mathcal{G}})\mathbf{K}_y^{1/2}(\mathbf{K}_y + \epsilon\mathbf{I})^{-1/2}. \tag{26c}$$

Furthermore, the optimal objective value (25a) *is the sum of the d largest singular values of* $\mathbf{C}$.

Remark 3. When the kernel functions needed to form $\mathbf{K}_x$ and $\mathbf{K}_y$ are not available, one may presume $\mathbf{K}_x := \sum_{m=1}^M \theta_m\mathbf{K}_m$ and $\mathbf{K}_y := \sum_{m=1}^M \delta_m\mathbf{K}_m$ for (25). Here, $\{\mathbf{K}_m\}_{m=1}^M$ are known kernel matrices for a preselected dictionary of kernels, while $\{\theta_m,\,\delta_m\}_{m=1}^M$ are unknown coefficients to be optimized along with the canonical vectors through (25). Such a data-driven approach is also known as multi-kernel learning, and it has been broadly studied; see for example, [26], [27].

In terms of computational cost, we summarize the complexities of gCCA, gdCCA, gKCCA, CCA, dCCA, and KCCA in Table I, where $D := \max(D_x, D_y)$. Note that gCCA incurs higher computational cost than standard CCA, due to the extra

Algorithm 2 Graph kernel canonical correlation analysis

```
1: Input: $\{\mathbf{x}_i\}_{i=1}^N$, $\{\mathbf{y}_i\}_{i=1}^N$, $\mathbf{W}$, $d$, $\gamma$, $\epsilon$, $\kappa_x(\cdot)$, and $\kappa_y(\cdot)$.
2: Construct $\mathbf{K}_x$ and $\mathbf{K}_y$ using (24).
3: Build $\mathbf{L}_{\mathcal{G}}$ using (7).
4: Perform SVD on $\mathbf{C} := \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ in (26c), where the diagonal elements of $\mathbf{\Sigma}$ are organized in descending order; $\mathbf{U} \in \mathbb{R}^{N\times N}$, $\mathbf{V} \in \mathbb{R}^{N\times N}$, and $\mathbf{\Sigma} \in \mathbb{R}^{N\times N}$.
5: Extract the first $d$ columns of $\mathbf{U}$ and $\mathbf{V}$ to form $\bar{\mathbf{A}}^* \in \mathbb{R}^{N\times d}$ and $\bar{\mathbf{B}}^* \in \mathbb{R}^{N\times d}$, respectively.
6: Compute $\mathbf{A}^* = \mathbf{K}_x^{-1/2}(\mathbf{K}_x + \epsilon\mathbf{I})^{-1/2}\bar{\mathbf{A}}^*$ and $\mathbf{B}^* = \mathbf{K}_y^{-1/2}(\mathbf{K}_y + \epsilon\mathbf{I})^{-1/2}\bar{\mathbf{B}}^*$.
7: Output: $\mathbf{A}^*$ and $\mathbf{B}^*$.
```

TABLE I: Computational complexity comparison

gCCA	$\mathcal{O}(\min(D_x, D_y)N^2)$
CCA	$\mathcal{O}(D^2N)$
gdCCA (dCCA)	$\mathcal{O}(DN^2)$
gKCCA (KCCA)	$\mathcal{O}(\max(D, N)N^2)$

multiplication term of $\mathbf{Y}\mathbf{L}_{\mathcal{G}}\mathbf{X}^T$ in gCCA. If $N \ll D$, then gCCA in its present form is not feasible, or suboptimal even if the pseudo-inverse or Tikhonov regularization is employed, at computational complexity $\mathcal{O}(D^3)$. In this case, gdCCA is computationally more attractive since its complexity grows only linearly with D. In terms of gKCCA, when $D \gg N$, evaluating the kernel matrices dominates the computational complexity, giving rise to $\mathcal{O}(DN^2)$. When $D \ll N$, Steps 4 and 6 in Alg. 2 dominate the complexity, incurring complexity of $\mathcal{O}(N^3)$.

VI. Numerical Tests

To showcase the merits of our novel approaches, several classification experiments using real data are reported in this section. Classification accuracies of our proposed gCCA, gdCCA and gKCCA are compared with competing alternatives.

A. Tests for gCCA

In this experiment, the AR face dataset [28], and the Extended Yale-B (EYB) face image dataset [29], were used to examine the classification performance of different schemes, including gCCA, CCA, graph (g) PCA [14], PCA, graph regularized multi-set (GrM) CCA [19], and the k-nearest neighbors (KNN) method.

The AR face database contains color face images of 100 individuals, each depicted in 26 images. These 26 images per person were taken under different lighting conditions, occlusions and expressions. Each image was cropped and resized to 40×30 pixels, converted to grayscale image, and vectorized to obtain a $1,200\times 1$ vector. The $1,200$ features of each image are unevenly split into two views, where one view consists of the first 300 features collected in one column of $\mathbf{X}_0 \in \mathbb{R}^{300\times 2,600}$ ($2,600$ columns for all the images) , while the remaining 900 features were used to form $\mathbf{Y}_0 \in \mathbb{R}^{900\times 2,600}$. Suppose that $N_{\rm tr}$ columns were randomly drawn from 26 columns of $\mathbf{X}_0$ and $\mathbf{Y}_0$ that correspond to one person, to form the training data $\mathbf{X} \in \mathbb{R}^{300\times 100N_{\rm tr}}$ and $\mathbf{Y} \in \mathbb{R}^{900\times 100N_{\rm tr}}$, respectively. For the remaining $(26 - N_{\rm tr})$ columns of $\mathbf{X}_0$ associated with each person, half of them will be used for tuning the hyper-parameters, and the other half for testing, which are collected in $\mathbf{X}_{\rm tu} \in \mathbb{R}^{300\times 100(13-0.5N_{\rm tr})}$ and $\mathbf{X}_{\rm te} \in \mathbb{R}^{300\times 100(13-0.5N_{\rm tr})}$ accordingly. Here, we consider the scenario where only one view, namely $\mathbf{X}_{\rm te}$, is available in the testing phase, which is of practical importance when one only has partial information about the testing images.

The EYB database consists of frontal face images of 38 individuals, each of which has around 65 color images of 192×168 pixels. All images are resized to 30×20 pixels and converted to grayscale before being vectorized to obtain a 600×1 vector. Then, the vector associated with each image is split into two subvectors (views) with $D_x = 250$ and $D_y = 350$. For each individual, $N_{\rm tr}$ images are randomly selected and the corresponding two views are used to construct the training datasets $\mathbf{X} \in \mathbb{R}^{D_x\times 38N_{\rm tr}}$ and $\mathbf{Y} \in \mathbb{R}^{D_y\times 38N_{\rm tr}}$. Among the remaining images, $(30 - 0.5N_{\rm tr})$ images per individual are used for tuning dataset $\mathbf{X}_{\rm tu} \in \mathbb{R}^{D_x\times 38(30-0.5N_{\rm tr})}$ and another $(30 - 0.5N_{\rm tr})$ for testing dataset $\mathbf{X}_{\rm te} \in \mathbb{R}^{D_x\times 38(30-0.5N_{\rm tr})}$, after following a similar process to build $\mathbf{X}$.

Letting $N := 100N_{\rm tr}$ for the AR data experiment ($N := 38N_{\rm tr}$ for EYB), we collected all common sources $\{\mathbf{s}_i\}_{i=1}^N$ into matrix $\mathbf{S}$, which was constructed using the training data as follows: $\mathbf{S} := [\mathbf{X}^\top\, \mathbf{Y}^\top]^\top = [\mathbf{s}_1 \cdots \mathbf{s}_N]$. Based on $\mathbf{S}$, matrix $\mathbf{W}$ is formed to have $(i,\, j)$-th entry given by

$$w_{ij} := \begin{cases} \frac{\mathbf{s}_i^\top \mathbf{s}_j}{\|\mathbf{s}_i\|_2\|\mathbf{s}_j\|_2} & \mathbf{s}_i \in \mathcal{N}_k(\mathbf{s}_j) \text{ or } \mathbf{s}_j \in \mathcal{N}_k(\mathbf{s}_i) \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

for $i,\, j = 1,\, 2,\, \ldots,\, N$, where $\mathcal{N}_k(\mathbf{s}_j)$ denotes the set of the k-nearest neighbors of $\mathbf{s}_j$ that belong to the same class (person) in $\mathbf{S}$. In this experiment, $k = N_{\rm tr} - 1$ was kept fixed.

In this experiment, 30 Monte Carlo (MC) simulations were run to assess the classification performance of gCCA, standard CCA, GrMCCA, gPCA, PCA, and KNN on the AR face dataset, as well as the EYB dataset. For fairness, the weight matrix $\mathbf{W}$ in (27) is used for gPCA. The classification accuracy is defined as the ratio between the number of correctly classified images and the total number of images tested. For gCCA, CCA, GrMCCA, gPCA, and PCA, 50 (100) canonical vectors for the AR (EYB) face dataset were found to obtain the low-dimensional representations of testing data, which were subsequently classified through the 10-nearest neighbors algorithm based on the Euclidean distance metric. The hyper-parameters in gCCA, gPCA, and GrMCCA were tuned among 30 logarithmically-spaced values between 10^{-3} and 10^3 to maximize the classification accuracies on 'tuning set' of images.

Figures 1 and 2 depict the classification accuracies of gCCA, CCA, GrMCCA, gPCA, PCA, and KNN on the AR data, and the EYB data, respectively, for a varying number of training samples. It is evident that the accuracies of all simulated schemes improve as $N_{\rm tr}$ grows, and our proposed gCCA outperforms alternatives for $N_{\rm tr} \geq 10$. This corroborates that incorporating the source graph that encodes dependencies among common sources, pays off.

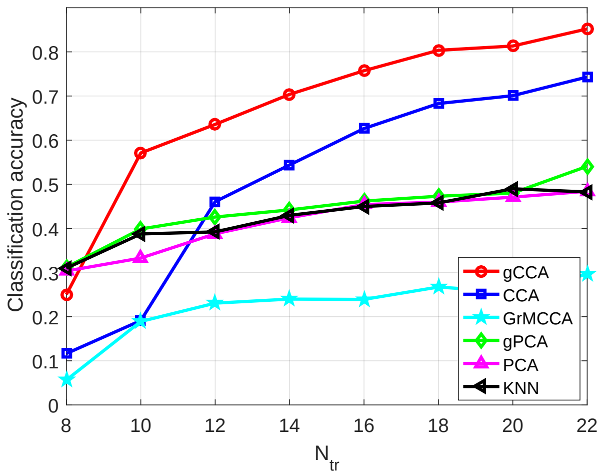

Fig. 1: Classification accuracy of gCCA on the AR face dataset [28].

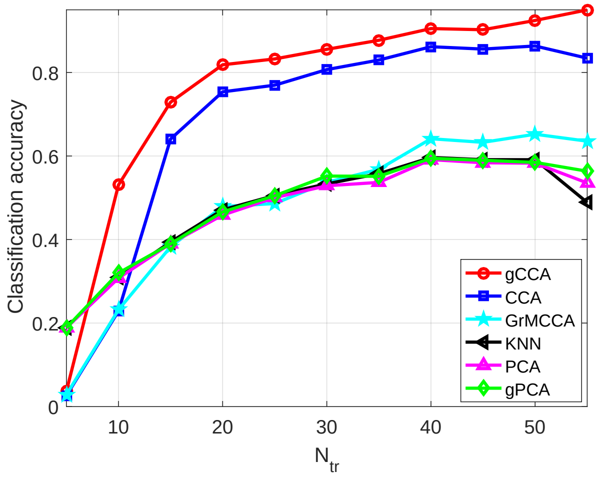

Fig. 2: Classification accuracy of gCCA on the EYB dataset [29].

B. Tests for gdCCA

The second experiment evaluates the capability of gdCCA for classification using again the AR face dataset and the EYB dataset. Per MC run on the AR face dataset, we collected all images of 10 randomly sampled people. For each selected person, N_{tr}, $(13-0.5N_{\mathrm{tr}})$, and $(13-0.5N_{\mathrm{tr}})$ images were randomly drawn for training, tunning, and testing, respectively. In the training phase, each image was first converted to a grayscale image, resized to 80×60 pixels, and subsequently lexicographically ordered to obtain a $4,800\times 1$ vector. To create the two views, this vector was partitioned into two subvectors of size $D_x=1,000$ for $\mathbf{X}\in\mathbb{R}^{D_x\times 10N_{\mathrm{tr}}}$ and of size $D_y=3,800$ for $\mathbf{Y}\in\mathbb{R}^{D_y\times 10N_{\mathrm{tr}}}$. Similarly, the training data $\mathbf{X}_{\mathrm{tu}}\in\mathbb{R}^{D_x\times 10(13-0.5N_{\mathrm{tr}})}$ and testing data $\mathbf{X}_{\mathrm{te}}\in\mathbb{R}^{D_x\times 10(13-0.5N_{\mathrm{tr}})}$ were generated.

Per realization on the EYB dataset, images of 10 individuals were randomly selected, and the two-view data $\mathbf{X}\in\mathbb{R}^{D_x\times 10N_{\mathrm{tr}}}$ and $\mathbf{Y}\in\mathbb{R}^{D_y\times 10N_{\mathrm{tr}}}$ were generated using the same procedure described for the AR data, except for $D_x=1,000$ and $D_y=7,000$. For both the tuning data $\mathbf{X}_{\mathrm{tu}}\in\mathbb{R}^{D_x\times 10(30-0.5N_{\mathrm{tr}})}$ and the testing data $\mathbf{X}_{\mathrm{te}}\in\mathbb{R}^{D_x\times 10(30-0.5N_{\mathrm{tr}})}$, a number of $(30-0.5N_{\mathrm{tr}})$ images were randomly chosen per person.

The two-view data in the training phase form $\mathbf{S}=[\mathbf{X}^\top\ \ \mathbf{Y}^\top]^\top$ and are further used to build $\mathbf{W}$ as in (27). For fairness, graph dual (gd) PCA [14] is tested with the same $\mathbf{W}$ as in gdCCA. Moreover, the two associated graph adjacency matrices in Laplacian regularized (Lr) CCA [18] are constructed via (27) after substituting $\mathbf{S}$ by $\mathbf{X}$ and $\mathbf{Y}$, respectively. We tune the hyper-parameters in gdCCA, dual (d) CCA, LrCCA and gdPCA among 30 logarithmically spaced values between 10^{-3} and 10^{3} to maximize the classification accuracy on data $\mathbf{X}_{\mathrm{tu}}$. Here, dCCA is implemented by gdCCA after assigning $\gamma=0$. In gdCCA, dCCA, LrCCA, gdPCA and dPCA [14], 20 and 100 projection vectors are used for obtaining lower-dimensional representations of $\mathbf{X}_{\mathrm{te}}$ for AR data and EYB data, respectively. Then, the K-NN rule with $K=10$ was applied to carry out the classification tasks.

Figures 3 and 4 present the averaged classification accuracies of gdCCA, dCCA, LrCCA, gdPCA, dPCA, and KNN for a varying number of training images per person over 30 MC realizations. Clearly, our gdCCA enjoys the best classification performance among all simulated schemes for different training samples.

There are two hyper-parameters, namely γ and ϵ in gdCCA. To understand how the hyper-parameters influence the classification performance, the gdCCA was simulated on the AR face dataset for a range of γ and ϵ values. For each person, 17 (9) images were employed for training (testing). Figure 5 plots the averaged classification accuracies over 30 MC runs, with γ varying from 10^{-3} to 10^{3} and ϵ from 10^{-5} to 10^{3}. For small γ values, the performance of gdCCA with small ϵ values outperforms that using large ϵ values. This is because with small γ, gdCCA approximates dCCA, and the Tikhonov regularization with excessively large ϵ values dominates the term for promoting uncorrelatedness between canonical variables. When ϵ is small, with γ increasing, the classification accuracy gradually increases by progressively exploiting the graph information, but subsequently decreases due to discarding the maximization of canonical correlations. Those observations confirm the assertion that with properly selected and nonzero γ and ϵ values, the performance of gdCCA reaches the best, in which case both maximizing the canonical correlations and exploiting the graph knowledge are in effect.

C. Tests for gKCCA

This last experiment assesses gKCCA for classification using the MNIST dataset [1]. There are 10 classes of handwritten 28×28 grayscale digit images in the MNIST, and each class (digit) consists of $7,000$ images. Per MC run, 5 classes of images were randomly sampled for classification. For each selected class, N_{tr}, $0.5N_{\mathrm{tr}}$, and $0.5N_{\mathrm{tr}}$ images are randomly sampled for training, parameter tuning, and testing, respectively. The two-view data were created as follows. The images were first resized to 20×20 pixels, followed by vectorization. Each vector was split to 2 subvectors of

[1] Downloaded from http://yann.lecun.com/exdb/mnist/.

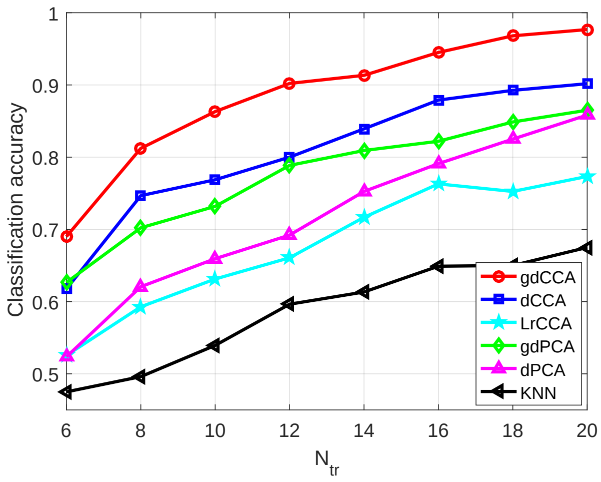

Fig. 3: Classification accuracy of gdCCA using dataset [28].

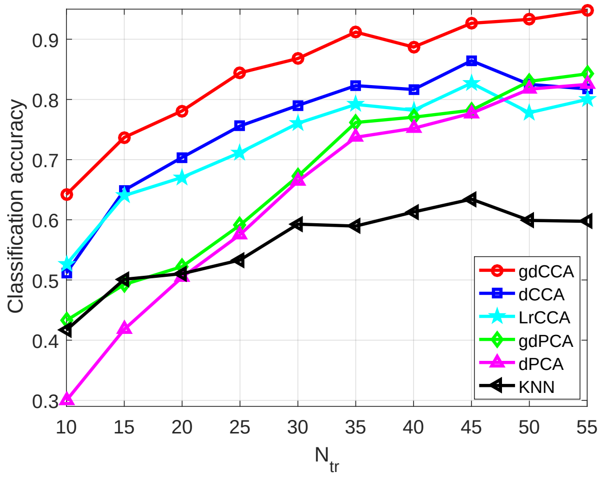

Fig. 4: Classification accuracy of gdCCA using dataset [29].

sizes D_x and $D_y = 400 - D_x$ for the two views. The first/second view of training data is denoted by training dataset $\mathbf{X} \in \mathbb{R}^{D_x \times 5N_{\rm tr}}/\mathbf{Y} \in \mathbb{R}^{D_y \times 5N_{\rm tr}}$. The tuning/testing dataset $\mathbf{X}_{\rm tu}/\mathbf{X}_{\rm te}$ are the first views of tuning/testing images.

Gaussian kernels were used for $\mathbf{X}$, $\mathbf{Y}$, and the common source $\mathbf{S} := [\mathbf{X}^\top \; \mathbf{Y}^\top]^\top$, whose bandwidth parameters were set as the medians of the corresponding Euclidean distances. The idea to generate the $\mathbf{W}$ in Sec. VI-A was adopted and adjusted for constructing the graph adjacency matrix, which was also denoted by $\mathbf{W}$ for notational simplicity. Obviously, the similarity between two sources in $\mathbf{S}$ can not be measured by the linear correlation coefficient, which instead can be represented by a corresponding element in the kernel matrix of $\mathbf{S}$, namely $\mathbf{K}_s$. Specifically,

$$w_{ij} := \begin{cases} \mathbf{K}_s(i,j) & \mathbf{s}_i \in \mathcal{M}_{k_1}(\mathbf{s}_j) \text{ or } \mathbf{s}_j \in \mathcal{M}_{k_1}(\mathbf{s}_i) \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

for $i, j = 1, 2, \ldots, 5N_{\rm tr}$, where $\mathbf{s}_i$ denotes the i-th source (i-th column) in $\mathbf{S}$, and $\mathcal{M}_{k_1}(\mathbf{s}_j)$ is the set containing the k_1-nearest neighbors of $\mathbf{s}_j$ from the same class. In the simulations of this subsection, $k_1 = N_{tr} - 1$. Further, graph kernel (gK) PCA [14] was simulated with the same $\mathbf{W}$ as in gKCCA. The graph Laplacian regularized (Lr) KCCA [18] was associated with two graph adjacency matrices, which were obtained by (28) after substituting $\mathbf{K}_s$ with $\mathbf{K}_x$ and $\mathbf{K}_y$ accordingly. For fairness, all the kernel-based methods, namely gKCCA, KCCA, LrKCCA, gKPCA, and KPCA, shared the same kernel $\mathbf{K}_x$ (and $\mathbf{K}_y$). When implementing the CCA-based and PCA-based subspace methods, 20 projection vectors were used for classification using the K-NN algorithm with $K = 10$. The hyper-parameters of gKCCA, KCCA, gdCCA, dCCA, LrKCCA, LrCCA, gKPCA, and gdPCA, were selected from 30 logarithmically spaced values between 10^{-3} and 10^3. For each algorithm, the parameters were selected with the best classification accuracy on the tuning dataset $\mathbf{X}_{\rm tu}$. In the following tests, the classification performance of all aforementioned algorithms was achieved after running 30 independent realizations.

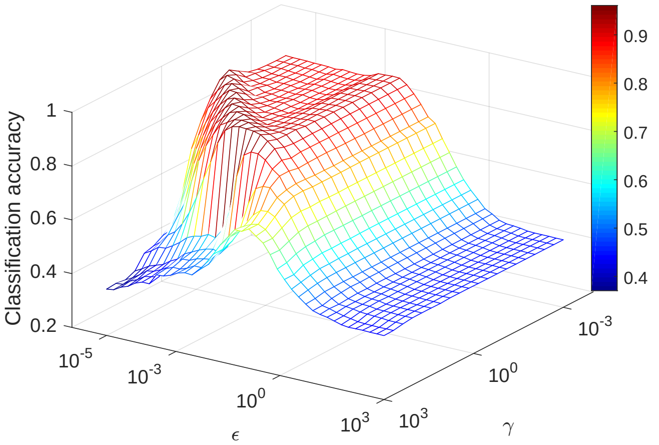

Fig. 5: Classification accuracy of gdCCA versus γ and ϵ.

In Fig. 6, the classification accuracies of simulated schemes for a variable number of training samples are reported, with $D_x = 120$ and $D_y = 280$. The plots validate the advantage of our gKCCA relative to the other 10 methods. Moreover, with extra training samples becoming available, the performance of all simulated schemes improves. Figure 7 depicts the classification accuracies of all methods for different D_x values, with $N_{\rm tr} = 30$ kept fixed. It is clear that gKCCA outperforms alternatives under different vector splittings. On the other hand, with D_x decreasing, it becomes more challenging to classify the testing data, so the classification accuracies of all schemes decrease. Interestingly, the performance gap between gKCCA and the others widens for smaller D_x values.

VII. Conclusions

Graph regularized CCA, dual CCA, as well as kernel CCA methods were revisited in this paper to exploit hidden low-dimensional common structures from two-view data of the same sources. Distinguishing itself from prior CCA contributions, our gCCA framework leverages additional information to improve the low-dimensional approximations through the canonical variables, by embedding the hidden common sources in a graph and invoking this graph prior knowledge as a CCA

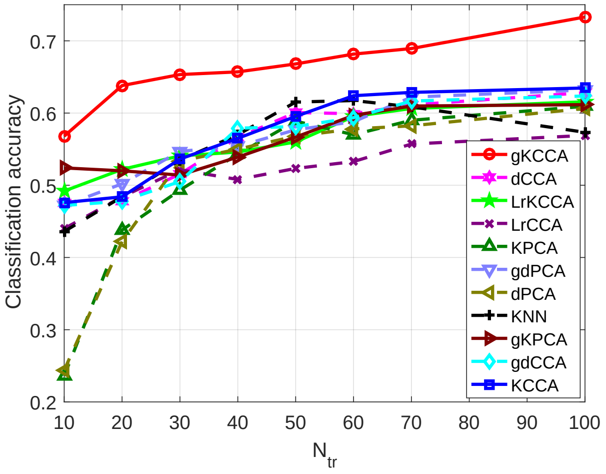

Fig. 6: Classification accuracy of gKCCA versus N_{tr}.

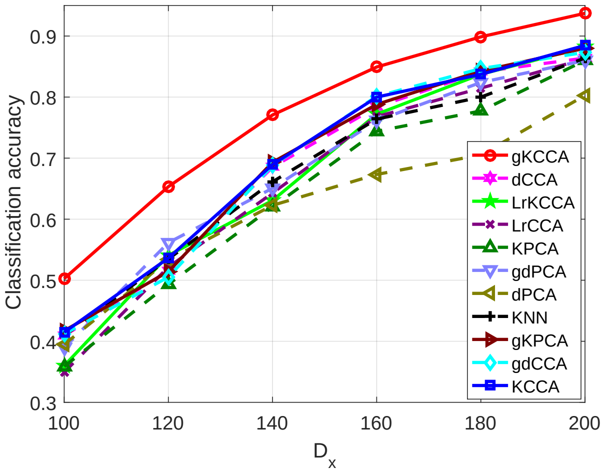

Fig. 7: Classification accuracy of gKCCA versus D_x.

regularizer. As such, canonical pairs that are able to capture the structural information between data vectors can be revealed. In certain practical setups where the number of data samples is small relative to the data vector dimensionality, our gCCA is not directly applicable, or leads to suboptimal performance and incurs high computational complexity. To bypass this, the dual model of gCCA, namely gdCCA, is put forth. To further account for nonlinear data dependencies, the graph kernel CCA is developed. Numerical tests on several real-world datasets are presented to demonstrate the merits of the novel approaches.

This paper opens up several intriguing directions for future research. To start, developing data-driven approaches to select the appropriate kernels (graphs) from a given or constructed dictionary of kernels (graphs) is timely and pertinent. To endow the proposed gCCA algorithms with scalability, distributed and online implementations are well-motivated for handling large-scale and/or high-dimensional streaming data. Generalizing our gCCA models to unpaired or multi-view datasets constitutes another interesting direction.

Appendix

A. Proof of Theorem 1

Letting

$$\bar{\mathbf{U}} := \mathbf{\Sigma}_x^{1/2}\mathbf{U} \in \mathbb{R}^{D_x \times d}, \quad \text{and} \quad \bar{\mathbf{V}} := \mathbf{\Sigma}_y^{1/2}\mathbf{V} \in \mathbb{R}^{D_y \times d}$$

the objective function (12a) can be rewritten as

$$\mathrm{Tr}(\bar{\mathbf{U}}^\top \mathbf{C}\bar{\mathbf{V}}) := \mathrm{Tr}(\bar{\mathbf{U}}^\top \mathbf{\Sigma}_x^{-1/2}(\mathbf{\Sigma}_{xy} - \gamma \mathbf{X}\mathbf{L}_{\mathcal{G}}\mathbf{Y}^\top)\mathbf{\Sigma}_y^{-1/2}\bar{\mathbf{V}})$$

and problem (12) boils down to

$$\max_{\bar{\mathbf{U}},\bar{\mathbf{V}}} \quad \mathrm{Tr}(\bar{\mathbf{U}}^\top \mathbf{C}\bar{\mathbf{V}}) \tag{29a}$$

$$\mathrm{s.to} \quad \bar{\mathbf{U}}^\top\bar{\mathbf{U}} = \mathbf{I}, \quad \text{and} \quad \bar{\mathbf{V}}^\top\bar{\mathbf{V}} = \mathbf{I}. \tag{29b}$$

Let $\bar{\mathbf{u}}_i \in \mathbb{R}^{D_x \times 1}$ and $\bar{\mathbf{v}}_i \in \mathbb{R}^{D_y \times 1}$ denote the i-th column of $\bar{\mathbf{U}}$ and $\bar{\mathbf{V}}$, respectively, with $i = 1, 2, \ldots, d$. The problem in (29) can be solved using d iterations with each iteration targeting the optimum over $\bar{\mathbf{u}}_i$ and $\bar{\mathbf{v}}_i$, namely

$$(\bar{\mathbf{u}}_i^*, \bar{\mathbf{v}}_i^*) := \arg\max_{\bar{\mathbf{u}}_i,\bar{\mathbf{v}}_i} \quad \bar{\mathbf{u}}_i^\top \mathbf{C}\bar{\mathbf{v}}_i \tag{30a}$$

$$\mathrm{s.to} \quad \bar{\mathbf{u}}_i^\top\bar{\mathbf{u}}_i = 1, \ \bar{\mathbf{v}}_i^\top\bar{\mathbf{v}}_i = 1 \tag{30b}$$

$$\bar{\mathbf{u}}_i^\top\bar{\mathbf{u}}_j = 0, \ \bar{\mathbf{v}}_i^\top\bar{\mathbf{v}}_j = 0 \tag{30c}$$

for all $j = 1, 2, \ldots, i-1$.

Since $\mathbf{C}^\top\mathbf{C}$ is symmetric, there exists orthonormal matrix $\mathbf{Z}^* \in \mathbb{R}^{d \times d}$ and diagonal matrix $\mathbf{\Lambda} \in \mathbb{R}^{d \times d}$ with diagonal entries $\lambda_1^2 \ge \lambda_2^2 \ge \cdots \ge \lambda_d^2$, such that

$$(\mathbf{Z}^*)^\top \mathbf{C}^\top\mathbf{C}\mathbf{Z}^* = \mathbf{\Lambda}. \tag{31}$$

The columns of $\mathbf{C}\mathbf{Z}^*$ are orthogonal, and have lengths equal to $\{\lambda_i \ge 0\}$. Concretely, with $\mathbf{z}_i^*$ denoting the ith column of $\mathbf{Z}^*$, it holds that

$$(\mathbf{C}\mathbf{z}_i^*)^\top(\mathbf{C}\mathbf{z}_j^*) = \begin{cases} \lambda_i^2, & j = i \\ 0, & j \neq i. \end{cases} \tag{32}$$

It follows readily that $\mathbf{Z}^* := [\mathbf{z}_1^* \ \cdots \ \mathbf{z}_d^*]$ is the optimizer of the following maximization problem

$$\max_{\mathbf{Z}} \quad \mathrm{Tr}(\mathbf{Z}^\top\mathbf{C}^\top\mathbf{C}\mathbf{Z})$$

$$\mathrm{s.to} \quad \mathbf{Z}^\top\mathbf{Z} = \mathbf{I}$$

which can be equivalently decomposed into d subproblems; that is,

$$\max_{\mathbf{z}_i} \quad \mathbf{z}_i^\top\mathbf{C}^\top\mathbf{C}\mathbf{z}_i \tag{33a}$$

$$\mathrm{s.to} \quad \mathbf{z}_i^\top\mathbf{z}_i = 1, \quad \text{and} \quad \mathbf{z}_i^\top\mathbf{z}_j = 0 \tag{33b}$$

for $j = 1, 2, \ldots, i-1$, and $i = 1, 2 \ldots, d$.

Now we focus on obtaining the first pair of canonical vectors by solving (30). After fixing $\bar{\mathbf{v}}_1$, the maximum value of $\bar{\mathbf{u}}_1^\top\mathbf{C}\bar{\mathbf{v}}_1$ over $\bar{\mathbf{u}}_1$ is obtained when $\bar{\mathbf{u}}_1$ is proportional to $\mathbf{C}\bar{\mathbf{v}}_1$, meaning

$$\bar{\mathbf{u}}_1^\top\mathbf{C}\bar{\mathbf{v}}_1 \le \|\mathbf{C}\bar{\mathbf{v}}_1\|_2 \tag{34}$$

where the equality is achieved when $\bar{\mathbf{u}}_1 = \mathbf{C}\bar{\mathbf{v}}_1/\|\mathbf{C}\bar{\mathbf{v}}_1\|_2$. It is clear from (33) that $\mathbf{z}_1^*$ maximizes $\mathbf{z}_1^\top\mathbf{C}^\top\mathbf{C}\mathbf{z}_1$. Thus, $\mathbf{z}_1^*$ also maximizes $\|\mathbf{C}\mathbf{z}_1\|_2$, and the maximum of $\|\mathbf{C}\bar{\mathbf{v}}_1\|_2$ is attained when $\bar{\mathbf{v}}_1 = \mathbf{z}_1^*$, yielding

$$\bar{\mathbf{u}}_1^\top\mathbf{C}\bar{\mathbf{v}}_1 \le \|\mathbf{C}\bar{\mathbf{v}}_1\|_2 \le \|\mathbf{C}\mathbf{z}_1^*\|_2 = \lambda_1. \tag{35}$$

When $\bar{\mathbf{u}}_1 = \mathbf{C}\mathbf{z}_1^*/\|\mathbf{C}\mathbf{z}_1^*\|_2$ and $\bar{\mathbf{v}}_1 = \mathbf{z}_1^*$, the first two inequalities in (35) hold as equalities, proving that $\bar{\mathbf{u}}_1^* = \mathbf{C}\mathbf{z}_1^*/\|\mathbf{C}\mathbf{z}_1^*\|_2$ and $\bar{\mathbf{v}}_1^* = \mathbf{z}_1^*$.

After finding the optimal $\bar{\mathbf{u}}_1^*$ and $\bar{\mathbf{v}}_1^*$, one can further search for $\bar{\mathbf{u}}_i^*$ and $\bar{\mathbf{v}}_i^*$ for $i \geq 2$ by solving (30). Without considering the constraints in (30c), we find $\bar{\mathbf{v}}_i^* = \mathbf{z}_i^*$ and $\bar{\mathbf{u}}_i^* = \mathbf{C}\mathbf{z}_i^*/\|\mathbf{C}\mathbf{z}_i^*\|_2$, which can be proved in the same way argued for $i = 1$. Next, we show that $\bar{\mathbf{v}}_i^*$ and $\bar{\mathbf{u}}_i^*$ satisfy constraints (30c). Obviously, $\bar{\mathbf{v}}_i^*$ is orthogonal to all vectors in the set $\{\bar{\mathbf{v}}_j^*\}_{j=1}^{i-1}$. Furthermore, $(\bar{\mathbf{u}}_i^*)^\top \bar{\mathbf{u}}_j^* = (\mathbf{C}\mathbf{z}_i^*)^\top \mathbf{C}\mathbf{z}_j^*/(\|\mathbf{C}\mathbf{z}_i^*\|_2\|\mathbf{C}\mathbf{z}_j^*\|_2)$, and (32) implies that $\bar{\mathbf{u}}_i^*$ is orthogonal to $\bar{\mathbf{u}}_j^*$ for $j = 1, 2, \ldots, i-1$.

Summarizing the two cases, we deduce that $\bar{\mathbf{u}}_i^* = \mathbf{C}\mathbf{z}_i^*/\|\mathbf{C}\mathbf{z}_i^*\|_2$ and $\bar{\mathbf{v}}_i^* = \mathbf{z}_i^*$ for $i = 1, 2, \ldots, d$, and $\bar{\mathbf{u}}_i^*$ and $\bar{\mathbf{v}}_i^*$ are the i-th left and right singular vector of $\mathbf{C}$ associated with the i-th largest singular value, that is λ_i.

In general, there may be zero eigenvalues. Suppose that the last positive eigenvalue is λ_k^2; in other words, it holds that $\lambda_1^2 \geq \cdots \geq \lambda_k^2 > \lambda_{k+1}^2 = \cdots = \lambda_d^2 = 0$. As such, the optimal solutions $\{\bar{\mathbf{u}}_i^*, \bar{\mathbf{v}}_i^*\}_{i=k+1}^d$ can be any set of d vectors satisfying constraints (30b) and (30c).

Once having computed $\bar{\mathbf{U}}^* = [\bar{\mathbf{u}}_1^* \; \cdots \; \bar{\mathbf{u}}_d^*]$ and $\bar{\mathbf{V}}^* = [\bar{\mathbf{v}}_1^* \; \cdots \; \bar{\mathbf{v}}_d^*]$, the optimal solutions $\mathbf{U}^*$ and $\mathbf{V}^*$ to problem (12) are obtained as $\mathbf{U}^* := \mathbf{\Sigma}_x^{-1/2}\bar{\mathbf{U}}^*$ and $\mathbf{V}^* := \mathbf{\Sigma}_y^{-1/2}\bar{\mathbf{V}}^*$. Moreover, the maximal value of (12a) becomes $\sum_{i=1}^d \lambda_i$.

B. Proof of Theorem 2

Upon defining

$$\bar{\mathbf{A}} := (\mathbf{K_x} + \epsilon\mathbf{I})^{1/2}\mathbf{K}_x^{1/2}\mathbf{A}$$
$$\bar{\mathbf{B}} := (\mathbf{K_y} + \epsilon\mathbf{I})^{1/2}\mathbf{K}_y^{1/2}\mathbf{B}$$

problem (25) can be rewritten as

$$(\bar{\mathbf{A}}^*, \bar{\mathbf{B}}^*) := \arg\max_{\bar{\mathbf{A}}, \bar{\mathbf{B}}} \; \mathrm{Tr}(\bar{\mathbf{A}}^\top \mathbf{C}\bar{\mathbf{B}}) \tag{36a}$$
$$\mathrm{s.\,to} \;\; \bar{\mathbf{A}}^\top\bar{\mathbf{A}} = \mathbf{I}, \text{ and } \bar{\mathbf{B}}^\top\bar{\mathbf{B}} = \mathbf{I}. \tag{36b}$$

Using the results in Appendix A, one readily concludes that the columns of optimizers $\bar{\mathbf{A}}^*$, $\bar{\mathbf{B}}^*$ consist of the d left and right singular vectors of $\mathbf{C}$ associated with the first d largest singular values, respectively, which leads to

$$\mathbf{A}^* = \mathbf{K}_x^{-1/2}(\mathbf{K}_x + \epsilon\mathbf{I})^{-1/2}\bar{\mathbf{A}}^*$$
$$\mathbf{B}^* = \mathbf{K}_y^{-1/2}(\mathbf{K}_y + \epsilon\mathbf{I})^{-1/2}\bar{\mathbf{B}}^*.$$

Likewise, the maximal value of (25a) is given by the sum of the d largest singular values of $\mathbf{C}$.

References

[1] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[2] I. T. Jolliffe, *Principal Component Analysis*. Wiley Online Library, 2002.

[3] F. R. S. Karl Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Phil. Mag. and J. of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[4] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, Jun. 2003.

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

[6] G. Wang, J. Chen, and G. B. Giannakis, "DPCA: Dimensionality reduction for discriminative analytics of multiple large-scale datasets," in *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Process.*, Calgary, Canada, April 15-20, 2018.

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

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

[9] S. VanVaerenbergh, J. Vía, and I. Santamaría, "Blind identification of SIMO wiener systems based on kernel canonical correlation analysis," *IEEE Trans. Signal Process.*, vol. 61, no. 9, pp. 2219–2230, May 2013.

[10] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Intl. Conf. on Mach. Learn.*, Atlanta, USA, Jun. 16-21 2013.

[11] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, Apr. 2009.

[12] J. Chen and I. D. Schizas, "Online distributed sparsity-aware canonical correlation analysis," *IEEE Trans. Signal Process.*, vol. 64, no. 3, pp. 688–703, Feb. 2016.

[13] B. Jiang, C. Ding, and J. Tang, "Graph-Laplacian PCA: Closed-form solution and robustness," in *Proc. of Intl. Conf. on Comput. Vision and Pattern Recognit.*, Portland, USA, Jun. 25-27, 2013.

[14] Y. Shen, P. Traganitis, and G. B. Giannakis, "Nonlinear dimensionality reduction on graphs," in *IEEE Intl. Wksp. Comput. Adv. in Multi-Sensor Adaptive Process.*, Curacao, Dutch Antilles, Dec. 10-13, 2017.

[15] N. Shahid, N. Perraudin, V. Kalofolias, G. Puy, and P. Vandergheynst, "Fast robust PCA on graphs," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 4, pp. 740–756, Feb. 2016.

[16] F. Shang, L. Jiao, and F. Wang, "Graph dual regularization non-negative matrix factorization for co-clustering," *Pattern Recognit.*, vol. 45, no. 6, pp. 2237–2250, Jun. 2012.

[17] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.

[18] M. B. Blaschko, J. A. Shelton, A. Bartels, C. H. Lampert, and A. Gretton, "Semi-supervised kernel canonical correlation analysis with application to human fMRI," *Pattern Recognit. Lett.*, vol. 32, no. 11, pp. 1572–1583, Aug. 2011.

[19] Y. Yuan and Q. Sun, "Graph regularized multiset canonical correlations with applications to joint feature extraction," *Pattern Recognit.*, vol. 47, no. 12, pp. 3907–3919, Dec. 2014.

[20] J. A. Brown, J. D. Rudie, A. Bandrowski, V. H. J. D., and S. Y. Bookheimer, "The UCLA multimodal connectivity database: A web-based platform for brain connectivity matrix sharing and analysis," *Front Neuroinform.*, vol. 6, p. 28, Nov. 2012.

[21] D. P. Bertsekas, *Nonlinear Programming*. 2nd ed., MA, USA: Athena Scientific, 1999.

[22] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Nov. 2006.

[23] A. J. Smola and R. Kondor, "Kernels and regularization on graphs," in *Learning Theory and Kernel Machines*. Springer, 2003, pp. 144–158.

[24] Y. Yamanishi, J.-P. Vert, A. Nakaya, and M. Kanehisa, "Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis," *Bioinformatics*, vol. 19, no. 1, pp. i323–i330, Jul. 2003.

[25] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, no. 3, pp. 337–404, 1950.

[26] F. R. Bach, G. R. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proc. of Intl. Conf. on Mach. Learn.*, New York, USA, Jul. 4-8, 2004.

[27] D. Romero, M. Ma, and G. B. Giannakis, "Kernel-based reconstruction of graph signals," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 764–778, Feb. 2017.

[28] A. R. Martinez and R. Benavente, "The AR face database, 1998," *Computer Vision Center, Technical Report*, vol. 3, p. 5, 2007.

[29] K. C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.