Graph Multiview Canonical Correlation Analysis

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

***Abstract*—Multiview canonical correlation analysis (MCCA) seeks latent low-dimensional representations encountered with multiview data of shared entities (a.k.a. common sources). However, existing MCCA approaches do not exploit the geometry of the common sources, which may be available *a priori*, or can be constructed using certain domain knowledge. This prior information about the common sources can be encoded by a graph, and be invoked as a regularizer to enrich the maximum variance MCCA framework. In this context, this paper's novel graph-regularized MCCA (GMCCA) approach minimizes the distance between the wanted canonical variables and the common low-dimensional representations, while accounting for graph-induced knowledge of the common sources. Relying on a function capturing the extent to which the low-dimensional representations of the multiple views are similar, a generalization bound of GMCCA is established based on Rademacher's complexity. Tailored for setups where the number of data pairs is smaller than the data vector dimensions, a graph-regularized dual MCCA approach is also developed. To further deal with nonlinearities present in the data, graph-regularized kernel MCCA variants are put forward too. Interestingly, solutions of the graph-regularized linear, dual, and kernel MCCA are all provided in terms of generalized eigenvalue decomposition. Several corroborating numerical tests using real datasets are provided to showcase the merits of the graph-regularized MCCA variants relative to several competing alternatives including MCCA, Laplacian-regularized MCCA, and (graph-regularized) PCA.**

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

I. Introduction

In SEVERAL applications, such as multi-sensor surveillance systems, multiple datasets are collected offering distinct views of the common information sources. With advances in data acquisition, it becomes easier to access heterogeneous data representing samples from multiple views in various scientific fields, including genetics, computer vision, data mining, and pattern recognition, to name a few. In genomics for instance, a patient's lymphoma data set consists of gene expression, SNP, and array CGH measurements [37]. In a journal's dataset, the title, keywords, and citations can be considered as three different views of a given paper [31]. Learning with heterogeneous data of different types is commonly referred to as multiview learning, and in different communities as information fusion or data integration from multiple feature sets. Multiview learning is an emerging field in data science with well-appreciated analytical tools and matching application domains [12], [30].

Canonical correlation analysis (CCA) is a classical tool for multiview learning [16]. Formally, CCA looks for latent low-dimensional representations from a paired dataset comprising two views of several common entities. Multiview (M) CCA generalizes two-view CCA and also principal component analysis (PCA) [19], to handle jointly datasets from multiple views [20]. In contrast to PCA that operates on vectors formed by multi-view sub-vectors, MCCA is more robust to outliers per view, because it ignores the principal components per view that are irrelevant to the latent common sources. Popular MCCA formulations include the sum of correlations (SUMCOR), maximum variance (MAXVAR) [15], sum of squared correlations, the minimum variance, and generalized variance methods [20]. With the increasing capacity of data acquisition and the growing demand for multiview data analytics, the research on MCCA has been re-gaining attention recently.

To capture nonlinear relationships in the data, extensions using (multi-)kernels and deep neural networks have also been developed; see e.g., [1], [10], [33], [35], [38], that have well-documented merits for (nonlinear) dimensionality reduction of multiview data, as well as for multiview feature extraction. Recent research efforts have also focused on addressing the scalability issues in (kernel) MCCA, using random Fourier features [22], or leveraging alternating optimization advances [18] to account for sparsity [8], [18], [32], [36] or other types of structure-promoting regularizers such as nonnegativity and smoothness [11], [23].

Lately, graph-aware regularizers have demonstrated promising performance in a gamut of machine learning applications, such as dimensionality reduction, data reconstruction, clustering, and classification [10], [13], [17], [25], [26], [28]. CCA with structural information induced by a common source graph has been reported in [10], but it is limited to analyzing two-views of data, and its performance has been tested only experimentally. Further, multigraph-encoded information provided by the underlying physics, or, inferred from alternative views of the information sources, has not been investigated.

Building on but considerably going beyond our precursor work in [10], this paper introduces a novel graph-regularized (G) MCCA approach, and develops a bound on its generalization error performance. Our GMCCA is established by minimizing the

Manuscript received November 26, 2018; revised March 9, 2019; accepted April 3, 2019. Date of publication April 11, 2019; date of current version April 26, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alexander Bertrand. This work was supported in part by the National Science Foundation under Grants 1500713, 1505970, 1514056, and 1711471. This paper was presented in part at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Brighton, U.K., May 2019 [7]. *(Corresponding author: Georgios B. Giannakis.)*

The authors are with the Digital Technology Center and the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: chen5625@umn.edu; gangwang@umn.edu; georgios@umn.edu).

Digital Object Identifier 10.1109/TSP.2019.2910475

difference between the low-dimensional representation of each view and the common representation, while also leveraging the statistical dependencies due to the common sources hidden in the views. These dependencies are encoded by a graph, which can be available from the given data, or can be deduced from correlations. A finite-sample statistical analysis of GMCCA is provided based on a regression formulation offering a meaningful error bound for unseen data samples using Rademacher's complexity.

GMCCA is operational when there are sufficient data samples (larger than the number of features per view). For cases where the data are insufficient, we develop a graph-regularized dual (GD) MCCA scheme that avoids this limitation at lower computational complexity. To cope with nonlinearities present in real data, we further put forward a graph-regularized kernel (GK) MCCA scheme. Interestingly, the linear, dual, and kernel versions of our proposed GMCCA admit simple analytical-form solutions, each of which can be obtained by performing a single generalized eigenvalue decomposition.

Different from [4], [39], where MCCA is regularized using multiple graph Laplacians separately per view, GMCCA here jointly leverages a single graph effected on the common sources. This is of major practical importance, e.g., in electric power networks, where besides the power, voltage, and current quantities observed, the system operator has also access to the network topology [34] that captures the connectivity between substations through power lines.

Finally, our proposed GMCCA approaches are numerically tested using several real datasets on different machine learning tasks, including e.g., dimensionality reduction, recommendation, clustering, and classification. Corroborating tests showcase the merits of GMCCA schemes relative to its completing alternatives such as MCCA, PCA, graph PCA, and the k-nearest neighbors (KNN) method.

Notation: Bold uppercase (lowercase) letters denote matrices (column vectors). Operators $\mathrm{Tr}(\cdot)$, $(\cdot)^{-1}$, $\mathrm{vec}(\cdot)$ and $(\cdot)^\top$ stand for matrix trace, inverse, vectorization, and transpose, respectively; $\|\cdot\|_2$ denotes the ℓ_2-norm of vectors; $\|\cdot\|_F$ the Frobenius norm of matrices; $\mathrm{diag}(\{a_m\}_{m=1}^M)$ is an $M\times M$ diagonal matrix holding entries of $\{a_m\}_{m=1}^M$ on its main diagonal; $\langle \mathbf{a},\ \mathbf{b}\rangle$ denotes the inner product of same-size vectors $\mathbf{a}$ and $\mathbf{b}$; vector $\mathbf{0}$ has all zero entries whose dimension is clear from the context; and $\mathbf{I}$ is the identity matrix of suitable size.

II. Preliminaries

Consider M datasets $\{\mathbf{X}_m\in\mathbb{R}^{D_m\times N}\}_{m=1}^M$ collected from $M\geq 2$ views of N common source vectors $\{\check{\mathbf{s}}_n\in\mathbb{R}^\rho\}_{n=1}^N$ stacked as columns of $\check{\mathbf{S}}\in\mathbb{R}^{\rho\times N}$, where D_m is the dimension of the m-th view data vectors, with possibly $\rho\ll \min_m\{D_m\}_{m=1}^M$. Vector $\mathbf{x}_{m,i}$ denotes the i-th column of $\mathbf{X}_m$, meaning the i-th datum of the m-th view, for all $i=1,\ldots,N$ and $m=1,\ldots,M$. Suppose without loss of generality that all per-view data vectors $\{\mathbf{x}_{m,i}\}_{i=1}^N$ have been centered. Two-view CCA works with datasets $\{\mathbf{x}_{1,i}\}_{i=1}^N$ and $\{\mathbf{x}_{2,i}\}_{i=1}^N$ from $M=2$ views. It looks for low-dimensional subspaces $\mathbf{U}_1\in\mathbb{R}^{D_1\times d}$ and $\mathbf{U}_2\in\mathbb{R}^{D_2\times d}$ with $d\leq\rho$, such that the Euclidean distance between linear projections $\mathbf{U}_1^\top\mathbf{X}_1$ and $\mathbf{U}_2^\top\mathbf{X}_2$ is minimized. Concretely, classical CCA solves the following problem [14]

$$\min_{\mathbf{U}_1,\mathbf{U}_2}\quad \left\|\mathbf{U}_1^\top\mathbf{X}_1-\mathbf{U}_2^\top\mathbf{X}_2\right\|_F^2 \tag{1a}$$

$$\text{s. to}\qquad \mathbf{U}_m^\top\left(\mathbf{X}_m\mathbf{X}_m^\top\right)\mathbf{U}_m=\mathbf{I},\quad m=1,\,2 \tag{1b}$$

where columns of $\mathbf{U}_m$ are called loading vectors of the data (view) $\mathbf{X}_m$; while projections $\{\mathbf{U}_m^\top\mathbf{X}_m\}_{m=1}^2$ are termed canonical variables; they satisfy (1b) to prevent the trivial solution; and, they can be viewed as low (d)-dimensional approximations of $\check{\mathbf{S}}$. Moreover, the solution of (1) is provided by a generalized eigenvalue decomposition [16].

When analyzing multiple (≥ 3) datasets, (1) can be generalized to a pairwise matching criterion [6]; that is

$$\min_{\{\mathbf{U}_m\}_{m=1}^M}\quad \sum_{m=1}^{M-1}\sum_{m'>m}^{M}\left\|\mathbf{U}_m^\top\mathbf{X}_m-\mathbf{U}_{m'}^\top\mathbf{X}_{m'}\right\|_F^2 \tag{2a}$$

$$\text{s. to}\qquad \mathbf{U}_m^\top\left(\mathbf{X}_m\mathbf{X}_m^\top\right)\mathbf{U}_m=\mathbf{I},\quad m=1,\ldots,M \tag{2b}$$

where (2b) ensures a unique nontrivial solution. The formulation in (2) is referred to as the sum-of-correlations (SUMCOR) MCCA, that is known to be NP-hard in general [24].

Instead of minimizing the distance between paired low-dimensional approximations, one can look for a shared low-dimensional representation of different views, namely $\mathbf{S}\in\mathbb{R}^{d\times N}$, by solving [20]

$$\min_{\{\mathbf{U}_m\}_{m=1}^M,\mathbf{S}}\quad \sum_{m=1}^{M}\left\|\mathbf{U}_m^\top\mathbf{X}_m-\mathbf{S}\right\|_F^2 \tag{3a}$$

$$\text{s. to}\qquad \mathbf{S}\mathbf{S}^\top=\mathbf{I} \tag{3b}$$

yielding the so-called maximum-variance (MAXVAR) MCCA formulation. Similarly, the constraint (3b) is imposed to avoid a trivial solution. If all per-view sample covariance matrices $\{\mathbf{X}_m\mathbf{X}_m^\top\}_m$ have full rank, then for a fixed $\mathbf{S}$, the $\mathbf{U}_m$-minimizers are given by $\{\hat{\mathbf{U}}_m=(\mathbf{X}_m\mathbf{X}_m^\top)^{-1}\mathbf{X}_m\mathbf{S}^\top\}_m$. Substituting $\{\hat{\mathbf{U}}_m\}_m$ into (3), the $\mathbf{S}$-minimizer can be obtained by solving the following eigenvalue decomposition problem

$$\hat{\mathbf{S}}:=\arg\max_{\mathbf{S}}\ \ \mathrm{Tr}\Big[\mathbf{S}\Big(\sum_{m=1}^{M}\mathbf{X}_m^\top\left(\mathbf{X}_m\mathbf{X}_m^\top\right)^{-1}\mathbf{X}_m\Big)\mathbf{S}^\top\Big] \tag{4a}$$

$$\text{s. to}\ \ \mathbf{S}\mathbf{S}^\top=\mathbf{I}. \tag{4b}$$

The columns of $\hat{\mathbf{S}}^\top$ are given by the first d principal eigenvectors of matrix $\sum_{m=1}^M\mathbf{X}_m^\top(\mathbf{X}_m\mathbf{X}_m^\top)^{-1}\mathbf{X}_m$. In turn, we deduce that $\{\hat{\mathbf{U}}_m=(\mathbf{X}_m\mathbf{X}_m^\top)^{-1}\mathbf{X}_m\hat{\mathbf{S}}^\top\}_{m=1}^M$.

A couple of comments are worth noting about (3) and (4).

Remark 1: Solutions of the SUMCOR MCCA in (2) and the MAXVAR MCCA in (3) are generally different. Specifically, for $M=2$, both admit analytical solutions that can be expressed in terms of distinct eigenvalue decompositions; but for $M>2$, the SUMCOR MCCA can not be solved analytically, while the MAXVAR MCCA still admits an analytical solution though at the price of higher computational complexity because it involves the extra matrix variable $\mathbf{S}$.

III. Graph-Regularized MCCA

In many applications, the common source vectors $\{\check{\mathbf{s}}_i\}_{i=1}^N$ may reside on, or their dependencies form a graph of N nodes. In ResearchIndex for example, networks, besides keywords, titles, Abstracts, and Introductions of collected articles, one has also access to the citation network capturing the connectivity among those papers. More generally, the graph of interdependent sources can be dictated by the underlying physics, or it can be a prior provided by an 'expert,' or, it can be learned from extra (e.g., historical) views of the data. This structural prior information can be leveraged along with multiview datasets to improve MCCA performance. Specifically, we will capture this extra knowledge here using a graph, and effect it in the low-dimensional common source estimates through a graph regularization term.

Consider representing the graph of the N common sources using the tuple $\mathcal{G} := \{\mathcal{N}, \mathcal{W}\}$, where $\mathcal{N} := \{1, \ldots, N\}$ is the vertex set, and $\mathcal{W} := \{w_{ij}\}_{(i,j)\in\mathcal{N}\times\mathcal{N}}$ collects all edge weights $\{w_{ij}\}$ over all vertex pairs (i, j). The so-termed weighted adjacency matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$ is formed with w_{ij} being its (i, j)-th entry. Undirected graphs for which $\mathbf{W} = \mathbf{W}^\top$ holds are considered in this work. Upon defining $d_i := \sum_{j=1}^N w_{ij}$ and $\mathbf{D} := \mathrm{diag}(\{d_i\}_{i=1}^N) \in \mathbb{R}^{N\times N}$, the Laplacian matrix of graph $\mathcal{G}$ is defined as

$$\mathbf{L}_\mathcal{G} := \mathbf{D} - \mathbf{W}. \tag{5}$$

Next, a neat link between canonical correlations and graph regularization will be elaborated. To start, let us assume that sources $\{\check{\mathbf{s}}_i\}_{i=1}^N$ are smooth over $\mathcal{G}$. This means that two sources $(\check{\mathbf{s}}_i, \check{\mathbf{s}}_j)$ residing on two connected nodes $i, j \in \mathcal{N}$ are also close to each other in Euclidean distance. As explained before, vectors $\mathbf{s}_i$ and $\mathbf{s}_j$ are accordingly the d-dimensional approximations of $\check{\mathbf{s}}_i$ and $\check{\mathbf{s}}_j$. Accounting for this fact, a meaningful regularizer is the weighted sum of distances between any pair of common source estimates $\mathbf{s}_i$ and $\mathbf{s}_j$ over $\mathcal{G}$

$$\mathrm{Tr}\big(\mathbf{S}\mathbf{L}_\mathcal{G}\mathbf{S}^\top\big) = \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N w_{ij}\,\|\mathbf{s}_i - \mathbf{s}_j\|_2^2. \tag{6}$$

Clearly, source vectors $\mathbf{s}_i$ and $\mathbf{s}_j$ residing on adjacent nodes $i, j \in \mathcal{N}$ having large weights w_{ij} will be forced to be similar to each other. To leverage such additional graph information of the common sources, the quadratic term (6) is invoked as a regularizer in the standard MAXVAR MCCA, yielding our novel graph-regularized (G) MCCA formulation

$$\min_{\{\mathbf{U}_m\}\atop\mathbf{S}} \quad \sum_{m=1}^M \big\|\mathbf{U}_m^\top\mathbf{X}_m - \mathbf{S}\big\|_F^2 + \gamma\mathrm{Tr}\left(\mathbf{S}\mathbf{L}_\mathcal{G}\mathbf{S}^\top\right) \tag{7a}$$

$$\text{s. to} \quad \mathbf{S}\mathbf{S}^\top = \mathbf{I} \tag{7b}$$

where the coefficient $\gamma \geq 0$ trades off minimizing the distance between the canonical variables and their corresponding common source estimates with promoting smoothness of common source estimates over the graph $\mathcal{G}$. Specifically, when $\gamma = 0$, GMCCA reduces to the classical MCCA in (3); and, as γ increases, GMCCA relies more heavily in this extra graph knowledge when finding the canonical variables.

Algorithm 1: Graph-regularized MCCA.

1: **Input:** $\{\mathbf{X}_m\}_{m=1}^M$, d, γ, and $\mathbf{W}$.
2: **Build** $\mathbf{L}_\mathcal{G}$ using (5).
3: **Construct** $\mathbf{C} = \sum_{m=1}^M \mathbf{X}_m^\top\left(\mathbf{X}_m\mathbf{X}_m^\top\right)^{-1}\mathbf{X}_m - \gamma\mathbf{L}_\mathcal{G}$.
4: **Perform** eigendecomposition on $\mathbf{C}$ to obtain the d eigenvectors associated with the d largest eigenvalues, which are collected as columns of $\hat{\mathbf{S}}^\top$.
5: **Compute** $\left\{\hat{\mathbf{U}}_m = \left(\mathbf{X}_m\mathbf{X}_m^\top\right)^{-1}\mathbf{X}_m\hat{\mathbf{S}}^\top\right\}_{m=1}^M$.
6: **Output:** $\{\hat{\mathbf{U}}_m\}_{m=1}^M$ and $\hat{\mathbf{S}}$.

If all per-view sample covariance matrices $\{\mathbf{X}_m\mathbf{X}_m^\top\}$ have full rank, equating to zero the partial derivative of the cost in (7a) with respect to each $\mathbf{U}_m$, yields the optimizer $\hat{\mathbf{U}}_m = (\mathbf{X}_m\mathbf{X}_m^\top)^{-1}\mathbf{X}_m\mathbf{S}^\top$. Substituting next $\mathbf{U}_m$ by $\hat{\mathbf{U}}_m$ and ignoring the constant term in (7a) give rise to the following eigenvalue problem (cf. (4))

$$\max_{\mathbf{S}} \quad \mathrm{Tr}\Big[\mathbf{S}\Big(\sum_{m=1}^M \mathbf{X}_m^\top\big(\mathbf{X}_m\mathbf{X}_m^\top\big)^{-1}\mathbf{X}_m - \gamma\mathbf{L}_\mathcal{G}\Big)\mathbf{S}^\top\Big] \tag{8a}$$

$$\text{s. to} \quad \mathbf{S}\mathbf{S}^\top = \mathbf{I}. \tag{8b}$$

Similar to standard MCCA, the optimal solution $\hat{\mathbf{S}}$ of (8) can be obtained by the d leading eigenvectors of the matrix

$$\mathbf{C} := \sum_{m=1}^M \mathbf{X}_m^\top(\mathbf{X}_m\mathbf{X}_m^\top)^{-1}\mathbf{X}_m - \gamma\mathbf{L}_\mathcal{G}. \tag{9}$$

At the optimum, it is easy to verify that the following holds

$$\sum_{m=1}^M \Big\|\hat{\mathbf{U}}_m^\top\mathbf{X}_m - \hat{\mathbf{S}}\Big\|_F^2 + \gamma\mathrm{Tr}\big(\hat{\mathbf{S}}\mathbf{L}_\mathcal{G}\hat{\mathbf{S}}^\top\big) = Md - \sum_{i=1}^d \lambda_i$$

where λ_i denotes the i-th largest eigenvalue of $\mathbf{C}$ in (9).

A step-by-step description of our proposed GMCCA scheme is summarized in Algorithm 1.

At this point, a few remarks are in order.

Remark 2: We introduced a two-view graph CCA scheme in [10] using the SUMCOR MCCA formulation. However, to obtain an analytical solution, the original cost was surrogated in [10] by its lower bound, which cannot be readily generalized for multiview datasets with $M \geq 3$. In contrast, our GMCCA in (7) can afford an analytical solution for any $M \geq 2$.

Remark 3: Generally speaking, when $N \gg D_m$, it is likely that $\mathbf{X}_m\mathbf{X}_m^\top$ has full rank. Even if it is not invertible, one can replace $\mathbf{X}_m\mathbf{X}_m^\top$ with $\mathbf{X}_m\mathbf{X}_m^\top + c_m\mathbf{I}$ to ensure invertibility, where $c_m > 0$ is a small arbitrary constant.

Remark 4: Different from our single graph regularizer in (7), the proposals in [4] and [39] rely on M different regularizers $\{\mathbf{U}_m^\top\mathbf{X}_m\mathbf{L}_{\mathcal{G}_m}\mathbf{X}_m^\top\mathbf{U}_m\}_m$ to exploit the extra graph knowledge, for view-specific graphs $\{\mathbf{L}_{\mathcal{G}_m}\}_m$ on data $\{\mathbf{X}_m\}_m$. However, the formulation in [39] does not admit an analytical solution,

and convergence of the iterative solvers for the resulting nonconvex problem can be guaranteed only to a stationary point. The approach in [4] focuses on semi-supervised learning tasks, in which cross-covariances of pair-wise datasets are not fully available. In contrast, the single graph Laplacian regularizer in (7) is effected on the common sources, to exploit the pair-wise similarities of the N common sources. This is of practical importance when one has prior knowledge about the common sources besides the M datasets. Moreover, our proposed GMCCA approach comes with simple analytical solutions.

Remark 5: With regards to selecting γ, two ways are feasible: i) cross-validation for supervised learning tasks, where labeled training data are given, and γ is fixed to the one that yields optimal empirical performance on the training data; and, ii) using a spectral clustering method that automatically chooses the best γ values from a given set of candidates; see e.g., [9].

IV. Generalization Bound of GMCCA

In this section, we will analyze the finite-sample performance of GMCCA based on a regression formulation [27, Ch. 6.5], which is further related to the alternating conditional expectations method in [5]. Our analysis will establish an error bound for unseen source vectors (a.k.a. generalization bound) using the notion of Rademacher's complexity.

Recall that the goal of MCCA is to find common low-dimensional representations of the M-view data. To measure how close the estimated M low-dimensional representations are to each other, we introduce the following error function

$$g(\check{\mathbf{s}}) := \sum_{m=1}^{M-1}\sum_{m'>m}^{M} \left\|\mathbf{U}_m^\top \boldsymbol{\psi}_m(\check{\mathbf{s}}) - \mathbf{U}_{m'}^\top \boldsymbol{\psi}_{m'}(\check{\mathbf{s}})\right\|_F^2 \tag{10}$$

where the underlying source vector $\check{\mathbf{s}} \in \mathbb{R}^\rho$ is assumed to follow some fixed yet unknown distribution $\mathcal{D}$, and the linear function $\boldsymbol{\psi}_m(\cdot)$ maps a source vector from space $\mathbb{R}^\rho$ to the m-the view in $\mathbb{R}^{D_m}$, for $m = 1, \ldots, M$.

To derive the generalization bound, we start by evaluating the empirical average of $g(\check{\mathbf{s}})$ over say, a number N of given training samples, as follows

$$\begin{aligned}\bar{g}_N(\check{\mathbf{s}}) &:= \frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M-1}\sum_{m'>m}^{M} \left\|\mathbf{U}_m^\top \boldsymbol{\psi}_m(\check{\mathbf{s}}_n) - \mathbf{U}_{m'}^\top \boldsymbol{\psi}_{m'}(\check{\mathbf{s}}_n)\right\|_F^2 \\ &= \frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M-1}\sum_{m'>m}^{M} \Big[\boldsymbol{\psi}_m^\top(\check{\mathbf{s}}_n)\mathbf{U}_m\mathbf{U}_m^\top\boldsymbol{\psi}_m(\check{\mathbf{s}}_n) - 2\boldsymbol{\psi}_m^\top(\check{\mathbf{s}}_n) \\ &\quad \times \mathbf{U}_m\mathbf{U}_{m'}^\top\boldsymbol{\psi}_{m'}(\check{\mathbf{s}}_n) + \boldsymbol{\psi}_{m'}^\top(\check{\mathbf{s}}_n)\mathbf{U}_{m'}\mathbf{U}_{m'}^\top\boldsymbol{\psi}_{m'}(\check{\mathbf{s}}_n)\Big].\end{aligned}$$

For the quadratic terms, it can be readily verified that

$$\begin{aligned}\boldsymbol{\psi}_m^\top(\check{\mathbf{s}}) \quad \mathbf{U}_m\mathbf{U}_m^\top\boldsymbol{\psi}_m(\check{\mathbf{s}}) &= \big\langle \mathrm{vec}(\mathbf{U}_m\mathbf{U}_m^\top), \\ &\quad \mathrm{vec}(\boldsymbol{\psi}_m(\check{\mathbf{s}})\boldsymbol{\psi}_m^\top(\check{\mathbf{s}}))\big\rangle \qquad (11)\\ \boldsymbol{\psi}_m^\top(\check{\mathbf{s}}) \quad \mathbf{U}_m\mathbf{U}_{m'}^\top\boldsymbol{\psi}_{m'}(\check{\mathbf{s}}) &= \big\langle \mathrm{vec}(\mathbf{U}_m\mathbf{U}_{m'}^\top), \\ &\quad \mathrm{vec}(\boldsymbol{\psi}_m(\check{\mathbf{s}})\boldsymbol{\psi}_{m'}^\top(\check{\mathbf{s}}))\big\rangle. \qquad (12)\end{aligned}$$

Define two $\sum_{m=1}^{M-1}\sum_{m'>m}^{M}(D_m^2 + D_{m'}^2 + D_mD_{m'}) \times 1$ vectors

$$\begin{aligned}\boldsymbol{\psi}(\check{\mathbf{s}}) &:= \left[\boldsymbol{\psi}_{11}^\top(\check{\mathbf{s}}) \cdots \boldsymbol{\psi}_{1M}^\top(\check{\mathbf{s}})\boldsymbol{\psi}_{23}^\top(\check{\mathbf{s}}) \cdots \boldsymbol{\psi}_{M,M-1}^\top(\check{\mathbf{s}})\right]^\top \\ \mathbf{u} &:= \left[\mathbf{u}_{11}^\top \cdots \mathbf{u}_{1M}^\top\ \mathbf{u}_{23}^\top \cdots \mathbf{u}_{M,M-1}^\top\right]^\top\end{aligned}$$

where the two $(D_m^2 + D_{m'}^2 + D_mD_{m'}) \times 1$ vectors $\boldsymbol{\psi}_{mm'}(\check{\mathbf{s}})$ and $\mathbf{u}_{mm'}$ are defined as

$$\begin{aligned}\boldsymbol{\psi}_{mm'} &:= \left[\mathrm{vec}^\top(\boldsymbol{\psi}_m\boldsymbol{\psi}_m^\top)\mathrm{vec}^\top(\boldsymbol{\psi}_{m'}\boldsymbol{\psi}_{m'}^\top)\sqrt{2}\mathrm{vec}^\top(\boldsymbol{\psi}_m\boldsymbol{\psi}_{m'}^\top)\right]^\top \\ \mathbf{u}_{mm'} &:= \left[\mathrm{vec}^\top(\mathbf{U}_m\mathbf{U}_m^\top)\mathrm{vec}^\top(\mathbf{U}_{m'}\mathbf{U}_{m'}^\top)\right. \\ &\quad \left. - \sqrt{2}\mathrm{vec}^\top(\mathbf{U}_m\mathbf{U}_{m'}^\top)\right]^\top\end{aligned}$$

for $m = 1, \ldots, M-1$ and $m' = 2, \ldots, M$.

Plugging (11) and (12) into (10), one can check that function $g(\check{\mathbf{s}})$ can be rewritten as

$$g(\check{\mathbf{s}}) = \langle \mathbf{u},\, \boldsymbol{\psi}(\check{\mathbf{s}})\rangle\,. \tag{13}$$

with the norm of $\mathbf{u}$ given by

$$\|\mathbf{u}\|_2^2 = \sum_{m=1}^{M-1}\sum_{m'>m}^{M} \left\|\mathbf{U}_m^\top\mathbf{U}_m + \mathbf{U}_{m'}^\top\mathbf{U}_{m'}\right\|_F^2.$$

Starting from (13), we will establish next an upper bound on the expectation of $g(\check{\mathbf{s}})$ by means of (13), which is important because the expectation involves not only the N training source samples, but also unseen samples.

Theorem 1: Assume that i) the N common source vectors $\{\check{\mathbf{s}}_n\}_{n=1}^N$ are drawn i.i.d. from some distribution $\mathcal{D}$; ii) the M transformations $\{\boldsymbol{\psi}_m(\cdot)\}_{m=1}^M$ of vectors $\{\check{\mathbf{s}}_n\}_{n=1}^N$ are bounded; and, iii) subspaces $\{\mathbf{U}_m \in \mathbb{R}^{D_m\times d}\}_{m=1}^M$ satisfy $\sum_{m=1}^{M-1}\sum_{m'>m}^{M}\|\mathbf{U}_m^\top\mathbf{U}_m + \mathbf{U}_{m'}^\top\mathbf{U}_{m'}\|_F^2 \le B^2$ $(B>0)$ and $\{\mathbf{U}_m\}_{m=1}^M$ are the optimizers of (7). If we obtain low-dimensional representations of $\{\boldsymbol{\psi}_m(\check{\mathbf{s}})\}_{m=1}^M$ specified by subspaces $\{\mathbf{U}_m \in \mathbb{R}^{D_m\times d}\}_{m=1}^M$, it holds with probability at least $1-p$ that

$$\begin{aligned}\mathbb{E}[g(\check{\mathbf{s}})] &\le \bar{g}_N(\check{\mathbf{s}}) + 3RB\sqrt{\frac{\ln(2/p)}{2N}} + \frac{4B}{N} \\ &\quad \sqrt{\sum\nolimits_{n=1}^{N}\sum\nolimits_{m=1}^{M-1}\sum\nolimits_{m'>m}^{M}\left[\kappa_m(\check{\mathbf{s}}_n,\,\check{\mathbf{s}}_n) + \kappa_{m'}(\check{\mathbf{s}}_n,\,\check{\mathbf{s}}_n)\right]^2}\end{aligned} \tag{14}$$

where $\kappa_m(\check{\mathbf{s}}_n, \check{\mathbf{s}}_n) := \langle \boldsymbol{\psi}_m(\check{\mathbf{s}}_n), \boldsymbol{\psi}_m(\check{\mathbf{s}}_n)\rangle$ for $n = 1, \ldots, N$, and $m = 1, \ldots, M$, while the constant R is given by

$$R := \max_{\check{\mathbf{s}}\sim\mathcal{D}}\sqrt{\sum\nolimits_{m=1}^{M-1}\sum\nolimits_{m'>m}^{M}\left[\kappa_m(\check{\mathbf{s}},\check{\mathbf{s}}) + \kappa_{m'}(\check{\mathbf{s}},\check{\mathbf{s}})\right]^2}.$$

Proof: Equation (13) suggests that $g(\check{\mathbf{s}})$ belongs to the function class

$$\mathcal{F}_B := \{\check{\mathbf{s}} \to \langle \mathbf{u},\, \boldsymbol{\psi}(\check{\mathbf{s}})\rangle : \|\mathbf{u}\| \le B\}\,.$$

Consider the function class

$$\mathcal{H} = \left\{h : \check{\mathbf{s}} \to 1/(RB)f(\check{\mathbf{s}}) \middle| f(\cdot) \in \mathcal{F}_B\right\} \subseteq \mathcal{A} \circ \mathcal{F}_B$$

where the function $\mathcal{A}$ is defined as

$$\mathcal{A}(x) = \begin{cases} 0, & \text{if } x \leq 0 \\ \frac{x}{RB}, & \text{if } 0 \leq x \leq RB \\ 1, & \text{otherwise} \end{cases}.$$

It can be checked that $\mathcal{A}(\cdot)$ is a Lipschitz function with Lipschitz constant $1/(RB)$, and that the range of functions in $\mathcal{H}$ is $[0,\, 1]$. Appealing to [27, Th. 4.9], one deduces that with probability at least $1-p$, the following holds

$$\begin{aligned} \mathbb{E}[h(\check{\mathbf{s}})] &\leq \frac{1}{N}\sum_{n=1}^{N} h(\mathbf{s}_n) + R_N(\mathcal{H}) + \sqrt{\frac{\ln 2/p}{2N}} \\ &\leq \frac{1}{N}\sum_{n=1}^{N} h(\check{\mathbf{s}}_n) + \hat{R}_N(\mathcal{H}) + 3\sqrt{\frac{\ln 2/p}{2N}} \end{aligned} \tag{15}$$

where $\mathbb{E}[h(\check{\mathbf{s}})]$ denotes the expected value of $h(\cdot)$ on a new common source $\check{\mathbf{s}}$; and the Rademacher complexity $R_N(\mathcal{H})$ of $\mathcal{H}$ along with its empirical version $\hat{R}_N(\mathcal{H})$ is defined as

$$R_N(\mathcal{H}) := \mathbb{E}_{\check{\mathbf{s}}}[\hat{R}_N(\mathcal{H})]$$

$$\hat{R}_N(\mathcal{H}) := \mathbb{E}_{\boldsymbol{\delta}}\left[\sup_{h\in\mathcal{H}}\left|\frac{2}{N}\sum_{n=1}^{N}\delta_n h(\check{\mathbf{s}}_n)\right| \,|\check{\mathbf{s}}_1,\, \check{\mathbf{s}}_2,\, \ldots,\, \check{\mathbf{s}}_N\right]$$

where $\boldsymbol{\delta} := \{\delta_n\}_{n=1}^N$ collects independent random variables drawn from the Rademacher distribution, meaning $\{\Pr(\delta_n = 1) = \Pr(\delta_n = -1) = 0.5\}_{n=1}^N$. Further, $\mathbb{E}_{\boldsymbol{\delta}}[\cdot]$ and $\mathbb{E}_{\check{\mathbf{s}}}[\cdot]$ denote the expectation with respect to $\boldsymbol{\delta}$ and $\check{\mathbf{s}}$, respectively.

Since $\mathcal{A}(\cdot)$ is a Lipschitz function with Lipschitz constant $1/(RB)$ satisfying $\mathcal{A}(0) = 0$, the result in [2, Th. 12] asserts that

$$\hat{R}_N(\mathcal{H}) \leq 2/(RB)\hat{R}_N(\mathcal{F}_B). \tag{16}$$

Applying [27, Th. 4.12] leads to

$$\hat{R}_N(\mathcal{F}_B) \leq 2B/N\sqrt{\mathrm{Tr}(\mathbf{K})} \tag{17}$$

where the $(i,\, j)$-th entry of $\mathbf{K} \in \mathbb{R}^{N\times N}$ is $\langle \boldsymbol{\psi}(\check{\mathbf{s}}_i), \boldsymbol{\psi}(\check{\mathbf{s}}_j)\rangle$, for $i,\, j = 1, \ldots, N$. One can also confirm that

$$\mathrm{Tr}(\mathbf{K}) = \sum_{n=1}^{N}\sum_{m=1}^{M-1}\sum_{m'>m}^{M}\left[\kappa_m(\check{\mathbf{s}}_n, \check{\mathbf{s}}_n) + \kappa_{m'}(\check{\mathbf{s}}_n, \check{\mathbf{s}}_n)\right]^2. \tag{18}$$

Substituting (17) and (18) to (16) yields

$$\hat{R}_N(\mathcal{H}) \leq \frac{4}{RN} \sqrt{\sum\nolimits_{n=1}^{N}\sum\nolimits_{m=1}^{M-1}\sum\nolimits_{m'>m}^{M}\left[\kappa_m(\check{\mathbf{s}}_n, \check{\mathbf{s}}_n) + \kappa_{m'}(\check{\mathbf{s}}_n, \check{\mathbf{s}}_n)\right]^2}.$$

Multiplying (15) by RB along with the last equation gives rise to (14). ■

Theorem 1 confirms that the empirical expectation of $g(\cdot)$, namely $\bar{g}_N(\check{\mathbf{s}})$, stays close to its ensemble one $\mathbb{E}(g(\check{\mathbf{s}}))$, provided that $\{\|\mathbf{U}_m\|_F\}_m$ can be controlled. For this reason, it is prudent to trade off maximization of correlations among the M datasets with the norms of the resultant loading vectors.

V. Graph-Regularized Dual MCCA

In practical scenarios involving high-dimensional data vectors with dimensions satisfying $\min_m D_m > N$, the matrices $\{\mathbf{X}_m\mathbf{X}_m^\top\}$ become singular – a case where GMCCA in (7) does not apply. For such cases, consider rewriting the $D_m \times d$ loading matrices $\mathbf{U}_m$ in terms of the data matrices $\mathbf{X}_m$ as $\mathbf{U}_m = \mathbf{X}_m\mathbf{A}_m$, where $\mathbf{A}_m \in \mathbb{R}^{N\times d}$ will be henceforth termed the dual of $\mathbf{U}_m$. Replacing $\mathbf{U}_m$ with $\mathbf{X}_m\mathbf{A}_m$ in the linear GMCCA formulation (7) leads to its dual formulation

$$\min_{\{\mathbf{A}_m\},\mathbf{S}} \quad \sum_{m=1}^{M}\left\|\mathbf{A}_m^\top\mathbf{X}_m^\top\mathbf{X}_m - \mathbf{S}\right\|_F^2 + \gamma\mathrm{Tr}\left(\mathbf{S}\mathbf{L}_{\mathcal{G}}\mathbf{S}^\top\right) \tag{19a}$$

$$\text{s. to} \qquad \mathbf{S}\mathbf{S}^\top = \mathbf{I}. \tag{19b}$$

If the $N \times N$ matrices $\{\mathbf{X}_m^\top\mathbf{X}_m\}_{m=1}^M$ are nonsingular, it can be readily confirmed that the $d \leq \rho$ columns of the optimizer $\hat{\mathbf{S}}^\top$ of (19) are the d principal eigenvectors of $M\mathbf{I} - \gamma\mathbf{L}_{\mathcal{G}}$, while the dual matrices can be estimated in closed form as $\hat{\mathbf{A}}_m = (\mathbf{X}_m^\top\mathbf{X}_m)^{-1}\hat{\mathbf{S}}^\top$. Clearly, such an $\hat{\mathbf{S}}$ does not depend on the data $\{\mathbf{X}_m\}_{m=1}^M$, and this estimate goes against our goal of extracting $\hat{\mathbf{S}}$ as the latent low-dimensional structure commonly present in $\{\mathbf{X}_m\}_{m=1}^M$. To address this issue, we mimic the dual CCA trick (see e.g., [14]), and introduce a Tikhonov regularization term on the loading vectors through the norms of $\left\{\|\mathbf{U}_m\|_F^2 = \mathrm{Tr}\left(\mathbf{A}_m^\top\mathbf{X}_m^\top\mathbf{X}_m\mathbf{A}_m\right)\right\}$. This indeed agrees with the observation we made following Theorem 1 that controlling $\{\|\mathbf{U}_m\|_F^2\}$ improves the generalization. In a nutshell, our graph-regularized dual (GD) MCCA is given as

$$\begin{aligned} \min_{\{\mathbf{A}_m\},\mathbf{S}} \quad & \sum_{m=1}^{M}\left\|\mathbf{A}_m^\top\mathbf{X}_m^\top\mathbf{X}_m - \mathbf{S}\right\|_F^2 + \gamma\mathrm{Tr}\left(\mathbf{S}\mathbf{L}_{\mathcal{G}}\mathbf{S}^\top\right) \\ & + \sum_{m=1}^{M}\epsilon_m\mathrm{Tr}\left(\mathbf{A}_m^\top\mathbf{X}_m^\top\mathbf{X}_m\mathbf{A}_m\right) \end{aligned} \tag{20a}$$

$$\text{s. to} \qquad \mathbf{S}\mathbf{S}^\top = \mathbf{I}. \tag{20b}$$

where $\{\epsilon_m > 0\}$ denote pre-selected weight coefficients.

As far as the solution is concerned, it can be deduced that the i-th column of the optimizer $\hat{\mathbf{S}}$ of (20) is the eigenvector of $\mathbf{C}_d := \sum_{m=1}^M(\mathbf{X}_m^\top\mathbf{X}_m + \epsilon_m\mathbf{I})^{-1}\mathbf{X}_m^\top\mathbf{X}_m - \gamma\mathbf{L}_{\mathcal{G}}$ associated with the i-th largest eigenvalue. Once $\hat{\mathbf{S}}$ is found, the optimal dual matrices can be obtained as $\{\hat{\mathbf{A}}_m = (\mathbf{X}_m^\top\mathbf{X}_m + \epsilon_m\mathbf{I})^{-1}\hat{\mathbf{S}}^\top\}_{m=1}^M$.Further, the optimal value of (20a) is $Md - \sum_{i=1}^d\lambda_i$, where λ_i is the i-th largest eigenvalue of $\mathbf{C}_d$. Yet, it is not clear whether this optimal cost increases or decreases with $\{\epsilon_m\}$. The steps of implementing GDMCCA are summarized in Alg. 2.

VI. Graph-Regularized Kernel MCCA

The GMCCA and GDMCCA approaches are limited to analyzing linear data dependencies. Nonetheless, complex nonlinear data dependencies are not rare in practice. To account for nonlinear dependencies, a graph-regularized kernel (GK)

Algorithm 2: Graph-regularized Dual MCCA.

1: **Input:** $\{\mathbf{X}_m\}_{m=1}^M$, $\{\epsilon_m\}_{m=1}^M$, γ, and $\mathbf{W}$.
2: **Build** $\mathbf{L}_{\mathcal{G}}$ using (5).
3: **Construct** $\mathbf{C}_d = \sum_{m=1}^M \left(\mathbf{X}_m^\top \mathbf{X}_m + \epsilon_m \mathbf{I}\right)^{-1} \mathbf{X}_m^\top \mathbf{X}_m - \gamma \mathbf{L}_{\mathcal{G}}$.
4: **Perform** eigenvalue decomposition on $\mathbf{C}_d$ to obtain the d eigenvectors associated with the d largest eigenvalues, which are collected as columns of $\hat{\mathbf{S}}^\top$.
5: **Compute** $\{\hat{\mathbf{A}}_m = \left(\mathbf{X}_m^\top \mathbf{X}_m + \epsilon_m \mathbf{I}\right)^{-1} \hat{\mathbf{S}}^\top\}_{m=1}^M$.
6: **Output:** $\{\hat{\mathbf{A}}_m\}_{m=1}^M$ and $\hat{\mathbf{S}}$.

Algorithm 3: Graph-regularized Kernel MCCA.

1: **Input:** $\{\mathbf{X}_m\}_{m=1}^M$, $\{\epsilon_m\}_{m=1}^M$, γ, $\mathbf{W}$, and $\{\kappa^m\}_{m=1}^M$.
2: **Construct** $\{\mathbf{K}_m\}_{m=1}^M$ using (21).
3: **Build** $\mathbf{L}_{\mathcal{G}}$ using (5).
4: **Form** $\mathbf{C}_g = \sum_{m=1}^M \left(\mathbf{K}_m + \epsilon_m \mathbf{I}\right)^{-1} \mathbf{K}_m - \gamma \mathbf{L}_{\mathcal{G}}$.
5: **Perform** eigendecomposition on $\mathbf{C}_g$ to obtain the d eigenvectors associated with the d largest eigenvalues, which are collected as columns of $\hat{\mathbf{S}}^\top$.
6: **Compute** $\{\hat{\mathbf{A}}_m = \left(\mathbf{K}_m + \epsilon_m \mathbf{I}\right)^{-1} \hat{\mathbf{S}}^\top\}_{m=1}^M$.
7: **Output:** $\{\hat{\mathbf{A}}_m\}_{m=1}^M$ and $\hat{\mathbf{S}}$.

MCCA formulation is pursued in this section to capture the nonlinear relationships in the M datasets $\{\mathbf{X}_m\}_m$ through kernel-based methods. Specifically, the idea of GKMCCA involves first mapping the data vectors $\{\mathbf{X}_m\}_m$ to higher (possibly infinite) dimensional feature vectors by means of M nonlinear functions, on which features we will apply GMCCA to find the shared low-dimensional canonical variables.

Let $\boldsymbol{\phi}_m$ be a mapping from $\mathbb{R}^{D_m}$ to $\mathbb{R}^{L_m}$ for all m, where the dimension L_m can be as high as infinity. Clearly, the data enter the GDMCCA problem (20) only via the similarity matrix $\mathbf{X}_m^\top \mathbf{X}_m$. Upon mapping all data vectors $\{\mathbf{x}_{m,i}\}_{i=1}^N$ into $\{\boldsymbol{\phi}_m(\mathbf{x}_{m,i})\}_{i=1}^N$, the linear similarities $\{\langle \mathbf{x}_{m,i}, \mathbf{x}_{m,j}\rangle\}_{i,j=1}^N$ can be replaced with the mapped nonlinear similarities $\{\langle \boldsymbol{\phi}_m(\mathbf{x}_{m,i}), \boldsymbol{\phi}_m(\mathbf{x}_{m,j})\rangle\}_{i,j=1}^N$. After selecting some kernel function κ^m such that $\kappa^m(\mathbf{x}_{m,i}, \mathbf{x}_{m,j}) := \langle \boldsymbol{\phi}_m(\mathbf{x}_{m,i}), \boldsymbol{\phi}_m(\mathbf{x}_{m,j})\rangle$, the (i, j)-th entry of the kernel matrix $\bar{\mathbf{K}}_m \in \mathbb{R}^{N\times N}$ is given by $\kappa^m(\mathbf{x}_{m,i}, \mathbf{x}_{m,j})$, for all i, j, and m. In the sequel, centering $\{\boldsymbol{\phi}_m(\mathbf{x}_{m,i})\}_{i=1}^N$ is realized by centering the kernel matrix for data $\mathbf{X}_m$ as

$$\mathbf{K}_m(i, j) := \bar{\mathbf{K}}_m(i,j) - \frac{1}{N}\sum_{k=1}^N \bar{\mathbf{K}}_m(k,j) - \frac{1}{N}\sum_{k=1}^N \bar{\mathbf{K}}_m(i,k) + \frac{1}{N^2}\sum_{i,j=1}^N \bar{\mathbf{K}}_m(i,j) \tag{21}$$

for $m = 1, \ldots, M$.

Replacing $\{\mathbf{X}_m^\top \mathbf{X}_m\}_m$ in the GDMCCA formulation (20) with centered kernel matrices $\{\mathbf{K}_m\}_m$ yields our GKMCCA

$$\min_{\{\mathbf{A}_m\},\mathbf{S}} \quad \sum_{m=1}^M \left\|\mathbf{A}_m^\top \mathbf{K}_m - \mathbf{S}\right\|_F^2 + \gamma \mathrm{Tr}\left(\mathbf{S}\mathbf{L}_{\mathcal{G}}\mathbf{S}^\top\right) + \sum_{m=1}^M \epsilon_m \mathrm{Tr}\left(\mathbf{A}_m^\top \mathbf{K}_m \mathbf{A}_m\right) \tag{22a}$$

$$\text{s. to} \quad \mathbf{S}\mathbf{S}^\top = \mathbf{I}. \tag{22b}$$

Selecting invertible matrices $\{\mathbf{K}_m\}_{m=1}^M$, and following the logic used to solve (20), we can likewise tackle (22). Consequently, the columns of the optimizer $\hat{\mathbf{S}}^\top$ are the first d principal eigenvectors of $\mathbf{C}_g := \sum_{m=1}^M (\mathbf{K}_m + \epsilon_m \mathbf{I})^{-1}\mathbf{K}_m - \gamma \mathbf{L}_{\mathcal{G}} \in \mathbb{R}^{N\times N}$, and the optimal $\hat{\mathbf{A}}_m$ sought can be obtained as $\hat{\mathbf{A}}_m = (\mathbf{K}_m + \epsilon_m \mathbf{I})^{-1}\hat{\mathbf{S}}^\top$. For implementation, GKMCCA is presented in step-by-step form as Algorithm 3.

Remark 6: When the (non)linear maps $\boldsymbol{\phi}_m(\cdot)$ needed to form the kernel matrices $\{\mathbf{K}_m\}_{m=1}^M$ in (22) are not given a priori, the multi-kernel methods are well motivated (see e.g., [29], [40]). Concretely, one presumes that each $\mathbf{K}_m$ is a linear combination of P kernel matrices, namely $\mathbf{K}_m = \sum_{p=1}^P \beta_m^p \mathbf{K}_m^p$, where $\{\mathbf{K}_m^p\}_{p=1}^P$ represent preselected view-specific kernel matrices for data $\mathbf{X}_m$. The unknown coefficients $\{\beta_m^p \geq 0\}_{m,p}$ are then jointly optimized with $\{\mathbf{A}_m\}_m$ and $\mathbf{S}$ in (22).

Remark 7: When more than one type of connectivity information on the common sources are available, our single graph-regularized MCCA schemes can be generalized to accommodate multiple or multi-layer graphs. Specifically, the single graph-based regularization term $\gamma \mathrm{Tr}(\mathbf{S}\mathbf{L}_{\mathcal{G}}\mathbf{S}^\top)$ in (7), (20), and (22) can be replaced with $\sum_{i=1}^I \gamma_i \mathrm{Tr}(\mathbf{S}\mathbf{L}_{\mathcal{G}i}\mathbf{S}^\top)$ with possibly unknown yet learnable coefficients $\{\gamma_i\}_i$, where $\mathbf{L}_{\mathcal{G}i}$ denotes the graph Laplacian matrix of the i-th graph, for $i = 1, \ldots, I$.

Remark 8: To obtain an out-of-sample extension of GKMCCA, one can center the new data vectors with respect to the existing data in their corresponding transformed feature spaces, and subsequently project the new centered data onto the learned subspaces. Consider the m-th view projection matrix $\bar{\mathbf{U}}_m := \bar{\boldsymbol{\Phi}}_m \mathbf{A}_m \in \mathbb{R}^{L_m \times d}$, where the i-th column of $\bar{\boldsymbol{\Phi}}_m \in \mathbb{R}^{L_m \times N}$ denotes the centered $\boldsymbol{\phi}_m(\mathbf{x}_{m,i})$ for $i = 1, \ldots, N$. Suppose that the T new data vectors are collected in $\mathbf{Z}_m := [\mathbf{z}_{m,1} \cdots \mathbf{z}_{m,T}]$. We first map $\{\mathbf{z}_{m,i}\}$ from $\mathbb{R}^{D_m}$ to $\mathbb{R}^{L_m}$ via $\boldsymbol{\phi}_m(\cdot)$, and subtract from $\{\boldsymbol{\phi}_m(\mathbf{z}_{m,i})\}$ the original mean $\boldsymbol{\mu}_m := \frac{1}{N}\sum_{i=1}^N \boldsymbol{\phi}_m(\mathbf{x}_{m,i})$. The low-dimensional representation of $\mathbf{Z}_m$ is $\bar{\mathbf{U}}_m^\top [\boldsymbol{\phi}_m(\mathbf{z}_{m,1}) - \boldsymbol{\mu}_m \cdots \boldsymbol{\phi}_m(\mathbf{z}_{m,T}) - \boldsymbol{\mu}_m] \in \mathbb{R}^{d\times T}$, which can be equivalently expressed as $\mathbf{A}_m^\top \mathbf{K}_{zx}$, where $\mathbf{K}_{zx} := \bar{\mathbf{K}}_{zx} - \frac{1}{N}\mathbf{1}_{T\times N}\bar{\mathbf{K}}_m - \frac{1}{N}\bar{\mathbf{K}}_{zx}\mathbf{1}_{N\times N} + \frac{1}{N^2}\mathbf{1}_{T\times N}\bar{\mathbf{K}}_m\mathbf{1}_{N\times N}$ with $\langle \boldsymbol{\phi}_m(\mathbf{z}_{m,i}), \boldsymbol{\phi}_m(\mathbf{x}_{m,j})\rangle$ specifying the (i, j)-th entry of $\bar{\mathbf{K}}_{zx}$ for $i = 1, \ldots, T$ and $j = 1, \ldots, N$, and $\mathbf{1}_{T\times N} \in \mathbf{R}^{T\times N}$ denoting the all-one matrix.

VII. Computational Complexity

Regarding computational complexity, recall that GMCCA, GDMCCA, GKMCCA, MCCA, DMCCA, and KMCCA all require finding the eigenvectors of matrices with different dimensionalities. Defining $D := \max_m D_m$, it can be checked that they incur correspondingly complexities $\mathcal{O}(N^2 \max(N, DM))$, $\mathcal{O}(N^2 DM)$, $\mathcal{O}(N^2 M \max(N, D))$, $\mathcal{O}(N^2 \max(N, DM))$, $\mathcal{O}(N^2 DM)$, and $\mathcal{O}(N^2 M \max(N, D))$. Interestingly, introducing graph-regularization to e.g., MCCA, DMCCA, as well

as KMCCA does not increase computational complexity. When $\{N \ll D_m\}_{m=1}^M$, GMCCA in its present form is not feasible, or suboptimal even though pseudo-inverse can be utilized at complexity $\mathcal{O}(MD^3)$. In contrast, GDMCCA is computationally preferable as its cost grows only linearly with D. When $N \gg D$, the complexity of GKMCCA is dominated by the computational burden of $\{(\mathbf{K}_m + \epsilon\mathbf{I})^{-1}\mathbf{K}_m\}_{m=1}^M$ requiring complexity in the order of $\mathcal{O}(N^3M)$. On the other hand, implementing GKMCCA when $N \ll D$ incurs complexity of order $\mathcal{O}(N^2MD)$, required to evaluate the M kernel matrices.

Performing PCA and graph (G) PCA [28] on the concatenated vectors of dimension $D_t := \sum_{m=1}^N D_m$ incurs computational complexity $\mathcal{O}(D_t^2 \max(D_t, N))$ and $\mathcal{O}(D_t \max(D_t^2, N^2))$, respectively. As such, in settings where $D_t \leq DM \leq N$, $D_t \leq N \leq DM$, or $N \leq D_t \leq DM$ and $N^2DM \geq D_t^3$, GMCCA is computationally heavier than (G)PCA; otherwise, it is computationally more affordable.

Our GMCCA, GDMCCA, and GKMCCA schemes entail eigendecomposition of an $N \times N$ matrix, which incurs complexity $\mathcal{O}(N^3)$, and thus is not scalable to large datasets. Possible remedies include parallelization and efficient decentralized algorithms capable of handling structured MCCA; e.g., along the lines of [18]. These go beyond the scope of the present paper, but constitute interesting future research directions.

VIII. Numerical Tests

In this section, numerical tests using real datasets are provided to showcase the merits of our proposed MCCA approaches in several machine learning applications, including user engagement prediction, friend recommendation, clustering, and classification.

A. User Engagement Prediction

Given multi-view data of Twitter users, the goal of the so-called user engagement prediction is to determine which topics a Twitter user is likely to tweet about, by using hashtag as a proxy. The first experiment entails six datasets of Twitter users, which include EgoTweets, MentionTweets, FriendTweets, FollowersTweets, FriendNetwork, and FollowerNetwork data,[1] where $\{D_m = 1,000\}_{m=1}^6$ and $N = 1,770$ users' data are randomly chosen from the database. Details in generating those multiview data can be found in [3]. In this experiment, we corroborate that effective graph priors can be extracted from existing views. Specifically, considering data $\{\mathbf{X}_m \in \mathbb{R}^{D_m \times N}\}_{m=1}^3$ from the first 3 views, we can construct three adjacency matrices $\{\mathbf{W}_m \in \mathbb{R}^{N \times N}\}_{m=1}^3$ as follows, whose (i, j)-th entries are

$$w_{ij}^m := \begin{cases} K_m^t(i,j), & i \in \mathcal{N}_{k_1}(j) \text{ or } j \in \mathcal{N}_{k_1}(i) \\ 0, & \text{otherwise} \end{cases} \tag{23}$$

where $\mathbf{K}_m^t$ is a Gaussian kernel matrix of $\mathbf{X}_m$ with bandwidth equal to the mean of the corresponding Euclidean distances, and $\mathcal{N}_{k_1}(j)$ the set of column indices of $\mathbf{K}_m^t$ containing the k_1-nearest neighbors of column j. Our graph adjacency matrix is built using $\mathbf{W} = \sum_{m=1}^3 \mathbf{W}_m$. To perform graph (G) PCA [28] and PCA, six different views of the data are concatenated to form a single dataset of 6,000-dimensional data vectors.

[1] http://www.cs.jhu.edu/~mdredze/datasets/multiview_embeddings/

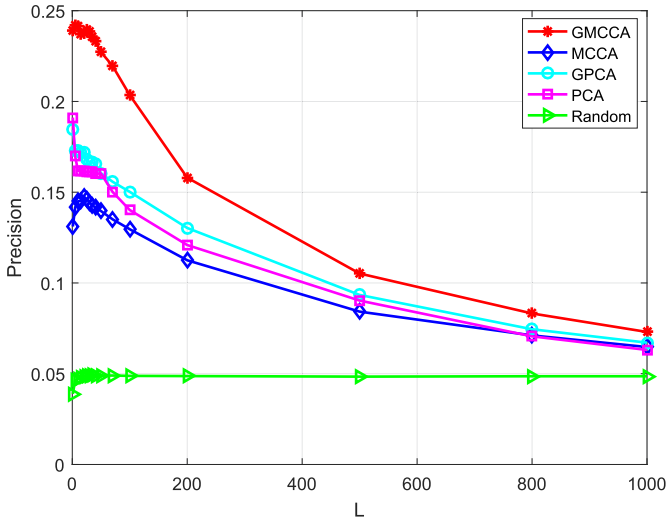


Fig. 1. Precision of user engagement prediction.

We selected 9 most frequently used hashtags. Per Monte Carlo (MC) run, 5 users who tweeted each selected hashtag were randomly chosen as exemplars of users that would employ this hashtag in the future. All other users that tweeted each hashtag were ranked by the cosine distance of their representations to the average representation of those 5 users, where the representation per user is either the corresponding estimate of the common source obtained by (G)MCCA or the principal components by (G)PCA. Before computing cosine distance, the d-dimensional representations were z-score normalized. In other words, each dimension has its mean removed, and subsequently scaled to have unit variance. The representations are learned on data collected pre-March 2015, while the association between hashtags and users is extracted in March 2015. This implies that the hashtags do not impact the representation learning. Pertinent hyper-parameters were set as $k_1 = 10$, $\gamma = 0.05$, and $d = 5$.

Prediction performance is evaluated using three metrics: precision, recall, and mean reciprocal rank (MRR), where a user is marked as correct if this user uses the hashtag. The precision is defined as the ratio of the number of correctly predicted users over the total number of predicted users considered. Recall is the ratio of the number of correctly predicted users over the total number of users that use the hashtag. MRR is the average inverse of the ranks of the first correctly predicted users.

Figures 1 and 2 present the average precision and recall of GMCCA, MCCA, GPCA, PCA, and a random ranking scheme over 100 MC realizations, with a varying number L of evaluated users per hashtag. Here, the random ranking is included as a baseline. Table I reports the prediction performance of simulated schemes with $L = 35$ being fixed. Clearly, GMCCA outperforms its competing alternatives in this Tweeter user engagement prediction task. Moreover, ranking through all approaches is consistent across precision, recall, and MRR.

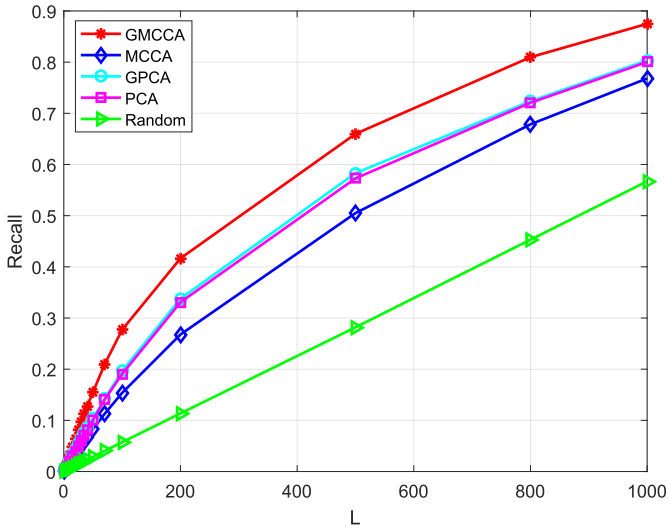


Fig. 2. Recall of user engagement prediction.

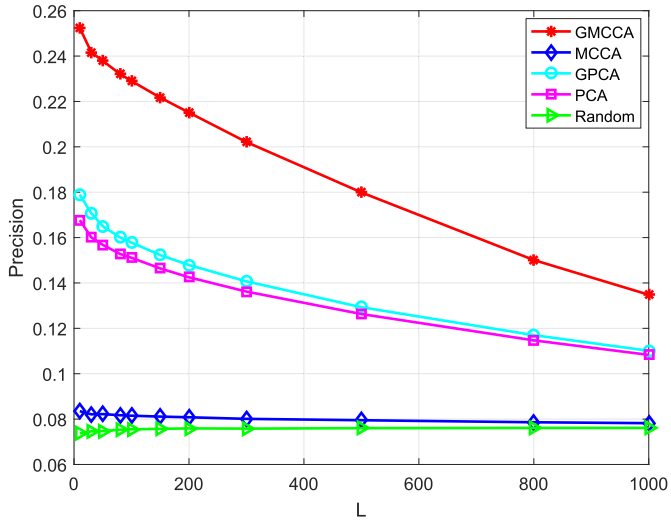


Fig. 3. Precision of friend recommendation.

TABLE I
USER ENGAGEMENT PREDICTION PERFORMANCE

Model	Precision	Recall	MRR
GMCCA	0.2357	0.1127	0.4163
MCCA	0.1428	0.0593	0.2880
GPCA	0.1664	0.0761	0.3481
PCA	0.1614	0.0705	0.3481
Random	0.0496	0.0202	0.1396

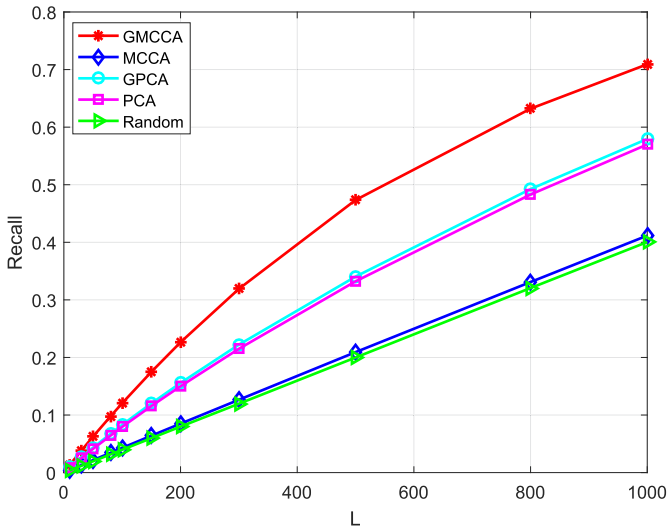


Fig. 4. Recall of friend recommendation.

B. *Friend Recommendation*

GMCCA is further examined for friend recommendation, where the graph can be constructed from an alternative view of the data, as we argued in Remark 4. Specifically for this test, 3 Twitter user datasets [3] from 2,506 users were used to form $\{\mathbf{X}_m \in \mathbb{R}^{1,000\times 2,506}\}_{m=1}^{3}$, which are EgoTweets, FollowersTweets, and FollowerNetwork data. Intuitively, the FriendTweets are helpful for the friend recommendation task. Thus, an alternative view, the FriendTweets data of the same group of users, was used to construct the common source graph. The weight matrix $\mathbf{W}$ is obtained following a similar way from forming $\mathbf{W}_m$ but replacing $\mathbf{K}_m^t$ in (23) with a Gaussian kernel matrix of FriendTweets data.

In the experiment, 20 most popular accounts were selected, which correspond to celebrities. Per realization, 10 users who follow each celebrity were randomly picked, and all other users were ranked by their cosine distances to the average of the 10 picked representations. We z-score normalize all representations before calculating the cosine distances. The same set of evaluation criteria as in user engagement prediction in Section VIII-A was adopted here, where a user is considered to be a correctly recommended friend if both follow the given celebrity. Hyper-parameters $k_1 = 50$, $\gamma = 0.05$, and $d = 5$ were simulated. The friend recommendation performance of GMCCA, MCCA, GPCA, PCA, and Random ranking is evaluated after averaging over 100 independent realizations.

In Figs. 3 and 4, the precision and recall of all simulated algorithms under an increasing number of recommended friends (L) are reported. Plots corroborate the advantages of our GMCCA relative to its simulated alternatives under different numbers of recommendations. Moreover, Table II compares the precision, recall, and MRR of simulated schemes for fixed $L = 100$. Regarding the results, we have the following observations: i) GMCCA is more attractive in the recommendation task than its alternatives; ii) precision and recall differences among approaches are consistent for different L values; and, iii) ranking achieved by these schemes is consistent across 3 metrics for fixed $L = 100$.

TABLE II
FRIEND RECOMMENDATION PERFORMANCE COMPARISON

Model	Precision	Recall	MRR
GMCCA	0.2290	0.1206	0.4471
MCCA	0.0815	0.0429	0.2225
GPCA	0.1578	0.0831	0.3649
PCA	0.1511	0.0795	0.3450
Random	0.0755	0.0397	0.2100

TABLE III
SIX SETS OF FEATURES OF HANDWRITTEN NUMERALS

mfeat-fou	76-dim. Fourier coeff. of character shapes features
mfeat-fac	216-dim. profile correlations features
mfeat-kar	64-dim. Karhunen-Love coefficients features
mfeat-pix	240-dim. pixel averages in 2 x 3 windows
mfeat-zer	47-dim. Zernike moments features
mfeat-mor	6-dim. morphological features

C. UCI Data Clustering

Handwritten digit data from the UCI machine learning repository[2] were called for to assess GMCCA for clustering. This dataset contains 6 feature sets of 10 classes corresponding to 10 digits from 0 to 9, as listed in Table III. There are 200 data per class (2,000 in total) per feature set. Seven clusters of data including digits 1, 2, 3, 4, 7, 8, and 9 were used to form the views $\{\mathbf{X}_m \in \mathbb{R}^{D_m \times 1,400}\}_{m=1}^{6}$ with $D_1 = 76$, $D_2 = 216$, $D_3 = 64$, $D_4 = 240$, $D_5 = 47$, and $D_6 = 6$. Similar to Section VIII-A, one existing view is used to build the graph. Specifically, the graph adjacency matrix is constructed using (23), after substituting $\mathbf{K}_m^t$ by the Gaussian kernel matrix of $\mathbf{X}_3$. GPCA and PCA were performed on the concatenated data vectors of dimension $\sum_{m=1}^{6} D_m$, while the K-means was performed using either $\hat{\mathbf{S}}$, or the principal components with $\gamma = 0.1$ and $d = 3$.

Clustering performance is evaluated in terms of two metrics, namely clustering accuracy and scatter ratio. Clustering accuracy is the percentage of correctly clustered samples. Scatter ratio is defined as $C_t / \sum_{i=1}^{7} C_i$, where C_t and C_i denote the total scatter value and the within-cluster scatter value, given correspondingly by $C_t := \|\hat{\mathbf{S}}\|_F^2$ and $C_i := \sum_{j \in \mathcal{C}_i} \|\hat{\mathbf{s}}_j - \frac{1}{|\mathcal{C}_i|} \sum_{\ell \in \mathcal{C}_i} \hat{\mathbf{s}}_\ell\|_2^2$; here, $\mathcal{C}_i$ is the set of data vectors belonging to the i-th cluster, and $|\mathcal{C}_i|$ is the cardinality of $\mathcal{C}_i$.

Table IV reports the clustering performance of MCCA, PCA, GMCCA, and GPCA for different k_1 values. Clearly, GMCCA yields the highest clustering accuracy and scatter ratio. Fixing $k_1 = 50$, Fig. 5 plots the first two dimensions of the common source estimates obtained by (G)MCCA along with the first two principal components of (G)PCA, with different colors signifying different clusters. As observed from the scatter plots, GMCCA separates the 7 clusters the best, in the sense that data points within clusters are concentrated but across clusters are far apart.

[2]http://archive.ics.uci.edu/ml/datasets/Multiple+Features.

TABLE IV
CLUSTERING PERFORMANCE COMPARISON

k_1	Clustering accuracy		Scatter ratio	
	GMCCA	GPCA	GMCCA	GPCA
10	0.8141	0.5407	9.37148	4.9569
20	0.8207	0.5405	11.6099	4.9693
30	0.8359	0.5438	12.2327	4.9868
40	0.8523	0.5453	12.0851	5.0157
50	0.8725	0.5444	12.1200	5.0640
MCCA	0.8007		5.5145	
PCA	0.5421		4.9495	

D. Generalization Bound Versus γ

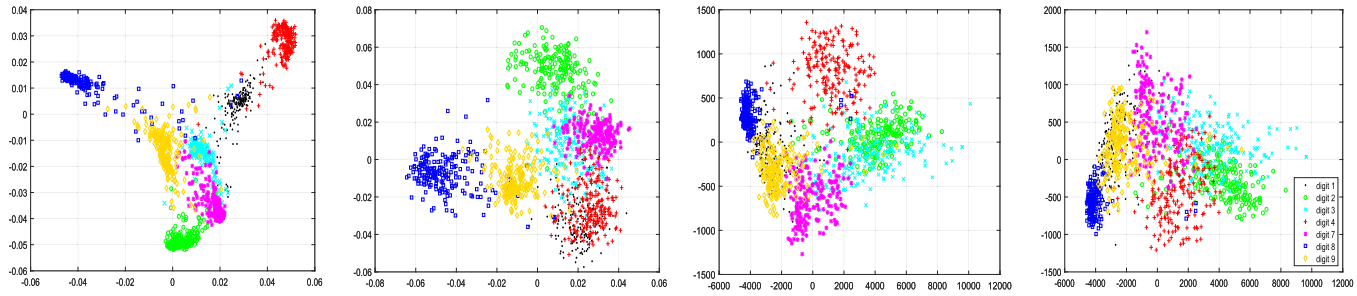
Here, we wish to demonstrate the usefulness of the generalization bound of GMCCA derived in Section IV. Specifically, we will test numerically the effect of γ on the generalization error bound defined on the right hand side (RHS) of (14).

In this experiment, 20 MC simulations were performed to evaluate the clustering performance of GMCCA using the UCI dataset described in Section VIII-C. Per MC realization, 200 samples per cluster were randomly and evenly divided to obtain training data $\{\mathbf{X}_m^{\rm tr} \in \mathbb{R}^{D_m \times 700}\}_{m=1}^{3}$ and testing data $\{\mathbf{X}_m^{\rm te} \in \mathbb{R}^{D_m \times 700}\}_{m=1}^{3}$. The same 7 digits in Section VIII-C and their first 3 views were employed. GMCCA was performed on the training data to obtain $\{\hat{\mathbf{U}}_m \in \mathbb{R}^{D_m \times 3}\}_{m=1}^{3}$. Subsequently, low-dimensional representations of the testing data were found as $\sum_{m=1}^{3} \hat{\mathbf{U}}_m^\top \mathbf{X}_m^{\rm te} \in \mathbb{R}^{3 \times 700}$, which were fed into the K-means for digit clustering. The generalization bound was evaluated utilizing the RHS of (14), where $p = 0.1$, and $B = \sqrt{\sum_{m=1}^{2} \sum_{m'=m+1}^{3} \|\hat{\mathbf{U}}_m^\top \hat{\mathbf{U}}_m + \hat{\mathbf{U}}_{m'}^\top \hat{\mathbf{U}}_{m'}\|_F^2}$.

Figure 6 depicts the average generalization error bound along with clustering accuracy on the test data for different γ values ranging from 0 to 500. Interestingly, at $\gamma = 0.01$, the bound attains its minimum, and at the same time, the clustering accuracy achieves its maximum. This indeed provides us with an effective way to select the hyper-parameter value for our GMCCA approaches.

E. Face Recognition

The ability of GDMCCA in face recognition is evaluated using the Extended Yale-B (EYB) face image database [21]. The EYB database contains frontal face images of 38 individuals, each having 65 images of 192×168 pixels. Per MC realization, we performed Coiflets, Symlets, and Daubechies orthonormal



(a) GMCCA (b) MCCA (c) GPCA (d) PCA

Fig. 5. Scatter plot of the first two rows of $\hat{\mathbf{S}}$ or principal components.

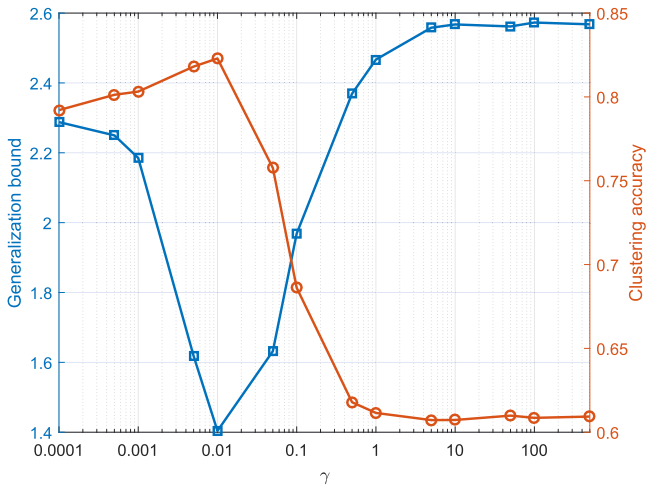


Fig. 6. Generalization bound versus γ.

wavelet transforms on 20 randomly selected individuals' images to form three feature datasets. Subsequently, three feature matrices of each image were further resized to 50×40 pixels, followed by vectorization to obtain three $2,000 \times 1$ vectors. For each individual, $N_{\rm tr}$ images were randomly chosen, and the corresponding three sets of wavelet transformed data were used to form the training datasets $\{\mathbf{X}_m \in \mathbb{R}^{2,000\times 20N_{\rm tr}}\}_{m=1}^3$. Among the remaining images, $(30-0.5N_{\rm tr})$ images per individual were obtained to form the validation datasets $\{\mathbf{X}_m^{\rm tu} \in \mathbb{R}^{2,000\times 20(30-0.5N_{\rm tr})}\}_{m=1}^3$, and another $(30-0.5N_{\rm tr})$ for testing $\{\mathbf{X}_m^{\rm te} \in \mathbb{R}^{2,000\times 20(30-0.5N_{\rm tr})}\}_{m=1}^3$, following a similar process to construct $\{\mathbf{X}_m\}_{m=1}^3$.

In order to achieve enhanced face recognition performance, the common information including both the original images and their labels will be leveraged to construct $\mathbf{W}$ here. This is in contrast with the methods adopted in Section VIII-A, VIII-B, and VIII-C, which were based on existing or alternative views. Specifically, the $20N_{\rm tr}$ original training images were resized to 50×40 pixels, and subsequently vectorized to obtain $2,000 \times 1$ vectors, collected as columns of $\mathbf{O} \in \mathbb{R}^{2,000\times 20N_{\rm tr}}$, which were further used to build $\mathbf{W} \in \mathbb{R}^{20N_{\rm tr}\times 20N_{\rm tr}}$. Per $(i,\, j)$-th entry of $\mathbf{W}$ is

$$w_{ij} := \begin{cases} \frac{\mathbf{o}_i^\top \mathbf{o}_j}{\|\mathbf{o}_i\|_2 \|\mathbf{o}_j\|_2}, & i \in \mathcal{M}_{k_2}(j) \text{ or } j \in \mathcal{M}_{k_2}(i) \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

where $\mathbf{o}_i$ is the i-th column of $\mathbf{O}$, and $\mathcal{M}_{k_2}(i)$ the set of the k_2 nearest neighbors of $\mathbf{o}_i$ belonging to the same individual.

In this experiment, $k_2 = N_{\rm tr} - 1$ was kept fixed. Furthermore, the three associated graph adjacency matrices in Laplacian regularized multi-view (LM) CCA [4] were built in a similar way to construct $\mathbf{W}$, after substituting $\mathbf{O}$ by $\{\mathbf{X}_m\}_{m=1}^3$ accordingly. The hyper-parameters in GDMCCA, DMCCA, GDPCA, LMCCA were tuned among 30 logarithmically spaced values between 10^{-3} and 10^3 to maximize the recognition accuracy on $\{\mathbf{X}_m^{\rm tu}\}_{m=1}^3$. After simulating GDMCCA, DMCCA, GDPCA, DPCA, and LMCCA, 10 projection vectors were employed to find the low-dimensional representations of $\{\mathbf{X}_m^{\rm te}\}_{m=1}^3$. Subsequently, the 1-nearest neighbor rule was applied for face recognition.

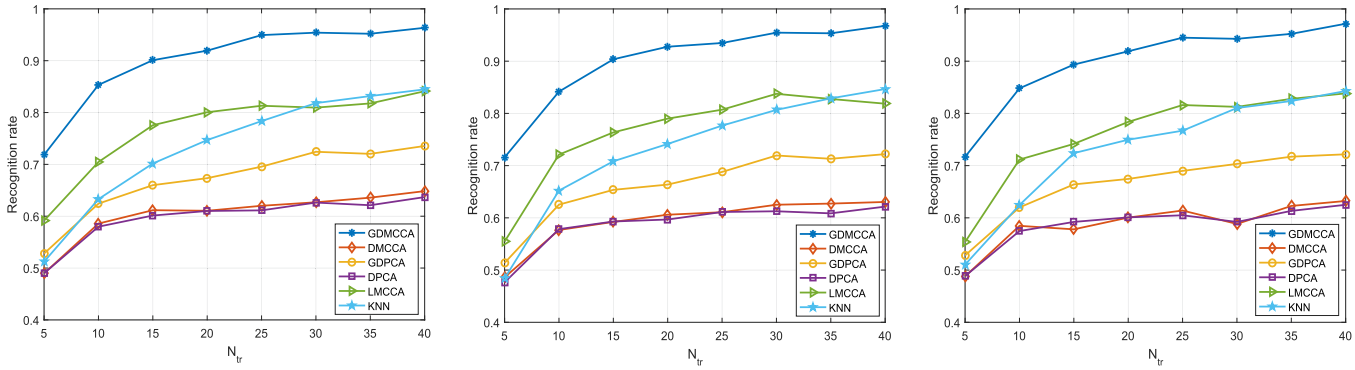
Figures 7(a), 7(b), and 7(c) describe the average recognition accuracies of GMDCCA, MDCCA, GDPCA, DPCA, LMCCA, and KNN, for testing data $\mathbf{X}_1^{\rm te}$, $\mathbf{X}_2^{\rm te}$, and $\mathbf{X}_3^{\rm te}$, respectively, and for a varying number $N_{\rm tr}$ of training samples over 30 MC realizations. It is clear that the recognition performance of all tested schemes improves as $N_{\rm tr}$ grows. Moreover, GDMCCA yields the highest recognition accuracy in all simulated settings.

F. Image Data Classification

The MNIST database[3] containing 10 classes of handwritten 28×28 digit images with 7,000 images per class, is used here to assess the merits of GKMCCA in classification. Per MC test, three sets of $N_{\rm tr}$ images per class were randomly picked for training, validation and testing, respectively. We followed the process of generating the three-view training, tuning, and testing data in Section VIII-E to construct $\{\mathbf{X}_m \in \mathbb{R}^{196\times 10N_{\rm tr}}\}_{m=1}^3$, $\{\mathbf{X}_m^{\rm tu} \in \mathbb{R}^{196\times 10N_{\rm tr}}\}_{m=1}^3$, and $\{\mathbf{X}_m^{\rm te} \in \mathbb{R}^{196\times 10N_{\rm tr}}\}_{m=1}^3$, except that each data sample per view was resized to 14×14 pixels.

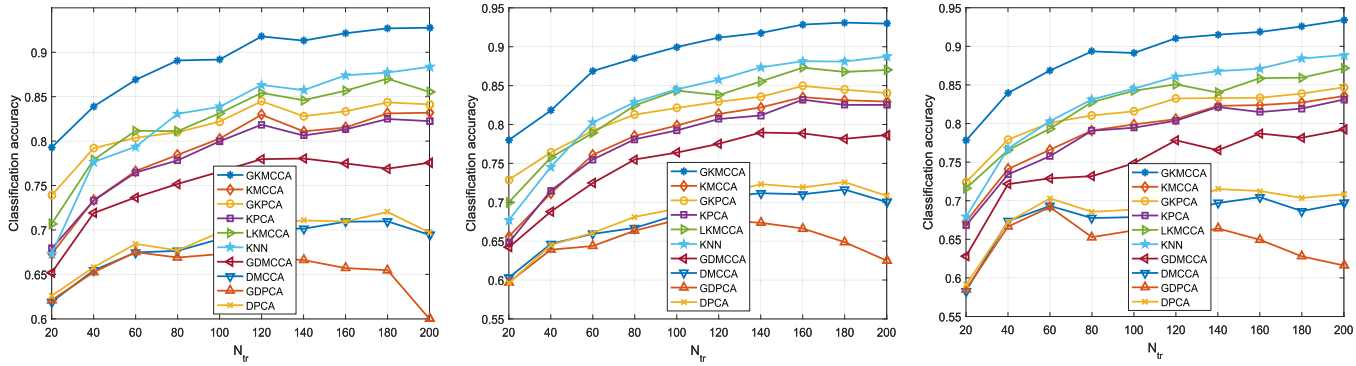
Gaussian kernels were used for $\{\mathbf{X}_m\}_{m=1}^3$, the resized, as well as the vectorized training images, denoted by $\{\mathbf{o}_i \in \mathbb{R}^{196\times 1}\}_{i=1}^{10N_{\rm tr}}$, where the bandwidth parameters were set equal to the mean of their corresponding Euclidean distances. Relying on the kernel matrix of $\{\mathbf{o}_i\}$, denoted by $\mathbf{K}_o \in \mathbb{R}^{10N_{\rm tr}\times 10N_{\rm tr}}$,

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



(a) 1st view (b) 2nd view (c) 3rd view

Fig. 7. Classification performance using YEB data.



(a) 1st view (b) 2nd view (c) 3rd view

Fig. 8. Classification results using MNIST data.

the graph adjacency matrix was constructed in the way depicted in (24) but with $\frac{\mathbf{o}_i^\top \mathbf{o}_j}{\|\mathbf{o}_i\|_2\|\mathbf{o}_j\|_2}$ and $20N_{\rm tr}$ replaced by the $(i,\,j)$-th entry of $\mathbf{K}_o$ and $10N_{\rm tr}$, respectively. The graph Laplacian regularized kernel multi-view (LKM) CCA [4] used three graph adjacency matrices, which were obtained by (24) after substituting $\frac{\mathbf{o}_i^\top \mathbf{o}_j}{\|\mathbf{o}_i\|_2\|\mathbf{o}_j\|_2}$ by the $(i,\,j)$-th entry of $\{\mathbf{K}_m\}_{m=1}^3$. To implement GDMCCA and GDPCA, the graph adjacency matrices were constructed via (24). In all tests of this subsection, we set $k_2 = N_{\rm tr} - 1$. The hyper-parameters of GKMCCA, KMCCA, GKPCA, LKMCCA, GDMCCA, DMCCA, and GDPCA were selected from 30 logarithmically spaced values between 10^{-3} and 10^3, that yields the best classification performance. Ten projection vectors are learned by GKMCCA, KMCCA, GKPCA, KPCA, LKMCCA, GDMCCA, DMCCA, GDPCA, and DPCA, which are further used to obtain the low-dimensional representations of $\{\mathbf{X}_m^{\rm te}\}_{m=1}^3$. Then, the 5-nearest neighbors rule is adopted for classification. The classification accuracies of all methods reported are averages over 30 MC runs.

In Figs. 8(a), 8(b), and 8(c), the classification accuracies of the 10-dimensional representations of $\mathbf{X}_1^{\rm te}$, $\mathbf{X}_2^{\rm te}$, and $\mathbf{X}_3^{\rm te}$ are plotted. The advantage of GKMCCA relative to other competing alternatives remains remarkable no matter which view of testing data is employed.

IX. Conclusion

In this work, CCA along with multiview CCA was revisited. Going beyond existing (M)CCA approaches, a novel graph-regularized MCCA method was put forth that leverages prior knowledge described by graph(s) that common information bearing sources belong to. By embedding the latent common sources in a graph and invoking this extra information as a graph regularizer, our GMCCA was developed to endow the resulting low-dimensional representations. Performance analysis of our GMCCA approach was also provided through the development of a generalization bound. To cope with data vectors whose dimensionality exceeds the number of data samples, we further introduced a dual form of GMCCA. To further account for nonlinear data dependencies, we generalized GMCCA to obtain a graph-regularized kernel MCCA scheme too. Finally, we showcased the merits of our proposed GMCCA approaches using extensive real-data tests.

This work opens up several interesting directions for future research. Developing efficient GMCCA algorithms for high-dimensional multiview learning is worth investigating. Generalizing our proposed GMCCA approaches to handle unaligned multiview datasets is also pertinent for semi-supervised learning as well. Incorporating additional structural forms regularization,

e.g., sparsity and non-negativity, into the novel GMCCA framework is meaningful too.

REFERENCES

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, Atlanta, GA, USA, Jun. 16–21, 2013, pp. 1247–1255.

[2] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, Nov. 2002.

[3] A. Benton, R. Arora, and M. Dredze, "Learning multiview embeddings of Twitter users," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, Berlin, Germany, Aug. 7–12, 2016, pp. 14–19.

[4] 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.

[5] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *J. Amer. Statist. Assoc.*, vol. 80, no. 391, pp. 580–598, Sep. 1985.

[6] J. D. Carroll, "Generalization of canonical correlation analysis to three or more sets of variables," in *Proc. Annu. Conv. Amer. Psychol. Assoc.*, vol. 3, 1968, pp. 227–228.

[7] J. Chen, G. Wang, and G. B. Giannakis, "Multiview canonical correlation analysis over graphs," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Brighton, U.K., May 12–17, 2019.

[8] J. Chen and I. D. Schizas, "Distributed efficient multimodal data clustering," in *Proc. Eur. Signal Process. Conf.*, Kos Island, Greece, Aug. 28–Sep. 2, 2017, pp. 2304–2308.

[9] J. Chen, G. Wang, and G. B. Giannakis, "Nonlinear dimensionality reduction for discriminative analytics of multiple datasets," *IEEE Trans. Signal Process.*, vol. 67, no. 3, pp. 740–752, May 2019.

[10] J. Chen, G. Wang, Y. Shen, and G. B. Giannakis, "Canonical correlation analysis of datasets with a common source graph," *IEEE Trans. Signal Process.*, vol. 66, no. 16, pp. 4398–4408, Aug. 2018.

[11] X. Chen, L. Han, and J. Carbonell, "Structured sparse canonical correlation analysis," in *Proc. Artif. Int. Statist.*, Mar. 2012, pp. 199–207.

[12] 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, Jun. 2010.

[13] G. B. Giannakis, Y. Shen, and G. V. Karanikolas, "Topology identification and learning over graphs: Accounting for nonlinearities and dynamics," *Proc. IEEE*, vol. 106, no. 5, pp. 787–807, May 2018.

[14] 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.

[15] P. Horst, "Generalized canonical correlations and their applications to experimental data," *J. Clin. Psych.*, vol. 17, no. 4, pp. 331–347, Oct. 1961.

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

[17] B. Jiang, C. Ding, and J. Tang, "Graph-Laplacian PCA: Closed-form solution and robustness," in *Proc. Int. Conf. Comput. Vision Pattern Recognit.*, Portland, USA, Jun. 25–27, 2013, pp. 3492–3498.

[18] C. I. Kanatsoulis, X. Fu, N. D. Sidiropoulos, and M. Hong, "Structured SUMCOR multiview canonical correlation analysis for large-scale data," *IEEE Trans. Signal Process.,*vol. 67, no. 2, pp. 306–319, Jan. 2019.

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

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

[21] 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.

[22] D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, and B. Schölkopf, "Randomized nonlinear component analysis," in *Proc. Int. Conf. Mach. Learn.*, Beijing, China, Jun. 21–26, 2014, pp. 1359–1367.

[23] P. Rastogi, B. Van Durme, and R. Arora, "Multiview LSA: Representation learning via generalized CCA," in *Proc. North Amer. Chap. Assoc. Comput. Linguistics: Human Lang. Tech.*, Denver, CO, USA, May 31–Jun. 5, 2015, pp. 556–566.

[24] J. Rupnik, P. Skraba, J. Shawe-Taylor, and S. Guettes, "A comparison of relaxations of multiset cannonical correlation analysis and applications," Feb. 2013, arXiv:1302.0974.

[25] 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.

[26] 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.

[27] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. 1st ed. Cambridge, U.K.: Cambridge Univ. Press, Jun. 2004.

[28] Y. Shen, P. Traganitis, and G. B. Giannakis, "Nonlinear dimensionality reduction on graphs," in *Proc. IEEE Int. Workshop Comput. Adv. Multi-Sensor Adaptive Process.*, Curacao, Dutch Antilles, Dec. 10–13, 2017, pp. 1–5.

[29] Y. Shen, T. Chen, and G. B. Giannakis, "Online ensemble multi-kernel learning adaptive to non-stationary and adversarial environments," in *Proc. Int. Conf. Artif. Intell. Stat.*, Lanzarote, Canary Islands, Apr. 9–11, 2018, pp. 2037–2046.

[30] S. Sun, "A survey of multi-view machine learning," *Neural Comput. Appl.*, vol. 23, no. 7/8, pp. 2031–2038, Dec. 2013.

[31] W. Tang, Z. Lu, and I. S. Dhillon, "Clustering with multiple graphs," in *Proc. Intel. Conf. Data Mining*, Miami, FL, USA, Dec. 6-9, 2009, pp. 1016–1021.

[32] A. Tenenhaus, C. Philippe, V. Guillemot, K.-A. Le Cao, J. Grill, and V. Frouin, "Variable selection for generalized canonical correlation analysis," *Biostatistics*, vol. 15, no. 3, pp. 569–583, Feb. 2014.

[33] G. Wang, G. B. Giannakis, and J. Chen, "Learning ReLU networks on linearly separable data: Algorithm, optimality, and generalization," *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2357–2370, May 2019.

[34] G. Wang, G. B. Giannakis, J. Chen, and J. Sun, "Distribution system state estimation: An overview of recent developments," *Front. Inf. Technol. Electron. Eng.*, vol. 20, no. 1, pp. 4–17, Jan. 2019.

[35] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. Int. Conf. Mach. Learn.*, Lille, France, Jul. 6–11, 2015, pp. 1083–1092.

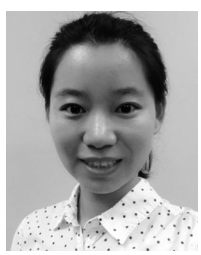
[36] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, Apr. 2009.

[37] D. M. Witten and R. J. Tibshirani, "Extensions of sparse canonical correlation analysis with applications to genomic data," *Statist. App. Genet. Mol. Bio.*, vol. 8, no. 1, pp. 1–27, Jan. 2009.

[38] 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.

[39] 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.

[40] L. Zhang, G. Wang, D. Romero, and G. B. Giannakis, "Randomized block Frank–Wolfe for convergent large-scale learning," *IEEE Trans. Signal Process.*, vol. 65, no. 24, pp. 6448–6461, Dec. 2019.



Jia Chen received the B.Sc. degree from the Southwest Jiaotong University, Chengdu, China, in 2009, and the M.Sc. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2012, both in electrical engineering, and the Ph.D. degree in electrical engineering from the University of Texas at Arlington, Arlington, TX, USA, in 2016. She is currently a Postdoctoral Research Associate with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA. Her research interests include signal processing, data analytics, and machine learning.

Gang Wang (M'18) received the B.Eng. degree in electrical engineering and automation in 2011 from the Beijing Institute of Technology, Beijing, China, and the Ph.D. degree in electrical and computer engineering in 2018 from the University of Minnesota, Minneapolis, MN, USA, where he is currently a Postdoctoral Associate in the Department of Electrical and Computer Engineering.

His research interests focus on the areas of statistical learning, optimization, and deep learning with applications to data science and smart grids. He received a National Scholarship in 2014, a Guo Rui Scholarship in 2017, and an Innovation Scholarship (first place) in 2017, all from China, as well as a Best Student Paper Award at the 2017 European Signal Processing Conference.

Georgios B. Giannakis (F'97) received the Diploma in electrical engineering from the National Technical University of Athens, Athens, Greece, in 1981, the M.Sc. degree in electrical engineering in 1983, the M.Sc. degree in mathematics in 1986, and the Ph.D. degree in electrical engineering in 1986 from the University of Southern California, Los Angeles, CA, USA. From 1982 to 1986, he was with the University of Southern California. He was a Faculty Member with the University of Virginia from 1987 to 1998, and since 1999 he has been a Professor with the University of Minnesota, Minneapolis, MN, USA, where he holds an ADC Endowed Chair, a University of Minnesota McKnight Presidential Chair in Electrical and Computer Engineering, and serves as the Director of the Digital Technology Center.

His general interests span the areas of statistical learning, communications, and networking—subjects on which he has published more than 450 journal papers, 750 conference papers, 25 book chapters, two edited books and two research monographs (h-index 142). His current research focuses on data science, and network science with applications to the Internet of Things, social, brain, and power networks with renewables. He is the (co-)inventor of 32 patents issued, and the (co-)recipient of 9 best journal paper awards from the IEEE Signal Processing (SP) and Communications Societies, including the G. Marconi Prize Paper Award in Wireless Communications. He also received Technical Achievement Awards from the SP Society in 2000, from EURASIP in 2005, a Young Faculty Teaching Award, the G. W. Taylor Award for Distinguished Research from the University of Minnesota, and the IEEE Fourier Technical Field Award (inaugural recipient) in 2015. He is a Fellow of EURASIP, and has served the IEEE in a number of posts, including that of a Distinguished Lecturer for the IEEE-SPS.