MULTIVIEW CANONICAL CORRELATION ANALYSIS OVER GRAPHS

Jia Chen, Gang Wang, and Georgios B. Giannakis

Dept. of ECE and Digital Tech. Center, Univ. of Minnesota, Mpls, MN 55455, USA
Emails: {chen5625, gangwang, georgios}@umn.edu

ABSTRACT

Multiview canonical correlation analysis (MCCA) looks for shared low-dimensional representations hidden in multiple transformations of common source signals. Existing MCCA approaches do not exploit the geometry of common sources, which can be either given *a priori*, or constructed from domain knowledge. In this paper, a novel graph-regularized (G) MCCA is developed to account for such geometry-bearing information via graph regularization in the classical maximum-variance MCCA model. GMCCA minimizes the distance between the sought canonical variables and the common sources, while incorporating the graph-induced prior of these sources. To capture nonlinear dependencies, GMCCA is further broadened to the graph-regularized kernel (GK) MCCA. Numerical tests using real datasets document the merits of G(K)MCCA in comparison with competing alternatives.

Index Terms— Dimensionality reduction, signal processing over graphs, Laplacian regularization, multiview learning

1. INTRODUCTION

Multiview data collected from different transformations of common signal(s) are typical in applications, such as multi-camera surveillance systems, where single-view data do not suffice for a comprehensive description of the common signal sources. In paper classification for instance, there are three views representing any given paper: the title, keywords, and its citations [1]. Learning with heterogeneous data from different domains is often referred to as *multiview learning*, which is an emerging direction in machine learning [2]. Canonical correlation analysis (CCA) is a learning tool with well-documented merits. It seeks linear transformations of two datasets so that the correlation between the transformed low-dimensional features is maximized [3]. Multiview (M) CCA generalizes the vanilla CCA to cope with data from more than two views [4], and enjoys popularity that grows with the heterogeneity of sensing devices.

Graph-aware subspace learning methods have been widely used in machine learning applications, such as dimensionality reduction, clustering, classification, and data reconstruction [5, 6]. Specifically, graph CCA accounts for the structural information present in a common source [7], but it is limited to analyzing two datasets. The geometric information of the common sources has not been leveraged in the context of MCCA.

Building upon but considerably going beyond our results in [7], a novel *graph-regularized* (G) MCCA framework is put forth here. GMCCA aims at minimizing the distance between the low-dimensional representations of each view and the common sources, while accounting for the statistical dependencies among these sources that are hidden in the multiple views. Such dependencies may be available from the given data, or can be deduced from correlations, which are encoded in a graph and we invoke as graph regularizers of standard MCCA. Going beyond linear transformations, we employ kernels along with a Tikhonov regularizer on the low-dimensional representations to develop a novel graph-regularized kernel (GK) MCCA tool. Interestingly, the solutions of GMCCA and GKMCCA can be analytically found by performing a single eigenvalue decomposition.

2. PRELIMINARIES

Consider $M \geq 2$ datasets $\{\mathbf{X}_m \in \mathbb{R}^{D_m \times N}\}_{m=1}^M$ collected from M views of the N common sources collected in the matrix $\check{\mathbf{S}} \in \mathbb{R}^{\rho \times N}$, with possibly $\rho \ll \min_m \{D_m\}_{m=1}^M$. Without loss of generality, assume that per dataset $\mathbf{X}_m$ has been centered. MCCA looks for low-dimensional subspaces $\{\mathbf{U}_m \in \mathbb{R}^{D_m \times d}\}_{m=1}^M$ with $d \leq \rho$, such that the difference between each pair of linear projections $\mathbf{U}_m^\top \mathbf{X}_m$ is minimized.

To reveal the underpinnings of our approach, we outline two popular MCCA formulations. The first, termed sum-of-correlations (SUMCOR) MCCA [4], matches the pairs by

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

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

where columns of $\mathbf{U}_m$ are known as the loading vectors of $\mathbf{X}_m$, and projections $\{\mathbf{U}_m^\top \mathbf{X}_m\}$ are the so-termed canonical variables, which can be viewed as low (d)-dimensional approximations of the hidden sources in $\check{\mathbf{S}}$. However, when $M \geq 3$, problem (1) is provably NP-hard [8].

Work in this paper was supported in part by NSF grants 1711471, 1514056, and the NIH grant no. 1R01GM104975-01.

Instead of minimizing the Euclidean distance between all low-dimensional representation pairs, one can also explicitly look for a common low-dimensional approximation of the common source matrix $\mathbf{S} \in \mathbb{R}^{d\times N}$, by solving [4]

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

$$\text{s. to} \quad \mathbf{S}\mathbf{S}^\top = \mathbf{I} \tag{2b}$$

which yields the so-termed maximum-variance (MAXVAR) MCCA formulation. If all sample covariance matrices $\{\mathbf{X}_m\mathbf{X}_m^\top\}_{m=1}^M$ have full rank, then the columns of the $\mathbf{S}$-minimizer are given by the first d principal eigenvectors of $\sum_{m=1}^M \mathbf{X}_m^\top(\mathbf{X}_m\mathbf{X}_m^\top)^{-1}\mathbf{X}_m$, while the $\mathbf{U}_m$-minimizers are found as $\{\hat{\mathbf{U}}_m = (\mathbf{X}_m\mathbf{X}_m^\top)^{-1}\mathbf{X}_m\hat{\mathbf{S}}^\top\}_{m=1}^M$ [9].

3. GRAPH-REGULARIZED MCCA

In a gamut of applications, the N common source columns $\{\check{\mathbf{s}}_i\}_{i=1}^N$ that form $\check{\mathbf{S}}$, may be nodal vectors residing on a graph $\mathcal{G}$ comprising N nodes. Besides the given data $\{\mathbf{X}_m\}$, such structural prior knowledge can be exploited to better estimate the canonical variables. Specifically for the present paper, this extra information is encoded in a graph $\mathcal{G}$ and embodied in the common low-dimensional approximation through a graph regularization term. This section deals precisely with graph-regularized MCCA.

Supposing that the graph $\mathcal{G}$ is undirected, its weighted adjacency matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$ is symmetric, that is $\mathbf{W} = \mathbf{W}^\top$. Letting $d_i := \sum_{j=1}^N w_{ij}$ with w_{ij} denoting the $(i,\, j)$-th entry of $\mathbf{W}$, and the diagonal matrix $\mathbf{D} := \mathrm{diag}(\{d_i\}_{i=1}^N) \in \mathbb{R}^{N\times N}$, the Laplacian of $\mathcal{G}$ is defined as $\mathbf{L}_\mathcal{G} := \mathbf{D} - \mathbf{W}$. Sources $\{\check{\mathbf{s}}_i\}_{i=1}^N$ are assumed smooth over $\mathcal{G}$, that is two vectors $(\check{\mathbf{s}}_i,\, \check{\mathbf{s}}_j)$ residing on connected nodes are also close in the Euclidean distance sense. As explained in Sec. 2, vectors $\mathbf{s}_i$ and $\mathbf{s}_j$ are d-dimensional approximations of $\check{\mathbf{s}}_i$ and $\check{\mathbf{s}}_j$, respectively. To capture this, a meaningful regularization is the weighted sum of Euclidean distances between all pairs of common source estimates $(\mathbf{s}_i, \mathbf{s}_j)$ over $\mathcal{G}$, given by

$$\mathrm{Tr}\big(\mathbf{S}\mathbf{L}_\mathcal{G}\mathbf{S}^\top\big) = \sum_{i=1}^N\sum_{j=1}^N w_{ij}\|\mathbf{s}_i - \mathbf{s}_j\|_2^2\,. \tag{3}$$

Evidently, minimizing (3) over $\mathbf{S}$ forces vectors $\mathbf{s}_i$ and $\mathbf{s}_j$ residing on adjacent nodes associated with large weights w_{ij} to be close to each other. To account for this prior on common sources, the quadratic term (3) is well motivated as a regularizer of the standard MAXVAR MCCA (cf. (2)), yielding our novel graph-regularized (G) MCCA as the solution of

$$\min_{\{\mathbf{U}_m\},\mathbf{S}} \quad \sum_{m=1}^{M} \left\|\mathbf{U}_m^\top \mathbf{X}_m - \mathbf{S}\right\|_F^2 + \gamma\mathrm{Tr}\big(\mathbf{S}\mathbf{L}_\mathcal{G}\mathbf{S}^\top\big) \tag{4a}$$

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

where the hyper-parameter $\gamma \geq 0$ balances minimizing the distance between canonical variables and common source estimates, and promoting smoothness of common source estimates over $\mathcal{G}$. Clearly, GMCCA reduces to MCCA in (2) when $\gamma = 0$; and as γ increases, GMCCA progressively leverages this additional graph-induced knowledge when seeking the common sources and canonical variables.

Taking the derivative of (4a) with respect to $\mathbf{U}_m$ and setting it to $\mathbf{0}$ lead to $\hat{\mathbf{U}}_m := (\mathbf{X}_m\mathbf{X}_m^\top)^{-1}\mathbf{X}_m\mathbf{S}^\top$. After substituting $\mathbf{U}_m$ by $\hat{\mathbf{U}}_m$ and ignoring the constant term in (4a), solving (4) boils down to maximizing $\mathrm{Tr}(\mathbf{S}\mathbf{C}\mathbf{S}^\top)$ subject to (4b), where $\mathbf{C} := \sum_{m=1}^M \mathbf{X}_m^\top(\mathbf{X}_m\mathbf{X}_m^\top)^{-1}\mathbf{X}_m - \gamma\mathbf{L}_\mathcal{G}$. It follows readily that rows of the $\hat{\mathbf{S}}$-optimizer are the d-principal eigenvectors of $\mathbf{C}$. Subsequently, the $\hat{\mathbf{U}}_m$-optimizer can be obtained as $\hat{\mathbf{U}}_m = (\mathbf{X}_m\mathbf{X}_m^\top)^{-1}\mathbf{X}_m\hat{\mathbf{S}}^\top$ for $m = 1, \ldots, M$.

Two remarks are worth making at this point.

Remark 1. *Distinct from the single graph Laplacian regularizer in our GMCCA, the related approaches in [10] and [11] 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 this extra graph information, for view-specific graphs $\{\mathbf{L}_{\mathcal{G}_m}\}_m$ on data $\{\mathbf{X}_m\}_m$. The approach in [11] however does not admit an analytical solution, while [10] copes with semi-supervised learning, where cross-covariances of pairwise datasets are not fully available. In contrast, our single graph regularizer in* (4) *is focused on the common sources. In practice, this is critical when one has prior information about the common sources along with the M views. In paper classification for instance, except for titles, keywords, and introductions of given articles, one may also have access to the citation network, capturing the similarities among papers. More generally, the graph knowledge of inter-dependent sources can be a prior given by an 'expert,' or, it can be dictated by the underlying physics (e.g., [12] in power networks), or, it can be learned from alternate views of the data. Finally, our GMCCA comes with simple analytical solutions.*

Remark 2. *In terms of selecting γ, two feasible methods are: i) cross-validation for supervised learning tasks, where γ is set to the one yielding the best empirical performance on the labeled training data; and, ii) a spectral clustering approach that automatically finds the best γ from a given set of candidates; see e.g., [13] for details.*

4. GRAPH-REGULARIZED KERNEL MCCA

In various practical setups, nonlinearly mapped data vectors are dependent and high-dimensional with $N \ll \min_m\{D_m\}$, while sample covariance matrices $\{\mathbf{X}_m\mathbf{X}_m^\top\}$ become singular. This renders GMCCA infeasible due to the following two reasons: i) GMCCA presumes M linear low-dimensional hyperplanes to project the M-view data vectors; and, ii) GMCCA incurs high computational complexity $\mathcal{O}(MD^3)$ with $D := \max_m\{D_m\}$. To address these issues, the linear GMCCA in (4) will be first re-expressed in its dual form, and the

M-view data will be then mapped to higher dimensional feature spaces through M nonlinear functions. Subsequently, the common low-dimensional representations can be obtained.

Toward this objective, we start by rewriting the loading vectors $\{\mathbf{U}_m\}$ as linear functions of associated datasets $\{\mathbf{X}_m\}$, yielding $\{\mathbf{U}_m := \mathbf{X}_m\mathbf{A}_m\}$, where $\{\mathbf{A}_m \in \mathbb{R}^{N\times d}\}$ are the unknown dual matrices. Substituting $\mathbf{U}_m$ by $\mathbf{X}_m\mathbf{A}_m$ in linear GMCCA (4) gives rise to its dual form

$$\min_{\{\mathbf{A}_m\},\mathbf{S}} \quad \sum_{m=1}^{M} \left\|\mathbf{A}_m^\top\mathbf{X}_m^\top\mathbf{X}_m - \mathbf{S}\right\|_F^2 + \gamma \mathrm{Tr}\left(\mathbf{S}\mathbf{L}_\mathcal{G}\mathbf{S}^\top\right) \quad (5a)$$

$$\text{s. to} \quad \mathbf{S}\mathbf{S}^\top = \mathbf{I}. \quad (5b)$$

Invoking kernels, (5) can be generalized to capture nonlinear dependencies among the M views. Specifically, assuming M nonlinear functions $\{\boldsymbol{\phi}_m\}$, data vectors $\{\mathbf{x}_{m,i}\}$ in space $\mathbb{R}^{D_m}$ (columns of $\mathbf{X}_m$) are mapped to $\{\boldsymbol{\phi}_m(\mathbf{x}_{m,i})\}$ in space $\mathbb{R}^{L_m}$ with possibly $L_m = \infty$. Interestingly, the dual in (5) depends on $\mathbf{X}_m$ only through $\mathbf{X}_m^\top\mathbf{X}_m$. Using the 'kernel trick,' we can thus replace $\{\langle \mathbf{x}_{m,i}, \mathbf{x}_{m,j}\rangle\}_{i,j=1}^N$ with $\{\langle \boldsymbol{\phi}_m(\mathbf{x}_{m,i}), \boldsymbol{\phi}_m(\mathbf{x}_{m,j})\rangle\}_{i,j=1}^N$.

Define a kernel matrix $\bar{\mathbf{K}}_m \in \mathbb{R}^{N\times N}$ for each $\mathbf{X}_m$, whose (i,j)-th entry is $\kappa_m(\mathbf{x}_{m,i}, \mathbf{x}_{m,j}) := \langle \boldsymbol{\phi}_m(\mathbf{x}_{m,i}), \boldsymbol{\phi}_m(\mathbf{x}_{m,j})\rangle$, where $\kappa_m(\cdot)$ is a so-termed kernel function. Similar to GMCCA, we first remove the means of all transformed data $\{\boldsymbol{\phi}_m(\mathbf{x}_{m,i})\}_{i=1}^N$ to effect centering

$$\mathbf{K}_m := \bar{\mathbf{K}}_m - \mathbf{1}\bar{\mathbf{K}}_m/N - \bar{\mathbf{K}}_m\mathbf{1}/N + \mathbf{1}\bar{\mathbf{K}}_m\mathbf{1}/N^2 \quad (6)$$

where $\mathbf{1} \in \mathbb{R}^{N\times N}$ is an all-one matrix. In the sequel, replacing $\{\mathbf{X}_m^\top\mathbf{X}_m\}$ in (5) with the centered kernel matrices $\{\mathbf{K}_m\}$, the nonlinear counterpart of (5) can be obtained as

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

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

Kernel matrices $\{\mathbf{K}_m\}$ are assumed to be nonsingular. Analogous to the process of solving GMCCA, one can confirm that rows of the $\hat{\mathbf{S}}$-optimizer of (7) coincide with the d principal eigenvectors of $M\mathbf{I} - \gamma\mathbf{L}_\mathcal{G}$ and that $\hat{\mathbf{A}}_m = \mathbf{K}_m^{-1}\hat{\mathbf{S}}^\top$. Clearly, the common source estimate $\hat{\mathbf{S}}$ does not depend on $\{\mathbf{X}_m\}$, which contradicts our goal of finding the shared low-dimensional representation in $\{\mathbf{X}_m\}$. To bypass this impasse, following kernel CCA (see e.g., [14]), we penalize the norms of $\left\{\|\mathbf{U}_m\|_F^2 = \mathrm{Tr}\left(\mathbf{A}_m^\top\mathbf{K}_m\mathbf{A}_m\right)\right\}$ by introducing a Tikhonov regularization term on each loading vector. This yields our graph-regularized kernel (GK) MCCA as

$$\min_{\{\mathbf{A}_m\},\mathbf{S}} \quad \sum_{m=1}^{M} \left\|\mathbf{A}_m^\top\mathbf{K}_m - \mathbf{S}\right\|_F^2 + \gamma \mathrm{Tr}\left(\mathbf{S}\mathbf{L}_\mathcal{G}\mathbf{S}^\top\right) + \sum_{m=1}^{M} \epsilon_m \mathrm{Tr}\left(\mathbf{A}_m^\top\mathbf{K}_m\mathbf{A}_m\right) \quad (8a)$$

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

Algorithm 1 Graph-regularized kernel MCCA.

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

where hyper-parameters $\{\epsilon_m \geq 0\}$ are predetermined penalty constants. Similar to the process of solving (4), optimizers of (8) can be readily obtained; see Alg. 1 for details.

MCCA, GMCCA, GKMCCA, and KMCCA incur respectively computational complexity of $\mathcal{O}(N^2\max(N, DM))$, $\mathcal{O}(N^2\max(N, DM))$, $\mathcal{O}(N^2M\max(N, D))$, and $\mathcal{O}(N^2M\max(N, D))$. When $N \ll D_m$ for some $m \in \{1, \ldots, M\}$, GMCCA is not applicable, or suboptimal even if pseudo-inverse is used at a computational cost of order $\mathcal{O}(MD^3)$. On the other hand, GKMCCA is computationally more affordable since its cost grows only linearly with D. Furthermore, when $D_m \gg N$ for all m, it can be readily verified that GMCCA is computationally more attractive than GKMCCA.

5. NUMERICAL TESTS

The UCI digit image database[1] is used to demonstrate the effectiveness of GMCCA in clustering. This database comprises 6 feature sets of 10 digits (classes), each having 200 data samples. Seven classes including digits $1, 2, 3, 4, 7, 8, 9$ were used to form the 6 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 $D_6 = 6$. Based on $\mathbf{X}_3$, the $\mathbf{W}$ was constructed having (i, j)-th entry

$$w_{ij} := \begin{cases} \mathbf{K}_3(i, j), & i \in \mathcal{N}_{k_1}(j) \text{ or } j \in \mathcal{N}_{k_1}(i) \\ 0, & \text{otherwise} \end{cases}$$

where $\mathbf{K}_3$ is a Gaussian kernel matrix of $\mathbf{X}_3$ with bandwidth equal to the mean of the corresponding Euclidean distances, and $\mathcal{N}_{k_1}(j)$ the set of column indices of $\mathbf{K}_3$ containing the k_1-nearest neighbors of column j. GPCA and PCA were run on the concatenated feature vectors, while the K-means was performed using either $\hat{\mathbf{S}}$, or the principal components (PCs) with parameters $\gamma = 0.1$ and $d = 3$.

Clustering performance is evaluated in terms of both clustering accuracy and scatter ratio defined in [13, Sec. VII-A]. Table 1 depicts the clustering performance of MCCA, PCA, GMCCA, and GPCA under different k_1 values. Evidently, GMCCA achieves the best clustering accuracy and scatter ratio. For $k_1 = 50$, Fig. 1 reports the first two rows of $\hat{\mathbf{S}}$ obtained by (G)MCCA along with the first two PCs of (G)PCA,

[1] Downloaded from http://archive.ics.uci.edu/ml/datasets/Multiple+Features

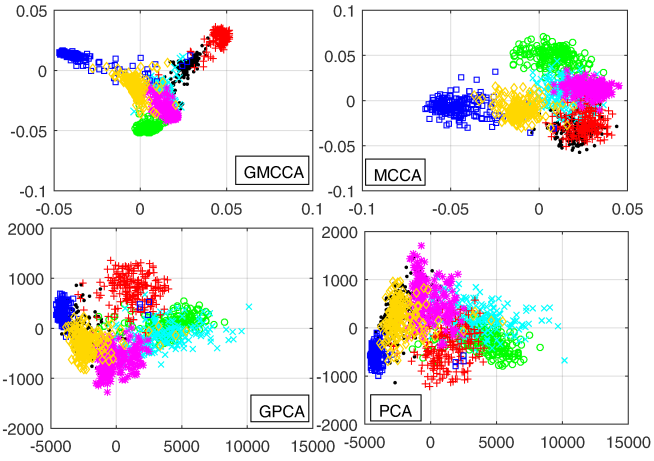

Fig. 1. Scatter plot of the first two rows of $\hat{\mathbf{S}}$ or PCs.

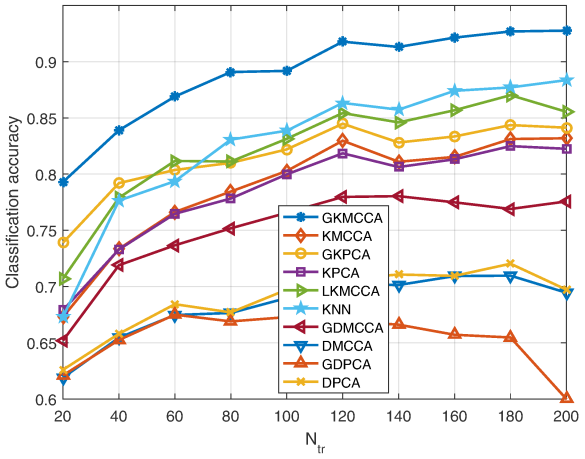

Fig. 2. Classification results of GKMCCA using MNIST data.

Table 1. 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	

where 7 different colors denote 7 classes. The scatter plots in Fig. 1 show that GMCCA separates the 7 clusters the best, implying that the data points within classes are more concentrated but between classes are farther apart.

The ability of GKMCCA in classification is assessed using the MNIST dataset[2], which consists of 10 classes of 28×28 handwritten images, each class (digit) having $7,000$ images. Per independent realization, we performed Coiflets, Symlets, and Daubechies orthonormal wavelet transforms on $3N_{\rm tr}$ randomly chosen images from each class to form the 3 views. Subsequently, the selected images were resized to 14×14 pixels, followed by vectorization to obtain 196×1 vectors, which were evenly divided into 3 groups ($N_{\rm tr}$ vectors per class per group) to construct the training data $\{\mathbf{X}_m \in \mathbb{R}^{196\times 10N_{\rm tr}}\}_{m=1}^3$, hyper-parameter tunning data $\{\mathbf{X}_m^{\rm tu} \in \mathbb{R}^{196\times 10N_{\rm tr}}\}_{m=1}^3$, and testing data $\{\mathbf{X}_m^{\rm te} \in \mathbb{R}^{196\times 10N_{\rm tr}}\}_{m=1}^3$. Gaussian kernels were used to build $\{\mathbf{K}_m\}$ for $\{\mathbf{X}_m\}$ with bandwidths set to the means of their associated distances. Similarly, the resized and vectorized training data were used to construct a Gaussian kernel matrix $\mathbf{K}_o \in \mathbb{R}^{10N_{\rm tr}\times 10N_{\rm tr}}$. Based on $\mathbf{K}_o$ and $k_1 = N_{\rm tr} - 1$, we formed $\mathbf{W}$ with $\mathbf{K}_3(i,j)$ substituted by the (i,j)-th entry of $\mathbf{K}_o$. When implementing graph Laplacian regularized kernel multi-view (LKM) CCA [10], the related three graph adjacency matrices were obtained with $\{\mathbf{K}_m\}_{m=1}^3$. Hyper-parameters of GKMCCA, KMCCA, GKPCA, KPCA, GDMCCA, DMCCA, GDPCA, and LKMCCA were chosen from 30 logarithmically spaced values in $[10^{-3}, 10^3]$ that yields the highest classification accuracy. Ten subspace vectors were obtained, and were further utilized to find the 10-dimensional representations of $\mathbf{X}_1$. The 5-nearest neighbors rule was employed for digit classification, and its accuracy is reported after averaging over 30 Monte Carlo runs.

Figure 2 depicts the classification accuracy of all considered approaches, and demonstrates that the performance gap between GKMCCA and any other competing alternative remains remarkably sizeable. See more numerical tests in the full version [15].

6. CONCLUSIONS

In this work, graph-regularized multiview (M) CCA and kernel MCCA were developed to uncover the latent low-dimensional structures commonly present in multiview data. Our distinct contributions relative to existing MCCA variants leverage extra geometrical knowledge of the common sources, encodes this dependency in a graph that is subsequently invoked as a regularizer in the standard MAXVAR MCCA framework. Numerical tests demonstrate the merits of our proposed approaches relative to state-of-the-art schemes in several machine learning tasks.

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

7. REFERENCES

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

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

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

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

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

[6] 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.

[7] 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.

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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, Feb. 2019.

[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] J. Chen, G. Wang, and G. B. Giannakis, "Graph multiview canonical correlation analysis," *IEEE Trans. Signal Process.*, submitted Nov. 2018. [Online]. Available: https://arxiv.org/abs/1811.12345.