Learning Canonical $\mathcal{F}$-Correlation Projection for Compact Multiview Representation

Yun-Hao Yuan[1], Jin Li[1], Yun Li[1], Jipeng Qiang[1], Yi Zhu[1], Xiaobo Shen[2], Jianping Gou[3]
[1]School of Information Engineering, Yangzhou University, Yangzhou, China
[2]School of Computer Science, Nanjing University of Science and Technology, Nanjing, China
[3]School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China
{yhyuan, liyun, jpqiang, zhuyi}@yzu.edu.cn
cvjinli@outlook.com, njust.shenxiaobo@gmail.com, goujianping@ujs.edu.cn

Abstract

Canonical correlation analysis (CCA) matters in multiview representation learning. But, CCA and its most variants are essentially based on explicit or implicit covariance matrices. It means that they have no ability to model the nonlinear relationship among features due to intrinsic linearity of covariance. In this paper, we address the preceding problem and propose a novel canonical $\mathcal{F}$-correlation framework by exploring and exploiting the nonlinear relationship between different features. The framework projects each feature rather than observation into a certain new space by an arbitrary nonlinear mapping, thus resulting in more flexibility in real applications. With this framework as a tool, we propose a correlative covariation projection (CCP) method by using an explicit nonlinear mapping. Moreover, we further propose a multiset version of CCP dubbed MCCP for learning compact representation of more than two views. The proposed MCCP is solved by an iterative method, and we prove the convergence of this iteration. A series of experimental results on six benchmark datasets demonstrate the effectiveness of our proposed CCP and MCCP methods.

1. Introduction

Canonical correlation analysis (CCA) [17], proposed by Hotelling, is a classic yet powerful technique in multidimensional data analysis for finding the relationship between two sets of random variables. CCA aims to seek pairs of linear projection vectors such that the projected variables are maximally correlated in the low-dimensional space. CCA has been applied to many applications such as multiview clustering [5], feature fusion [30], multilabel classification [29], and multiview representation learning (MRL) [23].

To date, there have been numerous useful variants of CCA, which can be broadly divided into three categories: supervised, semi-supervised, and unsupervised methods. Supervised variants of CCA take the label information of all the observations of different views into consideration during model training. For example, discriminative CCA [21, 32] was proposed by incorporating intraclass and interclass information of two view samples into CCA. Multilabel CCA [27,29] was presented by regarding one view as the data and the other view as class labels. To deal with multiview (more than two views) cases, some multiview supervised extensions of CCA have been proposed recently based on intraclass and interclass information of multiview observations; see, e.g., discriminative and labeled multiple CCA [11, 12].

Semi-supervised variants of CCA not only take advantage of unlabeled multiview data, but also labeled multiview data for learning latent representation. For example, Chen *et al.* [9] proposed a semi-supervised and semi-paired CCA method, which uses the global structure information of unlabeled data and local discriminative information of labeled data and meanwhile, exploits the limited paired data. Wan and Zhu [34] presented a cost-sensitive semi-supervised CCA approach, which carries out label propagation with CCA in a unified cost-sensitive learning framework.

In real-world applications, there are usually no shortage of unlabeled data but labels are expensive. Therefore, it is of great significance to develop unsupervised CCA's variants which can make full use of all of the multiview observations. In this paper, we consider the problem of CCA's generalization in unsupervised learning scenarios, which is a much harder problem owing to the absence of class labels that would guide the search for relevant information and compact multiview representation learning.

Much effort has been focused on unsupervised variants of CCA for compact multiview representation or feature learning. For example, in small sample size (SSS) problems where the dimensionality of feature vectors is larger than the number of observations, regularized CCA [16, 29] was proposed to prevent overfitting and the singularity of sample covariance matrices. Due to a large portion of features that are not informative to many multiview learning tasks, sparse CCA [3, 10, 15] was presented to learn pairs of projection directions with sparsity constraints for improving the interpretation ability of canonical projections. In addition, orthogonal CCA (OCCA) [41] was developed to preserve the covariance of the original data and the Euclidean metric structure of canonical subspaces. Probabilistic CCA [39] was proposed to provide a probabilistic interpretation for classical CCA. Recently, Xu and Li [37] proposed a truly alternating least-squares based CCA (TALSCCA) for performance improvement in practice.

For high-dimensional multiview data, there are usually very complex nonlinear associations in real world. From this perspective, unsupervised nonlinear extensions of CCA have attracted increasing attention in recent years. Early nonlinear generalizations of CCA mainly include kernel CCA (KCCA) [2, 16] and neural networks based CCA [18, 22]. The basic idea of KCCA is to implicitly project the input data of two views into higher-dimensional feature spaces by using two nonlinear mappings determined by some kernels. This makes it possible for the nonlinear relationship between two views in the original data spaces to become linear in feature spaces, thus allowing the use of classical CCA to learn the latent compact representations from two views. Recent kernel variants can be found in [14, 24, 33]. Neural networks based CCA combines CCA with neural networks for finding the nonlinear correlation between two views with different combination strategies. Due to the advantages of deep neural network (DNN), Andrew *et al.* [1] proposed a deep version of CCA, dubbed deep CCA (DCCA), which learns deep nonlinear representations of two view data by maximizing the linear correlation between the outputs of two individual DNNs. Moreover, there are some other deep extensions of CCA [13, 35, 40] that have been developed based on different DNN architectures. Another popular unsupervised nonlinear extensions of CCA are based on the theory of manifold learning; see, for example, locality preserving CCA [31] and graph CCA [7].

All the aforementioned unsupervised variants of CCA are only applicable to two view scenario. To handle multiview cases, multiset CCA (MCCA) [20,26] was proposed to simultaneously achieve multiple latent representation subspaces for multiple views. Recent years have witnessed an upsurge of research interests in multiview unsupervised extensions of CCA; see, for example, tensor CCA (TCCA) [25], deep probabilistic CCA [19], and TCCA network [38]. Another interesting multiview extensions are based on the idea of generalized CCA [4] that minimizes the difference between one common latent representation and individual representations, e.g., graph MCCA [6] and $L_{2,1}$-CCA [36].

Although CCA and its variants above have obtained impressive learning performance for MRL, most of them still face an intractable challenge. That is, CCA and its most variants (e.g., KCCA, DCCA, OCCA, and MCCA) are associated with the so-called *spectral* solvers, which are based on the top or bottom eigenvalues and eigenvectors of specially constructed data matrices. Such matrices specially constructed are essentially based on explicit or implicit covariance matrices, which evaluate the linear relationship of different features. This means that they have no ability to model the nonlinear relationship among *features* due to the intrinsic linearity of covariance measure.

In this paper, we address the preceding problem and propose a novel canonical $\mathcal{F}$-correlation framework by exploring and exploiting the nonlinear relationship between different features. By utilizing this framework as a tool, we propose a correlative covariation projection (CCP) method, where an explicit nonlinear mapping is used for the construction of $\mathcal{F}$-intraset and $\mathcal{F}$-interset covariation matrices. Further, we extend CCP and propose a multiset CCP (MCCP) approach (see Fig. 1 for illustration) for learning the uncorrelated compact representation of more than two views, which is desirable in many real applications. Extensive experimental results on six real-world datasets show that our proposed CCP and MCCP methods outperform related methods including the state-of-the-arts in classification and clustering tasks. It is worthwhile to highlight the contributions of this paper as follows:

1) To the best of our knowledge, canonical $\mathcal{F}$-correlation framework is novel in MRL, where an arbitrary nonlinear mapping can be used to project each *feature* rather than *observation* into a certain new space. Hence, our proposed framework is naturally capable of modeling the nonlinear relationship between different features, which can not be well modeled by traditional CCA-related methods.

2) With this framework, CCP is proposed by explicitly using a specific Gaussian kernel mapping, which not only has the great capability to discover the nonlinear feature relevance, but also to perform the MRL whether the scenario is an SSS problem or not.

3) To perform MRL beyond the limit of two views, MCCP is proposed by maximizing the accumulated correlations between any pair of views, which is solved by an iterative method. The convergence of this iteration is demonstrated theoretically.

2. Covariance and CCA

2.1. Covariance

Let $\mathbf{f}_1 \in \mathbb{R}^n$ and $\mathbf{f}_2 \in \mathbb{R}^n$ contain n observations of two feature variables. It is straightforward to compute their

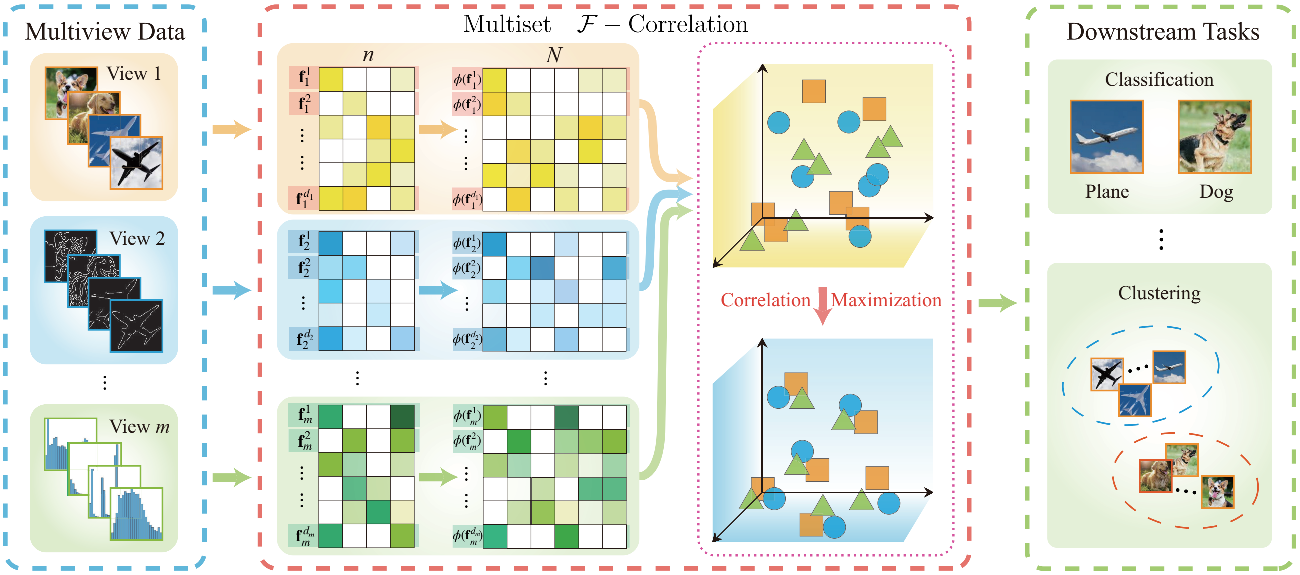

Figure 1. The flowchart of our framework. In the original data space, each feature $\mathbf{f}_i^t$ is mapped into a certain new space by using a nonlinear function $\phi(\cdot)$, where $i = 1, 2, \cdots, m$ and $t = 1, 2, \cdots, d_i$, m and d_i are, respectively, the number of views and dimension of i-th view. Then, the ϕ-transformed multiview data are maximally correlated with minimum redundancy within each view. The resulting compact representations are used for downstream multiview tasks such as classification and clustering.

covariance in the form of

$$\mathrm{Cov}(\mathbf{f}_1, \mathbf{f}_2) = \frac{1}{n}(\mathbf{f}_1 - \mu_1\mathbf{1})^T(\mathbf{f}_2 - \mu_2\mathbf{1}), \qquad (1)$$

where $\mathrm{Cov}(\cdot, \cdot)$ denotes the covariance operator, μ_1 and μ_2 are the mean of all n observations of $\mathbf{f}_1$ and $\mathbf{f}_2$, respectively, i.e., $\mu_1 = 1/n(\mathbf{f}_1^T\mathbf{1})$ and $\mu_2 = 1/n(\mathbf{f}_2^T\mathbf{1})$, and $\mathbf{1}$ is an all-one vector. As shown in (1), it is easy to find that the covariance is the inner product of two scaled vectors in essence. Hence, it actually measures the linear relationship between two feature variables.

2.2. CCA

Assume two-view data are given as $\mathbf{X}_1 \in \mathbb{R}^{d_1 \times n}$ and $\mathbf{X}_2 \in \mathbb{R}^{d_2 \times n}$, where d_1 and d_2 are the dimension of samples, and n is the number of samples. CCA aims to maximize the correlation between $\mathbf{w}_1^T\mathbf{X}_1$ and $\mathbf{w}_2^T\mathbf{X}_2$, which can be expressed as the following optimization problem:

$$\begin{aligned} &\max_{\mathbf{w}_1, \mathbf{w}_2} \ \mathbf{w}_1^T\mathbf{S}_{12}\mathbf{w}_2 \\ &s.t.\ \mathbf{w}_1^T\mathbf{S}_{11}\mathbf{w}_1 = \mathbf{w}_2^T\mathbf{S}_{22}\mathbf{w}_2 = 1, \end{aligned} \qquad (2)$$

where $\mathbf{w}_1 \in \mathbb{R}^{d_1}$ and $\mathbf{w}_2 \in \mathbb{R}^{d_2}$ are a pair of projection vectors, $\mathbf{S}_{12}$ is the between-set covariance matrix of $\mathbf{X}_1$ and $\mathbf{X}_2$, $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ are within-set covariance matrices of $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. It has been shown that optimization problem in (2) can be solved by the following eigenequation

$$\begin{bmatrix} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{S}_{11} & \\ & \mathbf{S}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix}, \qquad (3)$$

where λ is the eigenvalue corresponding to the eigenvector $[\mathbf{w}_1^T\ \mathbf{w}_2^T]^T$, and $\mathbf{S}_{21} = \mathbf{S}_{12}^T$. In CCA, multiple pairs of projection vectors can be obtained by computing the eigenvectors corresponding to the top eigenvalues of (3).

3. Proposed method

3.1. Motivation

As introduced in Section 2.2, CCA is based on within- and between-set covariance matrices. But, these covariance matrices have at least two common drawbacks, as follows:

- First, the covariance metric defined in (1) only assesses the linear variability of two random variables, as discussed in Section 2.1. Thus, covariance matrices in CCA naturally lack the capability to evaluate the nonlinear relationship among different features. Clearly, it limits the representation learning performance of CCA in real world.
- Second, in SSS cases, it is well-known that within-set covariance matrices in CCA are bound to be singular, thus making CCA unavailable for two-view representation learning. Although regularized CCA [16, 29] can tackle the singularity issue of covariance matrices, so far as we know, it is still not clear how to choose the optimal regularization parameters theoretically and practically.

To overcome the above shortcomings, we explore and exploit the nonlinear relationships between different features and thus propose a CCP method and its extension for MRL.

3.2. Formulation

Canonical $\mathcal{F}$-correlation framework. Let $\mathbf{f}_i^t \in \mathbb{R}^n$ be the feature vector corresponding to the t-th row of $\mathbf{X}_i$ in i-th view, $i = 1, 2$ and $t = 1, 2, \cdots, d_i$. Then, $\mathbf{X}_i$ can be rewritten as the form of $\mathbf{X}_i = [\mathbf{f}_i^1, \mathbf{f}_i^2, \cdots, \mathbf{f}_i^{d_i}]^T$. Let

$$\boldsymbol{\phi}(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}^N$$

be a nonlinear mapping function from $\mathbb{R}^n$ to $\mathbb{R}^N$, where N is the dimension of a certain new space. We use $\boldsymbol{\phi}(\cdot)$ to map all n-dimensional feature vectors $\{\mathbf{f}_i^t\}_{t=1}^{d_i}$ in the original input spaces into a new N-dimensional space and obtain

$$\boldsymbol{\phi}(\mathbf{f}_i^t) = \left[\phi^1(\mathbf{f}_i^t), \phi^2(\mathbf{f}_i^t), \cdots, \phi^N(\mathbf{f}_i^t)\right]^T. \tag{4}$$

Using (4), we can define the *$\mathcal{F}$-intraset and $\mathcal{F}$-interset covariation matrices* in $\mathbb{R}^N$ as:

$$\mathbf{C}_{ij}^{\mathcal{F}} = \left[\mathbf{C}_{ij}^{\mathcal{F}}(k, t)\right] = \left[\boldsymbol{\phi}(\mathbf{f}_i^k)^T \boldsymbol{\phi}(\mathbf{f}_j^t)\right] \in \mathbb{R}^{d_i \times d_j}, \tag{5}$$

where $\mathbf{C}_{ij}^{\mathcal{F}}(k, t)$ denotes the (k, t)-th entry of $\mathbf{C}_{ij}^{\mathcal{F}}$, $i, j = 1, 2$, $k = 1, 2, \cdots, d_i$, and $t = 1, 2, \cdots, d_j$.

Through using (5), our proposed canonical $\mathcal{F}$-correlation framework aims to seek a pair of linear transformations $\mathbf{W}_1 \in \mathbb{R}^{d_1 \times d}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_2 \times d}$ ($d \leq \min(d_1, d_2)$) by the following optimization problem:

$$\begin{aligned} &\max_{\mathbf{W}_1, \mathbf{W}_2} \ \operatorname{Tr}(\mathbf{W}_1^T \mathbf{C}_{12}^{\mathcal{F}} \mathbf{W}_2) \\ &s.t. \quad \mathbf{W}_1^T \mathbf{C}_{11}^{\mathcal{F}} \mathbf{W}_1 = \mathbf{I}_d, \ \mathbf{W}_2^T \mathbf{C}_{22}^{\mathcal{F}} \mathbf{W}_2 = \mathbf{I}_d, \end{aligned} \tag{6}$$

where $\operatorname{Tr}(\cdot)$ denotes the trace of a matrix and $\mathbf{I}_d \in \mathbb{R}^{d \times d}$ is the identity matrix.

Discussion: In contrast to conventional nonlinear variants of CCA such as KCCA and DCCA, which make use of the nonlinear mapping to project each observation in two views, our framework employs a nonlinear function $\boldsymbol{\phi}(\cdot)$ to map *each feature* rather than *each observation* into a certain new space, thus leading to the following four advantages: 1) In the proposed framework, our $\mathcal{F}$-intraset and $\mathcal{F}$-interset covariation matrices can better uncover the nonlinear relationship hidden in different features, while within- and between-set covariance matrices in, for instance, CCA and OCCA, fail to model this relationship due to their inherent linearity. 2) Our canonical $\mathcal{F}$-correlation framework projects each original feature vector of both views into an N-dimensional space. As $N > n$, it seems to mean that our framework naturally yields new observation components for learning compact multiview representations. It is obvious to benefit SSS cases. 3) The presented framework has no change in the dimensionality of two-view observations. In contrast, KCCA and its variants map original observations of two views into higher- or even infinite-dimensional feature spaces, where numerous large-scale learning problems are turned into SSS cases. 4) Due to the flexibility of use of nonlinear function $\boldsymbol{\phi}(\cdot)$, the proposed framework can be taken as a general platform, where new canonical correlation methods for compact multiview representation learning are further developed. In addition, our framework can incorporate CCA as a special case when $\boldsymbol{\phi}(\mathbf{f}_i^t) = \mathbf{f}_i^t$ and each feature $\mathbf{f}_i^t$ is centered.

CCP. The nonlinear function $\boldsymbol{\phi}(\cdot)$ in (6) can be chosen as kernel mappings, neural networks, deep networks, etc. Thus, different $\boldsymbol{\phi}(\cdot)$ will lead to different special methods. In this paper, we select $\boldsymbol{\phi}(\cdot)$ as Gaussian kernel mapping[1], thus resulting in optimization problem of CCP as follows:

$$\begin{aligned} &\max_{\mathbf{W}_1, \mathbf{W}_2} \ \operatorname{Tr}(\mathbf{W}_1^T \mathbf{K}_{12}^{\mathcal{F}} \mathbf{W}_2) \\ &s.t. \quad \mathbf{W}_1^T \mathbf{K}_{11}^{\mathcal{F}} \mathbf{W}_1 = \mathbf{I}_d, \ \mathbf{W}_2^T \mathbf{K}_{22}^{\mathcal{F}} \mathbf{W}_2 = \mathbf{I}_d, \end{aligned} \tag{7}$$

where $\mathbf{K}_{ij}^{\mathcal{F}} \in \mathbb{R}^{d_i \times d_j}$ and its (k, t)-th entry is computed by

$$\mathbf{K}_{ij}^{\mathcal{F}}(k, t) = \boldsymbol{\phi}(\mathbf{f}_i^k)^T \boldsymbol{\phi}(\mathbf{f}_j^t) = \ker(\mathbf{f}_i^k, \mathbf{f}_j^t) \tag{8}$$

with $\ker(\cdot, \cdot)$ as a Gaussian kernel function, i.e.,

$$\ker(\mathbf{f}_i^k, \mathbf{f}_j^t) = \exp\left(-\|\mathbf{f}_i^k - \mathbf{f}_j^t\|^2 / (2\sigma^2)\right), \tag{9}$$

$\sigma > 0$ is the width parameter of Gaussian kernel, $\|\cdot\|$ denotes the 2-norm of a vector, $i, j = 1, 2$, $k = 1, 2, \cdots, d_i$, and $t = 1, 2, \cdots, d_j$. Clearly, $\mathbf{K}_{11}^{\mathcal{F}}$ and $\mathbf{K}_{22}^{\mathcal{F}}$ are symmetric positive semi-definite matrices.

3.3. Solution of CCP

Theorem 1. Suppose $\{\mathbf{f}_i^t \in \mathbb{R}^n\}_{t=1}^{d_i}$ are a set of d_i distinct feature vectors in view i, where $i = 1, 2$. Then, $\mathcal{F}$-intraset covariation matrices, $\mathbf{K}_{11}^{\mathcal{F}}$ and $\mathbf{K}_{22}^{\mathcal{F}}$, must be nonsingular.

The proof of Theorem 1 can be found in the supplementary material. According to Theorem 1, it is clear that $\mathbf{K}_{11}^{\mathcal{F}}$ and $\mathbf{K}_{22}^{\mathcal{F}}$ are bound to be symmetric positive definite. Thus,

$$\mathbf{K}_{ii}^{\mathcal{F}} = (\mathbf{K}_{ii}^{\mathcal{F}})^{\frac{1}{2}} (\mathbf{K}_{ii}^{\mathcal{F}})^{\frac{1}{2}}, \ i = 1, 2, \tag{10}$$

must exist. Due to the nonsingularity of $\mathbf{K}_{11}^{\mathcal{F}}$ and $\mathbf{K}_{22}^{\mathcal{F}}$, let

$$\mathbf{W}_i = (\mathbf{K}_{ii}^{\mathcal{F}})^{-\frac{1}{2}} \tilde{\mathbf{W}}_i, \ i = 1, 2. \tag{11}$$

Let $\tilde{\mathbf{K}}_{12}^{\mathcal{F}} = (\mathbf{K}_{11}^{\mathcal{F}})^{-\frac{1}{2}} \mathbf{K}_{12}^{\mathcal{F}} (\mathbf{K}_{22}^{\mathcal{F}})^{-\frac{1}{2}}$. Together with (11), optimization problem in (7) can be equivalently reformulated as

$$\begin{aligned} &\max_{\tilde{\mathbf{W}}_1, \tilde{\mathbf{W}}_2} \ \operatorname{Tr}(\tilde{\mathbf{W}}_1^T \tilde{\mathbf{K}}_{12}^{\mathcal{F}} \tilde{\mathbf{W}}_2) \\ &s.t. \ \tilde{\mathbf{W}}_1^T \tilde{\mathbf{W}}_1 = \mathbf{I}_d, \ \tilde{\mathbf{W}}_2^T \tilde{\mathbf{W}}_2 = \mathbf{I}_d. \end{aligned} \tag{12}$$

For the solution to optimization problem in (12), we have the following important theorem:

[1] Kernel mapping possesses the remarkable advantage that the associated kernel matrix is positive semi-definite.

Theorem 2. Let singular value decomposition (SVD) of $\tilde{\mathbf{K}}_{12}^{\mathcal{F}}$ be $\tilde{\mathbf{K}}_{12}^{\mathcal{F}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \cdots, \sigma_r) \in \mathbb{R}^{r \times r}$, where $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_r$, $\{\sigma_i\}_{i=1}^r$ are the singular values in descending order, i.e., $\sigma_1 \geq \cdots \geq \sigma_r > 0$, and $r = \mathrm{rank}(\tilde{\mathbf{K}}_{12}^{\mathcal{F}}) \leq \min(d_1, d_2)$. Then,

$$\tilde{\mathbf{W}}_1 = \mathbf{U}(:, 1 : d) \text{ and } \tilde{\mathbf{W}}_2 = \mathbf{V}(:, 1 : d) \tag{13}$$

are a solution of (12), where $\mathbf{A}(:, 1 : d)$ denotes the matrix that consists of the first d columns of $\mathbf{A}$, and $d \leq r$.

The proof of Theorem 2 can be found in the supplementary material. Using (11), we can obtain the resulting projection matrices of CCP in the form of $\mathbf{W}_1 = (\mathbf{K}_{11}^{\mathcal{F}})^{-\frac{1}{2}}\mathbf{U}(:, 1 : d)$ and $\mathbf{W}_2 = (\mathbf{K}_{22}^{\mathcal{F}})^{-\frac{1}{2}}\mathbf{V}(:, 1 : d)$.

3.4. Out-of-sample projection

For an unseen sample $\mathbf{x}^T = [\mathbf{x}_1^T, \mathbf{x}_2^T]$ with $\mathbf{x}_i \in \mathbb{R}^{d_i}$, we can compute its latent representation in the form of $\mathbf{y}_i = \mathbf{W}_i^T\mathbf{x}_i$, $i = 1, 2$. For $\mathbf{y}_1$ and $\mathbf{y}_2$, we combine them by the following strategy: $\mathbf{y}^T = [\mathbf{y}_1^T, \mathbf{y}_2^T]$, which is used to represent sample $\mathbf{x}$ for downstream tasks.

3.5. Comparison with other methods

Comparison with CCA. CCP and CCA are both unsupervised two-view subspace learning methods. Differently, CCP has the capability to model the nonlinear relationships between different features of two views, while CCA does not. On the other hand, when $d_1 > n$ and $d_2 > n$, CCP can efficiently learn the latent low-dimensional representation due to the nonsingularity of $\mathcal{F}$-intraset covariation matrices. In this situation, CCA is not available owing to the singularity of within-set covariance matrices.

Comparison with KCCA. Both CCP and KCCA use the nonlinear mapping to project two view inputs into new spaces. They, however, are quite different. First, KCCA uses kernel functions to essentially measure the similarity between two samples, while our CCP to evaluate the similarity between any two features. Second, KCCA often suffers from a famous trivial learning problem [16]. Recall that, in KCCA, dual representation vector in one view is obtained by the following naive eigenvalue problem [16]:

$$\mathbf{I}_n\mathbf{u} = \lambda\mathbf{u} \tag{14}$$

when kernel matrix is invertible, where λ is the eigenvalue associated with the eigenvector $\mathbf{u}$. Clearly (14) has nothing to do with training data, thus providing useless results. In contrast, our CCP does not encounter this problem, which can be efficiently solved by an SVD; see Section 3.3. Third, KCCA needs to employ all training data when computing the representations of unseen observations. This makes the testing phase of KCCA heavily dependent on training data as well as time-consuming in large-scale learning problems. On the contrary, our CCP computes two explicit projection matrices $\mathbf{W}_1$ and $\mathbf{W}_2$, and directly applies them to unseen observations, as shown in Section 3.4. Besides, the size of kernel matrix in KCCA is $n \times n$, while our $\mathcal{F}$-interset covariation matrix is of size $d_1 \times d_2$. This suggests that CCP can learn more features than KCCA when $\min(d_1, d_2) > n$.

4. Extension

We extend CCP to learn compact representation from more than two views, which thus leads to a multiset CCP (MCCP) method. Given m-view training samples $\{\mathbf{X}_i = [\mathbf{f}_i^1, \mathbf{f}_i^2, \cdots, \mathbf{f}_i^{d_i}]^T \in \mathbb{R}^{d_i \times n}\}_{i=1}^m$, the optimization problem of MCCP can be formulated as

$$\max_{\mathbf{W}_1, \cdots, \mathbf{W}_m} \sum_{i=1}^m \sum_{j=1}^m \mathrm{Tr}(\mathbf{W}_i^T\mathbf{K}_{ij}^{\mathcal{F}}\mathbf{W}_j) \quad s.t. \quad \mathbf{W}_i^T\mathbf{K}_{ii}^{\mathcal{F}}\mathbf{W}_i = \mathbf{I}_d, \ i = 1, 2, \cdots, m, \tag{15}$$

where $\mathbf{K}_{ij}^{\mathcal{F}} \in \mathbb{R}^{d_i \times d_j}$ with (k, t)-th element as $\ker(\mathbf{f}_i^k, \mathbf{f}_j^t)$, $i, j = 1, 2, \cdots, m$, $k = 1, 2, \cdots, d_i$, and $t = 1, 2, \cdots, d_j$. Clearly, MCCP reduces to CCP when $m = 2$.

Since the orthogonal constraints are nonconvex, (15) is a nonconvex optimization problem. Except for the $m = 2$ case, the nonconvexity of this problem makes it intractable to solve. Thus, we resort to an iterative method to compute these variables.

Let $\tilde{\mathbf{w}}_i = (\mathbf{K}_{ii}^{\mathcal{F}})^{\frac{1}{2}}\mathbf{w}_i$ with $\mathbf{w}_i$ as k-th column of $\mathbf{W}_i$. Assume we have obtained the first $k-1$ sets of projection vectors, i.e., $\mathbf{Q}_i = [\tilde{\mathbf{w}}_i^1, \tilde{\mathbf{w}}_i^2, \cdots, \tilde{\mathbf{w}}_i^{k-1}] \in \mathbb{R}^{d_i \times (k-1)}$, $i = 1, 2, \cdots, m$ and $1 \leq k \leq d$. Then, optimization problem in (15) can be reformulated as

$$\max_{\tilde{\mathbf{w}}_1, \cdots, \tilde{\mathbf{w}}_m} \sum_{i=1}^m \sum_{j=1}^m \tilde{\mathbf{w}}_i^T\tilde{\mathbf{K}}_{ij}^{\mathcal{F}}\tilde{\mathbf{w}}_j \quad s.t. \quad \tilde{\mathbf{w}}_i^T\tilde{\mathbf{w}}_i = 1, \ \mathbf{Q}_i^T\tilde{\mathbf{w}}_i = \mathbf{0}, \ i = 1, 2, \cdots, m, \tag{16}$$

where $\tilde{\mathbf{K}}_{ij}^{\mathcal{F}} = (\mathbf{K}_{ii}^{\mathcal{F}})^{-\frac{1}{2}}\mathbf{K}_{ij}^{\mathcal{F}}(\mathbf{K}_{jj}^{\mathcal{F}})^{-\frac{1}{2}}$.

Let $\mathbf{P}_i = \mathbf{I}_{d_i} - \mathbf{Q}_i\mathbf{Q}_i^T$. It is obvious that each $\mathbf{P}_i$ is a projection operator which maps a vector to the space orthogonal to the range space of $\mathbf{Q}_i$. Through $\{\mathbf{P}_i\}_{i=1}^m$, optimization problem in (16) can be equivalently reformulated as

$$\max_{\tilde{\mathbf{w}}_1, \cdots, \tilde{\mathbf{w}}_m} \sum_{i=1}^m \sum_{j=1}^m \tilde{\mathbf{w}}_i^T\mathbf{P}_i^T\tilde{\mathbf{K}}_{ij}^{\mathcal{F}}\mathbf{P}_j\tilde{\mathbf{w}}_j \quad s.t. \quad \tilde{\mathbf{w}}_i^T\tilde{\mathbf{w}}_i = 1, \ i = 1, 2, \cdots, m. \tag{17}$$

Using the Lagrange multiplier technique, we can obtain the following updating rules:

$$\lambda_i \leftarrow \Big\| \sum_{j=1}^m \mathbf{P}_i^T\tilde{\mathbf{K}}_{ij}^{\mathcal{F}}\mathbf{P}_j\tilde{\mathbf{w}}_j \Big\|, \tag{18}$$

$$\tilde{\mathbf{w}}_i \leftarrow \frac{1}{\lambda_i} \sum_{j=1}^m \mathbf{P}_i^T\tilde{\mathbf{K}}_{ij}^{\mathcal{F}}\mathbf{P}_j\tilde{\mathbf{w}}_j. \tag{19}$$

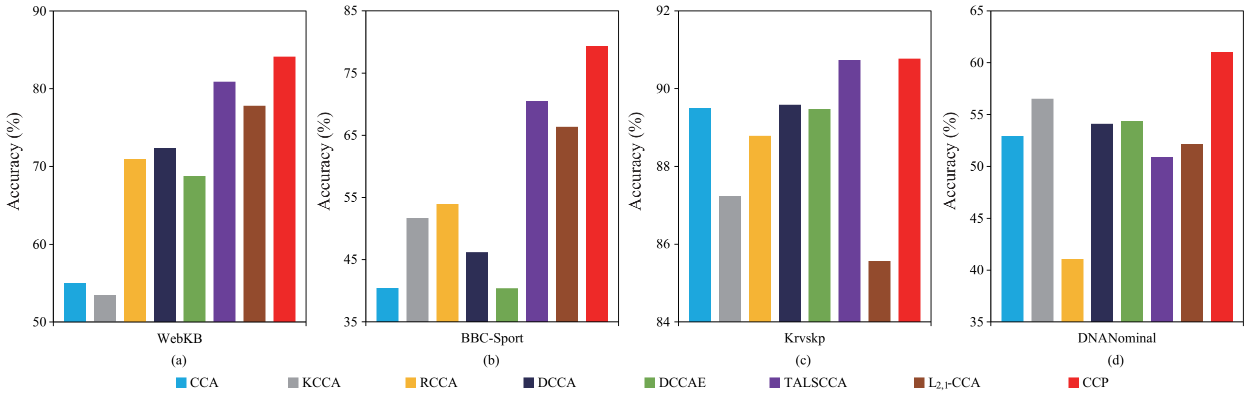

Figure 2. Classification accuracy of each method with two views on (a) WebKB, (b) BBC-Sport, (c) Krvskp, and (d) DNANominal.

Method	Yale-RG	Yale-RL	Yale-GL	COIL-20-RG	COIL-20-RL	COIL-20-GL
CCA	80.4 ± 4.70	73.8 ± 4.94	74.5 ± 3.52	58.2 ± 2.44	59.6 ± 1.89	47.5 ± 1.93
KCCA	76.8 ± 2.91	71.8 ± 3.31	78.3 ± 1.86	71.7 ± 2.76	59.6 ± 5.72	64.0 ± 4.11
RCCA	78.4 ± 2.79	73.4 ± 4.79	76.4 ± 3.77	69.3 ± 2.01	71.1 ± 2.04	63.3 ± 4.01
DCCA	79.6 ± 3.63	72.8 ± 4.03	79.5 ± 2.70	72.2 ± 2.34	70.5 ± 2.30	69.3 ± 1.68
DCCAE	79.8 ± 3.60	72.4 ± 3.59	$\underline{80.1 \pm 2.20}$	72.5 ± 2.31	70.1 ± 2.03	$\underline{69.4 \pm 3.35}$
TALSCCA	$\underline{80.8 \pm 4.10}$	75.9 ± 4.52	77.3 ± 3.19	$\underline{72.7 \pm 3.81}$	70.5 ± 3.12	68.1 ± 2.37
$L_{2,1}$-CCA	80.3 ± 4.50	$\underline{77.9 \pm 5.26}$	77.0 ± 4.29	61.1 ± 2.08	$\underline{72.4 \pm 1.76}$	65.2 ± 1.87
CCP	$\mathbf{82.9 \pm 4.60}$	$\mathbf{80.1 \pm 6.57}$	$\mathbf{82.9 \pm 3.05}$	$\mathbf{76.9 \pm 2.50}$	$\mathbf{75.9 \pm 2.49}$	$\mathbf{72.7 \pm 1.82}$

Table 1. Classification accuracy (%) of each method with two views on Yale and COIL-20 and corresponding standard deviations.

The detailed derivation of (18) and (19) can be found in the supplementary material. Regarding these two updating rules, we have the following convergence theorem:

Theorem 3. The objective function in (17) is nondecreasing and converges under the updating rules in (18) and (19).

The detailed proof of Theorem 3 can be found in the supplementary material.

5. Experiments

Extensive experiments on six real-world datasets are carried out to demonstrate the effectiveness of CCP and MCCP for learning compact multiview representation. We compare CCP and MCCP with existing related methods including the state-of-the-arts in classification and clustering.

5.1. Data preparation

The statistics of six datasets are summarized below: **WebKB** [28]: It contains 1051 samples of two classes (i.e., course and non-course), where there are 230 samples in course and 821 samples in non-course. Each sample has two views that are 1840-dimensional hyperlink feature and 3000-dimensional textual content feature. **BBC-Sport** [8]: It contains 544 documents of five topic classes collected from the BBC Sport website. Each document is divided into two sub-parts, thus yielding two different views with dimensions as 3183 and 3203, respectively. **Krvskp** [42]: It includes 3196 samples of two classes with 36 dimensions. We use the first 18 dimensions of each sample as one view and the rest as the other view for yielding two views. **DNANominal** [42]: It consists of 3186 gene samples with dimensionality as 60, which are classified into three classes: acceptor, donor, and non-splice. Likewise, the first 30 dimensions of each gene sequence are used as one view and the rest as the other view. **Yale** [8]: It consists of 165 grayscale images of 15 individuals with various lighting conditions and facial expressions. Each individual has 11 frontal facial images with size of 120×91. Three views are yielded by using raw images, Gabor, and local binary pattern (LBP) features, whose respective dimensions are 10920. **COIL-20** [13]: It contains 1440 grayscale images of 20 objects taken at pose intervals of 5 degrees. Each object has 72 images with size of 128×128. On this dataset, we employ raw images, Gabor, and LBP features to form three views, each with 16384 dimensions.

5.2. Compared methods

To demonstrate how the learning performance can be improved by our methods, we compare the following ten meth-

Dataset	Metric	CCA	KCCA	RCCA	DCCA	DCCAE	TALSCCA	MCCA	$L_{2,1}$-CCA	MCCP
Yale	ACC	76.3	72.9	73.6	77.0	76.4	78.8	77.9	82.2	**82.5**
	Std	4.60	3.77	2.61	1.72	2.49	3.45	3.67	3.73	3.17
COIL-20	ACC	59.6	62.2	58.6	75.3	74.5	70.1	56.6	66.3	**76.9**
	Std	2.30	3.44	3.87	2.60	2.61	2.73	2.37	2.15	1.89

Table 2. Classification accuracy (%) of each method with more than two views on Yale and COIL-20 and corresponding standard deviations.

ods: **CCA**, a classic yet powerful tool for two view representation learning. **KCCA**, a kernel extension of CCA. In our experiments, the kernel function is chosen as Gaussian kernel and optimal parameter is selected between 1 and 25 at sampled interval of 1. **Randomized CCA (RCCA)** [24], a randomized and nonlinear variant of CCA. **DCCA** [1], a deep nonlinear variant of CCA. We use the MATLAB code[2] to implement DCCA. **Deep canonically correlated autoencoders (DCCAE)** [35], which not only maximizes the canonical correlation between the learned deep representations, but also minimizes the reconstruction errors of two autoencoders. The MATLAB package[2] is adopted to perform DCCAE. **TALSCCA** [37], an alternating least-squares based CCA method. **MCCA** [26], a multiview extension of CCA that can learn the compact representations of more than two views. **$L_{2,1}$-CCA** [36], which takes $L_{2,1}$-norm constraints into consideration for compact MRL. **CCP** and **MCCP**, which are two new methods proposed in this paper. There is a parameter σ in our methods, as shown in (9).We empirically set it as $\sigma = 10^{i-6}$, $i = 0, 1, \cdots, 8$ and choose the value with the best CCP and MCCP performance. Note that all the foregoing ten methods involve a principal component analysis (PCA) phase. In this PCA phase, we keep more than 98% data energy of each view for all datasets.

5.3. Classification results

We randomly choose ten samples per class on WebKB and BBC-Sport, three samples per class on Yale and COIL-20, and 500 samples per class on Krvskp and DNANominal to generate the training sets, respectively, while the rest are used for testing. On each dataset, 10 independent tests are performed to test the performance. The nearest neighbor classifier is used and we explore the performance of each method on all possible feature dimensions and report the best results.

Two-view scenario. For WebKB, BBC-Sport, Krvskp, and DNANominal, we directly carry out two-view classification. Fig. 2 shows the average classification accuracy of each method. For Yale and COIL-20, there are three different pairwise view combinations in total, i.e., raw-Gabor, raw-LBP, and Gabor-LBP, which are denoted as X-RG, X-RL, and X-GL, respectively, and X denotes the dataset. The classification is performed on each pairwise combination. Table 1 lists the average classification results across ten runs of each method and the corresponding standard deviations.

From Fig. 2, we can see the following interesting points. First, on both WebKB and BBC-Sport, our CCP method obviously outperforms the other seven methods. Second, on Krvskp , CCP and TALSCCA perform comparably and better than the other six methods, and CCA, DCCA, and DCCAE perform comparably to one another. In addition, on this dataset, it is particularly worth noticing that $L_{2,1}$-CCA performs the worst, although it is a dedicated MRL method. Third, on DNANominal, CCP performs much better than other methods. From Table 1, we can clearly see that CCP consistently outperforms other methods, irrespective of pairwise view combinations. These results suggest that our CCP is a powerful tool for compact MRL and classification tasks.

More than two view scenario. For a fair comparison, we use the same preprocessing strategy as in [30] to transform three views into two views, and then CCA, KCCA, RCCA, DCCA, DCCAE, and TALSCCA are carried out based on the two new views, respectively. For MCCA, $L_{2,1}$-CCA, and our MCCP, we directly use them to learn compact multiview representation from three views. Table 2 summarizes the average classification accuracy (ACC) of these nine methods on Yale and COIL-20 and the corresponding standard deviations (Std). As can be seen, our proposed MCCP method is superior to the other eight methods on Yale and COIL-20. These results demonstrate that the proposed MCCP method is effective for multiview classification.

5.4. Clustering results

We evaluate the clustering performance of our proposed CCP and MCCP methods in multiview clustering tasks. For each dataset, the number of clusters is set to the class number. All the methods are, respectively, used for compact MRL. After MRL, we apply K-means algorithm for clustering. The accuracy (ACC) and normalized mutual information (NMI) are used to estimate the clustering quality. Note that the higher the values of ACC and NMI are, the better each method performs. Table 3 reports the clustering performance of each method with two views under two met-

[2] https://ttic.uchicago.edu/~wwang5/dccae.html

Dataset	Metric	CCA	KCCA	RCCA	DCCA	DCCAE	TALSCCA	$L_{2,1}$-CCA	CCP
WebKB	ACC	0.9610	0.9191	0.9458	0.9762	0.9686	0.9638	0.9743	**0.9800**
	NMI	0.7225	0.5155	0.6267	0.7986	0.7501	0.7371	0.7824	**0.8290**
BBC-Sport	ACC	0.8750	0.7261	0.8254	0.6691	0.7757	0.8621	0.6581	**0.9430**
	NMI	0.8012	0.6984	0.6652	0.4982	0.5367	0.7932	0.6870	**0.8352**
Krvskp	ACC	0.5563	0.5241	0.5701	0.5645	0.5350	0.5319	0.5917	**0.6834**
	NMI	0.0090	0.0143	0.0482	0.0561	0.0439	0.0036	0.0616	**0.1148**
DNANominal	ACC	0.5182	0.5436	0.3685	0.5063	0.5543	0.5712	0.5022	**0.6871**
	NMI	0.1296	0.0775	0.0309	0.1038	0.1445	0.1408	0.1299	**0.3022**
Yale-RG	ACC	0.5394	0.6788	0.6364	0.5576	0.6364	0.6667	**0.7273**	**0.7273**
	NMI	0.6430	0.7273	0.7508	0.7416	0.7574	0.7504	0.7656	**0.8112**
Yale-RL	ACC	0.5818	0.6364	0.5879	0.6424	0.6545	0.5697	0.5879	**0.7030**
	NMI	0.7059	0.6807	0.6892	0.6939	0.6979	0.7102	0.7221	**0.7408**
Yale-GL	ACC	0.6545	0.6182	0.6061	0.6424	0.5818	0.6061	0.6667	**0.7394**
	NMI	0.7481	0.7016	0.7035	0.7190	0.7119	0.7154	0.7810	**0.8013**
COIL-20-RG	ACC	0.5222	0.6368	0.4792	0.6396	0.6556	0.5486	0.4299	**0.6604**
	NMI	0.6725	0.7598	0.6142	0.7882	0.8148	0.7135	0.6216	**0.8257**
COIL-20-RL	ACC	0.6007	0.6778	0.5701	0.6757	0.6563	0.6181	0.5493	**0.6875**
	NMI	0.7396	0.7763	0.7011	0.7611	0.7835	0.7432	0.6960	**0.8120**
COIL-20-GL	ACC	0.5347	0.6333	0.6090	0.6646	0.6632	0.6035	0.4549	**0.6736**
	NMI	0.6730	0.7755	0.6789	0.7667	0.7696	0.7121	0.6334	**0.7814**

Table 3. Clustering performance of each method with two views under two metrics on different datasets.

Dataset	Metric	CCA	KCCA	RCCA	DCCA	DCCAE	TALSCCA	MCCA	$L_{2,1}$-CCA	MCCP
Yale	ACC	0.6485	0.6121	0.6485	0.6606	0.6000	0.6182	0.5939	0.6364	**0.6727**
	NMI	0.7129	0.6848	0.6659	0.7174	0.6883	0.7318	0.7069	0.7170	**0.7440**
COIL-20	ACC	0.6646	0.5396	0.5007	0.6681	0.6500	0.5708	0.5431	0.4875	**0.6826**
	NMI	0.7768	0.7019	0.5640	0.8082	0.7981	0.7530	0.6919	0.6148	**0.8135**

Table 4. Clustering performance of each method with more than two views under two metrics on Yale and COIL-20.

rics on different datasets. As we can see, the proposed CCP method consistently performs better than other methods on all datasets, whether the metric is ACC or NMI.

In addition, we also perform multiview clustering tests. The experimental settings are the same as those used in two-view clustering experiments. For methods with two-view inputs, we use again the same preprocessing strategy as that used in [30]. Table 4 records the clustering performance of each method with more than two views. As can be seen, our MCCP performs the best among all the methods, regardless of the metrics and datasets. In brief, these results demonstrate that our CCP and MCCP methods are also powerful for multiview clustering.

6. Conclusion

In this paper, we present a novel canonical $\mathcal{F}$-correlation framework by exploring and exploiting the nonlinear relationship between different features. With this framework, we propose a CCP method by using an explicit nonlinear mapping. Moreover, we extend CCP and propose an MCCP for learning compact representation of more than two views. MCCP is solved by an iterative method and the convergence of this iteration is proved in theory. A series of experimental results on six benchmark datasets demonstrate the effectiveness of our proposed CCP and MCCP methods in classification and clustering. A future interesting study is how to theoretically choose the optimal parameter of our methods.

Acknowledgments

This work is supported by the China Postdoctoral Science Foundation under grant 2020M670995, the Postgraduate Research and Practice Innovation Project Fund of Yangzhou University under grant KYCX21-3220, the National Natural Science Foundation of China under grants 62176126, 62076217, and 61906060, the Natural Science Foundation of Jiangsu Province of China under grant BK20190440, and the Yangzhou Science Project Foundation under grant YZ2020173.

References

[1] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, pages 1247–1255, 2013. 2, 7

[2] Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *JMLR*, 3:1–48, 2002. 2

[3] Jia Cai, Wei Dan, and Xiaowei Zhang. l_0-based sparse canonical correlation analysis with application to cross-language document retrieval. *Neurocomputing*, 329:32–45, 2019. 2

[4] J Douglas Carroll. Generalization of canonical correlation analysis to three or more sets of variables. In *Proceedings of the 76th Annual Convention of the American Psychological Association*, pages 227–228, 1968. 2

[5] Kamalika Chaudhuri, Sham M. Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *ICML*, pages 129–136, 2009. 1

[6] Jia Chen, Gang Wang, and Georgios B. Giannakis. Graph multiview canonical correlation analysis. *IEEE TSP*, 67(11):2826–2838, 2019. 2

[7] Jia Chen, Gang Wang, Yanning Shen, and Georgios B. Giannakis. Canonical correlation analysis of datasets with a common source graph. *IEEE TSP*, 66(16):4398–4408, 2018. 2

[8] Mansheng Chen, Ling Huang, Chang-Dong Wang, and Dong Huang. Multi-view clustering in latent embedding space. In *AAAI*, pages 3513–3520, 2020. 6

[9] Xiaohong Chen, Songcan Chen, Hui Xue, and Xudong Zhou. A unified dimensionality reduction framework for semi-paired and semi-supervised multi-view data. *Pattern Recognition*, 45(5):2005–2018, 2012. 1

[10] Delin Chu, Li-Zhi Liao, Michael K. Ng, and Xiaowei Zhang. Sparse canonical correlation analysis: New formulation and algorithm. *IEEE TPAMI*, 35(12):3050–3065, 2013. 2

[11] Lei Gao, Lin Qi, Enqing Chen, and Ling Guan. Discriminative multiple canonical correlation analysis for information fusion. *IEEE TIP*, 27(4):1951–1965, 2018. 1

[12] Lei Gao, Rui Zhang, Lin Qi, Enqing Chen, and Ling Guan. The labeled multiple canonical correlation analysis for information fusion. *IEEE TMM*, 21(2):375–387, 2019. 1

[13] Quanxue Gao, Huanhuan Lian, Qianqian Wang, and Gan Sun. Cross-modal subspace clustering via deep canonical correlation analysis. In *AAAI*, pages 3938–3945, 2020. 2, 6

[14] Xizhan Gao, Sijie Niu, and Quansen Sun. Two-directional two-dimensional kernel canonical correlation analysis. *IEEE SPL*, 26(11):1578–1582, 2019. 2

[15] David R. Hardoon and John Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83(3):331–353, 2011. 2

[16] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004. 1, 2, 3, 5

[17] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. 1

[18] William W. Hsieh. Nonlinear canonical correlation analysis by neural networks. *Neural Networks*, 13(10):1095–1105, 2000. 2

[19] Mahdi Karami and Dale Schuurmans. Deep probabilistic canonical correlation analysis. In *AAAI*, pages 8055–8063, 2021. 2

[20] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971. 2

[21] Tae-Kyun Kim, Josef Kittler, and Roberto Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE TPAMI*, 29(6):1005–1018, 2007. 1

[22] Pei Ling Lai and Colin Fyfe. Canonical correlation analysis using artificial neural networks. In *ESANN*, pages 363–368, 1998. 2

[23] Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE TKDE*, 31(10):1863–1883, 2019. 1

[24] David Lopez-Paz, Suvrit Sra, Alexander J. Smola, Zoubin Ghahramani, and Bernhard Schölkopf. Randomized nonlinear component analysis. In *ICML*, pages 1359–1367, 2014. 2, 7

[25] Yong Luo, Dacheng Tao, Kotagiri Ramamohanarao, Chao Xu, and Yonggang Wen. Tensor canonical correlation analysis for multi-view dimension reduction. *IEEE TKDE*, 27(11):3111–3124, 2015. 2

[26] Allan Aasbjerg Nielsen. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE TIP*, 11(3):293–305, 2002. 2, 7

[27] Xin Shu and Guoying Zhao. Scalable multi-label canonical correlation analysis for cross-modal retrieval. *Pattern Recognition*, 115:107905, 2021. 1

[28] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *ICML*, pages 824–831, 2005. 6

[29] Liang Sun, Shuiwang Ji, and Jieping Ye. Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis. *IEEE TPAMI*, 33(1):194–200, 2011. 1, 3

[30] Quan-Sen Sun, Sheng-Gen Zeng, Yan Liu, Pheng-Ann Heng, and De-Shen Xia. A new method of feature fusion and its application in image recognition. *Pattern Recognition*, 38(12):2437–2448, 2005. 1, 7, 8

[31] Tingkai Sun and Songcan Chen. Locality preserving CCA with applications to data visualization and pose estimation. *Image and Vision Computing*, 25(5):531–543, 2007. 2

[32] Tingkai Sun, Songcan Chen, Jing-Yu Yang, and Pengfei Shi. A novel method of combined feature extraction for recognition. In *ICDM*, pages 1043–1048, 2008. 1

[33] Viivi Uurtio, Sahely Bhadra, and Juho Rousu. Large-scale sparse kernel canonical correlation analysis. In *ICML*, pages 6383–6391, 2019. 2

[34] Jianwu Wan and Feng Zhu. Cost-sensitive canonical correlation analysis for semi-supervised multi-view learning. *IEEE SPL*, 27:1330–1334, 2020. 1

[35] Weiran Wang, Raman Arora, Karen Livescu, and Jeff A. Bilmes. On deep multi-view representation learning. In *ICML*, pages 1083–1092, 2015. 2, 7

[36] Meixiang Xu, Zhenfeng Zhu, Xingxing Zhang, Yao Zhao, and Xuelong Li. Canonical correlation analysis with $L_{2,1}$-norm for multiview data representation. *IEEE TCyb*, 50(11):4772–4782, 2020. 2, 7

[37] Zhiqiang Xu and Ping Li. Towards practical alternating least-squares for CCA. In *NeurIPS*, pages 14737–14746, 2019. 2, 7

[38] Xinghao Yang, Weifeng Liu, and Wei Liu. Tensor canonical correlation analysis networks for multi-view remote sensing scene recognition. *IEEE TKDE*, DOI: 10.1109/TKDE.2020.3016208. 2

[39] Xinghao Yang, Weifeng Liu, Wei Liu, and Dacheng Tao. A survey on canonical correlation analysis. *IEEE TKDE*, 33(6):2349–2368, 2021. 2

[40] Xinghao Yang, Weifeng Liu, Dapeng Tao, and Jun Cheng. Canonical correlation analysis networks for two-view image recognition. *Information Sciences*, 385:338–352, 2017. 2

[41] Lei-Hong Zhang, Li Wang, Zhaojun Bai, and Ren-Cang Li. A self-consistent-field iteration for orthogonal canonical correlation analysis. *IEEE TPAMI*, 44(2):890–904, 2022. 2

[42] Chengzhang Zhu, Longbing Cao, and Jianping Yin. Unsupervised heterogeneous coupling learning for categorical representation. *IEEE TPAMI*, 44(1):533–549, 2022. 6