Multimedia Tools and Applications (2023) 82:19539–19555
https://doi.org/10.1007/s11042-022-14237-5

A correlation analysis framework via joint sample and feature selection

Na Qiang[1] · Xiangjun Shen[2] · Ernest Domanaanmwi Ganaa[2] · Yang Yang[2] · Shengli Wu[2] · Zengmin Zhao[3] · Shucheng Huang[4]

Received: 25 March 2021 / Revised: 23 June 2022 / Accepted: 4 November 2022 / Published online: 24 December 2022

Abstract

Correlation Analysis is a popular technique for describing relationships between two datasets. In this paper, we proposed a correlation analysis framework via Joint Sample and Feature Selection (CAF-JSFS). Different from traditional correlation analysis where only feature selection is considered and each data point is treated equally, the significance of each data point is measured by a sample selection strategy in this framework. Considering that the principal projection is a feasible representation of data, the relationship between this principal projection and each data sample is recursively learnt through two sample selection strategies: cosine similarity and total distance metrics. In addition, CAF-JSFS solves the problem of feature redundancy caused by sample feature selection, and eliminates irrelevant features, thereby improving classification accuracy. This enhances the discriminative power of CAF-JSFS in noisy scenarios which makes better correlation projections achievable to improve performance. Extensive experiments on several datasets demonstrated the effectiveness of the proposed method compared to the state-of-the-art methods.

Keywords Correlation analysis · Sample selection · Feature selection · Cosine similarity metric · Total distance metric

1 Introduction

In the fields of machine learning and data mining, researchers are confronted with different forms of data, such as video, audio, image and text in very high dimensions which lead to the

✉ Xiangjun Shen
xjshen@ujs.edu.cn

[1] JingJiang College, JiangSu University, JiangSu, 212013 China

[2] School of Computer Science and Communication Engineering, JiangSu University, JiangSu, 212013 China

[3] ChangZhou Uzone Intelligent Technology Co.,Ltd, Jiangsu, China

[4] School of Computer Science, Jiangsu University of Science and Technology, JiangSu, 212003 China

problem of dimensionality curse [17]. Therefore, it is crucial to avoid this problem through dimensionality reduction (DR) [24] techniques, to improve the performance of subsequent processing such as classification and clustering. DR techniques can be broadly classified into feature selection [9] and subspace learning [33]. Feature selection techniques selects a subset of most representative or discriminative features from the input feature set, and while subspace learning methods transforms the original input features to a lower dimensional subspace.

Principal Component Analysis (PCA) [4], Linear Discriminant Analysis (LDA) [4], Locality Preserving Projection (LPP) [11] and Correlation Analysis (CA) [12] are perhaps the most popular DR methods. Despite the different motivations of these methods, they can all be interpreted by a unified graph embedding framework [33]. One major disadvantage of the above methods is that, the projections are a linear combination of all the original features. Thus, it is often difficult to interpret the results. Sparse subspace learning methods attempted to solve this problem. For example, Zou et al. proposed a sparse PCA algorithm based on L_2-norm and L_1-norm regularizations [36]. Mohammad et al. [20] proposed both exact and greedy algorithms for binary class sparse LDA as well as its spectral bound. Cai et al. proposed a unified sparse subspace learning (SSL) framework based on L_1-norm regularized Spectral Regression [5].

Among these DR algorithms, correlation analysis is a widely used technique for modeling the relationship between two datasets. There exist several variants of correlation analysis (CA) techniques. For instance, Magnus et al. proposed a unified approach to PCA, PLS, MLR, and Canonical Correlation Analysis(CCA) [2]. Discriminant CCA (DCCA) and local discriminant CCA (LDCCA) [29] were presented for fusing multi-feature information. Sun et al. [34] combined CCA with uncorrelated linear discriminant analysis and proposed a multi-view uncorrelated linear discriminant analysis (MULDA). It seeks discriminative correlations in the inter-view and intra-view data points simultaneously by dealing with new linear weighted combination methods for sparse ensembles.

Although all the above methods can attain a good performance on clean datasets, their performances degrade seriously when noisy data points are present in the datasets. This is because existing techniques concentrate on only useful feature selection and therefore, fail to effectively learn the correlation structure of datasets in the presence of noise. This leads to decreasing performances of models in classification and DR. Unfortunately, due to social media upsurge, these noisy or corrupt data points are prevalent these days. To address this problem, a Correlation Analysis Framework via Joint Sample and Feature Selection (CAF-JSFS) is proposed in this paper. In the propose model, in order to discriminate between noisy and relevant data points and suppress the impact of the former in pursuing projections, we introduced sample factors which impose penalties on each data point. To effectively suppress the effect of outliers, two sample selection strategies: cosine similarity and total distance metrics are used geometrically to iteratively learn the relationship between each sample and the principal projections in the feature space. In addition, feature selection is introduced into the proposed sample selection methods to obtain joint sample and feature selection methods to ensure that the proposed framework can classify data more accurately.

The main contributions of this paper are as follows:

1) We propose a novel framework by introducing sample factors into some traditional correlation analysis (CA) models to suppress the impact of outliers in order to obtain better correlation structures.
2) We further propose two sample selection strategies: cosine similarity and total distance metrics. These metrics iteratively evaluate the importance of each sample in pursuing

projections by learning the relationship between each sample and the principal projections in the feature space. This is to discriminate between authentic and corrupt data samples.

3) Finally, we introduced structured sparse L_2, 1-norm to eliminate feature redundancy in the process of sample selection and thus, propose a joint sample and feature selection framework (CAF-JSFS). CAF-JSFS can therefore learn a compact subspace resulting in better correlation structures in noisy datasets. Extensive experiments on many image datasets demonstrate the superiority of our method over state-of-the-art methods such as ALPCCA [31] and SPCA [18].

The rest of this paper is organized as follows. In Section 2, we present related work. Section 3 presents formulation of the proposed CAF-JSFS, experiments and result analyses is presented in Section 4, conclusion and future work are also presented in Section conclusion 5.

2 Related work

2.1 Dimension reduction through feature selection

DR has gained much attention in recent years due to the vital role it plays in machine learning. Many DR methods have therefore been proposed with the same focus on mapping high dimensional data to low dimensional spaces. In other words, given a problem of classification as in [33], with the training sample set $X = [x_1, x_2, \cdots, x_N]$,$x_i \in R^m$ with N samples and m dimensions for each sample. DR methods focus on finding a mapping function that transform the original data x_i to a low-dimensional representation $y_i \in R^d$ where $m >> d$.

Compared with subspace learning techniques which create new features, feature selection does not change the original representations of data variables. Consequently, many feature selection techniques have been proposed in the past few years. These feature selection methods are mainly put into two different categories: supervised and unsupervised. Since there is no label information in unsupervised feature selection methods, they are more difficult to implement than their supervised counterparts. Due to this, there are relatively fewer investigations dedicated to unsupervised techniques. Most unsupervised feature selection approaches are either based on filters [22], wrappers [26] or embeddings [8]. Although the performances of traditional unsupervised feature selection approaches are prominent in many cases, their efficiencies can still be improved since: (1) from the view of manifold learning [6], high dimensional data naturally lie on a low dimensional manifold. Traditional methods have not taken full considerations of data manifold structures. (2) Different from feature learning, traditional feature selection approaches only employ data statistical character to rank the features essentially. There is a lack of learning mechanism as in feature learning, which is proved to be powerful and widely used in many areas [23].

2.2 Correlation analysis

Correlation analysis is a well-known family of statistical tools for analyzing associations between variables or sets of variables. Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA), Partial Least Squares (PLS) and Multiple Linear Regression (MLR) are four efficient correlation analysis methods. Based on the least squares

Table 1 Four models of correlation analysis

Method	functions	constraint
PCA	$J(v) = \left\|X^T v\right\|_2^2$	$s.t.v^T v = I_r$
CCA	$J(v, w) = \left\|X^T v - Y^T w\right\|_2^2$	$s.t.v^T XX^T v = I_r, w^T YY^T w = I_r$
PLS	$J(v, w) = \left\|X^T v - Y^T w\right\|_2^2$	$s.t.v^T v = I_r, w^T w = I_r$
MLR	$J(v, w) = \left\|X^T v - Y^T w\right\|_2^2$	$s.t.v^T XX^T v = I_r, w^T w = I_r$

framework, we present four different correlation analysis objective functions of these four methods in Table 1.

Among these methods, PCA is the most popular correlation analysis technique. It can assist in understanding underlying data structures, clustering analysis, regression analysis, and many other tasks. Hu et al. [14] presented methodological, theoretical, and numerical studies on PCA in high-dimensional settings. In many practical studies, it is found that only a small subset of variables are relevant, while others are noise. To identify relevant variables and generate more interpretable results, a sparse PCA (SPCA) [18] technique that applies regularized estimation to generate sparse loadings has been developed. New PCA algorithms for graph embedding that incorporate data distribution and multiple penalty factors into the least squares framework regularized with multiple local graphs for multiview dimension reduction were proposed [3, 27]. Nie et al. [21] proposed to maximize the L21-norm based robust PCA objective, which is theoretically connected to the minimization of reconstruction error. More importantly, we propose the efficient non-greedy optimization algorithms to solve our objective and the more general L_2, 1-norm maximization problem with theoretically guaranteed convergence.

Proposed by H. Hotelling in 1936 [10], CCA finds basis vectors for two sets of variables such that the correlation between the projections of the variables onto these basis vectors are mutually maximized. In an attempt to increase the flexibility of feature selection, kernelization of CCA (KCCA) has been applied to map the hypotheses to a higher-dimensional feature space. KCCA has been applied in some preliminary work by Fyfe and Lai [15], Akaho [1] and recently by Vinokourov et al. [32] with improved results. Ping [25] proposed the label-wise orthogonal canonical correlation analysis (LOCCA), which constrains the label-based relationships and orthogonalizes correlation projection directions. In the method, the discriminative structures constrained by class labels are effectively preserved, and the correlation projection directions from LOCCA reduce the information redundancy by orthogonality criterion as much as possible. Chen [7] introduces four deep neural network (DNN) models that are suitable to combine with CCA, and the general form of DNN-CCA is given in detail. Then, the experimental comparison of these methods is conducted through three cases, so as to analyze the characteristics and distinctions of CCA aided by each DNN model. Finally, some suggestions on method selection are summarized, and the existed open issues in the current DNN-CCA form and future directions are discussed.

PLS is a multivariate technique that delivers an optimal basis in x-space for y on x regression. Reduction to a certain subset of the basis introduces a bias but reduces the variance. In general, PLS is based on a maximization of the covariance between $< v, x >$ and $< w, y >$, which are successive linear combinations in x and y spaces respectively. L. Hoegaerts et al. proposed kernel Partial Least Squares (KPLS), which fits naturally in a primal dual optimization class of kernel machines [30]. To model the function of non-linear relationships among videos in NDVR, KPLS maps the original video data into a Reproducing Kernel Hilbert Space (RKHS), and therefore it is able to efficiently handle high-dimensional videos in NFVs.

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables.

3 Proposed method

3.1 Sample selection induced correlation analysis

In this section, we present a framework of the proposed correlation analysis based on sample selection (CAF-JSS). From Table 1, the four correlation analysis(CA) algorithms can be summarized as the using the objective function as shown in (1), and PCA is a special case:

$$\arg\min \sum_{i,j} \left\| v^T x_i - w^T y_j \right\| \tag{1}$$

It can be seen that the CA algorithms use least square framework to minimize the sum distance between the original data set X and the reconstructed data set $v^T X$, the original data set Y and the reconstructed data set $w^T Y$.

This geometrical characteristic will force the projection vectors v and w to pass through the densest data points to minimize the sum distance, which is illustrated in Fig. 1, where v and w are the principal projection vectors. We consider the relationship between the projection vectors and the data samples. This geometrical relationship between data samples and projection vectors motivates us to evaluate the importance of each data sample in pursuing projections. Therefore, we reformulate formula (1) by introducing sample factors which impose penalties on the sample spaces to minimize the impact of corrupt data samples in (2) as follows:

$$\arg\min \sum_{i,j} \left\| v^T x_i d_{x_i} - w^T y_j d_{y_j} \right\|_2^2 \tag{2}$$

where d_{x_1} and d_{y_1} are sample factors that consider the contributions of data samples to projections.

Similarly, by introducing sample factors into the four traditional CA models presented in Table 1, we obtain our four proposed D-CA models as presented in Table 2. In our new model, a new data representation is presented as $\hat{X} = XD_X$ and $\hat{Y} = YD_Y$, therefore, $\hat{X}$ and $\hat{Y}$ are now obtained with the effect of corrupt data samples suppressed. D_X and D_Y are diagonal sample factor matrices where $D_x = diag\left(d_{x_1}, \cdots, d_{x_n}\right)$ and $D_y = diag\left(d_{y_1}, \cdots, d_{y_n}\right)$. By introducing the Lagrange multiplier (λ) into the DCA models and taking partial derivatives w.r.t. v and w, we obtain the eigenvector solutions presented in Table 3.

To demonstrate the effectiveness of our proposed models, we give a mathematical singular value decomposition (SVD) explanation to the CCA model. Mathematically, there is a direct relationship between CCA and SVD when CCA components are calculated from the covariance matrix [19]. The following demonstrate the singular value decomposition (SVD)

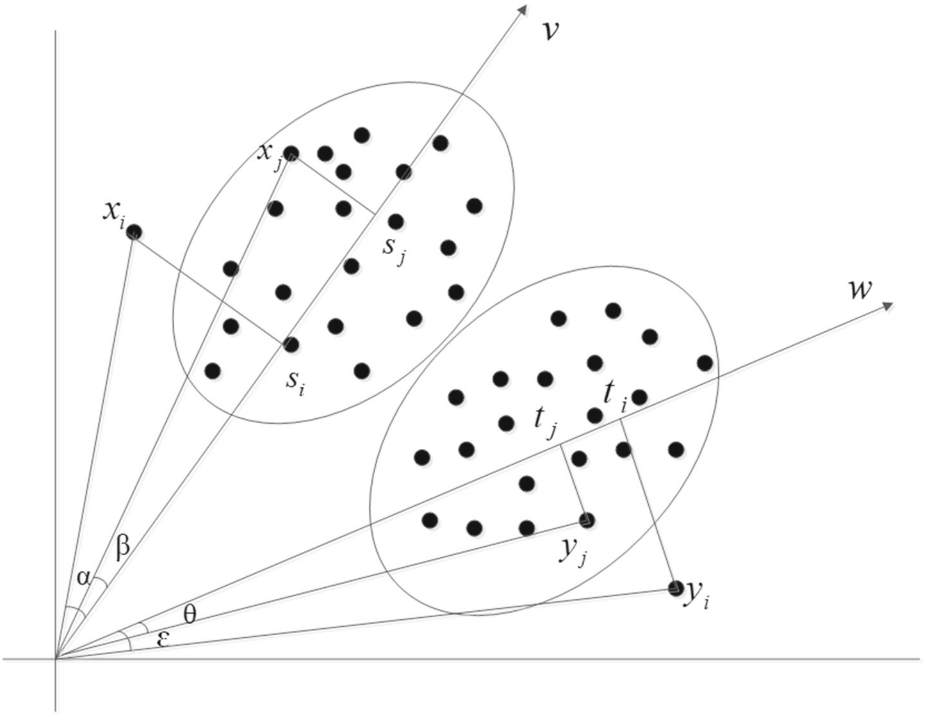


Fig. 1 Illustration of importance evaluation of data samples

of X and Y. However, we first introduce some notations, let $C_{xx} = X^T X$, $C_{yy} = Y^T Y$ and $C_{xy} = X^T Y$. For simplicity, assume C_{xx} and C_{yy} are full rank, and also let

$$\tilde{C}_{xy} = C_{xx}^{-\frac{1}{2}} C_{xy} C_{yy}^{-\frac{1}{2}} \tag{3}$$

Let $\tilde{C}_{xy} = V \Sigma W^T$ be the SVD of C_{xy} where v_i, w_j denote the left, right singular vectors and τ_i denotes the singular values. Then $X C_{xx}^{-\frac{1}{2}} v_i$, $Y C_{yy}^{-\frac{1}{2}} w_j$ are the canonical variables of the X, Y spaces respectively. In our proposed model, as $\hat{X} = X D_X$ and $\hat{Y} = Y D_Y$, therefore $X D_X C_{XX}^{-\frac{1}{2}} v_i$, $Y D_Y C_{yy}^{-\frac{1}{2}} w_j$ are the canonical variables respectively. In this way, with the imposition of the sample factors D_X and D_Y on the sample spaces, our proposed model can learn a better low dimensional subspace in corrupt data sets.

Table 2 The Four proposed canonical analysis methods

Method	functions	constraint
D-PCA	$argmin \sum_{i,j} \left\| d_{x_i} x_i^T v \right\|$	$s.t. v^T v = I_r$
D-CCA	$\arg\min \sum_{i,j} \left\| v^T x_i d_{x_i} - w^T y_j d_{y_j} \right\|_2^2$	$s.t. v^T X D_X^2 X^T v = I_r, w^T Y D_Y^2 Y^T w = I_r$
D-PLS	$\arg\min \sum_{i,j} \left\| v^T x_i d_{x_i} - w^T y_j d_{y_j} \right\|_2^2$	$s.t. v^T v = I_r, w^T w = I_r$
D-MLR	$\arg\min \sum_{i,j} \left\| v^T x_i d_{x_i} - w^T y_j d_{y_j} \right\|_2^2$	$s.t. v^T X D_X^2 X^T v = I_r, w^T w = I_r$

Table 3 Standard eigenvector solutions of the proposed D-CA models

Method	Derivation process
D-PCA	$XD_X^2X^Tv = \lambda v$
D-CCA	$(YD_Y^2Y^T)^{-1}(YD_YD_XX^T)(XD_X^2X^T)^{-1}(XD_XD_YY^T)w = \lambda w$ $(XD_X^2X^T)^{-1}(XD_XD_YY^T)(YD_Y^2Y^T)^{-1}(YD_YD_XX^T)v = \lambda v$
D-PLS	$(YD_Y^2Y^T + \lambda_2 I)^{-1}(YD_YD_XX^T)(XD_X^2X^T + \lambda_1 I)^{-1}(XD_XD_YY^T)w = \lambda w$ $(XD_X^2X^T + \lambda_1 I)^{-1}(XD_XD_YY^T)(YD_Y^2Y^T + \lambda_2 I)^{-1}(YD_YD_XX^T)v = \lambda v$
D-MLR	$(YD_Y^2Y^T + \lambda_1 I)^{-1}(YD_YD_XX^T)(XD_X^2X^T)^{-1}(XD_XD_YY^T)w = \lambda w$ $(XD_X^2X^T)^{-1}(XD_XD_YY^T)(YD_Y^2Y^T + \lambda_2 I)^{-1}(YD_YD_XX^T)v = \lambda v$

3.2 Obtaining sample factors d_x and d_y

In this subsection, we discuss how to model the relationship between data samples and the principal projections. Intuitively, the closer a sample to the projection vector v or w, the more important the sample is for calculating the projections. Based on this intuitive observation, we iteratively learn the relationship between data samples and the principal projections v and w using two strategies: total distance and cosine similarity metrics. This is to effectively distinguish between authentic and corrupt data samples based on how a data sample and the principal projection relate. Both can be obtained geometrically as shown in Fig. 1.

The first strategy uses total distance metric to iteratively learn the relationship between each sample and the principal projection. The total distance of an instance is the square sum of the distances between the coordinate of each instance and the coordinates of every other instance in the training set to the projections v or w. From Fig. 1, the coordinate (s_i) of data sample x_i to the projection v and the coordinate (t_i) of data sample y_i to the projection w are obtained based on (4) respectively as follows:

$$s_i = v^T x_i, t_i = w^T y_i \tag{4}$$

We then compute the total distance of data samples as follows:

$$d_{x_i} = \sum\nolimits_{i,j=1}^{n} (s_i - s_j)^2, d_{y_j} = \sum\nolimits_{i,j=1}^{n} (t_j - t_i)^2 \tag{5}$$

The total distance of a data sample is a natural way to evaluate its importance within the dataset in constructing projections. From Fig. 1, we can observe that the total distance of sample x_i or y_i which are outside the clusters will be relatively bigger than that of samples x_j and y_j within the clusters. Therefore, samples x_i and y_i are more likely to be outliers than samples x_j and y_j. Thus, the bigger d_{x_i} or d_{y_i}, the more likely x_i and y_i are noisy data samples and hence their relevance will be scaled accordingly to suppress their effects on the projections.

The second strategy uses the cosine similarity metric to build the sample factors d_{x_i} and d_{y_i}. This iteratively learns the angle relationship between each data sample in the training set and the principal projections v and w . In Fig. 1, the angle between data sample x_i and the projection v is α_i, and the angle between data sample y_i and the projection w is ε_i. The construction of sample penalty factor proposed can be obtained by normalizing (6):

$$d_{x_i} = cos\alpha_i = \frac{v^T x_i}{\|v\| \, \|x_i\|}, d_{y_j} = \cos\theta_j = \frac{w^T y_j}{\|w\| \, \|y_j\|} \tag{6}$$

In formula (6), a bigger $\cos\alpha_i$ implies a smaller angle α between sample x_i and the principal projection v and vice versa. Similarly, a bigger $\cos\varepsilon_i$ implies a smaller angle ε between sample y_i and the principal projection w. From Fig. 1, it can be seen that, the angle β between sample x_j and the principal projection v is smaller than the angle α of sample x_i and the principal projection v. Thus, x_i is considered less important in finding the best projections than x_j; likewise, we consider that y_i is less important than y_j.

Futhermore, d_{x_i} and d_{y_i} can now be obtained as follows :

$$d_{x_i} = cos\alpha_i + \eta, d_{y_j} = \cos\theta_j + \eta \tag{7}$$

where η is a adjust parameter to prevent d_{x_i} and d_{y_i} from approaching infinity.

3.3 CAF-JSFS

We further extend the proposed four models to feature selection in this subsection. After adding feature selection [13], the proposed D-CA model in (2) can now be written as:

$$\arg\min \sum\nolimits_{i,j} \left\| v^T x_i d_{x_i} - w^T y_j d_{y_j} \right\|_2^2 + \lambda_1 \left\| v \right\|_{2,1} + \lambda_2 \left\| w \right\|_{2,1} \tag{8}$$

Taking the D-CCA model as an example, the derivation of our proposed feature selection DQ-CCA model is as follows:

$$\begin{aligned} L &= 2w^T Y D_Y D_X X^T v - \lambda_1 v^T Q_x v - \lambda_2 w^T Q_y w \\ &\quad -\lambda_3 v^T X D_X^2 X^T v - \lambda_4 w^T Y D_Y^2 Y^T w \end{aligned} \tag{9}$$

where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are regularization parameters, and $Q_x \in R^{d\times d}$ is a diagonal matrix with (i, i) th element $Q_{ii}^x = (\gamma/2(\|v_i\|_2^2 + \varepsilon)^{1/2})(\varepsilon \to 0)$ and $v = \begin{bmatrix} v \\ \vdots \\ v_d \end{bmatrix} \in R^{d\times m}$, so is Q_y. And $\|v\|_{2,1}$, $\|w\|_{2,1}$ are based on ℓ_2-norm and ℓ_1 -norm regularizations [35].

By incorporating the feature selection into the D-CA models and taking partial derivatives w.r.t. v and w, according to Table 3 , we obtain the following eigenvector solutions of the proposed four DQ-CA models, which are presented in Table 4.

We further extend the proposed four models to feature selection in this subsection. After adding feature selection [13], the proposed D-CA model in (2) can now be written as:

$$\arg\min \sum\nolimits_{i,j} \left\| v^T x_i d_{x_i} - w^T y_j d_{y_j} \right\|_2^2 + \lambda_1 \left\| v \right\|_{2,1} + \lambda_2 \left\| w \right\|_{2,1} \tag{10}$$

Taking the D-CCA model as an example, the derivation of our proposed feature selection DQ-CCA model is as follows:

$$\begin{aligned} L &= 2w^T Y D_Y D_X X^T v - \lambda_1 v^T Q_x v - \lambda_2 w^T Q_y w \\ &\quad -\lambda_3 v^T X D_X^2 X^T v - \lambda_4 w^T Y D_Y^2 Y^T w \end{aligned} \tag{11}$$

where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are regularization parameters, and $Q_x \in R^{d\times d}$ is a diagonal matrix with (i, i) th element $Q_{ii}^x = (\gamma/2(\|v_i\|_2^2 + \varepsilon)^{1/2})(\varepsilon \to 0)$ and $v = \begin{bmatrix} v \\ \vdots \\ v_d \end{bmatrix} \in R^{d\times m}$, so is Q_y. And $\|v\|_{2,1}$, $\|w\|_{2,1}$ are based on ℓ_2-norm and ℓ_1 -norm regularizations [35].

By incorporating the feature selection into the D-CA models and taking partial derivatives w.r.t. v and w, according to Table 3 , we obtain the following eigenvector solutions of the proposed four DQ-CA models, which are presented in Table 4.

Table 4 Standard eigenvector solutions of the four proposed DQ-CA models

Method	Derivation process
DQ-PCA	$(XD^2X^T - \lambda_1 Q_x)v = \lambda_2 v$
DQ-CCA	$(XD_X^2X^T - \gamma Q_x)^{-1}(XD_XD_YY^T)(YD_Y^2Y^T - \sigma Q_y)^{-1}(YD_YD_XX^T)v = \lambda v$ $(YD_Y^2Y^T - \sigma Q_y)^{-1}(YD_YD_XX^T)(XD_X^2X^T - \gamma Q_x)^{-1}(XD_XD_YY^T)w = \lambda w$
DQ-PLS	$(\gamma Q_x + I)^{-1}(XD_XD_YY^T)(\sigma Q_y + I)^{-1}(YD_YD_XX^T)v = \lambda v$ $(\sigma Q_y + I)^{-1}(YD_YD_XX^T)(\gamma Q_x + I)^{-1}(XD_XD_YY^T)w = \lambda w$
DQ-MLR	$(\gamma Q_x + XD_X^2X^T)^{-1}(XD_XD_YY^T)(\sigma Q_y + I)^{-1}(YD_YD_XX^T)v = \lambda v$ $(\sigma Q_y + I)^{-1}(YD_YD_XX^T)(\gamma Q_x + XD_X^2X^T)^{-1}(XD_XD_YY^T)w = \lambda w$

The algorithm for the proposed CAF-JSFS is shown in Algorithm 1 as follows:

```
Input: Training set $X$,$Y$
Output: The projection vector $v$ , $w$
Initialize $D_x$ ,$D_y$ ,$Q_x$ ,$Q_y$ as identity matrices
While not converged do
Obtain $v$ and $w$ based on Table 3
Update $D_x$ and $D_y$ according to (5) and (10)
Update $Q_x$ and $Q_y$ based on Table 4
Compute loss based on (10)
End while
```

Algorithm 1 CAF-JSFS.

4 Experimental results

In this section, we first evaluate the performance of the proposed correlation analysis and feature selection methods against classical methods such as CCA, PCA, PLS and MLR. We further evaluate the performance of the proposed methods against state-of-the-art methods ALPCCA and SPCA.

4.1 Parameter settings and datasets description

For each dataset, we randomly sampled 60% and 40% for training and testing respectively in our experiments. We set the k-nearest-neighbors parameter K to 5 in the proposed D-CA and DQ-CA methods and all other comparative methods, in order to make a very fair comparison. Also, the parameters of ALPCCA, SPCA and LDA were set according to their literature. We finally make use of the K-nearest neighbor (KNN) classifier for classifications. The experiments are repeated 20 times and we record the average classification accuracies and standard deviations for the various methods. In our experiments, in order to fairly compare the performance among the KNN-graph models, graphs were constructed with the same neighbors N in CAF-JSFS correspondings to (10)

$$\min_{v} \sum_{i=1}^{n} \left(v^T x_i - \sum_{j=1}^{n} N_{ij} v^T x_j \right)^2 \tag{12}$$

As unsupervised constructions, v in CAF-JSFS corresponds to Table 4. In order to evaluate the performance of CAF-JSFS, we carry out a series of experiments on face and handwritten datasets.

4.2 Experiments on image datasets

Aside from the seven UCI datasets, we use several image datasets to test the proposed method's performance in this subsection. The image datasets include:

- **ORL**[1] face dataset contains 400 face image samples taken from 40 subjects, each with 10 face images. The face images per subject were taken by varying the lighting, facial expressions, and facial details at different times [28].
- **AR**[2] face database was created by Aleix Martinez and Robert Benavente in the Computer Vision Center (CVC) at the U.A.B. It contains over 4,000 color images corresponding to 126 people's faces (70 men and 56 women).
- **Extended YaleB**[3] dataset contains 165 face images from 15 subjects, each of which has 11 face images. The face images were taken by varying lighting conditions and facial expressions. In our experiment, each image is cropped and resized to 32 × 32, and the gray level values of each image are rescaled to (0, 1). That is, the dimensionality of each image sample is 1024 [28].
- **CMU-PIE**[4] face database contains more than 750,000 images of 337 people recorded in up to four sessions over the span of five months. Subjects were imaged under 15 view points and 19 illumination conditions while displaying a range of facial expressions. In addition, high resolution frontal images were acquired as well. In total, the database contains more than 305 GB of face data. The Content page describes the database in more detail.
- **USPS**[5] dataset contains a total of 9298 digit images of 0 through 9, each of which is of size 16 × 16 pixels, with 256 gray levels per pixel. In the experiment, each image is represented by a 256-dimensional vector [16].
- **MNIST**[6] dataset is constructed from the larger NISTs Special Database 3 and 1, which consist of binary images of handwritten digits. The images of each class (digit) are of size 28 × 28. Thus, each digit image is represented by a 784-dimensional vector [16].

The further detailed descriptions of the ORL, Extended YaleB, MINIST, USPS, COIL20, and CIFAR-10 datasets are presented in Table 5

4.3 Experimental analysis in no-noise scene

In this section we analyse and discuss the results obtained by each method on the different datasets used in our experiments.

[1] http://cam-orl.co.uk/facedatabase.html

[2] http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html

[3] http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html

[4] http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html

[5] https://www.kaggle.com/bistaumanga/usps-dataset

[6] http://yann.lecun.com/exdb/mnist/

Table 5 Summary of the six datasets

Dataset	Type	Samples	Features
ORL	Face	400	1024
AR	Face	793	1400
Extended YaleB	Face	1024	2414
CMU-PIE	Face	1360	4096
USPS	Handwritten	9298	256
MNIST	Handwritten	4000	784

4.3.1 Face recognition

In this section, we first demonstrate our proposed DQ-CCA, DQ-PCA, DQ-PLS and DQ-MLR have superior performances than the traditional CCA, PCA, PLS and MLR on face recognition. We further undertake experimental comparison with two state-of-the-art algorithms, ALPCCA and SPCA. We present results for each method on the ORL, AR, Extended YaleB and CMU-PIE datasets as shown in Table 6 with best results in bold in each case.

From Table 6, we can see that DQ-CCA, DQ-PCA, DQ-PLS and DQ-MLR all have superior performances than all the comparative methods on the ORL, AR, Extended YaleB and COMU-PIE face datasets. For the ORL face dataset, with an impressive recognition accuracy of 95.01%, DQ-CCA outperforms the traditional CCA by a significant margin of 3.01% and D-CCA by a small margin of 0.82%. The results show our proposed D-CCA and DQ-CCA both have improved results than the traditional CCA because they are able to significantly suppress the effect of corrupt data samples better than the traditional CCA. Also, the recognition accuracy of DQ-PCA is 87.70% in excess of 1.93% to the traditional PCA and 0.08% to D-PCA. Again, the proposed DQ-PCA and D-PCA show significant improvement in face recognition as compared to the traditional PCA due to their abilities to distinguish between authentic and corrupt data samples. DQ-PLS also performs 2.56%

Table 6 Mean Classification Accuracies ± standard deviations (%) of the various methods on the ORL, AR and extended YaleB data sets

CA methods	ORL	AR	Extended YaleB	CMU-PIE
CCA	92.00(± 0.0030)	86.66(± 0.0050)	71.42(± 0.0056)	89.00 (± 0.0041)
D-CCA	94.19(± 0.0025)	91.80(±0.0020)	72.05(± 0.0024)	91.87 (± 0.0022)
DQ-CCA	**95.01(± 0.0020)**	**92.50 (± 0.0012)**	**73.01(± 0.0010)**	**92.75(± 0.0017)**
PCA	85.77(± 0.0040)	81.74(± 0.0044)	79.04(± 0.0050)	93.35 (± 0.0044)
D-PCA	87.62(± 0.0020)	83.19(± 0.0023)	81.02(± 0.0023)	95.08(± 0.0032)
DQ-PCA	**87.70(± 0.0020)**	**84.00 (± 0.0005)**	**82.05(± 0.0010)**	**96.04(± 0.0020)**
PLS	86.74(± 0.0036)	74.84(± 0.0050)	79.05(± 0.0040)	87.60 (± 0.0050)
D-PLS	88.33(± 0.0020)	76.32(± 0.0030)	80.19(± 0.0030)	89.03 (± 0.0037)
DQ-PLS	**89.30(± 0.0020)**	**77.32 (± 0.0010)**	**81.90(± 0.0011)**	**90.05(± 0.0012)**
MLR	85.77(± 0.0040)	76.73(± 0.0064)	85.77(± 0.0064)	84.39 (± 0.0055)
D-MLR	91.65(± 0.0017)	78.40(± 0.0020)	87.75(± 0.0033)	87.06(± 0.0033)
DQ-MLR	**92.00(± 0.0017)**	**79.15 (± 0.0013)**	**88.97(± 0.0013)**	**87.98 (± 0.0020)**
ALPCCA	86.51(± 0.0090)	83.07(± 0.0070)	79.05(± 0.0070)	69.05(± 0.0040)
SPCA	87.09(± 0.0045)	82.05(± 0.0060)	80.07(± 0.0045)	81.12(± 0.0065)

more than the traditional PLS and 0.97% more than D-PLS. DQ-MLR also has a superior performance of 6.23% over the traditional MLR and 0.35% over D-MLR.

Also, for the AR dataset, DQ-CCA outperforms D-CCA by 0.70% and the traditional CCA by a significant margin of 8.54%. DQ-PLS also has a superior performance of 1.00% over D-PLS and 2.48% more than the traditional PLS. DQ-PCA also proves to be 1.03% better than D-PCA and 3.01% better than the traditional PCA for the Extended YaleB dataset. Still on the Extended YaleB data set, DQ-MLR also outperforms D-MLR by 1.22% and the traditional MLR by 3.20%. For CMU-PIE dataset, the accuracies of ALPCCA and SPCA are not as good as all the CA algorithms proposed in this paper.

4.3.2 Handwriting recognition

To further evaluate the effectiveness of the proposed methods on handwritten digits recognition, we run experiments on the USPS and MNIST datasets. The results for the various methods for these datasets are presented in Table 7.

It is apparent from Table 7 that, the proposed methods once again demonstrate their superiority over the traditional techniques in handwritten digit recognition. For the USPS dataset, with a digit recognition accuracy of 72.56%, DQ-CCA outperforms D-CCA by 1.54% and the traditional CCA by 3.56%. DQ-PCA also has superior digit recognition of 1.69% over D-PCA and 3.58% over the standard PCA. DQ-PLS also outperforms D-PLS and the traditional PLS by 0.96% and 3.00%, respectively. The proposed methods have superior performances than the traditional methods because they are able to discover the intrinsic data structure and also suppress the impact of corrupt data samples. On the MNIST dataset, our method is superior to the traditional CA methods, ALPCCA and SPCA.

4.4 Experimental analysis in noisy scenarios

In this section, we add different degrees of salt and pepper noise to the ORL and USPS datasets to verify whether the performances of the proposed algorithms are better than the traditional CA methods in noisy scenarios so as to prove their superior noise suppression

Table 7 Mean Classification Accuracies ± standard deviations (%) of the various methods on USPS and MNIST data sets

methods	USPS	MNIST
CCA	69.00 (± 0.0052)	81.98 (± 0.0041)
D-CCA	71.02 (± 0.0023)	84.70 (± 0.0035)
DQ-CCA	**72.56(± 0.0012)**	**85.71 (± 0.0018)**
PCA	72.15 (± 0.0034)	83.03 (± 0.0050)
D-PCA	74.04(± 0.0023)	85.98 (± 0.0041)
DQ-PCA	**75.73(± 0.0015)**	**87.08 (± 0.0017)**
PLS	67.01 (± 0.0070)	86.25 (± 0.0045)
D-PLS	69.05 (± 0.0030)	88.03 (± 0.0037)
DQ-PLS	**70.01(± 0.0010)**	**88.97 (± 0.0011)**
MLR	68.05 (± 0.0053)	85.58 (± 0.0043)
D-MLR	70.08 (± 0.0027)	87.88 (± 0.0024)
DQ-MLR	**71.35(± 0.0013)**	**87.93 (± 0.0012)**
ALPCCA	71.10(± 0.0017)	72.25(± 0.0005)
SPCA	70.55(± 0.0028)	73.01(± 0.0016)

Table 8 Mean Classification Accuracies (%) of the various methods on noisy ORL dataset

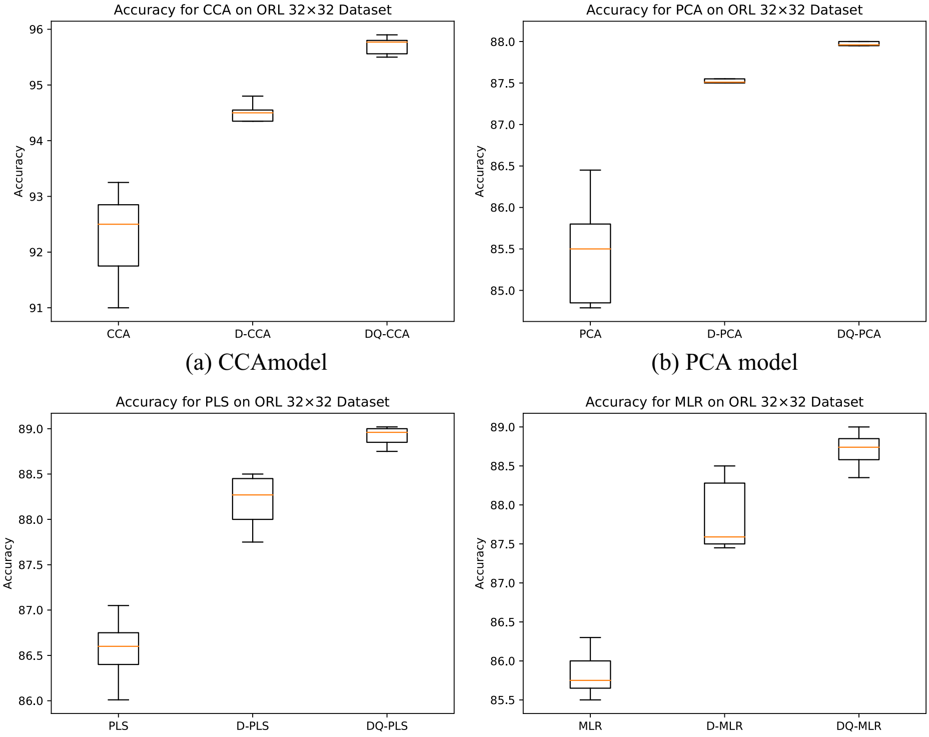
Level	PCA	D-PCA	DQ-PCA	CCA	D-CCA	DQ-CCA	PLS	D-PLS	DQ-PLS	MLR	D-MLR	DQ-MLR
0%	91.98	94.70	**95.50**	93.03	95.98	**96.04**	87.25	88.03	**88.06**	85.59	87.03	**88.25**
5%	78.50	80.06	**80.16**	80.75	82.20	**83.25**	75.59	77.01	**78.96**	80.94	80.75	**81.75**
15%	69.50	72.75	**73.50**	75.67	78.75	**79.75**	72.00	75.40	**76.00**	75.20	76.12	**77.03**

abilities. The noise added to the experimental datasets in this section is divided into three levels: 0%, 5%, and 15%. Best results are bolded in each case.

4.4.1 Face image denoising analysis

The average classification accuracies of the various methods on a noisy ORL dataset are recorded in Table 8.

It can be seen from Table 8 that, the proposed correlation methods have higher classification accuracies than the traditional methods in the absence of noise. When the noise level is 5%, the classification accuracy of DQ-PCA is 1 to 2% higher than DQ-CCA, DQ-PLS, and DQ-MLR. With an increase in the noise level, the proposed CAF-JSFS framework shows good robustness in classification performance. Thus, when the noise level is 15%, DQ-PCA has the best classification accuracy among all the correlation analysis methods. In addition,



(a) CCAmodel

(b) PCA model

(c) PLSmodel

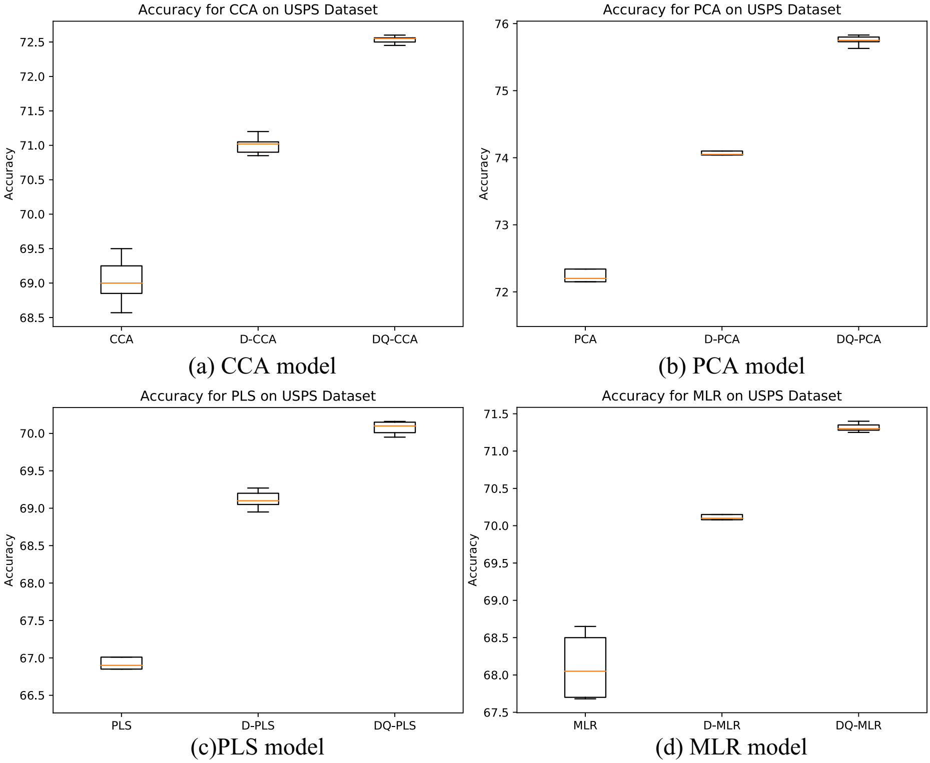
(d) MLR model

Fig. 2 Box diagram of four correlation analysis models in noisy ORL dataset (80% training)

Table 9 Mean Classification Accuracies (%) of the various methods on noisy USPS dataset

Level	PCA	D-PCA	DQ-PCA	CCA	D-CCA	DQ-CCA	PLS	D-PLS	DQ-PLS	MLR	D-MLR	DQ-MLR
0%	69.00	71.02	**72.56**	72.15	74.04	**75.73**	67.01	69.05	**70.01**	68.05	70.08	**71.35**
5%	65.86	66.26	**67.47**	68.69	70.25	**71.20**	61.83	62.25	**63.88**	63.49	64.75	**65.03**
15%	60.75	61.75	**62.17**	60.05	61.43	**62.65**	57.71	57.99	**59.03**	60.01	60.89	**61.58**

it can be seen that, with increasing noise levels, CCA model has the worst performance. Unlike the other methods, the accuracies of the proposed methods decrease at a slower rate with increasing noise levels. Generally, the proposed methods show significant improvements in face recognition as compared to the traditional techniques due to the abilities of the proposed methods to distinguish between authentic and corrupt data samples. The proposed methods also prove their consistency in performances due to the lower variances they obtain in all cases as compared to the traditional techniques. Figure 2 shows the experimental box diagrams.



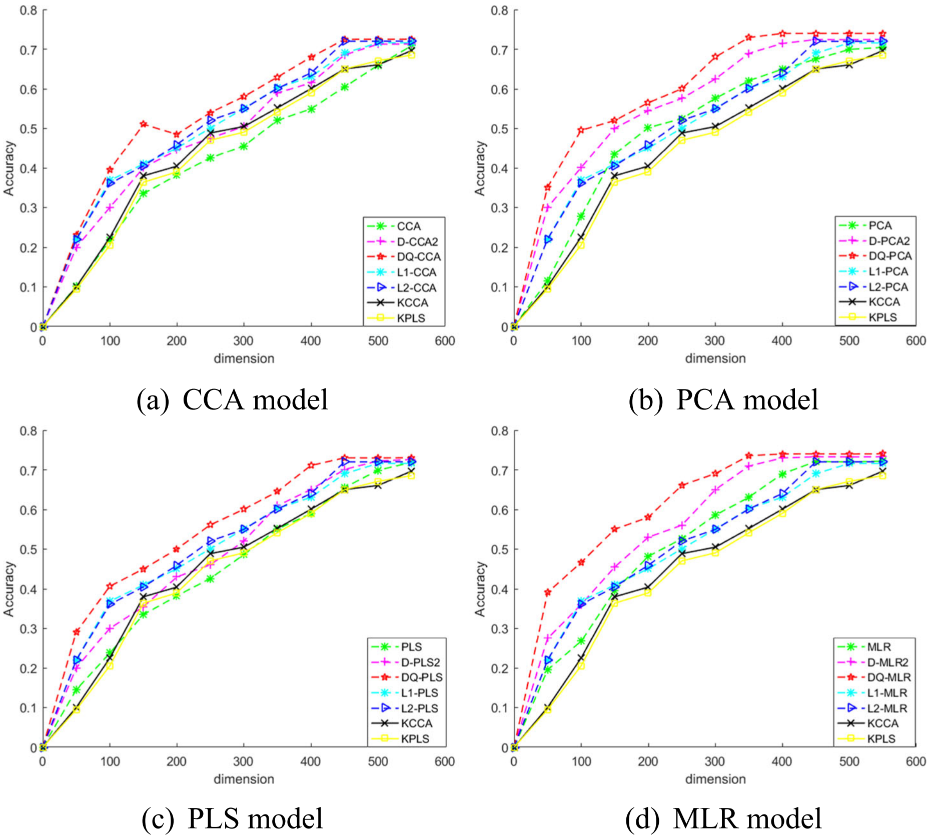
(a) CCA model

(b) PCA model

(c)PLS model

(d) MLR model

Fig. 3 Box diagram of four correlation analysis models in noisy USPS dataset (80% training)



(a) CCA model

(b) PCA model

(c) PLS model

(d) MLR model

Fig. 4 Classification results of the four CA models under the COIL20 dataset (70% training)

4.4.2 Experiments on noisy USPS dataset

Table 9 shows the classification results on a noisy USPS dataset. Similar to the results in Table 8, the proposed CAF-JSFS framework has the best facial recognition accuracy among all methods.

Also, from Fig. 3, it is evident that, the proposed methods have been stable in performance since they obtain lower variances in digit recognition.

4.5 Dimensionality analysis

In this subsection, we test our proposed method on object recognition using Columbia Object Image (COIL-20) dataset. We use 70% training and 30% testing samples in this section. To intuitively see whether the proposed method is effective in expressing data features in low dimensional spaces, we draw graphs showing classification accuracies in varying dimensions in Fig. 4. It can be seen that, the proposed CAF-JSFS framework has better classification accuracy and stability of low dimensional subspace representation data characteristics. From Fig. 4, it can also be observed that, the DQ-CA models have leading performances and their classification performances seem to stabilize in lower-dimension.

5 Conclusion

In this paper, a correlation analysis framework via joint sample and feature selection (CAF-JSFS) is proposed. Different from other variants of correlation analysis, sample factors that impose penalties on the sample spaces are introduced to suppress the impact of noise in pursuing projections. Two strategies, cosine similarity and total distance metrics are used geometrically to iteratively learn the relationships between each sample and the principal projections. This enables our framework to discriminate between authentic and corrupt data samples in order to suppress the impact of the latter. We further combined our sample selection idea with feature selection to obtain a joint sample and feature selection methods. With these ideas combined in our models, better correlation projections are achievable by sample and feature selection jointly. Our CAF-JSFS can learn better correlation projections in a noisy scenario, with the effect of noisy data points being suppressed. Extensive experiments on ORL, AR, extended YaleB and USPS datasets demonstrate CAF-JSFS achieves superior classification performance over state-of-the-art correlation analysis methods. This is because the proposed methods are able to effectively distinguish between authentic and corrupt data samples, thereby minimizing the impact of the latter. In the future, we will extend our proposed framework to low-rank representation and graph embedding.

Declarations

Conflict of Interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Akaho S. (2006) A kernel method for canonical correlation analysis. https://doi.org/10.48550/arXiv.cs/0609071
2. Arenasgarcia J, Petersen K, Hansen LK (2013) Kernel multivariate analysis framework for supervised subspace learning: a tutorial on linear and kernel multivariate methods. IEEE Signal Proc Mag 30(4):16–29
3. Apasiba Abeo T, Shen X-J, Bao B-K, Zha Z-J, Fan J A generalized multi-dictionary least squares framework regularized with multi-graph embeddings, vol 90
4. Belhumeur PN, Kriegman DJ (1996) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. In: European conference on computer vision
5. Cai D, He X, Han J (2007) Spectral regression: A unified approach for sparse subspace learning. In: Seventh IEEE international conference on data mining (ICDM 2007), IEEE, 2007, pp. 73–82
6. Cai D, Zhang C, He X (2010) Unsupervised feature selection for multi-cluster data. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 333–342
7. Chen Z, Liang K, Ding SX, Yang C, Peng T, Yuan X (2022) A comparative study of deep neural network-aided canonical correlation analysis-based process monitoring and fault detection methods. IEEE Trans Neural Netw Learn Syst 33(11):6158–6172
8. Dy JG, Brodley CE (2004) Feature selection for unsupervised learning. J Mach Learn Res 5(4):845–889
9. Guyon I, Elisseeff A, Kaelbling LP (2003) An introduction to variable and feature selection. J Mach Learn Res 3(6):1157–1182
10. Hardoon DR, Szedmak S, Shawetaylor J (2004) Canonical correlation analysis: an overview with application to learning methods. Neural Comput 16(12):2639–2664
11. He X, Cai D, Min W (2005) Statistical and computational analysis of locality preserving projection. In: Proceedings of the 22nd international conference on machine learning, 2005, pp. 281–288
12. He X, Cai D, Yan S, Zhang H-J (2005) Neighborhood preserving embedding. In: Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1, Vol. 2, IEEE, 2005, pp. 1208–1213

13. Hou C, Nie F, Yi D, Wu Y (2011) Feature selection via joint embedding learning and sparse regression. In: Twenty-Second international joint conference on Artificial Intelligence
14. Hu Z, Gang P, Wang Y, Wu Z (2016) Sparse principal component analysis via rotation and truncation. IEEE Trans Neural Netw Learn Syst 27(4):875
15. Lai PL, Fyfe C (2000) Kernel and nonlinear canonical correlation analysis. Int J Neural Syst 10(05):365–377
16. Liu J, Chen Y, Zhang J, Xu Z (2014) Enhancing low-rank subspace clustering by manifold regularization. IEEE Trans Image Process 23(9):4022–4030
17. Liu W, Tao D (2013) Multiview hessian regularization for image annotation. IEEE Trans Image Process 22(7):2676–2687
18. Liu Y, Zeng J, Xie L, Luo S, Su H (2018) Structured joint sparse principal component analysis for fault detection and isolation. IEEE Trans Ind Inf PP(99):1–1
19. Lu Y, Foster DP (2014) Large scale canonical correlation analysis with iterative least squares, Advances in Neural Information Processing Systems 27
20. Moghaddam B, Weiss Y, Avidan S (2006) Generalized spectral bounds for sparse lda. In: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 641–648
21. Nie F, Tian L, Huang H, Ding C (2016) Non-greedy l21-norm maximization for principal component analysis, IEEE Transactions on Image Processing 30 (2021) 5277–5286.
22. Nie F, Xiang S, Jia Y, Zhang C, Yan S (2008) Trace ratio criterion for feature selection. In: AAAI, Vol. 2, 2008, pp. 671–676
23. Nie F, Xu D, Tsang IW, Zhang C (2010) Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction. IEEE Trans Image Process 19(7):1921–1932
24. Passalis N, Tefas A (2018) Dimensionality reduction using similarity-induced embeddings. IEEE Trans Neural Netw Learn Syst 29(8):3429–3441
25. Ping X, Zhao Q, Su S (2021) Label-wise Orthogonal Canonical Correlation Analysis and Its Application to Image Recognition. ICAIIS 2021: 2021 2nd International Conference on Artificial Intelligence and Information Systems, May 2021. Article No. 87. Pages 1–5. https://doi.org/10.1145/3469213.3470288
26. Roth V, Lange T (2003) Feature selection in clustering problems, Advances in neural information processing systems 16
27. Sakar CO, Kursun O, Gurgen F (2012) A feature selection method based on kernel canonical correlation analysis and the minimum redundancy–maximum relevance filter method. Expert Syst Appl 39(3):3432–3437
28. Shen X-J, Ni C, Wang L, Zha Z-J (2021) Sliker: sparse loss induced kernel ensemble regression. Patt Recogn 109:107587
29. Sun S, Xie X (2016) Semisupervised support vector machines with tangent space intrinsic manifold regularization. IEEE Trans Neural Netw Learn Syst 27(9):1827–1839
30. Tao JL, Zhang JM, Wang LJ, Shen XJ, Zha ZJ (2019) Near-Duplicate Video Retrieval Through Toeplitz Kernel Partial Least Squares. In: Kompatsiaris, I., Huet, B., Mezaris, V., Gurrin, C., Cheng, WH., Vrochidis, S. (eds) MultiMedia Modeling. MMM 2019. Lecture Notes in Computer Science, vol 11296. pp 352–364
31. Wang F, Zhang D (2013) A new locality-preserving canonical correlation analysis algorithm for multi-view dimensionality reduction. Neural Process Lett 37(2):135–146
32. Wang S, Zhuang F, Jiang S, Huang Q, Qi T (2015) Cluster-sensitive structured correlation analysis for web cross-modal retrieval. Neurocomputing 168:747–760
33. Yan S, Xu D, Zhang B, Zhang HJ, Yang Q, Lin S (2007) Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans Patt Analy Mach Intell 29(1):40
34. Yuan YH, Sun QS, Ge HW (2014) Fractional-order embedding canonical correlation analysis and its applications to multi-view dimensionality reduction and recognition. Pattern Recogn 47(3):1411–1424
35. Zhang R, Nie F, Li X (2017) Self-weighted supervised discriminative feature selection. IEEE Trans Neural Netw Learn Syst 29(8):3913–3918
36. Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. J Comput Graph 15(2):265–286

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.