Feature Selective Projection with Low-Rank Embedding and Dual Laplacian Regularization

Chang Tang, *Member, IEEE*, Xinwang Liu, *Member, IEEE*, Xinzhong Zhu, Jian Xiong, *Member, IEEE*, Miaomiao Li, Jingyuan Xia, Xiangke Wang, *Senior Member, IEEE*, and Lizhe Wang, *Senior Member, IEEE*

Abstract—Feature extraction and feature selection have been regarded as two independent dimensionality reduction methods in most of the existing literature. In this paper, we propose to integrate both approaches into a unified framework and design an unsupervised linear feature selective projection (FSP) for feature extraction with low-rank embedding and dual Laplacian regularization, with the aim to exploit the intrinsic relationship among data and suppress the impact of noise. Specifically, a projection matrix with an $l_{2,1}$-norm regularization is introduced to project original high dimensional data points into a new subspace with lower dimension, where the $l_{2,1}$-norm regularization can endow the projection with good interpretability. We deploy a coefficient matrix with low rank constraint to reconstruct the data points and the $l_{2,1}$-norm is imposed to regularize the data reconstruction errors in the low-dimensional subspace and make FSP robust to noise. Furthermore, a dual graph Laplacian regularization term is imposed on the low dimensional data and data reconstruction matrix for preserving the local manifold geometrical structure of data. Finally, an alternatively iterative algorithm is carefully designed for solving the proposed optimization model. Theoretical convergence and computational complexity analysis of the algorithm are also provided. Comprehensive experiments on various benchmark datasets have been carried out to evaluate the performance of the proposed FSP. As indicated, our algorithm significantly outperforms other state-of-the-art methods for feature extraction.

Index Terms—Dimensionality reduction, feature extraction, feature selection, subspace learning, low rank representation, graph Laplacian regularization

1 Introduction

In many practical applications of data mining and machine learning, the original data is often represented as high dimensional features [1], [2], [3], [4], [4], [5], [6], [7], [8], [9], [10]. Since redundant and noisy features are inevitably mixed in high-dimensional data, directly dealing with them not only takes intensive memory and computational cost, but also degenerates the performance of learning tasks such as clustering and classification [11], [12], [13], [14], [15], [16], [17], [18]. This is well known as "curse of dimensionality" [19], [20]. As a consequence, dimensionality reduction has been intensively studied and achieved fruitful research results in the past few decades [2], [3], [4], [8], [21], [22], [23], [24], [25], [26]. Its aim is to represent the original high dimensional data in a lower and more discriminative dimensional space, in which the intrinsic structure of original data can be better revealed.

There are mainly two kinds of methods for dimensionality reduction: feature selection based methods and feature extraction based methods [27], [28], [29]. Feature selection techniques do not alter the original features, but aim at obtaining a subset of them. Thus the original semantics of features can be well preserved. Large amounts of feature selection methods including unsupervised [30], [31], [32], [33], [34], [35], [36], [37], [38], semi-supervised [39], [40], and supervised [41], [42] ones have been proposed. Among these methods, unsupervised ones, which aim to select a feature subset from original features without using the labels of data samples, have received much attention recently due to their practicality since it is laborious and expensive to obtain the data labels in many practical applications.

Different from feature selection based methods, feature extraction based ones aim to learn a projection to project original high-dimensional feature space into a new subspace with lower dimension. Thus, the features in the new space are different to original ones. Among this kind of approaches, the most representative ones including principal component analysis (PCA) [43], which maximizes the variance of samples during the projection process, linear discriminant analysis (LDA) [44], which maximizes the inter-class distances and minimizes the intra-class distances of data points

- *C. Tang and L. Wang are with the School of Computer Science, China University of Geosciences, Wuhan 430074, China. E-mail: tangchang@cug.edu.cn, lizhe.wang@gmail.com.*
- *X. Liu and M. Li are with the School of Computer, National University of Defense Technology, Changsha 410073, China. E-mail: xinwangliu@nudt.edu.cn, miaomiaolinudt@gmail.com.*
- *X. Zhu is with the College of Mathematics, Physics, and Information Engineering, Zhejiang Normal University, Jinhua, Zhengjiang 321004, China. E-mail: zxz@zjnu.edu.cn.*
- *J. Xiong is with the School of Business Administration, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China. E-mail: xiongjian2017@swufe.edu.cn.*
- *J. Xia is with the Department of Electric and Electronic Engineering, Imperial College London, London SW7 2AZ, United Kingdom. E-mail: j.xia16@imperial.ac.uk.*
- *X. Wang is with the College of Mechatronics and Automation, National University of Defense Technology, Changsha 410073, P.R. China. E-mail: xkwang@nudt.edu.cn.*

Manuscript received 31 July 2018; revised 16 Feb. 2019; accepted 9 Apr. 2019. Date of publication 17 Apr. 2019; date of current version 5 Aug. 2020. (Corresponding author: Xinwang Liu.) Recommended for acceptance by L. B. Holder. Digital Object Identifier no. 10.1109/TKDE.2019.2911946

during the projection process and their variants such as scatter difference discriminant criterion [45], kernelized LDA [46], model-based discriminant analysis [47] bilateral PCA [48], regularized LDA [49], to name just a few.

In recent literatures, researchers demonstrate that high dimensional data in practical applications substantially lies in or approximates a smooth nonlinear manifold with lower dimension. To this end, manifold learning based dimensionality reduction approaches have been proposed to uncover the intrinsic manifold geometric structure of data during the projection process. Among these approaches, representative ones include isometric feature mapping (ISOMAP) [50], Laplacian eigenmaps (LE) [51], [52] and Locally Linear Embedding (LLE) [53]. In the previous work [54], Yan et al. claimed that LLE, ISOMAP and LE can be unified into a general graph embedding based dimensionality reduction framework. Though these manifold learning methods have earned great success in revealing the inherent nonlinear structure hidden in data, they cannot handle the "out-of-sample" problem. In order to address this problem, some other representative approaches such as locality preserving projections (LPP) [55], neighborhood preserving projections (NPP) [56], neighborhood preserving embedding (NPE) [57], sparsity preserving projections (SPP) [58] and isometric projection (IsoP) [59] have been developed. As a linear version of LE, LPP learns a projection which preserves the local manifold geometry of the original data by an affinity graph regularizer. Then the learned projection can be directly used to project new data points into new lower-dimensional space. Based on the basic model of LPP, some extensions have also been proposed to improve the performance, e.g., semi-supervised LPP [60], orthogonal LPP (OLPP) [26], [61], [62], [63], [64]. Similar to LPP, NPE aims to preserve the local neighborhood structure of data, while the affinity graph [65] in NPE is constructed by using a local squares residual regularization term. NPP learns the global structure by utilizing the local neighborhood relations. Different to LPP and NPE, SPP [58] preserves the sparse representation relationship of the data points by using a ℓ_1-norm induced sparsity regularization term. Although the manifold learning based methods demonstrate promising performance, the intrinsic local manifold geometrical structure of data can be easily effected by various kinds of noises such as illumination change, corruptions and occlusion.

Recently, the low-rank representation (LRR) [66], [67], which can construct robust graph for many data processing tasks, has gained much attention due to its robustness to noises and corrupted data. LRR is under the assumption that high dimensional data points often intrinsically lie on a low dimensional subspace, thus the rank constraint is imposed on the coefficient matrix for data representation. During last few years, a various of LRR methods have been introduced for learning tasks such as robust PCA (RPCA) [68], non-negative low rank and sparse graph (NNLRS) [69]. By considering the manifold structure of data, the Laplacian regularized LRR was also built [70], [71]. Considering that traditional LRR cannot handle the data that lies in joint subspaces, Tang et al. proposed a structure LRR model by combining dense block regularization and sparse representation [72], [73]. You et al. [74] presented a scalable sparse subspace clustering by orthogonal matching pursuit, their method can perform well no matter the subspaces are independent or disjoint. Peng et al. [75], [76] proposed a method for clustering both dependent and disjoint subspaces by eliminating the effect of the errors from the projection space.

Although the LRR based methods work well by learning a representation low rank matrix for subspace clustering, they also suffer the "out-of-sample" problem. To solve this problem, Bao et al. [77] proposed a robust projection learning model named inductive RPCA (IRPCA). In recent years, a variety of projection learning methods based on LRR have been proposed to make the learned projection be robust to noises mixed in data points such as occlusions and corruptions. Lu et al. [78] proposed to learn a feature projection with low-rank being well preserved (LRPP) for image classification task. In LRPP, a projection matrix is learned to project original data into a low dimensional subspace, during the projection process, the low rank property of the data representation matrix is preserved. In [25], Wong et al. integrated the projection learning into traditional low rank representation and proposed a low-rank embedded projection (LRE). In such a manner, the learned projection is robust to noises such as data corruptions and occlusions. In order to reduce the complexity and sensitivity to dimensions, We et al. [79] proposed a low-rank preserving projection learning based on graph regularized reconstruction (LRPP_GRR), they combined a data reconstruction matrix and feature projection matrix together to perform the learning process.

As discussed above, preservation of the intrinsic manifold geometrical structure of data and robustness to noises are two critical issues for projection based feature extraction. In this work, we introduce an unsupervised linear feature selective projection (referred to as FSP briefly) for feature extraction with low-rank embedding and dual Laplacian regularization. Different from previous projection based feature extraction methods which lack interpretability during the projection process, i.e., they only extract the low dimensional features from original data but neither interpret how the projection works nor reflect the importance of different features, our method possesses the ability for simultaneously feature selection and feature extraction. On one hand, we impose the row sparsity on the projection matrix to enable the model to jointly select the key features from all of original features for composing the low dimensional subspace, i.e., the learned projection is more interpretable. On the other hand, we use the low-rank representation to enable the model be robust to noises. In addition, a dual graph Laplacian regularization term is integrated into our model for preserving the local manifold geometrical structure of original data. In brief, we summarize the major contributions of this work as follows:

1) Proposing an unsupervised linear feature selective projection (FSP) for feature extraction with low-rank embedding and dual Laplacian regularization, which can select important features for composing the low dimensional subspace.
2) The proposed FSP integrates dimensionality reduction, feature selection and feature extraction as well as low-rank representation learning into a unified framework.

3) An alternatively iterating algorithm is well designed for solving the optimization problem of FSP, and theoretical analysis of its convergence and computational complexity are also provided.
4) Comprehensive experiments have been carried out to validate the efficacy of the proposed FSP, and experimental results demonstrate the superiority of FSP when compared with other state-of-the-art methods.

We organize the rest of this paper as follows. Some related works will be reviewed in Section 2. Then we present the proposed FSP model in Section 3, the optimization algorithm for solving the objective function is also presented in this section. Theoretical convergence and computational complexity analysis of the algorithm for solving FSP and connection with previous works are presented in Section 4. Experimental results and comparison are shown in Section 5. Finally, we draw the conclusion in Section 6.

2 Related Works

Since our work focuses on projection based feature extraction, we first briefly review some related projection learning works. Throughout this paper, scalars are denoted as lowercase letters. Matrices and vectors are denoted by bold uppercase letters and bold lowercase letters, respectively. A data matrix is denoted as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$, which contains n samples with m dimension. The ith row and jth column of matrix $\mathbf{M}$ are denoted as $\mathbf{m}^i$ and $\mathbf{m}_j$, respectively. For a square matrix M, $Tr(\mathbf{M})$ represents its trace. The transpose of matrix $\mathbf{M}$ is written as $\mathbf{M}^T$. For any two matrices, their standard inner product is denoted as $\langle \mathbf{A}, \mathbf{B} \rangle$. $\mathbf{I}_m$ represents an identity matrix with size $m \times m$ (we use $\mathbf{I}$ instead if the size is obviously known). For matrix $\mathbf{M}$, its $\ell_{2,1}$-norm is defined as $||\mathbf{M}||_{2,1} = \sum_{j=1}^{n} ||\mathbf{m}_j|| = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m} \mathbf{M}_{ij}{}^2}$. $||\mathbf{M}||_*$ is the nuclear norm and $||\mathbf{M}||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{M}_{ij}{}^2}$ is the well-known Frobenius norm. Projection based feature extraction aims to learn an optimal projection coefficient matrix $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_d] \in \mathbb{R}^{m \times d}$ with $d < m$, which can transform original m dimensional data into a d dimensional subspace.

As mentioned in the introduction section, the earliest projection based feature extraction methods include PCA, LDA, etc. For capturing the local manifold geometrical structure of data, a variety of manifold learning based methods have been proposed, i.e., NPE, LPP and LLE. However, there often exists noises in the original data, which limits the performance of previous projection learning methods. In recent years, thanks to the development of LRR models, some LRR based feature extraction approaches have been put forward, these methods are more robust to noise than traditional ones. Here, we introduce two LRR based feature extraction methods (LRPP and LRE) which we think to be the most related to our work.

2.1 Low-Rank Preserving Projections (LRPP)

As a combination of LPP and LRR, LRPP [78] uses the low rank data representation coefficient matrix to construct an affinity graph for local manifold geometrical structure preservation, the projection matrix and data representation matrix are learned simultaneously by regularizing the low rank of data representation and sparsity of noises. Compared to traditional LPP which constructs the affinity graph by using the pair-wise euclidean distances, LRPP jointly learns the affinity graph and projection matrix. The objective function of LRPP is constructed as follows:

$$\begin{aligned} \min_{\mathbf{P},\mathbf{C},\mathbf{E}} & \frac{1}{2} \sum_{i,j=1}^{n} (\mathbf{C}_{ij} + \mathbf{C}_{ji}) ||\mathbf{P}^T \mathbf{x}_i - \mathbf{P}^T \mathbf{x}_j||_2^2 + \alpha ||\mathbf{C}||_* \\ & + \beta ||\mathbf{E}||_{2,1} \\ & s.t. \ \ \mathbf{X} = \mathbf{X}\mathbf{C} + \mathbf{E}, \end{aligned} \tag{1}$$

where $\mathbf{P}$ and $\mathbf{C}$ are the projection matrix and data representation matrix need to be learned, and $\mathbf{E}$ is the noise matrix under sparsity assumption. The first term in Eq. (1) is used to preserve the local manifold geometrical structure of data in the projected subspace, the nuclear norm imposed on $\mathbf{C}$ is used to regularize the low rankness of the data representation coefficient matrix, the last term imposes the sparsity on the noise matrix. As can be seen, LRPP can effectively learn a robust projection matrix with the local manifold geometrical structure of data being well preserved.

2.2 Low-Rank Embedding (LRE)

By considering that the LRR is robust to the noise, corruptions and occlusions, Wong et al. [25] integrated the LRR and feature projection together to formulate a low-rank embedding (LRE) model for feature extraction. LRE is also under the LRR based data representation assumption and its objective function is formulated as follows:

$$\begin{aligned} & \arg\min_{\mathbf{C},\mathbf{P}} \ rank(\mathbf{C}) + \lambda \left\| \mathbf{P}^T \mathbf{X} - \mathbf{P}^T \mathbf{X}\mathbf{C} \right\|_{2,1}, \\ & s.t. \ \ \mathbf{P}^T \mathbf{P} = \mathbf{I}. \end{aligned} \tag{2}$$

In Eq. (2), the $\ell_{2,1}$-norm used for regularizing the reconstruction term can make the model more robust than the Frobenius norm in cope with the sample-specific corruptions and outliers [42]. Since minimizing the rank constrained problem is NP-hard, the LRE model can be transferred to the following approximated problem by using nuclear norm instead of the rank function [80]

$$\begin{aligned} & \arg\min_{\mathbf{C},\mathbf{P}} \ ||\mathbf{Z}||_* + \lambda \left\| \mathbf{P}^T \mathbf{X} - \mathbf{P}^T \mathbf{X}\mathbf{C} \right\|_{2,1}, \\ & s.t. \ \ \mathbf{P}^T \mathbf{P} = \mathbf{I}. \end{aligned} \tag{3}$$

Eq. (3) can be solved by using the argument Lagrangian multiplier method [25].

3 Feature Selective Projection

In this section, we present the detailed elaboration of our proposed method, i.e., feature selective projection (FSP). The motivation of FSP will be explained first, and the objective function and solutions are given successively.

3.1 The Motivation of FSP

Due to its powerful ability to explore the embedded low dimensional subspace structure of data, LRR has attracted more and more attention. It should be noted that the LRR can still handle the case that the data contains noises or outliers, which demonstrates the robustness of LRR for subspace clustering [67]. However, LRR lacks the functionality for

dimensionality reduction, which results in the deficiency of LRR in feature extraction. Thus, LRE [25] and LRPP [78] integrate LRR and feature projection learning into a unified framework to make feature extraction more robust to noises. Since the extracted low-dimensional features can be regarded as a linear combination of original features, one drawback of previous methods lies in that the learned projection matrix $\mathbf{P}$ lacks interpretability, i.e., it only projects original data into a lower dimensional subspace but cannot reflect which features are critical for combining the low-dimensional features. In other words, they cannot perform feature extraction and feature selection jointly. In the following we use a detailed equation to elaborate this problem. Let $\mathbf{Y} \in \mathbb{R}^{d\times n}$ with $d < m$ denotes the low-dimensional projection of original data $\mathbf{X}$, i.e., $\mathbf{Y} = \mathbf{P}^T\mathbf{X}$, then we have $\mathbf{Y}^T = \mathbf{X}^T\mathbf{P}$ each row of $\mathbf{Y}$ can be expanded as following form:

$$\begin{aligned}
\mathbf{y}^1 &= \mathbf{x}^1\mathbf{p}_{1,1} + \mathbf{x}^2\mathbf{p}_{1,2} + \cdots + \mathbf{x}^m\mathbf{p}_{1,m} \\
\mathbf{y}^2 &= \mathbf{x}^1\mathbf{p}_{2,1} + \mathbf{x}^2\mathbf{p}_{2,2} + \cdots + \mathbf{x}^m\mathbf{p}_{2,m} \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\mathbf{y}^d &= \mathbf{x}^1\mathbf{p}_{d,1} + \mathbf{x}^2\mathbf{p}_{d,2} + \cdots + \mathbf{x}^m\mathbf{p}_{d,m}.
\end{aligned} \tag{4}$$

As can be seen from Eq. (4), the low-dimensional representation of $\mathbf{X}$ is linearly combined by its original features, and each row of the projection matrix $\mathbf{P}$ can be used to measure the importance of corresponding feature dimension for combining the low-dimensional representation. This motivates us to add some constrains on $\mathbf{P}$ to select discriminative features for dimension reduction.

Another important issue is that LRR does not take the non-linear manifold geometric structure implied in data into consideration, thus it cannot capture the locality among data points during the learning process. To this end, we impose a dual graph based Laplacian regularization term for preserving the local structure of data in the projected lower dimensional subspace. The dual graph based Laplacian regularization term preserves the locality among data from following two aspects:

1) If two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are close to each other in original space, their corresponding mappings in the projected lower dimensional space, i.e., $\mathbf{y}_i$ and $\mathbf{y}_j$ should be also close to each other;
2) If two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are close to each other in original space, their representation coefficients should also be similar to each other.

3.2 Formulation of FSP

In our work, feature selection and feature extraction are organically integrated into a uniform framework, and the feature selection can serve to feature extraction. Thus, we aim to learn a feature selective projection for dimension reduction. Similar to other LRR methods, in order to make the proposed method robust to noises, the $\ell_{2,1}$-norm is leveraged to regularize the data reconstruction errors in the projected subspace. Consequently, we formulate the mathematical model of our FSP as follows:

$$\begin{aligned}
&\min_{\mathbf{P},\mathbf{Z}} ||\mathbf{P}^T\mathbf{X} - \mathbf{P}^T\mathbf{X}\mathbf{Z}||_{2,1} + \lambda||\mathbf{Z}||_* + \beta\mathcal{R}(\mathbf{P}) + \gamma\mathcal{L}(\mathbf{P},\mathbf{Z}), \\
&s.t.\ \ \mathbf{P}^T\mathbf{X}\mathbf{X}^T\mathbf{P} = \mathbf{I},
\end{aligned} \tag{5}$$

where the first term is used to measure the reconstruction errors. Compared to the Frobenious norm, the $\ell_{2,1}$-norm used here is more robust to noises such as sample outliers and sample-specific corruptions [25], [42], [67]. The second term constrains that the projected data points in the lower dimensional subspace can be linearly reconstructed by themselves by using a representation coefficient matrix with low rank constraint. The third term $\mathcal{R}(\mathbf{P})$ is a constraint used to enable the feature selection functionality of projection matrix $\mathbf{P}$, and the forth term $\mathcal{L}(\mathbf{P},\mathbf{Z})$ is the dual Laplacian regularization term for locality preservation. During the feature projection process, we aim to reduce the redundancies of original data to the maximum extent, thus the PCA-like constraint $\mathbf{P}^T\mathbf{X}\mathbf{X}^T\mathbf{P} = \mathbf{I}$ is used to conduct subspace learning [81], [82], which intends to make the low-dimensional features more discriminative.

Eq. (5) aims to find an optimal low rank reconstruction matrix in the low-dimensional data space and a feature selective projection matrix which projects original data points into the lower dimensional subspace. During the projection and reconstruction process, the dual Laplacian regularization term is used to preserve the local manifold geometrical structure of original data.

As depicted by Eq. (4), the low-dimensional representation of $\mathbf{X}$ is linearly combined by its original features, and each row of $\mathbf{P}$ weighs the importance of corresponding feature dimension for combining the low-dimensional representation. Thus the discriminative features should be assigned with larger weights while unimportant features should be suppressed. By considering this point, we impose the $\ell_{2,1}$-norm regularization on $\mathbf{P}^T$ to enable the row sparsity of $\mathbf{P}$. Therefore, we have

$$\mathcal{R}(\mathbf{P}) = ||\mathbf{P}^T||_{2,1}. \tag{6}$$

For the dual Laplacian regularization term, we constrain that if two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are close to each other in original data space, their mappings and representation coefficients in the projected lower dimensional subspace should also be similar to each other. To this end, we construct the following model:

$$\mathcal{L}(\mathbf{P},\mathbf{Z}) = \frac{1}{2}\sum_{i,j=1}^{n} ||\mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j||^2\mathbf{S}_{ij} + \frac{1}{2}\sum_{i,j=1}^{n} ||\mathbf{z}_i - \mathbf{z}_j||^2\mathbf{S}_{ij}, \tag{7}$$

where $\mathbf{S}$ denotes the affinity matrix of data points. The ijth element of $\mathbf{S}$ is often defined by using following Gaussian kernel function (other distance measuring functions can be also used instead)

$$\mathbf{S}_{ij} = \begin{cases} \exp\left(\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{-2\sigma^2}\right), & \mathbf{x}_i \in \mathcal{N}_\epsilon(\mathbf{x}_j)\ or\ \mathbf{x}_j \in \mathcal{N}_\epsilon(\mathbf{x}_i); \\ 0, & otherwise, \end{cases} \tag{8}$$

where $\mathcal{N}_\epsilon(\mathbf{x}_i)$ represents the data points set of ϵ nearest neighbors of $\mathbf{x}_i$ and σ is the kernel width. Eq. (7) can be regarded as the graph embedding term [83] and it is easy to be reformulated as the following trace form:

$$\mathcal{L}(\mathbf{P},\mathbf{Z}) = Tr(\mathbf{P}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P}) + Tr(\mathbf{Z}\mathbf{L}\mathbf{Z}^T), \tag{9}$$

where $\mathbf{L} \in \mathbb{R}^{n\times n}$ represents the corresponding Laplacian matrix with $\mathbf{L} = \mathbf{D} - \mathbf{S}$, where $\mathbf{D}$ is a diagonal degree matrix with $\mathbf{D}_{ii} = \sum_j \mathbf{S}_{ij}$.

By combining Eqs. (5), (6) and (9) together, we have the final objective function of our FSP as following:

$$\begin{aligned} &\min_{\mathbf{P},\mathbf{Z}} ||\mathbf{P}^T\mathbf{X} - \mathbf{P}^T\mathbf{X}\mathbf{Z}||_{2,1} + \lambda||\mathbf{Z}||_* + \beta||\mathbf{P}^T||_{2,1} \\ &+ \gamma\{Tr(\mathbf{P}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P}) + Tr(\mathbf{Z}\mathbf{L}\mathbf{Z}^T)\}, \\ &s.t.\ \ \mathbf{P}^T\mathbf{X}\mathbf{X}^T\mathbf{P} = \mathbf{I}. \end{aligned} \tag{10}$$

As can be seen from Eq. (10), FSP jointly integrates feature extraction, feature selection and low rank representation into a unified framework, in which the redundancies among original data can be efficiently reduced for learning tasks. Therefore, the proposed FSP can work for both feature selection and feature extraction, and the feature selection and feature extraction can boost each other during the optimization process. In contrast, both LRPP and LRE only work for feature extraction.

3.3 Optimal Solution of FSP

In this section, we give the detailed optimization steps of FSP. Since it is difficult to simultaneously optimize the two variables in (10), we develop an alternatively iterative algorithm to solve it. Specifically, we first optimize $\mathbf{Z}$ with fixed $\mathbf{P}$ and then optimize $\mathbf{P}$ with fixed $\mathbf{Z}$.

3.3.1 Optimize $\mathbf{Z}$ with Fixed $\mathbf{P}$

When $\mathbf{P}$ is fixed, solving $\mathbf{Z}$ can be transformed to minimize following object function:

$$\min_{\mathbf{Z}} ||\mathbf{P}^T\mathbf{X} - \mathbf{P}^T\mathbf{X}\mathbf{Z}||_{2,1} + \lambda||\mathbf{Z}||_* + \gamma Tr(\mathbf{Z}\mathbf{L}\mathbf{Z}^T), \tag{11}$$

We use the linearized alternating direction method with adaptive penalty (LADMAP) [84] to solve Eq. (11). In order to make the objective function separable, an auxiliary variable $\mathbf{J}$ is first introduced to convert Eq. (11) to the following equivalent problem:

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{J}} ||\mathbf{P}^T\mathbf{X} - \mathbf{P}^T\mathbf{X}\mathbf{Z}||_{2,1} + \lambda||\mathbf{J}||_* + \gamma Tr(\mathbf{Z}\mathbf{L}\mathbf{Z}^T), \\ &s.t.\ \ \mathbf{Z} = \mathbf{J}. \end{aligned} \tag{12}$$

Eq. (12) can be solved by using the augmented Lagrangian method (ALM), and the corresponding augmented Lagrangian function can be written as follows:

$$\begin{aligned} \mathcal{H}(\mathbf{Z},\mathbf{J},\mathbf{M},\mu) = &||\mathbf{P}^T\mathbf{X} - \mathbf{P}^T\mathbf{X}\mathbf{Z}||_{2,1} + \lambda||\mathbf{J}||_* + \gamma Tr(\mathbf{Z}\mathbf{L}\mathbf{Z}^T) \\ &+ < \mathbf{M}, \mathbf{Z} - \mathbf{J} > + \frac{\mu}{2}||\mathbf{Z} - \mathbf{J}||_F^2, \end{aligned} \tag{13}$$

where $\mathbf{M}$ is the introduced Lagrange multiplier, $\mu > 0$ is a penalty parameter. Eq. (13) is an unconstrained problem. $\mathbf{Z}$ and $\mathbf{J}$ can be updated iteratively by fixing each other.

The optimal $\mathbf{J}$ can be obtained by

$$\min_{\mathbf{J}} \lambda||\mathbf{J}||_* + \frac{\mu}{2}||\mathbf{Z} - \mathbf{J} + \frac{\mathbf{M}}{\mu}||_F^2, \tag{14}$$

of which the solution can be obtained by using the Singular Value Thresholding (SVT) operator [85].

Then $\mathbf{Z}$ can be solved by

$$\min_{\mathbf{Z}} ||\mathbf{P}^T\mathbf{X} - \mathbf{P}^T\mathbf{X}\mathbf{Z}||_{2,1} + \frac{\mu}{2}||\mathbf{Z} - \mathbf{J} + \frac{\mathbf{M}}{\mu}||_F^2 + \gamma Tr(\mathbf{Z}\mathbf{L}\mathbf{Z}^T). \tag{15}$$

For solving $\mathbf{Z}$, we deploy the iterative reweighted least-squares (IRLS) algorithm [86]. By taking the derivative of Eq. (15) w.r.t $\mathbf{Z}$, and setting the derivative to zero, then we get

$$\begin{aligned} &\mathbf{X}^T\mathbf{P}\mathbf{P}^T\mathbf{X}\mathbf{Z} + \frac{\mu}{2}\mathbf{Z}\mathbf{G}^{-1} + \gamma\mathbf{Z}\mathbf{L}\mathbf{G}^{-1} \\ &- \mathbf{X}^T\mathbf{P}\mathbf{P}^T\mathbf{X} + (\frac{\mathbf{M}}{2} - \frac{\mu\mathbf{J}}{2})\mathbf{G}^{-1} = 0, \end{aligned} \tag{16}$$

where $\mathbf{G}$ is a diagonal matrix and its ith diagonal entry is calculated as

$$\mathbf{G}_{i,i} = \frac{1}{2||\mathbf{q}_i||_2}, \tag{17}$$

where $\mathbf{q}_i$ denotes the ith column of $\mathbf{P}^T\mathbf{X} - \mathbf{P}^T\mathbf{X}\mathbf{Z}$. Eq. (16) is a Sylvester equation [87] with the form

$$\mathbf{A}\mathbf{Z} + \mathbf{Z}\mathbf{B} = \mathbf{C}, \tag{18}$$

with

$$\begin{aligned} &\mathbf{A} = \mathbf{X}^T\mathbf{P}\mathbf{P}^T\mathbf{X}; \\ &\mathbf{B} = \frac{\mu}{2}\mathbf{G}^{-1} + \gamma\mathbf{L}\mathbf{G}^{-1}; \\ &\mathbf{C} = \mathbf{X}^T\mathbf{P}\mathbf{P}^T\mathbf{X} + (\frac{\mu\mathbf{J}}{2} - \frac{\mathbf{M}}{2})\mathbf{G}^{-1}. \end{aligned} \tag{19}$$

Since $\mathbf{A}$ can be ensured to be strictly positive definite, the Sylvester Eq. (18) has a unique solution.

Then the multiplier and the parameter at each step can be updated as follows:

$$\mathbf{M} \leftarrow \mathbf{M} + \mu(\mathbf{Z} - \mathbf{J}), \mu \leftarrow \min(\rho\mu, \mu_{\max}), \tag{20}$$

where $\rho > 0$ is manually set parameter.

In summary, the detailed steps for solving $\mathbf{Z}$ can be described by Algorithm 1.

Algorithm 1. Iterative Algorithm for Solving $\mathbf{Z}$

Input: Data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, projection matrix $\mathbf{P}$, parameters α and γ.
Initialization: $\mathbf{Z} = \mathbf{0}$, $\mathbf{J} = \mathbf{0}$, $\mathbf{M} = \mathbf{0}$ and $\mathbf{G} = \mathbf{I}$, $\mu = 10^{-6}$, $\mu_{max} = 10^6$, and $\rho = 1.1$.
while not converged do
1. Update $\mathbf{J}$ using (14);
  while not converged do
    2.1 Update $\mathbf{Z}$ by solving Eq. (18);
    2.2 Update $\mathbf{G}$ using Eq. (17);
  end while
3. Update $\mathbf{M}$ and μ using Eq. (20);
end while
Output: $\mathbf{Z}$ and $\mathbf{J}$.

3.3.2 Optimize $\mathbf{P}$ with Fixed $\mathbf{Z}$

When $\mathbf{Z}$ is given, the problem for solving $\mathbf{P}$ becomes

$$\begin{aligned} &\min_{\mathbf{P}} ||\mathbf{P}^T\mathbf{X} - \mathbf{P}^T\mathbf{X}\mathbf{Z}||_{2,1} + \beta||\mathbf{P}^T||_{2,1} \\ &+ \gamma Tr(\mathbf{P}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P}), \\ &s.t.\ \ \mathbf{P}^T\mathbf{X}\mathbf{X}^T\mathbf{P} = \mathbf{I}. \end{aligned} \tag{21}$$

We also use the diagonal matrix $\mathbf{G}$ in the step for solving $\mathbf{Z}$ and introduce a new diagonal matrix Λ and its ith diagonal element is defined as

$$\Lambda_{i,i} = \frac{1}{2||\mathbf{p}^i||_2}. \tag{22}$$

By integrating Λ into Eq. (21), we convert the problem for solving $\mathbf{P}$ to the following form:

$$\begin{aligned} &\min_{\mathbf{P}} Tr(\mathbf{P}^T\mathbf{X}(\mathbf{I}-\mathbf{Z})\mathbf{G}(\mathbf{I}-\mathbf{Z})^T\mathbf{X}^T\mathbf{P}) + \beta Tr(\mathbf{P}^T\Lambda\mathbf{P}) \\ &+ \gamma Tr(\mathbf{P}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P}) \\ &s.t.\ \ \mathbf{P}^T\mathbf{X}\mathbf{X}^T\mathbf{P} = \mathbf{I}. \end{aligned} \tag{23}$$

Eq. (23) can be solved by eigen-decomposition. By solving the following minimum eigenvalues problem as described by Eq. (24), each column of $\mathbf{P}$, i.e., $\mathbf{p}_i$ can easily obtained

$$[\mathbf{X}(\mathbf{I}-\mathbf{Z})\mathbf{G}(\mathbf{I}-\mathbf{Z})^T\mathbf{X}^T + \beta\Lambda + \gamma\mathbf{X}\mathbf{L}\mathbf{X}^T]\mathbf{p}_i = \xi\mathbf{X}\mathbf{X}^T\mathbf{p}_i. \tag{24}$$

Let $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_d]$ be the solution of (24). Column vectors $\mathbf{p}_i (i = 1, \ldots, d)$ correspond to the eigenvectors corresponding to the first d smallest eigenvalues. The details for solving $\mathbf{P}$ can be summaried by Algorithm 2

Algorithm 2. Iterative Algorithm for Solving $\mathbf{P}$

Input: Data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, representation coefficient matrix $\mathbf{Z}$, parameter β.
Initialization: $\mathbf{G}$ and Λ.
while not converged do
1. Update $\mathbf{P}$ by solving (23);
2. Update Λ using (22);

end while
Output: $\mathbf{P}$.

In a nutshell, the whole iterative optimization process of FSP is depicted by Algorithm 3.

Algorithm 3. The Optimization Algorithm of FSP

Input: Data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$.
while not converged do
1. Update $\mathbf{Z}$ using Algorithm. (1);
2. Update $\mathbf{P}$ using Algorithm. (2);

end while
Output: $\mathbf{Z}$ and $\mathbf{P}$.

4 Theoretical Algorithm Analysis

In this section, we theoretically analyze the convergence and computational complexity of the optimization algorithm for solving the proposed FSP model. The connections between FSP and previous works are also discussed.

4.1 Convergence Analysis

In the step of solving $\mathbf{Z}$, the exact LADMAP algorithm can converge well, which has been generally proven in [84]. In Algorithm 1, there are two blocks need to be updated, i.e., $\mathbf{Z}$ and $\mathbf{J}$, while $\mathbf{J}$ is the only one auxiliary variable. Thus, the convergence of Algorithm 1 can be well guaranteed. For solving $\mathbf{P}$, the IRLS algorithm with convergence guarantee [86] is used to iteratively update $\mathbf{P}$ and Λ. Therefore, the convergence of the optimization algorithm for solving FSP can be well reached. In the experiments section, we will also plot the values of the object function (10) with iteration times on real datasets to empirically validate the convergence property of Algorithm (3).

4.2 Computational Complexity Analysis

We analyze the computational complexity of the optimization algorithms for solving FSP in this section. In Algorithm 1, the main computation cost consists of updating $\mathbf{J}$ and $\mathbf{Z}$. For computing $\mathbf{J}$, the main computational complexity comes from the SVD which needs $\mathcal{O}(n^3)$. For Solving $\mathbf{Z}$, the classical Bartels Stewart algorithm is used for the solving the Sylvester equation, whose complexity is $\mathcal{O}(n^3)$. In Algorithm 2, the main computation cost comes from computing $\mathbf{P}$ by solving Eq. (23), whose complexity is $\mathcal{O}(m^3)$.

4.3 Connections with Previous Works

LPP obtains widely attentions in the last decade due to its popularity for dimensionality reduction. In LPP, the local manifold geometrical structure of the data can be well preserved, it has been used in many learning models [26], [71]. In our proposed FSP model (10), the first Laplacian regularization term can be also regarded as the LPP model.

LRE simultaneously learns the feature projection matrix and the data representation coefficient matrix in the projected lower dimensional subspace. Thus, the learned projection is robust to noises. Compared to LRE, the projection matrix learned by FSP can measure the importance of original features for generating lower dimensional features. In addition, the local manifold geometrical structure of original data is preserved by using the dual Laplacian regularization term. It should be noted that when we set $\beta = 0$ and $\gamma = 0$, the main part of the FSP model (10) degenerates into the LRE model (3). In other words, FSP can degenerate to LRE by dropping the $l_{2,1}$-norm regularization on $\mathbf{P}^T$ and the dual Laplacian regularization term.

5 Experimental Results

In this section, we carry out a series of experiments to demonstrate the efficacy of FSP model for image feature extraction and classification in terms of various noises and occlusions. We compare the proposed FSP with the classical subspace learning method, i.e., RPCA, some of manifold learning regularized methods, i.e., LPP, SPP, and NPE, the low-rank representation (LRR) [66], and the recently proposed low-rank preserving projections (LRPP) and low-rank embedding feature extraction (LRE).

5.1 Datasets

We use eight different publicly released datasets in our experiments, including ORL [88], USPS [89], COIL20 [90], CMU PIE (Pose29, with light and illumination change) [91], FERET [92], Yale [93], AR [94] and MNIST [95]. Following we present the detailed description of the datasets.

For ORL dataset, it consists of 40 distinct subjects and there are 10 different face images for each subject. The images are taken at different times with varying the lighting, facial expressions and facial details. All of the images are captured

against a dark homogeneous background and the subjects are in an upright, frontal position (with tolerance for some side movement).

The COIL20 dataset consists of 20 object images. For each object, 72 gray images are taken from different view directions.

The USPS is a handwritten digital image dataset, which contains 11,000 images in total. The digits include "0" to "9", and each digit has 1,100 examples. 100 images for each digital were randomly select from this dataset for our experiments.

There are 41,368 face images captured from 68 subjects in CMU PIE dataset. The images captured are under 43 different illumination conditions, 13 different poses and four different expressions for each subject. In our experiments, we choose the subsets named "C29", "C05", "C07", "C09", and illumination indexed as 08 and 11 which involves variations in pose for our experiment.

The FERET dataset contains a total of 14,126 images pertaining to 1,199 individuals along with 365 duplicate sets of images that were taken on different days. Following [78], we also randomly choose 70 people and six images for each individual to construct a subset for conducting our experiments.

As to the Yale face dataset, there are 165 images of 15 individuals. For each individual, 11 images with different facial expression or configuration are captured.

In AR face dataset, there are over 4,000 color face images captured from 126 people which contain 70 men and 56 women. The face images were taken with different conditions, including different facial expressions, lighting conditions, and occlusions. The face images of most individuals were captured in two sessions which are separated by two weeks. In our experiments, we construct a subset by randomly selecting images of 50 men and 50 women.

The MNIST dataset is a large dataset of handwritten digits which consists of digital numbers from "0" – "9". It contains 60,000 training samples and 10,000 testing samples.

In our experiments, we use the ORL, COIL20 and USPS datasets for evaluating the performance of FSP with data containing random pixel corruptions. For CMU PIE, FERET and Yale datasets, we use them to test the performance of FSP with data corrupted by block occlusions. As to AR dataset, we use it for testing the performance of FSP with face images corrupted by sunglasses and scarf occlusions. Finally, we use MNIST dataset to demonstrate the efficacy of FSP in terms of handwritten digit recognition.

For ORL, USPS, CMU PIE, FERET and Yale datasets, half of the images per class are selected as training samples and the rest are left as testing samples. As to COIL20 dataset, 30 images per class are selected as training samples and the rest are used for testing. For AR face dataset, we randomly select three neutral images and three images with sunglasses/scarf from session 1 as training samples. The testing samples are constructed by seven neutral images plus three images with sunglasses/scarf from session 2 [78]. The images in USPS dataset and MNIST dataset are normalized to 16×16 and 28×28 pixels, respectively, and the images in other six datasets are normalized to 32×32 pixels. We reshape each normalized image to a vector for constructing the feature matrix.

5.2 Experiment Setup

We evaluate the performance of FSP under various corruptions by using the above mentioned datasets and compare it with some previous typical feature extraction methods including RPCA, LPP, NPE, SPP, LRR, LRPP and LRE. Among these methods, RPCA is the classical subspace learning method. LPP, NPE and SPP are some manifold learning regularized approaches. LRR aims to find the low-rank feature representation of data. LRPP can reduce the dimensionality of data with the global data structure being well preserved, it can also reduce the disturbance of noises in the data by learning a low rank weight matrix. LRE is the most recent proposed low rank embedding feature extraction method which jointly learns the low rank representation and projection of data. In our experiments, for each dataset, we randomly select half of images from each subject as training samples and the rest are used for testing.

The neighborhood size in LPP, NPE and FSP is fixed to 5 in our experiments. The regularization parameters of all the methods are adjusted based on grid search and the optimal combination is determined from $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$. The numbers of final subspace dimensions for all the datasets are varied from 5 to 150 with step 5. We use two common classifiers, i.e., random forest (RF) and the 1-nearest neighbor (1-NN) to evaluate the final classification accuracy.

For each parameter combination, we independently run all the algorithms 5 times and the averaged classification accuracy is reported.

5.3 Experimental Results and Analysis

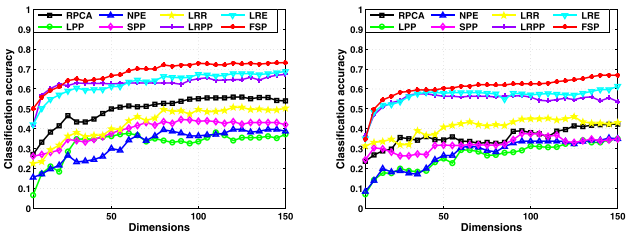
In this section, we will first plot the classification accuracies obtained by different methods using different extracted feature dimensions on different datasets under different occlusions/noises. Due to the space limitation, we only plot the classification results by using the 1NN classifier. Then we will summarize the best averaged classification accuracies and the corresponding extracted feature dimensions of different methods on all of the datasets by using both the 1NN and RF classifiers.

5.3.1 Classification Results with Random Pixel Corruptions

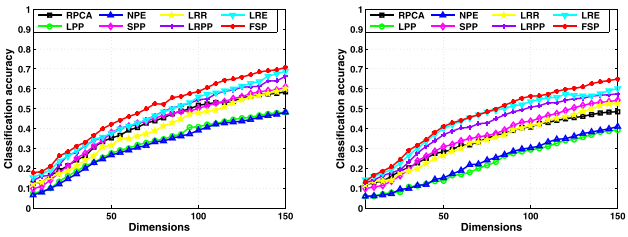
For evaluating the robustness of FSP to the data with random pixel corruptions, we add the salt and pepper noise with two different densities (0.1 and 0.15) to the images of ORL, COIL20 and USPS datasets. In the supplementary, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2019.2911946, we give some examples of the original and corrupted images under two different densities of the salt and pepper noise from the three datasets. The classification accuracies of different methods on this dataset with different extracted feature dimensions are plotted in Fig. 1. As can be seen, FSP steadily outperforms other methods under different extracted feature dimensions.

5.3.2 Classification Results with Block Occlusions

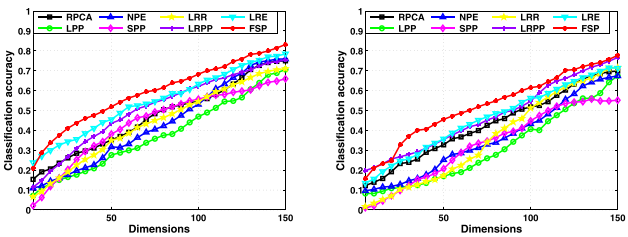
For evaluating the robustness of FSP to block occlusions, some black blocks are randomly added to different locations of the images in the CMU PIE, FERET and Yale datasets.



(a) ORL (Noisy density=0.1) (b) ORL (Noisy density=0.15)

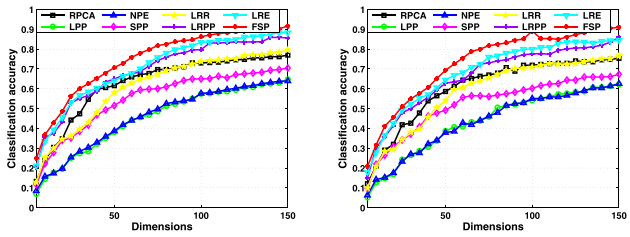


(c) COIL20 (Noisy density=0.1) (d) COIL20 (Noisy density=0.15)

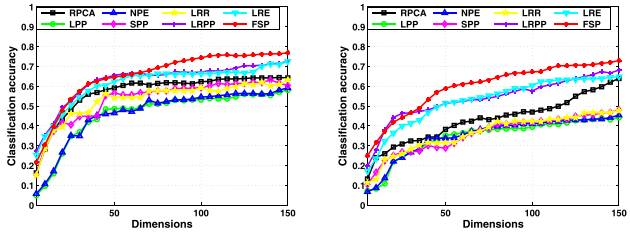


(e) USPS (Noisy density=0.1) (f) USPS (Noisy density=0.15)

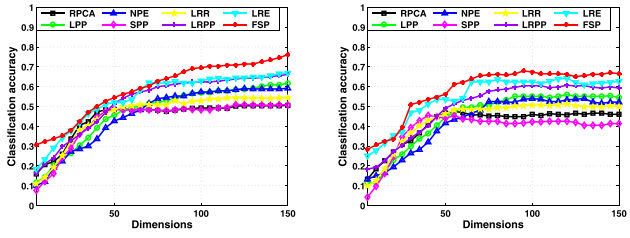
Fig. 1. Classification accuracies of different methods with different extracted feature dimensions on the on the ORL, COIL20, and USPS datasets under different densities of the salt and pepper noise.



(a) CMU PIE (Occlusion size 6×6) (b) CMU PIE (Occlusion size 8×8)



(c) FERET (Occlusion size 6×6) (d) FERET (Occlusion size 8×8)



(e) Yale (Occlusion size 6×6) (f) Yale (Occlusion size 8×8)

Fig. 2. Classification accuracies of different methods with different extracted feature dimensions on the CMU PIE, FERET, and Yale datasets with different block occlusions.

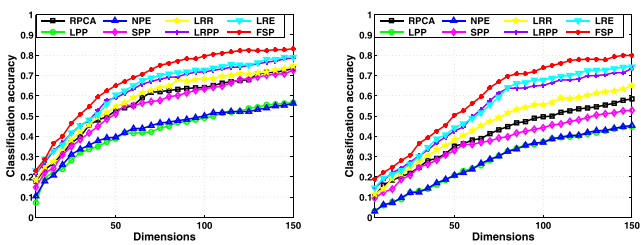
The added blocks are set with two different sizes: 6×6 and 8×8. In the supplementary, available online, we show some examples of original images and the images corrupted by different sizes of blocks. We plot the classification accuracies obtained from different methods on this dataset with different extracted feature dimensions in Fig. 2. The results also demonstrate the superiority of the proposed FSP, which demonstrates its robustness to block occlusions.

5.3.3 Classification Results with Sunglasses/Scarf Occlusions

In reality, human faces are usually occluded by sunglasses or scarf. In this experiment, we also test the robustness of the proposed FSP to these two kinds of occlusions. The AR face dataset is used to conduct this experiment. In the supplementary, available online, we give some examples of clean images and images occluded by sunglasses and scarf. We plot the classification accuracies obtained from different methods on this dataset with different extracted feature dimensions in Fig. 3. As can be seen, FSP has the highest classification accuracy when compared to other previous methods, which demonstrates that FSP is more robust than the other methods in terms of sunglasses and scarf occlusion.

5.3.4 Classification Results with Sample-Specific Corruptions

For testing the robustness of FSP to sample-specific corruptions, we randomly select half of the images from the "C29"



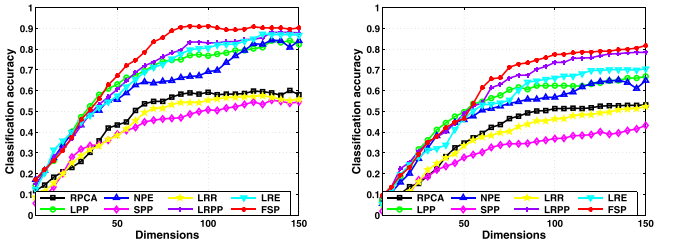
(a) AR (Occluded by sunglasses) (b) AR (Occluded by scarf)

Fig. 3. Classification accuracies of different methods with different extracted feature dimensions on the AR face dataset with sunglasses and scarf occlusions.

subset of CMU PIE dataset, and add the baboon face image with different intensities to the selected images. In our experiments, the intensities of the baboon image added into the images are set to 0.2 and 0.4. In the supplementary, available online, we show some samples of clean images and the images corrupted by the baboon face image with different intensities. The classification accuracies obtained by different methods with various extracted dimensional features on this dataset are plot in Fig. 4, it shows that FSP has the highest classification accuracies when compared to other methods. Thus, FSP works better in the case that the images are mixed with the sample-specific corruptions.

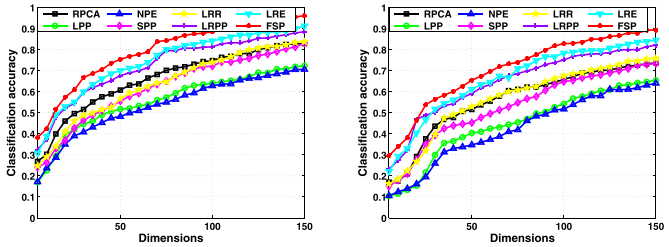
5.3.5 Handwritten Digits Recognition Results

In order to test the performance of FSP in handwritten digits recognition. We conduct experiments on the MNIST dataset



(a) CMU PIE (Noisy density=0.2) (b) CMU PIE (Noisy density=0.4)

Fig. 4. Classification accuracies of different methods with different extracted feature dimensions on the CMU PIE dataset with different densities of sample-specific corruptions.



(a) MNIST (Noisy density=0.1) (b) MNIST (Noisy density=0.15)

Fig. 5. Classification accuracies of different methods with different extracted feature dimensions on the MNIST dataset with different densities of salt and pepper noise.

to classify each image into one of the ten digits by using FSP. There are 60,000 training samples and 10,000 testing samples in this dataset. Similar to [78], 3,000 images from the 60,000 training samples are randomly selected to construct the training set and 5,000 images from the 10,000 testing samples are randomly selected to construct the test set. We also add the salt and pepper noise with different densities to the images for testing the robustness of FSP. In the supplementary, available online, we show some clean images and the images corrupted by different densities of salt and pepper noise. Fig. 5 plots the classification accuracies of different methods on the MNIST dataset with the salt and pepper noises. The results also verify the superiority of our proposed FSP.

5.3.6 Best Classification Accuracies of Different Methods on Different Datasets

In this section, we summarize the best averaged classification accuracies and the corresponding extracted feature dimensions of different methods on different datasets. Due to the space limitation, the results are shown in the supplementary, available online. The results obtained by the two classifiers, i.e., RF and 1NN are reported. As can be seen, our proposed FSP performs better than other compared methods in terms of classification when the datasets are corrupted with various occlusions or noises. As aforementioned, the LRR is robust to a certain extent to the noisy/corrupted data. Therefore, the LRR based methods, i.e., LRPP, LRE and FSP perform better than the rest methods. In addition, manifold learning based dimensionality reduction approaches can well uncover the intrinsic geometric manifold structure (especially the local structure) of data during the projection process. Thus, FSP can obtain better results than LRE, and LPP and NPE outperform RPCA in most cases.

As to the ORL, Yale and CMU PIE face datasets, it should be noted that the high classification accuracies of most of the methods do not monotonically increase with the feature dimensions. This is due to the inherent characteristics of the human face, i.e., the facial features including eyes, ears, nose, mouth and eyebrows can be characterized by features with a certain dimension, faces belong to different subjects can be well distinguished by using these features. When the feature dimension increases, it dose not contribute significantly to the classification accuracy. As to the COIL20 and USPS datasets, there are no certain dimensional features can uniformly describe different digital objects and handwritten letters. Therefore, the higher dimension of the features, better classification accuracies can be obtained for these two datasets.

5.3.7 Non-Parametric Statistical Test

In order to verify whether the improvement of the proposed FSP in term of classification performance is statistically significant, we statistically validate the classification results of different methods on different datasets under various noisy conditions. Following [96], we also conduct the non-parametric pair-wised Wilcoxon test on the classification accuracies. In our experiment, the level of significance is set to 0.05. The p-values of FSP against other compared methods with respect to classification accuracy are shown in the supplementary, available online. Note that a smaller p-value means that the corresponding model is more statistically significant. As can be seen from the results, our FSP achieves the smallest p-values on different datasets, which validates that FSP achieves statistically significant improvements.

5.4 Parameters Sensitivity Analysis

There are three regularization parameters in FSP model, i.e., λ, β and γ. In our experiments, we choose them from $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ by a grid search manner. In order to analyze the parameter effect on the final classification results, for each dataset, we show the classification accuracy obtained by the 1NN classifier versus one of the parameters with other two fixed. Meanwhile, the extracted feature dimension for each dataset is set as the optimal value as shown in the supplementary, available online. In Fig. 6, we give the classification accuracies of FSP with different parameters on different datasets. As can be seen, the performance of the parameters' variations on the eight datasets are very similar. For different datasets, the best performance of FSP can be obtained when $\lambda = 0.1$ in most cases. For β, good results can be expected when it is set to 1. As to γ, when it varies between 10 to 100, best classification accuracy can be reached.

5.5 The Feature Selection Property of FSP

5.5.1 Intuitive Results

As we discussed in Section 3.1, the low-dimensional representation of original data is linearly combined by the original features, and each row of matrix $\mathbf{P}$ acts as the combination coefficients which can measure the importance of corresponding feature dimension. In order to give an intuitive

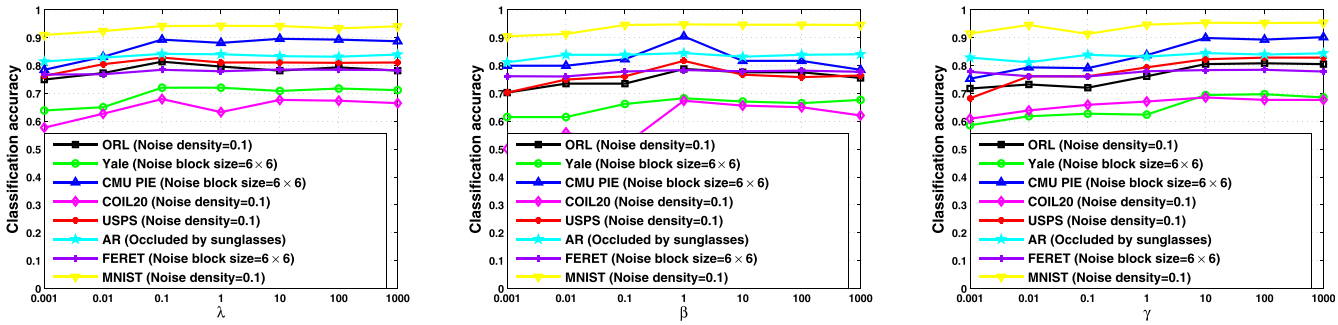


Fig. 6. The classification accuracies versus the parameter (a) λ with $\beta = \gamma = 1$, (b) β with $\lambda = \gamma = 1$, and (c) γ with $\lambda = \beta = 1$ on different databases.

interpretation, we use the ORL dataset to learn a projection matrix $\mathbf{P} \in \mathbb{R}^{m \times d}$, and then calculated the l_2-norm of each row of $\mathbf{P}$ to show the feature weights. The projection matrix and feature weight matrix are shown in Fig. 7. As can be seen, the learned feature projection matrix is clearly with row sparsity. Meanwhile, Fig. 7b can highlight the most important features on the human face, i.e., the feature points on the mouth, nose and eyes are with larger weights than other non-characteristic region. Therefore, the projection matrix learned from FSP is more interpretable and it can select discriminative features to combine the low-dimensional representation of original data.

5.5.2 Quantitative Results

In order to give more convincing results for validating the feature selection property of FSP, we compare the feature selection results of FSP with some other modern feature selection methods. Four datasets including ORL, Yale, COIL20 and USPS are used for comparison. Similar to the standard experiment settings in previous feature selection methods [34], [37], [97], [98], for each dataset, the top K features are selected based on their importance measured by a certain feature selection method. Then the selected features are used to perform k-means clustering. Two widely used evaluation metrics including accuracy (ACC) and normalized mutual information (NMI) are employed to evaluate the performance of clusters. The larger ACC and NMI represent better performance. More detailed definition of ACC and NMI can be found in previous works [34], [97]. We compare our FSP model with following representative unsupervised feature selection algorithms:

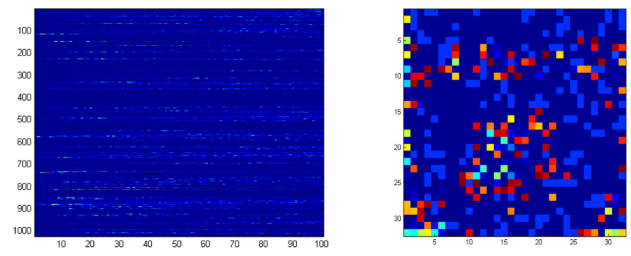


Fig. 7. An intuitive show of the learned projection matrix and feature weights on ORL face image database. (a) The learned projection matrix with $d = 100$. (b) The feature weight matrix reshaped by the l_2-norm of rows of the projection matrix.

- *Baseline*: All of the original features are adopted;
- *LS*: Laplacian Score [32], in which features are selected with the most consistency with Gaussian Laplacian matrix;
- *MCFS*: Multi-cluster feature selection [99], it uses the l_1-norm to regularize the feature selection process as a spectral information regression problem;
- *RSR*: Regularized self-representation feature selection method [100], which uses the $l_{2,1}$-norm to measure the fitting error and also promotes sparsity;
- *GLoSS*: Global and local structure preserving sparse subspace learning model for unsupervised feature selection [101], which can simultaneously realize feature selection and subspace learning.
- *GSR_SFS*: Graph self-representation sparse feature selection [102], in which the traditional fixed similarity graph is used to preserve the local geometrical structure of data.
- *DSRMR*: An efficient method for robust unsupervised feature selection which uses dual self-representation and manifold regularization [97].

Several parameters need to be set in previous methods. For LS, MCFS, SOGFS, SCUFS and RJGSC, we fixed the neighborhood size to 5 for all the datasets. In order to make fair comparison of different unsupervised feature selection methods, we tuned the hyper-parameters for all methods by a "grid-search" strategy from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. Because the optimal number of selected features is unknown, we set different numbers of selected features for all datasets, and the best clustering results from the optimal parameters are reported for all the algorithms. The selected feature number was tuned from $\{20, 30, \ldots, 90, 100\}$. After completing the feature selection process, we use the k-means algorithm to cluster the samples using the selected features. Since the performance of k-means depends on the initial point, we run it 20 times with random starting points and report the average value. The final results are shown in Table 1. As can be seen, the features selected by using our method can obtain higher ACC and NMI values, which demonstrates that the proposed FSP can also select discriminative features for learning tasks.

5.6 Convergence Study

In Section 4, theoretical analysis demonstrates that the algorithm for solving FSP converges well to the local optimum. In this section, we experimentally study the speed of

TABLE 1
Clustering Results (ACC% ± std% and NMI% ± std%) of Different Feature Selection Algorithms on Different Datasets

Dataset	Metrics	Baseline	LS	MCFS	RSR	GLoSS	GSR_SFS	DSRMR	FSP
ORL	ACC	55.79 ± 2.69	48.55 ± 2.30	50.01 ± 2.42	54.04 ± 2.66	54.19 ± 1.94	55.88 ± 2.63	57.69 ± 3.17	**58.93 ± 2.72**
	NMI	72.35 ± 1.21	71.59 ± 1.13	72.73 ± 1.32	72.25 ± 1.29	73.05 ± 1.65	71.18 ± 1.25	73.95 ± 1.42	**74.74 ± 1.61**
Yale	ACC	40.19 ± 3.38	40.76 ± 2.41	43.97 ± 3.33	39.15 ± 1.66	37.85 ± 1.74	39.76 ± 2.08	45.06 ± 1.28	**47.10 ± 2.07**
	NMI	48.51 ± 2.65	48.50 ± 1.97	49.69 ± 2.36	45.28 ± 1.32	47.21 ± 1.34	47.81 ± 1.28	51.09 ± 1.72	**52.33 ± 1.34**
COIL20	ACC	58.32 ± 5.40	49.86 ± 3.66	59.22 ± 3.14	60.68 ± 3.68	58.87 ± 1.46	60.56 ± 5.55	61.16 ± 3.07	**62.78 ± 2.92**
	NMI	74.01 ± 2.97	64.80 ± 1.69	70.89 ± 1.35	72.76 ± 1.19	66.07 ± 1.83	72.78 ± 2.02	74.27 ± 2.12	**75.69 ± 2.03**
USPS	ACC	66.34 ± 2.14	62.52 ± 2.03	72.02 ± 2.89	72.49 ± 4.45	67.67 ± 3.46	73.52 ± 4.81	74.73 ± 5.25	**76.12 ± 4.88**
	NMI	61.26 ± 1.62	59.50 ± 1.60	66.23 ± 0.44	66.16 ± 1.47	61.32 ± 1.82	67.03 ± 2.02	67.21 ± 2.02	**68.04 ± 1.51**

The best results are highlighted in bold.

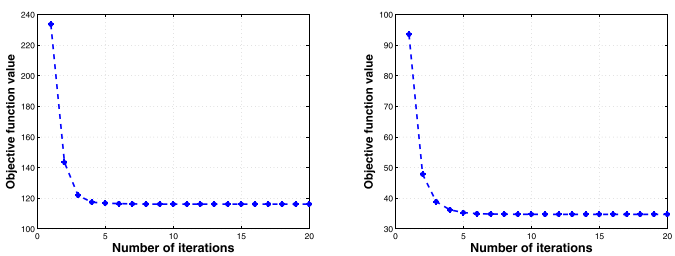
TABLE 2
Running Time (in Seconds) of Different Methods on Different Datasets

Datasets	RPCA	LPP	NPE	SPP	LRR	LRPP	LRE	FSP
ORL	1.321	1.525	3.914	5.482	22.774	53.984	59.377	85.522
COIL20	4.393	4.801	12.237	16.447	72.016	163.401	184.791	265.088
USPS	0.696	0.865	2.249	2.824	13.571	29.759	30.889	47.672
CMU PIE	3.184	3.240	9.769	11.889	52.675	116.114	133.813	175.177
FERET	1.412	1.725	4.979	7.298	25.221	64.286	79.072	113.842
Yale	0.402	0.493	1.237	1.634	7.107	15.647	18.741	25.735
AR	3.793	3.962	10.378	13.346	62.514	141.894	147.958	222.538
MNIST	34.277	38.102	86.680	118.382	520.904	1271.490	1311.573	2114.880

its convergence. The convergence curves of the objective value on ORL and USPS datasets are shown in Fig. 8. It can be seen that the proposed algorithm converges very fast and almost within 5 iterations.

5.7 Running Time Comparison

In order to evaluate the computational complexity of our method more intuitively, we compare the running time of FSP with other methods on the eight benchmark datasets used in our experiments. All of the algorithms are tested on a work station with 12 processors (1.70 GHz for each) and 32.0 GB RAM memory by MATLAB implementations with MATLAB R2016a. By considering that the running process may be affected by other possible application activities, for each algorithm with certain fixed parameters and each dataset, we run it 10 times and the averaged running time are reported in Table 2. As can be seen, although our proposed FSP is not the most efficient, its computational complexity is at the same level as LRPP and LRE.



(a) (b)

Fig. 8. The convergence curves of the proposed algorithm on (a) ORL database and (b) USPS database.

6 Conclusion

In this work, we propose an unsupervised linear feature selective projection method, named FSP for image feature extraction and classification by integrating feature selection and feature extraction into a unified framework. The low-rank and dual Laplacian regularization terms are embedded into the model for robustness to noises and preservation of the intrinsic local manifold geometrical structure of data, respectively. An $l_{2,1}$-norm regularization is imposed on the projection matrix to make it capable of selecting important features for composing the low dimensional subspace. Extensive experiments are conducted on five well-known databases to demonstrate the excellent performance of FSP against other state-of-the-art projection methods in robust image classification with various kinds of noises.

Acknowledgments

This work was supported in part by the National Science Foundation of China under Grant (NSFC 61701451 and 61773392), and in part by the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) under Grant CUG170654 and the Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing under Grant KLIGIP-2017B04.

References

[1] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang, "The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data," *Nature Rev. Cancer*, vol. 8, no. 1, pp. 37–49, 2008.

[2] M. Xu, H. Chen, and P. K. Varshney, "Dimensionality reduction for registration of high-dimensional data sets," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3041–3049, Aug. 2013.

[3] T. Zhou and D. Tao, "Double shrinking sparse dimension reduction," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 244–257, Jan. 2013.

[4] X. Jiang, J. Gao, T. Wang, and D. Shi, "TPSLVM: A dimensionality reduction algorithm based on thin plate splines," *IEEE Trans. Cybern.*, vol. 44, no. 10, pp. 1795–1807, Oct. 2014.

[5] Q. Cheng, H. Zhou, J. Cheng, and H. Li, "A minimax framework for classification with applications to images and high dimensional data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2117–2130, Nov. 2014.

[6] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2227–2240, Nov. 2014.

[7] Y. Chen, F. Li, J. Chen, B. Du, K. K. R. Choo, and H. Hassan, "EPLS: A novel feature extraction method for migration data clustering," *J. Parallel Distrib. Comput.*, vol. 103, pp. 96–103, 2016.

[8] C. Zhang, H. Fu, Q. Hu, P. Zhu, and X. Cao, "Flexible multi-view dimensionality co-reduction," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 648–659, Feb. 2017.

[9] M. Rahmani and G. K. Atia, "High dimensional low rank plus sparse matrix decomposition," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 2004–2019, Apr. 2017.

[10] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci, "Visualizing high-dimensional data: Advances in the past decade," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 3, pp. 1249–1268, Mar. 2017.

[11] Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang, and X. Huang, "Robust subspace clustering for multi-view data by exploiting correlation consensus," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3939–3949, Nov. 2015.

[12] Y. Wang, W. Zhang, L. Wu, X. Lin, M. Fang, and S. Pan, "Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 2153–2159.

[13] H. Yan, J. Yang, and J. Yang, "Robust joint feature weights learning framework," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 5, pp. 1327–1339, May 2016.

[14] K. Allab, L. Labiod, and M. Nadif, "A semi-NMF-PCA unified framework for data clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 2–16, Jan. 2017.

[15] D. Kaur, G. S. Aujla, N. Kumar, A. Zomaya, C. Perera, and R. Ranjan, "Tensor-based big data management scheme for dimensionality reduction problem in smart grid systems: SDN perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 10, pp. 1985–1998, Oct. 2018.

[16] Y. Wang and L. Wu, "Beyond low-rank representations: Orthogonal clustering basis reconstruction with optimized graph structure for multi-view spectral clustering," *Neural Netw.*, vol. 103, pp. 1–8, 2018.

[17] N. Zhao, L. Zhang, B. Du, Q. Zhang, J. You, and D. Tao, "Robust dual clustering with adaptive manifold regularization," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 11, pp. 2498–2509, Nov. 2017.

[18] P. Rathore, D. Kumar, J. C. Bezdek, S. Rajasegarar, and M. S. Palaniswami, "A rapid hybrid clustering algorithm for large volumes of high dimensional data," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 4, pp. 641–654, Apr. 2019.

[19] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[20] E. Keogh and A. Mueen, "Curse of dimensionality," in *Encyclopedia of Machine Learning*. Berlin, Germany: Springer, 2011, pp. 257–258.

[21] F. Nie, D. Xu, I. W.-H. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1921–1932, Jul. 2010.

[22] X. Zhu, Z. Huang, H. T. Shen, J. Cheng, and C. Xu, "Dimensionality reduction by mixed kernel canonical correlation analysis," *Pattern Recognit.*, vol. 45, no. 8, pp. 3003–3016, 2012.

[23] X. Zhu, Z. Huang, Y. Yang, H. T. Shen, C. Xu, and J. Luo, "Self-taught dimensionality reduction on the high-dimensional small-sized data," *Pattern Recognit.*, vol. 46, no. 1, pp. 215–229, 2013.

[24] X. Zhao, F. Nie, S. Wang, J. Guo, P. Xu, and X. Chen, "Unsupervised 2d dimensionality reduction with adaptive structure learning," *Neural Comput.*, vol. 29, no. 5, pp. 1352–1374, 2017.

[25] W. K. Wong, Z. Lai, J. Wen, X. Fang, and Y. Lu, "Low-rank embedding for robust image feature extraction," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2905–2917, Jun. 2017.

[26] R. Wang, F. Nie, R. Hong, X. Chang, X. Yang, and W. Yu, "Fast and orthogonal locality preserving projections for dimensionality reduction," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 5019–5030, Oct. 2017.

[27] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen, "Tensor canonical correlation analysis for multi-view dimension reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 11, pp. 3111–3124, Nov. 2015.

[28] B. Chen, H. Zhang, X. Zhang, W. Wen, H. Liu, and J. Liu, "Max-margin discriminant projection via data augmentation," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 7, pp. 1964–1976, Jul. 2015.

[29] S. Kanakam and C. A. Murthy, "Bridging feature selection and extraction: Compound feature generation," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 4, pp. 757–770, Apr. 2017.

[30] M. Banerjee and N. R. Pal, "Unsupervised feature selection with controlled redundancy (UFeSCoR)," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 12, pp. 3390–3403, Dec. 2015.

[31] P. Mitra, C. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 301–312, Mar. 2002.

[32] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2005, pp. 507–514.

[33] Z. Li and J. Tang, "Unsupervised feature selection via nonnegative spectral analysis and redundancy control," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5343–5355, Dec. 2015.

[34] F. Nie, Z. Wei, and X. Li, "Unsupervised feature selection with structured graph optimization," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 1302–1308.

[35] Y. Yuan, X. Zheng, and X. Lu, "Discovering diverse subset for unsupervised hyperspectral band selection," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 51–64, Jan. 2017.

[36] C. Tang, X. Zhu, J. Chen, P. Wang, and X. Liu, "Robust graph regularized unsupervised feature selection," *Expert Syst. Appl.*, vol. 96, pp. 64–76, 2018.

[37] C. Tang, X. Liu, M. Li, P. Wang, J. Chen, L. Wang, and W. Li, "Robust unsupervised feature selection via dual self-representation and manifold regularization," *Knowl.-Based Syst.*, vol. 145, pp. 109–120, 2018.

[38] C. Hou, F. Nie, H. Tao, and D. Yi, "Multi-view unsupervised feature selection with adaptive similarity and view weight," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 9, pp. 1998–2011, Sep. 2017.

[39] W. Liao, A. Pizurica, P. Scheunders, W. Philips, and Y. Pi, "Semisupervised local discriminant analysis for feature extraction in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 184–198, Jan. 2013.

[40] K. Benabdeslem and M. Hindawi, "Efficient semi-supervised feature selection: Constraint, relevance, and redundancy," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1131–1143, May 2014.

[41] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 1151–1157.

[42] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $l_{2,1}$-norms minimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.

[43] I. T. Jolliffe, "Principal component analysis and factor analysis," in *Principal Component Analysis*. Berlin, Germany: Springer, 1986, pp. 115–128.

[44] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces versus Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[45] F. Song, D. Zhang, D. Mei, and Z. Guo, "A multiple maximum scatter difference discriminant criterion for facial feature extraction," *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)*, vol. 37, no. 6, pp. 1599–1606, Dec. 2007.

[46] Q. Liu, H. Lu, and S. Ma, "Improving kernel fisher discriminant analysis for face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 42–49, Jan. 2004.

[47] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *J. Amer. Statistical Assoc.*, vol. 97, no. 458, pp. 611–631, 2002.

[48] W. Zuo, D. Zhang, J. Yang, and K. Wang, "BDPCA plus LDA: A novel fast feature extraction technique for face recognition," *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)*, vol. 36, no. 4, pp. 946–953, Aug. 2006.

[49] Y. Pang, S. Wang, and Y. Yuan, "Learning regularized LDA by clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2191–2201, Dec. 2014.

[50] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Sci.*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[51] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2002, pp. 585–591.

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

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

[54] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[55] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2004, pp. 153–160.

[56] Y. Pang, N. Yu, H. Li, R. Zhang, and Z. Liu, "Face recognition using neighborhood preserving projections," in *Proc. Pacific-Rim Conf. Multimedia*, 2005, pp. 854–864.

[57] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. Int. Conf. Comput. Vis.*, 2005, pp. 1208–1213.

[58] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognit.*, vol. 43, no. 1, pp. 331–341, 2010.

[59] D. Cai, X. He, J. Han, et al., "Isometric projection," in *Proc. AAAI Conf. Artif. Intell.*, 2007, pp. 528–533.

[60] Z. Ji, Y. Pang, Y. He, and H. Zhang, "Semi-supervised LPP algorithms for learning-to-rank-based visual search reranking," *Inf. Sci.*, vol. 302, pp. 83–93, 2015.

[61] D. Cai, X. He, J. Han, and H.-J. Zhang, "Orthogonal laplacianfaces for face recognition," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3608–3614, Nov. 2006.

[62] E. Kokiopoulou and Y. Saad, "Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2143–2156, Dec. 2007.

[63] F. Nie, S. Xiang, Y. Song, and C. Zhang, "Orthogonal locality minimizing globality maximizing projections for feature extraction," *Opt. Eng.*, vol. 48, no. 1, pp. 017 202–017 202, 2009.

[64] X. Fang, Y. Xu, X. Li, Z. Lai, S. Teng, and L. Fei, "Orthogonal self-guided similarity preserving projection for classification and clustering," *Neural Netw.*, vol. 88, pp. 1–8, 2017.

[65] X. Liu, J. Yin, Z. Feng, J. Dong, and L. Wang, "Orthogonal neighborhood preserving embedding for face recognition," in *Proc. IEEE Int. Conf. Image Process.*, 2007, pp. I–133.

[66] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 663–670.

[67] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[68] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2009, pp. 2080–2088.

[69] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu, "Non-negative low rank and sparse graph for semi-supervised learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2328–2335.

[70] J. Liu, Y. Chen, J. Zhang, and Z. Xu, "Enhancing low-rank subspace clustering by manifold regularization," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4022–4030, Sep. 2014.

[71] M. Yin, J. Gao, and Z. Lin, "Laplacian regularized low-rank representation and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 504–517, Mar. 2016.

[72] K. Tang, R. Liu, Z. Su, and J. Zhang, "Structure-constrained low-rank representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2167–2179, Dec. 2014.

[73] K. Tang, D. B. Dunson, Z. Su, R. Liu, J. Zhang, and J. Dong, "Subspace segmentation by dense block and sparse representation," *Neural Netw.*, vol. 75, pp. 66–76, 2016.

[74] C. You, D. Robinson, and R. Vidal, "Scalable sparse subspace clustering by orthogonal matching pursuit," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3918–3927.

[75] X. Peng, Z. Yi, and H. Tang, "Robust subspace clustering via thresholding ridge regression," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 3827–3833.

[76] X. Peng, Z. Yu, Z. Yi, and H. Tang, "Constructing the L2-graph for robust subspace learning and subspace clustering," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1053–1066, Apr. 2017.

[77] B.-K. Bao, G. Liu, C. Xu, and S. Yan, "Inductive robust principal component analysis," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3794–3800, Aug. 2012.

[78] Y. Lu, Z. Lai, Y. Xu, X. Li, D. Zhang, and C. Yuan, "Low-rank preserving projections," *IEEE Trans. Cybern.*, vol. 46, no. 8, pp. 1900–1913, Aug. 2016.

[79] J. Wen, N. Han, X. Fang, L. Fei, K. Yan, and S. Zhan, "Low-rank preserving projection via graph regularized reconstruction," *IEEE Trans. Cybern.*, vol. 49, no. 4, pp. 1279–1291, Apr. 2019.

[80] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, 2011, Art. no. 11.

[81] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 5, pp. 971–989, Sep./Oct. 2016.

[82] X. Zhu, Y. Zhu, S. Zhang, R. Hu, and W. He, "Adaptive hypergraph learning for unsupervised feature selection," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 3581–3587.

[83] S. Yan, D. Xu, B. Zhang, H. J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extension: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[84] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 612–620.

[85] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optimization*, vol. 20, no. 4, pp. 1956–1982, 2008.

[86] K. Lange, D. R. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *J. Comput. Graphical Statist.*, vol. 9, no. 1, pp. 1–20, 2000.

[87] R. H. Bartels and G. W. Stewart, "Solution of the matrix equation AX + XB = C," *Commun. ACM*, vol. 15, no. 9, pp. 820–826, 1972.

[88] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. 2nd IEEE Workshop Appl. Comput. Vis.*, 1994, pp. 138–142.

[89] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.

[90] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-20)," Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-005–96, 1996.

[91] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.

[92] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image Vis. Comput.*, vol. 16, no. 5, pp. 295–306, 1998.

[93] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[94] A. Martnez and R. Benavente, "The AR face database," Centre de Visio per Computador, Univ. Auton. Barcelona, Barcelona, Spain, Tech. Rep. 24, Jun. 1998.

[95] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[96] K. Zhan, X. Chang, J. Guan, C. Ling, Z. Ma, and Y. Yi, "Adaptive structure discovery for multimedia analysis using multiple features," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1826–1834, May 2019.

[97] C. Tang, X. Liu, M. Li, P. Wang, J. Chen, L. Wang, and W. Li, "Robust unsupervised feature selection via dual self-representation and manifold regularization," *Knowl.-Based Syst.*, vol. 145, pp. 109–120, 2018.

[98] T. Chang, C. Jiajia, L. Xinwang, L. Miaomiao, W. Pichao, W. Minhui, and L. Peng, "Consensus learning guided multi-view unsupervised feature selection," *Knowl.-Based Syst.*, vol. 160, pp. 49–60, 2018.

[99] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2010, pp. 333–342.
[100] P. Zhu, W. Zuo, L. Zhang, Q. Hu, and S. C. K. Shiu, "Unsupervised feature selection by regularized self-representation," *Pattern Recognit.*, vol. 48, no. 2, pp. 438–446, 2015.
[101] N. Zhou, Y. Xu, H. Cheng, J. Fang, and W. Pedrycz, "Global and local structure preserving sparse subspace learning: An iterative approach to unsupervised feature selection," *Pattern Recognit.*, vol. 53, pp. 87–101, 2016.
[102] W. He, X. Zhu, D. Cheng, R. Hu, and S. Zhang, "Unsupervised feature selection for visual classification via feature-representation property," *Neurocomput.*, vol. 236, pp. 5–13, 2017.

Chang Tang (M'16) received the PhD degree from Tianjin University, Tianjin, China, in 2016. He joined the AMRL Lab, University of Wollongong, between Sep. 2014 and Sep. 2015. He is now an associate professor with the School of Computer Science, China University of Geosciences, Wuhan, China. He has published more than 20 peer-reviewed papers, including those in highly regarded journals and conferences such as the *IEEE Transactions on Human-Machine Systems*, the *IEEE Signal Processing Letters*, ICCV, CVPR, ACMM, etc. He served on the technical program committees of IJCAI 2018, ICME 2018, AAAI 2019, ICME 2019, IJCAI 2019, and CVPR 2019. His current research interests include machine learning and data mining. He is a member of the IEEE.

Xinwang Liu received the PhD degree from the National University of Defense Technology (NUDT), China. He is now an assistant researcher with the School of Computer, NUDT. His current research interests include kernel learning and unsupervised feature learning. He has published more than 60 peer-reviewed papers, including those in highly regarded journals and conferences such as the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *IEEE Transactions on Knowledge and Data Engineering*, the *IEEE Transactions on Image Processing* the *IEEE Transactions on Neural Networks and Learning Systems*, the *IEEE Transactions on Multimedia*, the *IEEE Transactions on Information Forensics and Security*, ICCV, CVPR, AAAI, IJCAI, etc. He is a member of the IEEE.

Xinzhong Zhu received the PhD degree from Xidian University, China. He is a professor with the College of Mathematics and Computer Science, Zhejiang Normal University, and also the president of the Research Institute of Ningbo Cixing Co. Ltd., PR, China. His research interests include machine learning, computer vision, manufacturing informatization, robotics and system integration, and intelligent manufacturing. He is a member of the ACM and certified as a CCF senior member. He has published more than 30 peer-reviewed papers, including those in highly regarded journals and conferences such as the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *IEEE Transactions on Multimedia*, the *IEEE Transactions on Knowledge and Data Engineering*, CVPR, AAAI, IJCAI, etc.

Jian Xiong (M'13) received the BS degree in engineering, and the MS and PhD degrees in management from the National University of Defense Technology, Changsha, China, in 2005, 2007, and 2012, respectively. He is an associate professor with the School of Business Administration, Southwestern University of Finance and Economics. His research interests include data mining, multi objective evolutionary optimization, multiobjective decision making, project planning, and scheduling. He is a member of the IEEE.

Miaomiao Li received the PhD degree from the National University of Defense Technology, China. She is now a lecturer at Changsha College, Changsha, China. Her current research interests include kernel learning and multi-view clustering. She has published several peer-reviewed papers in conferences and journals such as AAAI, IJCAI, the *Neurocomputing*, etc. She served on the technical program committees of IJCAI 2017 and 2018.

Jingyuan Xia received the BSc and MSc degrees from the National University of Defense Technology, Hunan, China. He is currently working toward the PhD degree in the Department of Electrical and Electronic Engineering, Imperial College London at London, United Kingdom. His current research interests include bi-linear convex optimization, low-rank matrix completion, sparse signal processing, and intelligent transportation estimation.

Xiangke Wang (M'07-SM'12) received the BS, MS, and PhD degrees in control science and engineering from the National University of Defense Technology, Changsha, China, in 2004, 2006, and 2012, respectively. Since 2014, he has been an associate professor with the College of Mechatronic Engineering and Automation, National University of Defense Technology. He has published a number of journal papers, and his current research interests include coordinated decision and control of multiple UAV systems, nonlinear control, and visual servo control. He is a senior member of the IEEE.

Lizhe Wang received the BE and ME degrees from Tsinghua University, Beijing, China, and the DrEng (magna cum laude) degree from the University of Karlsruhe, Karlsruhe, Germany. He is currently a ChuTian chair professor with the School of Computer Science, China University of Geosciences, Beijing, and also a professor with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing. His research interests include high-performance computing, eScience, and remote sensing image processing. He is a fellow of the IET and the British Computer Society. He serves as an associate editor of the *IEEE Transactions on Parallel and Distributed Systems*, the *IEEE Transactions on Cloud Computing*, and the *IEEE Transactions on Sustainable Computing*. He is a senior member of the IEEE.