Unsupervised Feature Selection via Orthogonal Basis Clustering and Local Structure Preserving

Xiaochang Lin, Jiewen Guan, Bilian Chen, and Yifeng Zeng

***Abstract*—Due to the "curse of dimensionality" issue, how to discard redundant features and select informative features in high-dimensional data has become a critical problem, and there are many researches dedicated to solving this problem. Unsupervised feature selection technique, which doesn't require any prior category information to conduct with, has gained a prominent place in pre-processing high-dimensional data among all feature selection techniques, and it has been applied to many neural networks and learning systems related applications, e.g., pattern classification. In this paper, we propose an efficient method for unsupervised feature selection via Orthogonal basis Clustering and reliable Local Structure Preserving, which is referred to OCLSP briefly. Our OCLSP method consists of an orthogonal basis clustering together with an adaptive graph regularization, which realize the functionality of simultaneously achieving excellent cluster separation and preserving the local information of data. Besides, we exploit an efficient alternative optimization algorithm to solve the challenging optimization problem of our proposed OCLSP method, and we perform a theoretical analysis of its computational complexity and convergence. Eventually, we conduct comprehensive experiments on nine real-world datasets to test the validity of our proposed OCLSP method, and the experimental results demonstrate that our proposed OCLSP method outperforms many state-of-the-art unsupervised feature selection methods in terms of clustering accuracy and normalized mutual information, which indicates that our proposed OCLSP method has a strong ability in identifying more important features.**

***Index Terms*—Unsupervised feature selection, orthogonal basis clustering, locality preserving.**

I. Introduction

WITH the rapid development of information technology, data are often represented by high-dimensional feature vectors in many fields, such as computer vision [1], computational biology [2] and pattern classification [3], etc. However, the high-dimensional data not only increase computational complexities and memory requirements of learning algorithms, but also deteriorate the performance of learning algorithms due to the irrelevant, redundant, and noisy features [4], [5]. To this end, dimensionality reduction algorithms are proposed to solve this problem by reducing the dimensionality of data, which could make the learned model more compact and generalized [6].

Generally, dimensionality reduction can be roughly divided into two categories: feature extraction [7]–[9] and feature selection [10]–[12]. Feature extraction aims to map the high-dimensional features into a new low-dimensional space. The new low-dimensional feature space is usually a linear or non-linear combination of the original features. Correspondingly, feature selection seeks to select the optimal feature subset from the original feature set using some certain criteria. Although feature extraction methods have been demonstrated to have promising performance, they transform and compress the original features, which not only distorts the original data but also impairs the efficiency of processing [5]. On the contrast, feature selection methods have better interpretability because they retain the semantic meanings of original features. Moreover, the cost of collecting features for learning algorithms can be reduced through feature selection, because we only need to collect those features that are selected according to the feature selection methods rather than to utilize all the original features for projection as what feature extraction methods do [13].

Based on the availability of label information, feature selection methods can be further divided into three categories of supervised methods, semi-supervised methods and unsupervised methods [14]. When there are adequate labeled data, supervised approaches, which leverage label information to guide the feature selection process, are the first choice due to their high classification accuracy and reliability. However, labeled data are uncommon since it would cost a great deal of human resources to label data manually. Besides, labeled data may be polluted intentionally since the owners are not willing to share them. Hence, on the other hand, when there are not sufficient labeled data for us, semi-supervised approaches and unsupervised approaches are necessary [15].

Since discrimination information is often encoded in class labels, it is relatively easier for the supervised feature selection methods to find the discriminative features using label information. However, as illustrated above, the large scale data obtained in real life are usually unlabeled. Hence, researches in unsupervised feature selection have significant practical meanings. In this paper, we focus on the unsupervised feature selection problem which is a more challenging problem due to the lack of label information that could help select discriminative features. There are several strategies proposed to solve the unsupervised feature selection problem. Since the label information is absent in the unsupervised feature selection problem, the most typical strategy is to assign pseudo labels to data samples so as to transform an unsupervised

The work was supported in part by Youth Innovation Fund of Xiamen (Grant no. 3502Z20206049) and National Natural Science Foundation of China (Grant Nos. 61772442 and 61836005). *(Corresponding author: Bilian Chen and Yifeng Zeng.)*

Xiaochang Lin, Jiewen Guan and Bilian Chen are with the Department of Automation, Xiamen University, Xiamen 361005, China, and also with the Xiamen Key Laboratory of Big Data Intelligent Analysis and Decision-making, Xiamen 361005, China (e-mail: 13263396762@163.com; 23220191151283@stu.xmu.edu.cn; blchen@xmu.edu.cn).

Yizeng Zeng is with the Department of Computer and Information Science, Northumbria University, Newcastle-upon-Tyne NE1 8ST, U.K. (e-mail: yifeng.zeng@northumbria.ac.uk).

feature selection problem to a supervised counterpart, and then to solve the transformed supervised feature selection problem accordingly. One of the methods that unsupervised feature selection methods have adopted to generate pseudo labels is to utilize the intrinsic structure of data. Lately, feature selection methods in recent work are mainly based on matrix factorization technique which generates data pseudo labels together with a cluster center indicator matrix by learning a set of bases, and those matrix factorization based methods could obtain excellent performance in estimating the latent subspace of data.

In this paper, we propose a novel unsupervised feature selection method, namely unsupervised feature selection via Orthogonal basis Clustering and Local Structure Preserving (OCLSP), which achieves the functionality of simultaneously selecting discriminative features and performing orthogonal basis clustering while preserving the local structure of data points. Specifically, we decompose the target matrix into two matrices, which are regarded as the latent cluster center indicator and the sparse representations of different classes, respectively. In addition, an orthogonal constraint is imposed on the latent cluster center indicator matrix in order to ensure that the estimated centers are close to the ground-truths and to keep the sparse representations of different classes as far as possible. Besides, the orthogonal constraint would also make the feature selection matrix a better projection matrix, which is favorable for selecting more discriminative features. Meanwhile, in the process of iterative optimization, the local structure of data points, which is proven to be effective and discriminative [16]–[18], could be adaptively learned from the results of feature selection, and then the learned local structure could be used to reselect informative features to preserve such a structure. Fig. 1 illustrates the framework of our proposed OCLSP method.

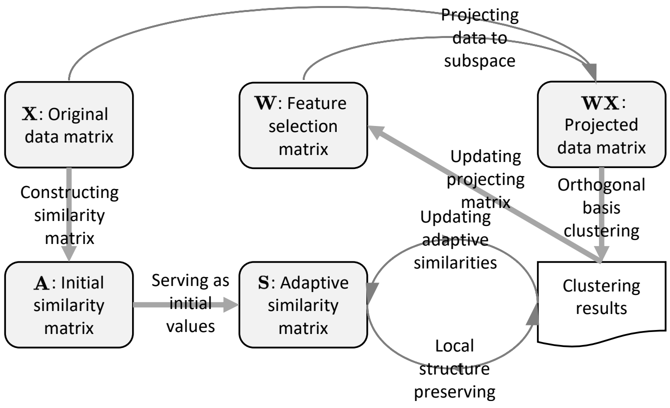

Projecting data to subspace
$\mathbf{X}$: Original data matrix
$\mathbf{W}$: Feature selection matrix
$\mathbf{WX}$: Projected data matrix
Constructing similarity matrix
Updating projecting matrix
Orthogonal basis clustering
Updating adaptive similarities
$\mathbf{A}$: Initial similarity matrix
Serving as initial values
$\mathbf{S}$: Adaptive similarity matrix
Clustering results
Local structure preserving

Fig. 1. Framework of proposed OCLSP method.

In summary, the contributions of this paper are highlighted as follows:

1) We combine orthogonal basis clustering with adaptive structure learning in the low-dimensional manifold, so that the learned manifold projection matrix whose quality is enhanced by orthogonal basis clustering can guide the process of local geometrical structure learning in the subspace, while the learned local geometrical structure in the subspace can further guide the process of manifold learning (feature selection) and orthogonal basis clustering.
2) We develop a simple but effective iterative updating algorithm to solve the optimization problem of our proposed OCLSP method.
3) Experimental results on nine real-world datasets demonstrate that our proposed OCLSP method could outperform the state-of-the-art unsupervised feature selection methods in most cases.

The remainder of this paper is arranged as follows. In Section II, the related work on the topic of unsupervised feature selection is reviewed. In Section III, we propose our OCLSP method. In Section V, we provide an effective but simple algorithm for solving the optimization problem of our proposed OCLSP method. In Section VI, we show and analyze results of comparative experiments on nine real-world datasets. In Section VII, overall conclusion is stated, and we provide possible future work.

II. Related Work

For the purpose of handling high-dimensional data, many unsupervised feature selection methods are proposed. Due to the absence of label information, unsupervised feature selection methods have to resort to select features that could preserve the intrinsic structure of data well without leveraging any label information. The earliest methods usually define a evaluation score and then rank features based on that. The most representative methods involving such a procedure are maximum variance method [19] and Laplacian score method [20]. However, the biggest deficiency of those methods is that the potential interaction between features is neglected. For the sake of addressing this issue, many methods proposed recently select features by simultaneously exploiting the intrinsic structure information of data and considering the correlation between features. The graph is the most popular data structure to represent such intrinsic structure information of data, and the graph based feature selection methods can be roughly categorized into two classes: one is to use a pre-defined graph, while the other is to learn an adaptive graph.

The pre-defined graph based unsupervised feature selection methods usually use some certain criteria to construct a predefined graph and then select features that could preserve such a graph structure well. For example, Unsupervised Discriminative Feature Selection (UDFS) [21] selects the most informative features by capturing the manifold structure, Nonnegative Discriminative Feature Selection (NDFS) [22] aims to select features and analyze non-negative spectrum at the same time, and Robust Unsupervised Feature Selection (RUFS) [23] performs robust clustering and robust feature selection simultaneously. However, the main drawback of these methods is that the graph is constructed in the original feature space, where large quantities of noises and redundant features exist, and this makes the pre-defined graph unreliable and eventually impairs the effectiveness of the selected features.

Distinct from the first class where the local geometrical structure remains unchanged throughout the learning process, the methods from the second class learn an adaptive graph in the procedure of feature selection, i.e., the structure of graph changes with the selected features during the iterative optimization process. For example, Du *et al.* [24] learned an adaptive graph for feature selection by preserving the global and local structures of data, Nie *et al.* [25] determined the similarity matrix from the results of feature selection adaptively, and Li *et al.* [26] proposed the uncorrelated regression model which performs feature selection and manifold learning simultaneously. Luo *et al.* [27] learned the optimal reconstruction graph and selective matrix simultaneously, instead of using a predetermined graph, which is very close to our work. However, different from [27], in our OCLSP method, the projected data points are further decomposed by orthogonal basis clustering, which improves the quality of the projection matrix (which is also the feature selection matrix). Besides, we generate a prior similarity graph to avoid unreasonable local geometrical structures.

One key issue in unsupervised feature selection is how to generate accurate class labels for data samples, i.e., how to cluster samples. Clustering methods applied in unsupervised feature selection are mainly variants of two prototypes: k-means clustering [28] and spectral clustering [29]. For k-means based methods, the most widely used techniques are the Non-negative Matrix Factorization (NMF) and its variants [30]–[33]. For example, Zhang *et al.* [33] adopts NMF and symmetric NMF to deal with constrained clustering problems. For spectral clustering based methods, they preserve the local geometrical structure by different ways [34]–[36]. For example, Wang *et al.* [36] proposed to use structured low-rank representations to capture local manifold structure of multi-view data. However, there are two main differences of our method from these work, because (1) we force clustering center matrix and clustering membership matrix to be both orthogonal, so as to get more precise pseudo clustering labels, and (2) we factorize the projected data samples $\mathbf{WX}$, i.e., cluster the data samples in the subspace, to get rid of not only redundancy but noises which lie in the original feature space.

Recently, matrix factorization technique has attracted more and more attention from machine learning and pattern recognition researches, and some matrix factorization based feature selection approaches are henceforth proposed and they could obtain excellent performance. Wang *et al.* [37] treated the process of feature selection as matrix factorization by introducing a subspace distance, and proposed an iterative updating algorithm which is based on non-negative matrix factorization [38] and concept factorization [39]. Han *et al.* [40] introduced Simultaneous Orthogonal basis Clustering Feature Selection (SOCFS) by decomposing the target matrix into two orthogonal matrices. However, these methods neglect the intrinsic structure of data, which is unfavorable in unsupervised feature selection.

Different from the previous work, our proposed OCLSP selects informative features by aggregating feature selection, matrix factorization and adaptive graph learning into a unified framework.

III. Preliminary

A. Notations

Throughout this paper, boldface capital letters represent matrices, whereas boldface lower case letters represent vectors, and italic lower case letters represent scalar values. For an arbitrary matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, m_{ij} denotes its (i,j)th entry, while $\mathbf{m}^i$ and $\mathbf{m}_j$ represent the ith row and jth column of $\mathbf{M}$, respectively. Besides, the $\ell_{2,1}$ norm of matrix $\mathbf{M}$ is defined as $\|\mathbf{M}\|_{2,1} = \sum_{i=1}^{m}\sqrt{\sum_{j=1}^{n} m_{ij}^2}$, and the Frobenius norm of matrix $\mathbf{M}$ is defined as $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} m_{ij}^2} = \sqrt{\mathrm{Tr}\left(\mathbf{M}\mathbf{M}^T\right)}$. $\mathrm{Tr}(\mathbf{M})$ denotes the trace of matrix $\mathbf{M}$ if $\mathbf{M}$ is square, and $\mathbf{M}^T$ denotes the transpose of $\mathbf{M}$.

B. Simultaneous Orthogonal Basis Clustering Feature Selection (SOCFS)

Assume that $\mathbf{X} \in \mathbb{R}^{m\times n}$ represents the data matrix, where m and n represent the number of features and the number of samples, respectively. Each row of $\mathbf{X}$ represents one feature dimension and each column of $\mathbf{X}$ represents a sample. Given a target matrix $\mathbf{T} \in \mathbb{R}^{d\times n}$, according to the methods proposed in [22] and [23], the unsupervised feature selection problem can be formulated as a multi-output regression problem

$$\min_{\mathbf{W}} \quad \mathcal{L}(\mathbf{W}^T\mathbf{X} - \mathbf{T}) + \eta\mathcal{R}(\mathbf{W}),$$

where $\mathbf{W} \in \mathbb{R}^{m\times d}$ is the feature weight matrix, $\mathcal{L}(\mathbf{W}^T\mathbf{X}-\mathbf{T})$ is the loss term, $\mathcal{R}(\mathbf{W})$ is the regularization term imposed on feature weight matrix $\mathbf{W}$, and η is a positive regularization parameter to control the sparsity of feature selection matrix $\mathbf{W}$. In this framework, the matrix factorization part has a role to cluster samples, while the regularization part is responsible for selecting features, and these two objectives are conducted simultaneously. It is crucial for the target matrix $\mathbf{T}$ to have the capability to discriminate projected clusters, hence we allow $\mathbf{T}$ to have extra degrees of freedom [40], by decomposing it into two other matrices $\mathbf{B} \in \mathbb{R}^{d\times c}$ and $\mathbf{E} \in \mathbb{R}^{n\times c}$ as $\mathbf{T} = \mathbf{B}\mathbf{E}^T$ with additional constraints in the following

$$\begin{aligned} \min_{\mathbf{W},\mathbf{B},\mathbf{E}} \quad & \left\|\mathbf{W}^T\mathbf{X} - \mathbf{B}\mathbf{E}^T\right\|_F^2 + \eta\|\mathbf{W}\|_{2,1} \\ \text{s.t.} \quad & \mathbf{B}^T\mathbf{B} = \mathbf{I}, \mathbf{E}^T\mathbf{E} = \mathbf{I}, \mathbf{E} \geq 0. \end{aligned}$$

The orthogonal constraint exerted on matrix $\mathbf{B}$ guarantees that each column of $\mathbf{B}$ is independent. That is, $\mathbf{B}$ is composed by the orthogonal bases of the projected sample space $\mathbf{W}^T\mathbf{X}$. Besides, the columns of $\mathbf{B}$ can be regarded as the directions of the corresponding cluster centers. On the other hand, $\mathbf{E}$ is the cluster indicator matrix, which shows the membership degrees of different samples belonging to different clusters. In addition, the non-negative and the orthogonal constraints exerted on $\mathbf{E}$ make each row of $\mathbf{E}$ has only one non-zero element [41]. Therefore, $\mathbf{T} = \mathbf{B}\mathbf{E}^T$ can be utilized to find latent cluster centers so as to achieve an excellent cluster separation [40].

C. Graph Regularization

Another key point in unsupervised feature selection problems is how to preserve the local geometrical structure of data reliably. Graph Laplacian is a typical technique that is widely employed to preserve the structure [42], [43]. A natural assumption could be that, if two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are close in the intrinsic graph of data, then the projected data points $\mathbf{W}^T\mathbf{x}_i$ and $\mathbf{W}^T\mathbf{x}_j$ should also be close enough in the projection subspace. This assumption could be obtained by defining the following embedding function

$$\min_{\mathbf{W}} \quad \frac{1}{2}\left\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\right\|_2^2 s_{ij}, \tag{1}$$

where the similarity s_{ij} between two samples i and j is usually calculated by a Gaussian kernel function defined as follows

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

where $\mathcal{N}_k\left(\mathbf{x}_i\right)$ denotes the set of k nearest neighbors of $\mathbf{x}_i$ and σ is the Gaussian kernel width.

After simple mathematical transformation, (1) becomes

$$\min_{\mathbf{W}} \quad \mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}),$$

where $\mathbf{L} = \mathbf{P} - (\mathbf{S}^T + \mathbf{S})/2 \in \mathbb{R}^{n\times n}$ is the graph Laplacian which is based on the similarity matrix $\mathbf{S} = [s_{ij}] \in \mathbb{R}^{n\times n}$, and the degree matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ is a diagonal matrix defined as below

$$\mathbf{P} = \mathrm{diag}\left(\sum_{j=1}^{n}\frac{s_{1j}+s_{j1}}{2}, \sum_{j=1}^{n}\frac{s_{2j}+s_{j2}}{2}, \ldots, \sum_{j=1}^{n}\frac{s_{nj}+s_{jn}}{2}\right).$$

IV. Orthogonal Basis Clustering and Reliable Local Structure Preserving Method

Although the aforementioned schemes could attain good performance in many scenarios, there are still flaws. For example, as for the SOCFS model, it neglects the local structure information of data, which is proven to be important in many literature [42], [43]. Besides, a variety of models, like [21]–[23], leverage such local structure information of data by constructing a pre-defined similarity graph, where similarities are computed using samples in the original sample space. However, samples in the original sample space contain a big deal of redundancy and noises, which leads to inaccurate similarity values, and eventually impairs the efficiency of feature selection. For the sake of mitigating the above problems, in this paper, we propose an adaptive similarity graph model which learns similarity information between samples in a cleaner subspace adaptively so as to capture the local structure information of data more precisely. Besides, we propose an unified framework which performs local structure information learning and orthogonal basis clustering simultaneously, where local structure information is learned adaptively from the results of feature selection, which are obtained by the results of orthogonal basis clustering, and the most informative features are then reselected to preserve the learned structure. Our OCLSP method not only inherits merits of the SOCFS model, but it learns local structure information of data adaptively, which leads to better results of selected features.

It is noted that in the existing graph regularization based methods, the stage of constructing graph and the stage of learning feature weight matrix are independent, i.e., the similarity based graph is first constructed by leveraging Gaussian kernel function, and then the graph is used for preserving local geometrical structure of data by optimization problem (1). Obviously, the quality of the constructed graph would be affected by the Gaussian kernel width σ and the noisy features in $\mathbf{X}$, and the unreliable constructed graph would further lead to a suboptimal result, which is unfavorable. Therefore, we present an unsupervised feature selection method via orthogonal basis clustering and local structure preserving. In our method, the similarity based graph and feature selection matrix can be mutually restricted and they could be jointly optimized. In this way, the quality of selected features and the reliability of learned local structure information can both be improved. The optimization problem of our OCLSP method is formulated as follows

$$\begin{aligned} \min_{\mathbf{S},\mathbf{B},\mathbf{E},\mathbf{W}} \quad & \left\|\mathbf{W}^T\mathbf{X} - \mathbf{B}\mathbf{E}^T\right\|_F^2 + \eta\left\|\mathbf{W}\right\|_{2,1} \\ & + \gamma\left(\mathrm{Tr}\left(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}\right) + \beta\|\mathbf{S} - \mathbf{A}\|_F^2\right) \\ \text{s.t.} \quad & \sum_j s_{ij} = 1, 0 \le s_{ij} \le 1, \text{for all } i, \\ & \mathbf{E}^T\mathbf{E} = \mathbf{I}, \mathbf{E} \ge 0, \mathbf{B}^T\mathbf{B} = \mathbf{I}, \end{aligned}$$

where γ is a positive coefficient to adjust the weight of structure learning, $\mathbf{A} \in \mathbb{R}^{n\times n}$ is the initial similarity matrix computed by (2), and β is a positive parameter used to control the degree that similarity matrix $\mathbf{S}$ changes from $\mathbf{A}$. In our scheme, $\|\mathbf{W}\|_{2,1}$ forces the rows of the feature selection matrix $\mathbf{W}$ to be sparse, thus, we can filter out features corresponding to small values in $\mathbf{W}$. There are two main differences of our method from previous adaptive structure learning schemes [16]–[18]. First, we generate a pre-defined graph $\mathbf{A}$ and constrain the learned local structure $\mathbf{S}$ to be close to $\mathbf{A}$ so as to avoid unreasonable local geometrical structures. Second, we perform adaptive structure learning and orthogonal basis clustering simultaneously based on a common projection matrix $\mathbf{W}$, so that the learned local geometrical structure contains the information from latent clusters produced by orthogonal basis clustering.

V. Optimization and Theoretical Analysis

A. Solution Method

Alternatively, we propose an equivalent formulation of the optimization problem aforementioned as follows

$$\begin{aligned}\min_{\mathbf{S},\mathbf{B},\mathbf{E},\mathbf{W},\mathbf{Z}} \quad & \left\|\mathbf{W}^T\mathbf{X}-\mathbf{B}\mathbf{E}^T\right\|_F^2+\eta\|\mathbf{W}\|_{2,1} \\ & +\gamma\left(\operatorname{Tr}\left(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}\right)+\beta\|\mathbf{S}-\mathbf{A}\|_F^2\right) \\ \text{s.t.} \quad & \sum_j s_{ij}=1, 0\le s_{ij}\le 1, \text{for all } i, \\ & \mathbf{E}^T\mathbf{E}=\mathbf{I}, \mathbf{Z}=\mathbf{E}, \mathbf{Z}\ge 0, \mathbf{B}^T\mathbf{B}=\mathbf{I},\end{aligned} \tag{3}$$

where $\mathbf{Z}\in\mathbb{R}^{n\times c}$ is an auxiliary matrix with an additional constraint of $\mathbf{Z}=\mathbf{E}$. This reformulation has the ability to detach the non-negative constraint from $\mathbf{E}$ and assign that constraint to a new matrix $\mathbf{Z}$. Through the additional constraint $\mathbf{Z}=\mathbf{E}$, $\mathbf{Z}$ has a role to bring non-negativity to $\mathbf{E}$ while $\mathbf{E}$ is responsible for keeping $\mathbf{Z}$ orthogonal. By rewriting (3), we present our final optimization problem as follows

$$\begin{aligned}\min_{\mathbf{S},\mathbf{B},\mathbf{E},\mathbf{W},\mathbf{Z}} \quad & \left\|\mathbf{W}^T\mathbf{X}-\mathbf{B}\mathbf{E}^T\right\|_F^2+\eta\|\mathbf{W}\|_{2,1}+\alpha\|\mathbf{Z}-\mathbf{E}\|_F^2 \\ & +\gamma\left(\operatorname{Tr}\left(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}\right)+\beta\|\mathbf{S}-\mathbf{A}\|_F^2\right) \\ \text{s.t.} \quad & \sum_j s_{ij}=1, 0\le s_{ij}\le 1, \text{for all } i, \\ & \mathbf{E}^T\mathbf{E}=\mathbf{I}, \mathbf{Z}\ge 0, \mathbf{B}^T\mathbf{B}=\mathbf{I},\end{aligned} \tag{4}$$

where $\alpha>0$ controls the degree of affinity between $\mathbf{Z}$ and $\mathbf{E}$.

In this manner, the hybridity between $\mathbf{Z}$ and $\mathbf{E}$ is removed, and we could henceforth optimize our objective by Alternating Direction Method of Multipliers (ADMM) as follows.

- **Update B**

When $\mathbf{B}$ is minimized, we fix $\mathbf{S}$, $\mathbf{E}$, $\mathbf{W}$, $\mathbf{Z}$. The subproblem that only relates to $\mathbf{B}$ becomes

$$\min_{\mathbf{B}^T\mathbf{B}=\mathbf{I}} \quad \|\mathbf{B}\mathbf{E}^T-\mathbf{W}^T\mathbf{X}\|_F^2, \tag{5}$$

and according to [44], the solution of (5) is obtained as

$$\mathbf{B}=\mathbf{V_B}\mathbf{I}_{d\times c}\mathbf{U}_\mathbf{B}^T, \tag{6}$$

where $\mathbf{U_B}$ and $\mathbf{V_B}$ are composed of the left and right eigenvectors of $\mathbf{E}^T\mathbf{X}^T\mathbf{W}$, computed by singular value decomposition, respectively.

- **Update W**

When updating $\mathbf{W}$, we fix $\mathbf{S}$, $\mathbf{E}$, $\mathbf{B}$, $\mathbf{Z}$, and the subproblem that is only related to $\mathbf{W}$ becomes

$$\begin{aligned}\min_{\mathbf{W}} \quad & \left\|\mathbf{W}^T\mathbf{X}-\mathbf{B}\mathbf{E}^T\right\|_F^2+\eta\|\mathbf{W}\|_{2,1} \\ & +\gamma\operatorname{Tr}\left(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}\right),\end{aligned} \tag{7}$$

and similar to [45], we set the derivative of (7) with respect to $\mathbf{W}$ as zero, and we henceforth have

$$\mathbf{X}\mathbf{X}^T\mathbf{W}-\mathbf{X}\mathbf{E}\mathbf{B}^T+\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}+\eta\mathbf{D}\mathbf{W}=0, \tag{8}$$

where $\mathbf{D}$ is an $m\times m$ diagonal matrix with diagonal elements $d_{ii}=\frac{1}{2\|\mathbf{w}^i\|_2}$. It is noted that the derivative of $\|\mathbf{W}\|_{2,1}$ with respect to $\mathbf{W}$ is computed to be $2\mathbf{D}\mathbf{W}$. Solving (8), we could obtain

$$\mathbf{W}=(\mathbf{X}\mathbf{X}^T+\gamma\mathbf{X}\mathbf{L}\mathbf{X}^T+\eta\mathbf{D})^{-1}\mathbf{X}\mathbf{E}\mathbf{B}^T. \tag{9}$$

- **Update S**

Let $\mathbf{Y}=\mathbf{W}^T\mathbf{X}\in\mathbb{R}^{d\times n}$, then the subproblem that is merely related to $\mathbf{S}$ can be converted to

$$\begin{aligned}\min_{\mathbf{S}} \quad & \operatorname{Tr}\left(\mathbf{Y}\mathbf{L}\mathbf{Y}^T\right)+\beta\|\mathbf{S}-\mathbf{A}\|_F^2 \\ \text{s.t.} \quad & \sum_j s_{ij}=1, 0\le s_{ij}\le 1, \text{for all } i.\end{aligned}$$

This problem is independent for different values of i. Hence, we can solve the following problem separately for each value of i

$$\min_{\sum_j s_{ij}=1, s_{ij}\ge 0} \quad \sum_j \left(s_{ij}-a_{ij}\right)^2+\frac{1}{2\beta}\sum_j\|\mathbf{y}_i-\mathbf{y}_j\|_2^2\, s_{ij}.$$

Denoting $\mathbf{h}_i\in\mathbb{R}^n$ as a vector whose j-th element is equal to $\|\mathbf{y}_i-\mathbf{y}_j\|_2^2$ (and similarly for $\mathbf{s}_i\in\mathbb{R}^n$ and $\mathbf{a}_i\in\mathbb{R}^n$), the problem above can be reformulated in vector form as

$$\min_{\mathbf{s}_i^T\mathbf{1}=1, \mathbf{s}_i\ge 0} \quad \left\|\mathbf{s}_i-\left(\mathbf{a}_i-\frac{1}{4\beta}\mathbf{h}_i\right)\right\|_2^2, \tag{10}$$

and this problem can be solved by an efficient iterative algorithm [46]. To be specific, let $\mathbf{r}=\left(\mathbf{a}_i-\frac{1}{4\beta}\mathbf{h}_i\right)$, then (10) is equivalent to

$$\min_{\mathbf{s}_i^T\mathbf{1}=1, \mathbf{s}_i\ge 0} \quad \|\mathbf{s}_i-\mathbf{r}\|_2^2. \tag{11}$$

We now write the Lagrangian function of (11) as

$$L=\frac{1}{2}\|\mathbf{s}_i-\mathbf{r}\|_2^2-\lambda(\mathbf{s}_i^T\mathbf{1}-1)-\boldsymbol{\zeta}^T\mathbf{s}_i,$$

where λ is a scalar and $\boldsymbol{\zeta}$ is a Lagrangian coefficient vector. Suppose the optimal solution of (11) is $\mathbf{s}_i^*$, and the associated Lagrangian coefficients are λ^* and $\boldsymbol{\zeta}^*$, then according to the KKT condition [47], we have the following equations

$$\begin{cases}\forall j, & s_{ij}^*-r_j-\lambda^*-\zeta_j^*=0, \\ \forall j, & s_{ij}^*\ge 0, \\ \forall j, & \zeta_j^*\ge 0, \\ \forall j, & s_{ij}^*\zeta_j^*=0,\end{cases}$$

where s_{ij}^* is the j-th element of vector $\mathbf{s}_i^*$. Since $\mathbf{s}_i^T\mathbf{1}=1$ and $s_{ij}^*-r_j-\lambda^*-\zeta_j^*=0$, we could obtain

$$\lambda^*=\frac{1-\mathbf{1}^T\mathbf{r}-\mathbf{1}^T\boldsymbol{\zeta}^*}{n}.$$

Plugging the above equation back into the KKT equations, we could get

$$\mathbf{s}_i^*=\left(\mathbf{r}-\frac{\mathbf{1}\mathbf{1}^T}{n}\mathbf{r}+\frac{1}{n}\mathbf{1}-\frac{\mathbf{1}^T\boldsymbol{\zeta}^*}{n}\mathbf{1}\right)+\boldsymbol{\zeta}^*.$$

Denoting $\bar{\zeta}^*=\frac{\mathbf{1}^T\boldsymbol{\zeta}^*}{n}$ and $\mathbf{u}=\mathbf{r}-\frac{\mathbf{1}\mathbf{1}^T}{n}\mathbf{r}+\frac{1}{n}\mathbf{1}$, for any j, we have $s_{ij}^*=u_j+\zeta_j^*-\bar{\zeta}^*$. According to the KKT conditions,

$s_{ij}^* = u_j + \zeta_j^* - \bar{\zeta}^* = \left(u_j - \bar{\zeta}^*\right)_+$, so we could obtain s_{ij}^* once we've got $\bar{\zeta}^*$. Applying the KKT conditions, we could rewrite $\zeta_j^* = (\bar{\zeta}^* - u_j)_+$, and since $\mathbf{r}$ is an n dimensional vector, we have $\zeta^* = \frac{1}{n-1}\sum_{j=1}^{n-1}(\bar{\zeta}^* - u_j)$. We define a function as

$$f(\bar{\zeta}) = \frac{1}{n-1}\sum_{j=1}^{n-1}\left(\bar{\zeta} - u_j\right)_+ - \bar{\zeta},$$

and it is easy to verify that $\bar{\zeta}^*$ is a root of $f(\bar{\zeta}) = 0$. We now could henceforth adopt optimization methods like Newton method to find the root of $f(\zeta) = 0$ efficiently so as to obtain $\bar{\zeta}^*$.

- **Update E**

The subproblem that is only relevant to $\mathbf{E}$ is

$$\begin{aligned} \min_{\mathbf{E}} \quad & \left\|\mathbf{W}^T\mathbf{X} - \mathbf{B}\mathbf{E}^T\right\|_F^2 + \alpha\|\mathbf{Z} - \mathbf{E}\|_F^2 \\ \text{s.t.} \quad & \mathbf{E}^T\mathbf{E} = \mathbf{I}, \end{aligned}$$

and the subproblem above can be rewritten as

$$\min_{\mathbf{E}^T\mathbf{E} = \mathbf{I}} \quad \left\|\mathbf{E} - \left(\mathbf{X}^T\mathbf{W}\mathbf{B} + \alpha\mathbf{Z}\right)\right\|_F^2. \tag{12}$$

Similar to updating $\mathbf{B}$, according to [44], we could obtain solution of this subproblem as

$$\mathbf{E} = \mathbf{V}_{\mathbf{E}}\mathbf{I}_{n\times c}\mathbf{U}_{\mathbf{E}}^T, \tag{13}$$

where $\mathbf{U}_{\mathbf{E}}$ and $\mathbf{V}_{\mathbf{E}}$ are composed of the left and right eigenvectors of $\mathbf{B}^T\mathbf{W}^T\mathbf{X} + \alpha\mathbf{Z}^T$, computed by singular value decomposition, respectively.

- **Update Z**

The subproblem that exclusively relates to $\mathbf{Z}$ is

$$\min_{\mathbf{Z} \geq 0} \quad \|\mathbf{Z} - \mathbf{E}\|_F^2, \tag{14}$$

and the solution of the above subproblem is straightforward, as

$$\mathbf{Z} = \max(\mathbf{E}, 0). \tag{15}$$

Based on the above analysis, we summarize the detailed optimization procedures in Algorithm 1.

B. Computational Complexity Analysis

The computational complexity of our proposed OCLSP method is analyzed here. Recall that n represents the total number of data samples, m represents the number of features in original data, c is the number of latent clusters, d is the dimension of the projected subspace, and t is the total number of iterations. Updating $\mathbf{B}$ needs to compute $\mathbf{E}^T\mathbf{X}^T\mathbf{W}$ and perform singular value decomposition on it, and their computational complexities are $\mathcal{O}(cnm + cmd)$ and $\mathcal{O}(c^2d + cd^2)$, respectively. Moreover, the computation complexity of calculating $\mathbf{V}_{\mathbf{B}}\mathbf{I}_{d\times c}\mathbf{U}_{\mathbf{B}}^T$ is $\mathcal{O}(cd^2 + c^2d)$. Thus, the computational complexity of updating $\mathbf{B}$ is $\mathcal{O}(cnm + cmd + c^2d + cd^2)$. Updating $\mathbf{W}$ involves many matrix manipulations, and its computational complexity is $\mathcal{O}(m^2n + mn^2 + m^3 + mnc + mcd)$. Updating $\mathbf{D}$ needs to compute all row Euclidean norms of $\mathbf{W}$, so its computational complexity is $\mathcal{O}(md)$. Besides, the cost of updating $\mathbf{S}$ is $\mathcal{O}(n\log n)$, the cost of calculating $\mathbf{L}$ is $\mathcal{O}(n^2)$, and it takes $\mathcal{O}(cmd + cmn + c^2n + cn^2)$ to update $\mathbf{E}$, $\mathcal{O}(nc)$ to update $\mathbf{Z}$. Suppose $c \ll n$, $c \ll m$, and $d \ll n$, $d \ll m$, which are usually established in reality, then the total computational complexity of our proposed OCLSP method is $\mathcal{O}\left(\left(mn^2 + nm^2 + m^3\right)\cdot t\right)$ approximately.

Algorithm 1 The optimization algorithm for OCLSP.

```
Input: Data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$; Regularization parameters: $\eta$,
    $\gamma$, $\beta$ and a large enough $\alpha$; Number of latent clusters
    $c$; Number of selected features $p$;
Output: $p$ features for the data set;
Use $k$-means to initialize $\mathbf{E} \in \mathbb{R}^{n\times c}$; Set $t = 0$ and
    $\mathbf{D}_t \in \mathbb{R}^{m\times m}$ as an identity matrix; Use (2) to construct
    the initial similarity matrix $\mathbf{A}$;
repeat
    Update $\mathbf{B}_t$ by (6);
    Update $\mathbf{W}_t$ by (9);
    Update $\mathbf{D}_t$ as $\begin{bmatrix} \frac{1}{2\|\mathbf{w}_t^1\|_2} & & \\ & \ddots & \\ & & \frac{1}{2\|\mathbf{w}_t^m\|_2} \end{bmatrix}$;
    Update $\mathbf{S}_t$ by (10), then calculate $\mathbf{L}_t = \mathbf{P}_t - \frac{\mathbf{S}_t^T + \mathbf{S}_t}{2}$;
    Update $\mathbf{E}_t$ by (13);
    Update $\mathbf{Z}_t$ by (15);
    $t = t + 1$;
until Convergence criterion satisfied
Sort all $m$ features according to $\left\|\mathbf{w}_t^i\right\|_2$ $(i = 1, 2, \ldots, m)$
    in descending order and select the top-$p$ ranked ones.
```

C. Convergence Analysis

In this section, we prove the convergence of Algorithm 1 for solving our proposed OCLSP method. For convenience, let us denote the objective function part of (4) as

$$\begin{aligned} \mathscr{L}\left(\mathbf{W}, \mathbf{B}, \mathbf{E}, \mathbf{Z}, \mathbf{S}\right) = & \left\|\mathbf{W}^T\mathbf{X} - \mathbf{B}\mathbf{E}^T\right\|_F^2 \\ & + \eta\|\mathbf{W}\|_{2,1} + \alpha\|\mathbf{Z} - \mathbf{E}\|_F^2 \\ & + \gamma\left(\mathrm{Tr}\left(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}\right) + \beta\|\mathbf{S} - \mathbf{A}\|_F^2\right). \end{aligned} \tag{16}$$

Theorem 1: The objective function (16) monotonically decreases in each iteration by running Algorithm 1, and the algorithm will converge.

Proof: According to (16), the part of $\mathscr{L}\left(\mathbf{W}, \mathbf{B}, \mathbf{E}, \mathbf{Z}, \mathbf{S}\right)$ that only relates to $\mathbf{W}$ can be separated as

$$\min_{\mathbf{W}} \quad \left\|\mathbf{W}^T\mathbf{X} - \mathbf{B}\mathbf{E}^T\right\|_F^2 + \eta\|\mathbf{W}\|_{2,1} + \gamma\,\mathrm{Tr}\left(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}\right).$$

For the $\mathbf{W}$ update by (9), $\mathbf{W}_{t+1}$ is also the solution of the following problem with fixed $\mathbf{D}_t$ as

$$\begin{aligned} \mathbf{W}_{t+1} = \arg\min_{\mathbf{W}} & \left\|\mathbf{W}^T\mathbf{X} - \mathbf{B}_t\mathbf{E}_t^T\right\|_F^2 \\ & + \gamma\,\mathrm{Tr}\left(\mathbf{W}^T\mathbf{X}\mathbf{L}_t\mathbf{X}^T\mathbf{W}\right) + \eta\,\mathrm{Tr}\left(\mathbf{W}^T\mathbf{D}_t\mathbf{W}\right), \end{aligned}$$

which indicates that

$$\begin{aligned}&\left\|\mathbf{W}_{t+1}^T\mathbf{X}-\mathbf{B}_t\mathbf{E}_t^T\right\|_F^2+\gamma\operatorname{Tr}\left(\mathbf{W}_{t+1}^T\mathbf{X}\mathbf{L}_t\mathbf{X}^T\mathbf{W}_{t+1}\right)\\&+\eta\operatorname{Tr}\left(\mathbf{W}_{t+1}^T\mathbf{D}_t\mathbf{W}_{t+1}\right)\\\leq&\left\|\mathbf{W}_t^T\mathbf{X}-\mathbf{B}_t\mathbf{E}_t^T\right\|_F^2+\gamma\operatorname{Tr}\left(\mathbf{W}_t^T\mathbf{X}\mathbf{L}_t\mathbf{X}^T\mathbf{W}_t\right)\\&+\eta\operatorname{Tr}\left(\mathbf{W}_t^T\mathbf{D}_t\mathbf{W}_t\right).\end{aligned}\tag{17}$$

According to the theorem proposed in the literature [45], for any nonzero row vectors $\mathbf{u},\mathbf{u}_t\in\mathbb{R}^c$, we have the inequality as

$$\|\mathbf{u}\|_2-\frac{\|\mathbf{u}\|_2^2}{2\|\mathbf{u}_t\|_2}\leq\|\mathbf{u}_t\|_2-\frac{\|\mathbf{u}_t\|_2^2}{2\|\mathbf{u}_t\|_2}.$$

Applying the above inequality with $\mathbf{u}=\mathbf{w}_{t+1}^i$ and $\mathbf{u}_t=\mathbf{w}_t^i$, we get

$$\|\mathbf{W}_{t+1}\|_{2,1}-\sum_{i=1}^m\frac{\left\|\mathbf{w}_{t+1}^i\right\|_2^2}{2\left\|\mathbf{w}_t^i\right\|_2}\leq\|\mathbf{W}_t\|_{2,1}-\sum_{i=1}^m\frac{\left\|\mathbf{w}_t^i\right\|_2^2}{2\left\|\mathbf{w}_t^i\right\|_2}.$$

We now rewrite the above inequality in matrix form as

$$\begin{aligned}&\|\mathbf{W}_{t+1}\|_{2,1}-\operatorname{Tr}\left(\mathbf{W}_{t+1}^T\mathbf{D}_t\mathbf{W}_{t+1}\right)\\\leq&\|\mathbf{W}_t\|_{2,1}-\operatorname{Tr}\left(\mathbf{W}_t^T\mathbf{D}_t\mathbf{W}_t\right).\end{aligned}\tag{18}$$

By combining (17) and (18), we have

$$\begin{aligned}&\left\|\mathbf{W}_{t+1}^T\mathbf{X}-\mathbf{B}_t\mathbf{E}_t^T\right\|_F^2+\gamma\operatorname{Tr}\left(\mathbf{W}_{t+1}^T\mathbf{X}\mathbf{L}_t\mathbf{X}^T\mathbf{W}_{t+1}\right)\\&+\eta\|\mathbf{W}_{t+1}\|_{2,1}\\\leq&\left\|\mathbf{W}_t^T\mathbf{X}-\mathbf{B}_t\mathbf{E}_t^T\right\|_F^2+\gamma\operatorname{Tr}\left(\mathbf{W}_t^T\mathbf{X}\mathbf{L}_t\mathbf{X}^T\mathbf{W}_t\right)\\&+\eta\|\mathbf{W}_t\|_{2,1},\end{aligned}$$

which implies that

$$\mathscr{L}\left(\mathbf{W}_{t+1},\mathbf{B}_t,\mathbf{E}_t,\mathbf{Z}_t,\mathbf{S}_t\right)\leq\mathscr{L}\left(\mathbf{W}_t,\mathbf{B}_t,\mathbf{E}_t,\mathbf{Z}_t,\mathbf{S}_t\right).\tag{19}$$

Then, since $\mathbf{B}$ is updated according to (5), we have

$$\begin{aligned}\mathbf{B}_{t+1}&=\arg\min_{\mathbf{B}}\|\mathbf{B}\mathbf{E}_t^T-\mathbf{W}_{t+1}^T\mathbf{X}\|_F^2\\&=\arg\min_{\mathbf{B}}\mathscr{L}\left(\mathbf{W}_{t+1},\mathbf{B},\mathbf{E}_t,\mathbf{Z}_t,\mathbf{S}_t\right),\end{aligned}$$

which indicates that

$$\mathscr{L}\left(\mathbf{W}_{t+1},\mathbf{B}_{t+1},\mathbf{E}_t,\mathbf{Z}_t,\mathbf{S}_t\right)\leq\mathscr{L}\left(\mathbf{W}_{t+1},\mathbf{B}_t,\mathbf{E}_t,\mathbf{Z}_t,\mathbf{S}_t\right).\tag{20}$$

In addition, since $\mathbf{E}$ is updated by (12), we could obtain

$$\begin{aligned}\mathbf{E}_{t+1}&=\arg\min_{\mathbf{E}}\left\|\mathbf{E}-\left(\mathbf{X}^T\mathbf{W}_{t+1}\mathbf{B}_{t+1}+\alpha\mathbf{Z}_t\right)\right\|_F^2\\&=\arg\min_{\mathbf{E}}\mathscr{L}\left(\mathbf{W}_{t+1},\mathbf{B}_{t+1},\mathbf{E},\mathbf{Z}_t,\mathbf{S}_t\right),\end{aligned}$$

and this implies that

$$\mathscr{L}\left(\mathbf{W}_{t+1},\mathbf{B}_{t+1},\mathbf{E}_{t+1},\mathbf{Z}_t,\mathbf{S}_t\right)\leq\mathscr{L}\left(\mathbf{W}_{t+1},\mathbf{B}_{t+1},\mathbf{E}_t,\mathbf{Z}_t,\mathbf{S}_t\right).\tag{21}$$

Moreover, owing to that $\mathbf{Z}$ is updated based on (14), we get

$$\begin{aligned}\mathbf{Z}_{t+1}&=\arg\min_{\mathbf{Z}}\|\mathbf{Z}-\mathbf{E}_{t+1}\|_F^2\\&=\arg\min_{\mathbf{Z}}\mathscr{L}\left(\mathbf{W}_{t+1},\mathbf{B}_{t+1},\mathbf{E}_{t+1},\mathbf{Z},\mathbf{S}_t\right),\end{aligned}$$

which means

$$\begin{aligned}&\mathscr{L}\left(\mathbf{W}_{t+1},\mathbf{B}_{t+1},\mathbf{E}_{t+1},\mathbf{Z}_{t+1},\mathbf{S}_t\right)\\\leq&\mathscr{L}\left(\mathbf{W}_{t+1},\mathbf{B}_{t+1},\mathbf{E}_{t+1},\mathbf{Z}_t,\mathbf{S}_t\right).\end{aligned}\tag{22}$$

Finally, on account of that we use (10) to update $\mathbf{S}$, we have

$$\begin{aligned}\mathbf{S}_{t+1}&=\arg\min_{\mathbf{S}}\left\|\mathbf{S}-\left(\mathbf{A}-\frac{1}{4\beta}\mathbf{H}_{t+1}\right)\right\|_F^2\\&=\arg\min_{\mathbf{S}}\mathscr{L}\left(\mathbf{W}_{t+1},\mathbf{B}_{t+1},\mathbf{E}_{t+1},\mathbf{Z}_{t+1},\mathbf{S}\right),\end{aligned}$$

and this means

$$\begin{aligned}&\mathscr{L}\left(\mathbf{W}_{t+1},\mathbf{B}_{t+1},\mathbf{E}_{t+1},\mathbf{Z}_{t+1},\mathbf{S}_{t+1}\right)\\\leq&\mathscr{L}\left(\mathbf{W}_{t+1},\mathbf{B}_{t+1},\mathbf{E}_{t+1},\mathbf{Z}_{t+1},\mathbf{S}_t\right).\end{aligned}\tag{23}$$

Based on (19), (20), (21), (22) and (23), we can now draw a conclusion that

$$\mathscr{L}\left(\mathbf{W}_{t+1},\mathbf{B}_{t+1},\mathbf{E}_{t+1},\mathbf{Z}_{t+1},\mathbf{S}_{t+1}\right)\leq\mathscr{L}\left(\mathbf{W}_t,\mathbf{B}_t,\mathbf{E}_t,\mathbf{Z}_t,\mathbf{S}_t\right),$$

which indicates that the objective function (16) monotonically decreases in each iteration. Furthermore, since function (16) is convex in each variable, the algorithm with update rules converges. □

VI. Experiments

In this section, we evaluate the performance of our proposed OCLSP method on nine real-world datasets and compare it with several state-of-the-art unsupervised feature selection methods. The k-means clustering algorithm [28], which is a common and basic clustering method with a wide range of applications, is adopted to validate the effectiveness of feature selection methods. All experiments are implemented in MATLAB R2020b, and codes are run on an Ubuntu server with 3.70-GHz i9-10900K CPU, 128-GB main memory.

A. Datasets

The evaluation is conducted on nine real-world datasets, including six image datasets and three biological datasets. The statistics of these datasets are summarized in Table I.

B. Comparison Algorithms

In order to verify the validity of our proposed OCLSP method, we compare it with eight unsupervised feature selection methods, they are:

1) ALLfea: The baseline method. ALLfea performs k-means clustering [28] using all original features.
2) LapScore [20]: Laplacian score evaluates the features separately according to their abilities to preserve the local geometric information of data.
3) UDFS [21]: Unsupervised discriminant feature selection uses $\ell_{2,1}$ norm regularization to exploit local discriminative information of data and excavate correlation between features simultaneously.
4) NDFS [22]: Nonnegative discriminative feature selection selects features by leveraging a joint framework of nonnegative spectral analysis and $\ell_{2,1}$ norm regularization.

TABLE I
THE STATISTICS OF DATASETS

Dataset	# Instances	# Original Features	# Classes	# Selected Features	Type
MSRA25	1799	256	12	[5,10,,50]	image
PalmData25	2000	256	100	[5,10,,50]	image
ECOLI	336	343	8	[5,10,,50]	biological
UMIST	575	644	20	[5,10,,50]	image
JAFFE	213	676	10	[5,10,,50]	image
COIL20	1440	1024	20	[5,10,,50]	image
warpPIE10p	210	2420	10	[50,100,150,200,250,300]	image
Lung	203	3312	5	[50,100,150,200,250,300]	biological
GLIOMA	50	4434	4	[50,100,150,200,250,300]	biological

5) FSASL [24]: Unsupervised feature selection with adaptive structure learning adaptively updates feature selection matrix based on global and local structure learning.
6) SOCFS [40]: Simultaneous orthogonal basis clustering feature selection performs orthogonal basis clustering by utilizing a new type of target matrix.
7) SOGFS [25]: Structured optimal graph feature selection conducts local structure learning with ideal neighbor assignment to select features.
8) URAFS [26]: Generalized uncorrelated regression with adaptive graph for feature selection aims to learn uncorrelated yet discriminative features with a closed-form solution through an improved sparse representation model.

Similar to previous work, we evaluate the performance of unsupervised feature selection methods by two widely employed evaluation metrics, i.e., accuracy (ACC) and normalized mutual information (NMI) [21], and the larger ACC and NMI indicate better performance. ACC is defined as follows

$$\mathrm{ACC}=\frac{\sum_{i=1}^{n}\delta\left(\operatorname{map}\left(r_i\right),l_i\right)}{n},$$

where l_i is the true label of $\mathbf{x}_i$ and r_i is the clustering result of $\mathbf{x}_i$, n is the total number of samples, $\delta(x,y)=1$ if $x=y$, otherwise $\delta(x,y)=0$, and $\operatorname{map}(\cdot)$ is the best permutation mapping function that permutes clustering labels to match the true labels using the KuhnMunkres algorithm [48]. Given two variables P and Q, NMI is defined as

$$\mathrm{NMI}(P,Q)=\frac{I(P,Q)}{\sqrt{H(P)H(Q)}},$$

where P and Q are the true labels and clustering results, respectively, $I(P,Q)$ is the mutual information between P and Q, and $H(P)$ and $H(Q)$ are the entropy of P and Q separately.

There are several parameters need to be tuned in our proposed OCLSP method and other unsupervised feature selection methods. First, for all unsupervised feature selection methods, each must specify the number of neighbors k, which is used in (2), and we set $k=5$ in our experiments. Besides, for NDFS and OCLSP methods, we fix $\alpha=10^4$ to guarantee the orthogonality of cluster indicator matrix. At last, in order to make our experiments fair enough, we tune the regularization parameters with a grid search strategy where parameters are ranging from $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$.

In the following experiments, we perform k-means clustering algorithm [28] in different subspaces, each of which is constructed by features selected by one of unsupervised feature selection methods aforementioned, respectively. Since the clustering results would be affected by the choice of the initial clustering seeds, we repeat the experiments 20 times with random initialization of clustering seeds, and report the average results of ACC and NMI. Besides, the parameter sensitivity of our OCLSP method and convergence of Algorithm 1 are also discussed.

C. Experimental Results and Analysis

To compare the performance of nine different methods comprehensively, the ACC and NMI values of these methods with best parameters on nine real-world datasets are shown in Table II and Table III respectively, and the values in bold represent the best values. Besides, the curves of the ACC and NMI values of different methods with varying number of features on nine real-world datasets are shown in Fig. 2 and Fig. 3 respectively, where the horizontal axis represents the number of selected features while the vertical axis indicates the ACC or NMI.

It can be seen from Table II, Table III, Fig. 2 and Fig. 3 that our proposed OCLSP method can obtain the best ACC and NMI in most cases compared with other unsupervised feature selection methods, which fully demonstrates the superiority of our proposed OCLSP method. In Table II, we could find that our OCLSP method outperforms all other methods in terms of ACC, and in Table III, our OCLSP performs best in terms of NMI except on PalmData25 dataset, where the baseline algorithm, ALLfea, gets the highest NMI. Part of the reason is that, our OCLSP method, which uses an adaptive graph to capture the local geometrical structure information of data points and leverages an orthogonal basis clustering to achieve an excellent cluster separation, has made up the drawbacks presented in previous methods. Besides, in Fig. 2, the experimental results show that our OCLSP method outperforms all other unsupervised feature selection methods on these nine real-world datasets in terms of ACC given a proper number interval of features, especially on the ECOLI dataset, where the ACC of our OCLSP method exceeds other methods dramatically. And Fig. 3 shows similar results that our OCLSP method achieves the best NMI on these nine real-world datasets provided a favorable number interval of features. All of these evidences have demonstrated the excellence of our proposed OCLSP method.

D. Parameter Sensitivity Analysis

For the proposed OCLSP method, we need to tune the regularization parameters η, γ and β.

We first discuss the effect of the parameter β on the results. In this experiment, the value of β is adjusted in the range of $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$ and other parameters are fixed as $\eta = 0.1$ and $\gamma = 0.1$. Fig. 4 shows the clustering results of the OCLSP method with different β values on nine real-world datasets, where the vertical axis indicates the clustering accuracy, and the horizontal axis represents the value of parameter β. As can be seen from the experimental results, different data need different β values to achieve the best ACC.

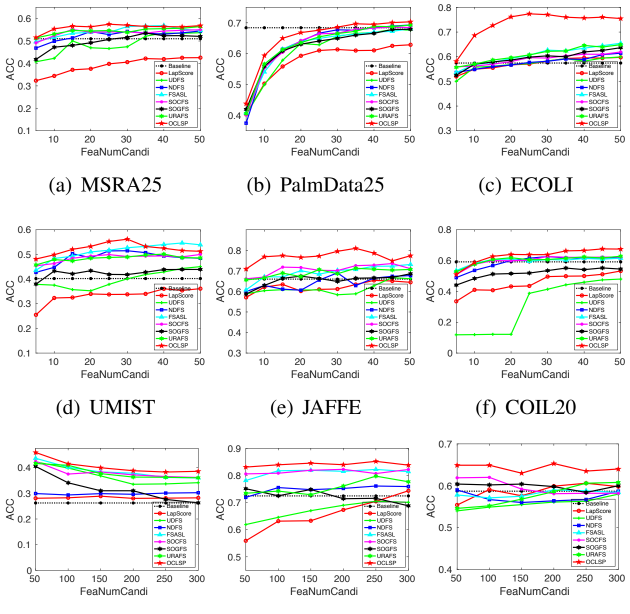
Next we focus on the parameters η and γ. With β fixed at 1, η and γ are searched in the range of $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$, and the ACC and NMI values are obtained under the combination of each pair of parameters η and γ. The three-dimensional histograms of ACC and NMI values on nine real-world datasets are shown in Fig. 5 and Fig. 6 respectively. It can be seen from Fig. 5 and Fig. 6 that when the parameters η and γ are varying, the ACC and NMI values keep nearly unchanged in most cases, which shows the robustness of our OCLSP method to some extent. However, in rare cases, e.g., on ECOLI dataset, the ACC and NMI decrease dramatically when γ is diminishing. Besides, in most cases, the ACC and NMI show a trend from ascending to descending with respect to both γ and η. Part of the reason is that, a small value of parameter is nearly equivalent to removing that term from optimization, which leads to a simpler method that can not work well, but a large enough value of parameter would make other terms in the objective function negligible, which twists the purpose of feature selection, and finally impairs the performance.

E. Convergence Analysis

We have already proven the convergence of Algorithm 1 for optimizing the objective function of our OCLSP method in the previous section, and now we experimentally study the speed of its convergence. The convergence curves of the objective value on nine real-world datasets are shown in Fig. 7, and the parameters for testing the convergence of Algorithm 1 are $\beta = 0.01$, $\eta = 1$ and $\gamma = 1$. We could observe from the figure that, our Algorithm 1 converges within five iterations on all datasets, which indicates that Algorithm 1 converges very fast. The fast convergence of Algorithm 1 ensures the speed of the whole proposed method.

F. Running Time Analysis

We here analyze running time of our proposed OCLSP method. Fig. 8(a) shows the running time of different unsupervised feature selection algorithms on the COIL20 dataset. As shown, all methods take less than 35 seconds to select features. Besides, although our proposed OCLSP method is the most time consuming, which is mainly attributed to the inversion involved in updating $\mathbf{W}$, the improvement it brings to the quality of selected features is remarkable. Fig. 8(b) shows the running time of our proposed OCLSP method on different datasets. We can observe that training on the GLIOMA dataset, which contains the largest number of features among all datasets, costs the most time. This phenomenon meets our theoretical complexity analysis that our method scales cubically with the

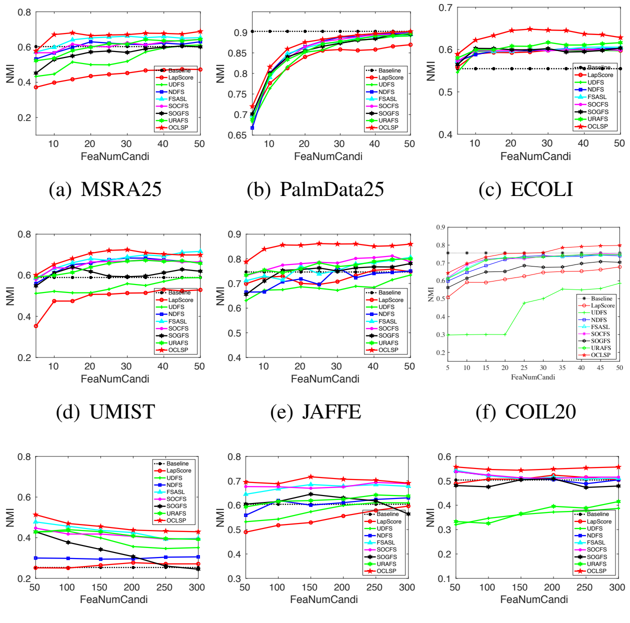


(a) MSRA25 (b) PalmData25 (c) ECOLI

(d) UMIST (e) JAFFE (f) COIL20

(g) warpPIE10P (h) LUNG (i) GLIOMA

Fig. 2. Clustering ACC of different feature selection algorithms with different number of selected features on nine real-world datasets.



(a) MSRA25 (b) PalmData25 (c) ECOLI

(d) UMIST (e) JAFFE (f) COIL20

(g) warpPIE10P (h) LUNG (i) GLIOMA

Fig. 3. Clustering NMI of different feature selection algorithms with different number of selected features on nine real-world datasets.

TABLE II

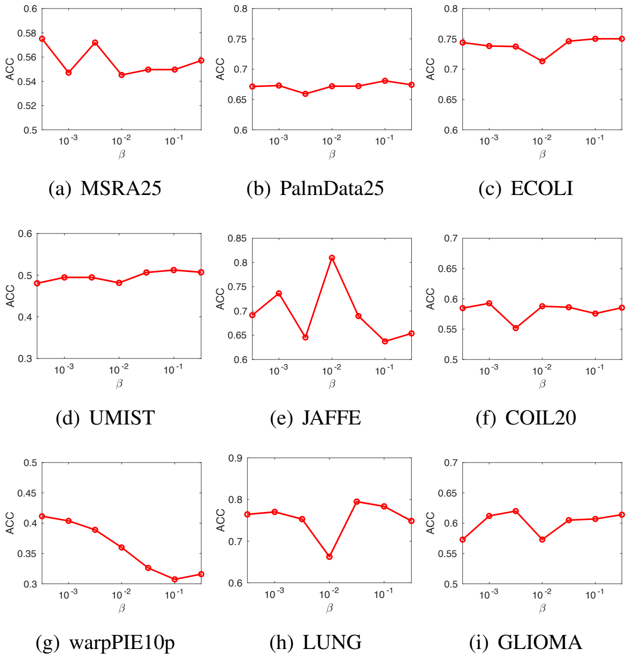
CLUSTERING RESULTS (ACC%±STD) OF DIFFERENT FEATURE SELECTION ALGORITHMS ON NINE REAL-WORLD DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Data set	ALLfea	LapS	UDFS	NDFS	FSASL	SOCFS	SOGFS	URAFS	OCLSP
MSRA25	51.02 ±5.25	42.59 ±1.64	53.73 ±4.38	54.46 ±4.95	56.91 ±5.24	55.18 ±4.58	53.56 ±3.53	56.45 ±1.62	**57.51 ±4.42**
PalmData25	68.41 ±2.29	62.88 ±2.07	67.79 ±2.58	69.29 ±1.61	68.08 ±2.22	68.60 ±2.43	67.90 ±1.43	69.29 ±1.78	**70.33 ±2.18**
ECOLI	57.44 ±8.26	59.75 ±8.15	59.93 ±8.15	61.43 ±9.82	65.46 ±6.40	61.93 ±6.96	63.66 ±7.72	64.88 ±10.60	**77.40 ±2.86**
UMIST	40.22 ±2.20	36.11 ±1.67	45.04 ±3.24	51.49 ±3.44	54.63 ±3.81	49.98 ±3.67	43.91 ±1.99	50.12 ±3.16	**56.17 ±2.35**
JAFFE	65.99 ±6.07	65.35 ±5.61	66.69 ±6.75	69.01 ±6.25	73.12 ±7.94	73.57 ±7.50	68.59 ±6.48	71.27 ±2.83	**80.96 ±8.48**
COIL20	59.17 ±3.98	53.25 ±4.04	48.01 ±2.95	62.61 ±4.45	61.56 ±4.82	62.70 ±4.35	55.22 ±3.02	62.84 ±4.09	**67.59 ±4.27**
warpPIE10p	26.24 ±2.03	28.88 ±1.96	42.24 ±3.56	30.24 ±2.660	43.64 ±3.68	42.45 ±3.92	40.48 ±6.05	41.62 ±5.11	**45.90 ±4.67**
Lung	72.46 ±10.20	74.36 ±7.21	70.49 ±12.58	76.11 ±6.62	83.00 ±7.29	82.27 ±5.07	75.12 ±10.57	79.75 ±7.70	**85.22 ±3.27**
GLIOMA	58.7 ±6.63	60.60 ±4.26	58.00 ±7.48	59.00 ±7.36	59.00 ±9.50	62.00 ±9.50	62.00 ±8.94	60.80 ±9.30	**65.30 ±7.55**

TABLE III

CLUSTERING RESULTS (NMI%±STD) OF DIFFERENT FEATURE SELECTION ALGORITHMS ON NINE REAL-WORLD DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Data set	ALLfea	LapS	UDFS	NDFS	FSASL	SOCFS	SOGFS	URAFS	OCLSP
MSRA25	60.24 ±4.06	47.41 ±1.93	61.15 ±1.92	62.99 ±2.56	66.17 ±1.84	64.33 ±3.07	60.56 ±3.14	64.36 ±0.94	**68.87 ±2.04**
PalmData25	**90.24 ±0.71**	86.99 ±0.75	89.14 ±0.91	89.89 ±0.52	89.48 ±0.65	89.77 ±0.82	89.53 ±0.45	89.74 ±0.73	90.20 ±0.56
ECOLI	55.48 ±2.92	60.16 ±3.17	59.86 ±3.00	60.26 ±3.47	60.57 ±2.81	60.02 ±2.95	60.35 ±2.36	61.68 ±3.55	**64.82 ±1.20**
UMIST	58.91 ±1.58	53.06 ±1.71	58.82 ±1.61	68.36 ±1.62	71.6 ±1.47	67.61 ±2.03	60.28 ±1.80	67.27 ±2.29	**72.42 ±1.87**
JAFFE	74.65 ±3.34	75.79 ±2.07	73.42 ±4.31	76.51 ±3.43	80.58 ±3.83	81.17 ±2.99	78.09 ±2.87	79.98 ±2.32	**86.18 ±2.80**
COIL20	75.58 ±1.64	67.80 ±1.62	58.68 ±1.02	74.39 ±1.62	74.72 ±2.42	75.27 ±2.33	70.79 ±1.00	74.67 ±1.23	**79.81 ±1.93**
warpPIE10p	25.36 ±3.18	27.72 ±2.01	43.24 ±4.08	30.58 ±3.44	47.77 ±2.67	44.74 ±4.31	42.92 ±4.99	43.97 ±2.72	**51.32 ±4.49**
Lung	60.37 ±5.38	59.63 ±5.40	61.04 ±6.09	63.02 ±5.16	68.47 ±4.10	69.33 ±4.78	64.52 ±3.05	64.20 ±5.99	**71.63 ±5.31**
GLIOMA	50.32 ±4.30	52.30 ±3.06	38.69 ±7.32	53.88 ±4.11	54.27 ±4.24	53.79 ±3.91	50.84 ±5.96	41.50 ±10.33	**55.68 ±5.43**

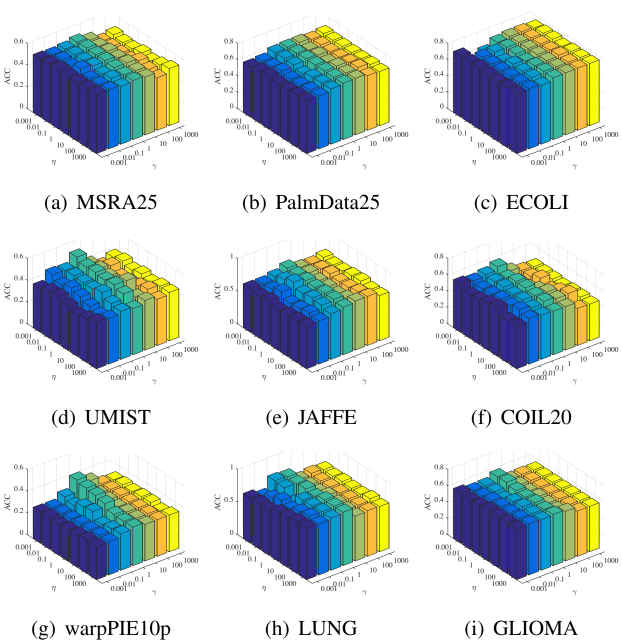


(a) MSRA25 (b) PalmData25 (c) ECOLI

(d) UMIST (e) JAFFE (f) COIL20

(g) warpPIE10p (h) LUNG (i) GLIOMA

Fig. 4. Clustering ACC of OCLSP with different β values on nine real-world datasets.



(a) MSRA25 (b) PalmData25 (c) ECOLI

(d) UMIST (e) JAFFE (f) COIL20

(g) warpPIE10p (h) LUNG (i) GLIOMA

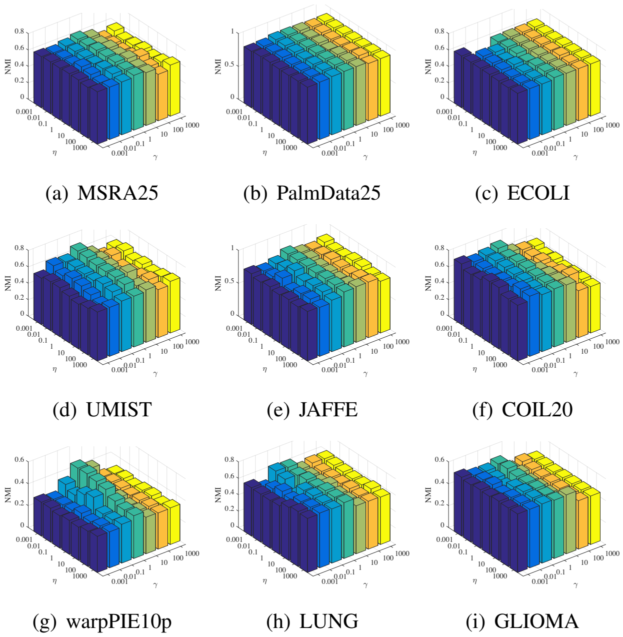
Fig. 5. Clustering ACC of OCLSP on nine real-world datasets with different γ and η values.

number of features, while training on datasets of moderate size spends nearly the same but not much time.

VII. Conclusion and Future Work

In this paper, a novel unsupervised feature selection method OCLSP, which performs feature selection and orthogonal basis clustering simultaneously under a joint framework, is proposed, and the objective function for our OCLSP method could be optimized efficiently by the advised optimization procedure. Extensive experiments on nine real-world datasets have verified the efficacy of the proposed OCLSP method.

Future work could be done in many aspects. First and foremost, how to decrease the computational complexity is a critical point. Although our OCLSP model could obtain the best effectiveness in most cases, the total computational complexity of our OCLSP method is $\mathcal{O}\left(\left(mn^2+nm^2+m^3\right)\cdot t\right)$ approximately, which is time consuming. Finally yet importantly, how to extend unsupervised feature selection method to handle tensorial data is also crucial. Since in reality, many data, e.g., colored images, are represented as tensors naturally.

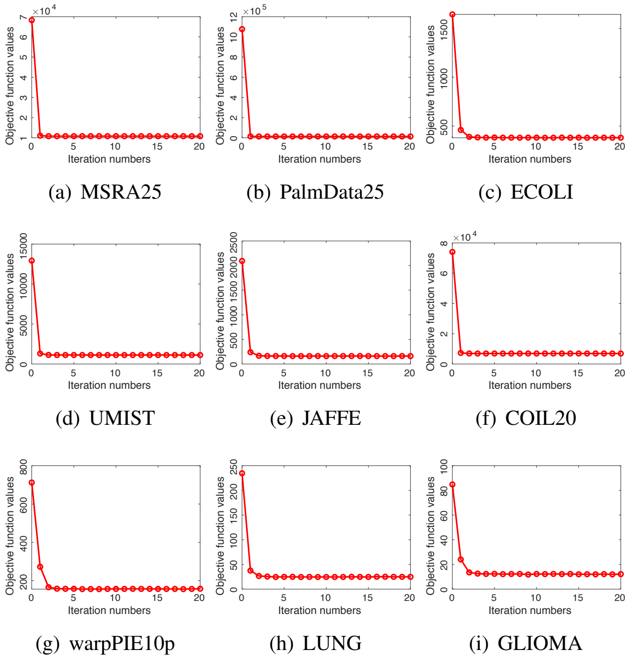


(a) MSRA25 (b) PalmData25 (c) ECOLI

(d) UMIST (e) JAFFE (f) COIL20

(g) warpPIE10p (h) LUNG (i) GLIOMA

Fig. 6. Clustering NMI of OCLSP on nine real-world datasets with different γ and η values.



(a) MSRA25 (b) PalmData25 (c) ECOLI

(d) UMIST (e) JAFFE (f) COIL20

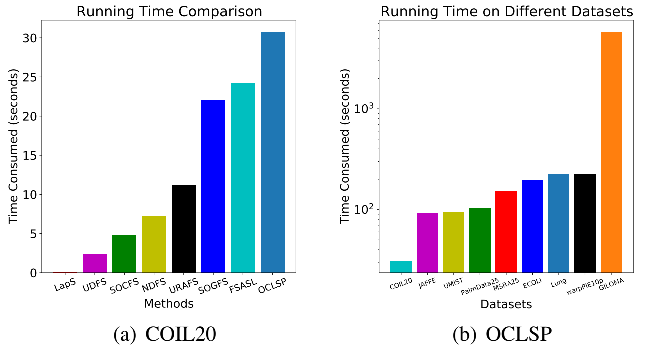
(g) warpPIE10p (h) LUNG (i) GLIOMA

Fig. 7. The convergence curves of our objective function on nine real-world datasets.

Constructing learning models and designing fast algorithms for solving such models for tensorial data is a significant research topic.

References

[1] S. Li, L. Yang, J. Huang, X.-S. Hua, and L. Zhang, "Dynamic anchor feature selection for single-shot object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6609–6618.



(a) COIL20 (b) OCLSP

Fig. 8. Running time analysis.

[2] Z. Chen, M. Pang, Z. Zhao, S. Li, R. Miao, Y. Zhang, X. Feng, X. Feng, Y. Zhang, M. Duan, L. Huang, and F. Zhou, "Feature selection may improve deep neural networks for the bioinformatics problems," *Bioinformatics*, vol. 36, pp. 1542–1552, 2020.

[3] G. He, J. Ji, H. Zhang, Y. Xu, and J. Fan, "Feature selection-based hierarchical deep network for image classification," *IEEE Access*, vol. 8, pp. 15 436–15 447, 2020.

[4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[5] K. Henni, N. Mezghani, and C. Gouin-Vallerand, "Unsupervised graph-based feature selection via subspace and pagerank centrality," *Expert Systems with Applications*, vol. 114, pp. 46–53, 2018.

[6] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *Acm Computing Surveys*, vol. 50, pp. 1–45, 2016.

[7] I. M. Guyon, S. R. Gunn, M. Nikravesh, and L. Zadeh, "Feature extraction, foundations and applications," *Studies in Fuzziness & Soft Computing*, vol. 205, pp. 68–84, 2006.

[8] Y. A. Ghassabeh, F. Rudzicz, and H. A. Moghaddam, "Fast incremental lda feature extraction," *Pattern Recognition*, vol. 48, pp. 1999–2012, 2015.

[9] W. Lin, J. Huang, C. Y. Suen, and L. Yang, "A feature extraction model based on discriminative graph signals," *Expert Systems with Applications*, vol. 139, p. 112861, 2020.

[10] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. Springer Science & Business Media, 2012.

[11] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using joint mutual information maximisation," *Expert Systems with Applications*, vol. 42, pp. 8520–8532, 2015.

[12] L. Hu, W. Gao, K. Zhao, P. Zhang, and F. Wang, "Feature selection considering two types of feature relevancy and feature interdependency," *Expert Systems with Applications*, vol. 93, pp. 423–434, 2018.

[13] T. Zhang, B. Ding, X. Zhao, and Q. Yue, "A fast feature selection algorithm based on swarm intelligence in acoustic defect detection," *IEEE Access*, vol. 6, pp. 28 848–28 858, 2018.

[14] 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 transactions on computational biology and bioinformatics*, vol. 13, pp. 971–989, 2015.

[15] J. Miao and L. Niu, "A survey on feature selection," *Procedia Computer Science*, vol. 91, pp. 919–926, 2016.

[16] M. Luo, X. Chang, L. Nie, Y. Yang, A. G. Hauptmann, and Q. Zheng, "An adaptive semisupervised feature analysis for video semantic recognition," *IEEE Transactions on Cybernetics*, vol. 48, no. 2, pp. 648–660, 2017.

[17] R. Zhou, X. Chang, L. Shi, Y.-D. Shen, Y. Yang, and F. Nie, "Person reidentification via multi-feature fusion with adaptive graph learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1592–1601, 2019.

[18] K. Zhan, X. Chang, J. Guan, L. Chen, Z. Ma, and Y. Yang, "Adaptive structure discovery for multimedia analysis using multiple features," *IEEE Transactions on Cybernetics*, vol. 49, no. 5, pp. 1826–1834, 2018.

[19] W. J. Krzanowski, "Selection of variables to preserve multivariate data structure, using principal components," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 36, pp. 22–33, 1987.

[20] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proceedings of the 18th International Conference on Neural Information Processing Systems*, 2006, pp. 507–514.

[21] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "L2,1-norm regularized discriminative feature selection for unsupervised learning," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011, pp. 1589–1594.

[22] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012, pp. 1026–1032.

[23] M. Qian and C. Zhai, "Robust unsupervised feature selection," in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013, pp. 1621–1627.

[24] L. Du and Y.-D. Shen, "Unsupervised feature selection with adaptive structure learning," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 209–218.

[25] F. Nie, W. Zhu, and X. Li, "Unsupervised feature selection with structured graph optimization," in *Thirtieth AAAI conference on artificial intelligence*, 2016, pp. 1302–1308.

[26] X. Li, H. Zhang, R. Zhang, Y. Liu, and F. Nie, "Generalized uncorrelated regression with adaptive graph for unsupervised feature selection," *IEEE Transactions on Neural Networks*, vol. 30, pp. 1587–1595, 2019.

[27] M. Luo, F. Nie, X. Chang, Y. Yang, A. G. Hauptmann, and Q. Zheng, "Adaptive unsupervised feature selection with structure regularization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 4, pp. 944–956, 2017.

[28] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[29] A. Y. Ng, M. I. Jordan, Y. Weiss *et al.*, "On spectral clustering: Analysis and an algorithm," in *Proceedings of the 15rd International Conference on Neural Information Processing Systems*, 2002, pp. 849–856.

[30] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[31] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, 2006, pp. 126–135.

[32] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proceedings of the 2013 SIAM International Conference on Data Mining*, 2013, pp. 252–260.

[33] X. Zhang, L. Zong, X. Liu, and J. Luo, "Constrained clustering with nonnegative matrix factorization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 7, pp. 1514–1526, 2015.

[34] 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 *Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2016.

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

[36] Y. Wang, L. Wu, X. Lin, and J. Gao, "Multiview spectral clustering via structured low-rank matrix factorization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4833–4843, 2018.

[37] S. Wang, W. Pedrycz, Q. Zhu, and W. Zhu, "Subspace learning for unsupervised feature selection via matrix factorization," *Pattern Recognition*, vol. 48, pp. 10–19, 2015.

[38] D. Cai, X. He, J. Han, and T. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 1548 – 1560, 2011.

[39] D. Cai, X. He, and J. Han, "Locally consistent concept factorization for document clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, pp. 902–913, 2010.

[40] D. Han and J. Kim, "Unsupervised simultaneous orthogonal basis clustering feature selection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5016–5023.

[41] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *Proceedings of the 2005 SIAM international conference on data mining*, 2005, pp. 606–610.

[42] S. Wang and H. Wang, "Unsupervised feature selection via low-rank approximation and structure learning," *Knowledge Based Systems*, vol. 124, pp. 70–79, 2017.

[43] R. Hu, X. Zhu, D. Cheng, W. He, Y. Yan, J. Song, and S. Zhang, "Graph self-representation method for unsupervised feature selection," *Neurocomputing*, vol. 220, pp. 130–137, 2017.

[44] T. Viklands, "Algorithms for the weighted orthogonal procrustes problem and other least squares problems," Ph.D. dissertation, Datavetenskap, 2006.

[45] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint l2,1-norms minimization," in *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, 2010, pp. 1813–1821.

[46] J. Huang, F. Nie, and H. Huang, "A new simplex sparse learning model to measure data similarity for clustering," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015, pp. 3569–3575.

[47] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[48] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

XIAOCHANG LIN received his master degree in automation from the school of aerospace engineering in Xiamen University in 2020. His research interests include feature selection and semi-supervised graph clustering.

JIEWEN GUAN received his bachelor degree in electronic information engineering from the school of information engineering in Zhejiang University of Technology in 2019. Now he is a master candidate in Xiamen University. His research interests include feature selection and recommendation system.

BILIAN CHEN received her Ph.D. degree from The Chinese University of Hong Kong in 2012. Now she is an associate professor in Xiamen University. Her research interests include machine learning, optimization theory and recommendation system. Her publications appear in SIAM Journal on Optimization, Journal of Global Optimization, Information Sciences, and so on.

YIFENG ZENG received his Ph.D. degree from National University of Singapore, Singapore, in 2006. He is a Professor with the Department of Computer and Information Science, Northumbria University, Newcastle-upon-Tyne, U.K. His research interests include intelligent agents, decision making, social networks, and computer games. Most of his publications appear in the most prestigious international academic journals and conferences, including Journal of Artificial Intelligence Research, Journal of Autonomous Agents and Multi-Agent Systems, International Conference on Autonomous Agents and Multi-Agent Systems, International Joint Conference on Artificial Intelligence, and Association for the Advancement of Artificial Intelligence.

From: **IEEE Transactions on Neural Networks and Learning Systems** onbehalfof@manuscriptcentral.com
Subject: [TNNLS EiC] Review Decision (minor revision): TNNLS-2020-P-15360.R1
Date: 31 March 2021 at 18:31
To: blchen@xmu.edu.cn
Cc: ieeetnnls@gmail.com, 13263396762@163.com, 23220191151283@stu.xmu.edu.cn, blchen@xmu.edu.cn, yifeng.zeng@northumbria.ac.uk

CAUTION: This email originated from outside of the University. Do not click links or open attachments unless you recognise the sender and know the content is safe.

31-Mar-2021

Dear Dr. Chen:

Manuscript ID TNNLS-2020-P-15360.R1 entitled "Unsupervised Feature Selection via Orthogonal Basis Clustering and Local Structure Preserving" which you submitted to the IEEE Transactions on Neural Networks and Learning Systems, has been reviewed. The comments of the reviewer(s) are included at the bottom of this email-letter.

I am pleased to inform you that your paper has received a positive recommendation making it conditionally acceptable for publication subject to modifications indicated in the reviews. The comments and suggestions of the Reviewers and Associate Editor should be rigorously addressed.

I invite you to respond to the reviewer(s)' comments and revise your manuscript. The revised manuscript will be reviewed by the Editor-in-Chief and possibly by the Associate Editor and some of the Reviewers. Please note that if you do not fully address all the comments from the reviewer(s) and Associate Editor, it is still possible that you paper may be rejected.

Please follow the following procedure for resubmitting your manuscript. First, revise your manuscript using a word processing program (e.g., Latex or Word) and save it on your computer. Once you have revised your manuscript, log into https://eur02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fmc.manuscriptcentral.com%2Ftnnls&amp;data=04%7C01%7Cyifeng.zeng%40northumbria.ac.uk%7C85472c23a4a14d95475b08d8f46ac52a%7Ce757cfdd1f354457af8f7c9c6b1437e3%7C0%7C0%7C637528086709362671%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&amp;sdata=6v1yeRgoYxAydmN%2Fa5jUbCQYheBemJo%2BG0henII8LGI%3D&amp;reserved=0 and enter your Author Center, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been appended to denote a revision. In addition to your revised manuscript, please also include a point-by-point "Summary of Changes" that describes the changes and explains how individual comments and suggestions of the Reviewers were incorporated into the revised manuscript. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the reviewer(s). Please submit one PDF file containing your revised manuscript and the Summary of Changes.

*** Please submit your manuscript in singled-spaced, double column, standard IEEE published format. See detailed author instructions at the IEEE Author Center Journals:

https://eur02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fjournals.ieeeauthorcenter.ieee.org%2F&amp;data=04%7C01%7Cyifeng.zeng%40northumbria.ac.uk%7C85472c23a4a14d95475b08d8f46ac52a%7Ce757cfdd1f354457af8f7c9c6b1437e3%7C0%7C0%7C637528086709362671%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&amp;sdata=0bFk%2BDcqJJWeJj5hk9kLIS8kHLGid%2Bc7wJktO0Q3DpE%3D&amp;reserved=0

*** Please observe the following page limits: 10 pages for a full paper, 15 pages for a survey paper, 6 pages for a brief paper, and 3 pages for a comment paper, to avoid over-length page charges. Also note that we only accept PDF files. See detailed IEEE TNNLS information at the following link:

https://eur02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fcis.ieee.org%2Fpublications%2Ft-neural-networks-and-learning-systems&amp;data=04%7C01%7Cyifeng.zeng%40northumbria.ac.uk%7C85472c23a4a14d95475b08d8f46ac52a%7Ce757cfdd1f354457af8f7c9c6b1437e3%7C0%7C0%7C637528086709362671%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&amp;sdata=wgyu5aLzwMf%2BLzPI4%2FnHCn2b30Laau%2BtkaU9OwGTqDY%3D&amp;reserved=0

Because we are trying to facilitate timely publication of manuscripts submitted to the IEEE Transactions on Neural Networks and Learning Systems, your revised manuscript should be uploaded as soon as possible, but no later than six weeks from the date of this email-letter. If it is not possible for you to submit your revision in a reasonable amount of time, we may have to consider your paper as a new submission.

** VERY IMPORTANT **
When you submit your revised paper, please make sure that you remove all old files related to this paper so that the TNNLS editorial office and our reviewers will only see the newest edition of your manuscript. It has happened a few times before that reviewers got confused about the revised paper and the old paper since both of them were in the system. One single PDF file containing your revised paper and a list of summary of changes is what we need. All other files should be removed from your author’s account. Please do not submit WORD files.

Thank you for submitting your manuscript to the IEEE Transactions on Neural Networks and Learning Systems and I look forward to receiving your revision.

Sincerely,

Haibo He
Editor-in-Chief
IEEE Transactions on Neural Networks and Learning Systems

Robert Haas Endowed Chair Professor
Department of Electrical, Computer, and Biomedical Engineering
University of Rhode Island
Kingston, RI 02881, USA

Email: ieeetnnls@gmail.com
https://eur02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fcis.ieee.org%2Fpublications%2Ft-neural-networks-and-learning-systems&amp;data=04%7C01%7Cyifeng.zeng%40northumbria.ac.uk%7C85472c23a4a14d95475b08d8f46ac52a%7Ce757cfdd1f354457af8f7c9c6b1437e3%7C0%7C0%7C637528086709362671%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&amp;sdata=wgyu5aLzwMf%2BLzPI4%2FnHCn2b30Laau%2BtkaU9OwGTqDY%3D&amp;reserved=0

BEGINNING OF COMMENTS TO THE AUTHOR(S)
++

Recommended Decision by Associate Editor: Recommendation #1: Minor Revision

Comments to Author(s) by Associate Editor:
Associate Editor
Comments to the Author:
The paper was significantly improved and now it requires only some final editorial effort to improve readability and remove remaining typos

+++++++++++++++++++++

Individual Reviews:

Reviewer(s)' Comments to Author(s):

Reviewer: 1

Comments to the Author
The authors have carefully addressed all of my concerns. The paper is in good shape now. I would like to accept this paper as it is.

Reviewer: 2

Comments to the Author
The authors have well addressed the comments by discussing the suggested related work. Overall, I am basically satisfied with the revisions. While suggesting discuss more related work published on TNNLS, such as

- Multiview Spectral Clustering via Structured Low-Rank Matrix Factorization. IEEE TNNLS, 2018.

I recommend the minor revision for another proof reading and address the above.

++
END OF COMMENTS TO THE AUTHOR(S)