IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 16, 2019, accepted November 23, 2019, date of publication November 26, 2019,
date of current version December 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2955971

Graph-Based Clustering via Group Sparsity and Manifold Regularization

JIANYU MIAO[1], TIEJUN YANG[1], JUNWEI JIN[1], AND LINGFENG NIU[2,3]
[1]College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China
[2]School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China
[3]Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China

Corresponding author: Lingfeng Niu (niulf@ucas.ac.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 11671379 and Grant 11331012 and in part by the High-Level Talent Fund Project of the Henan University of Technology under Grant 31401155.

ABSTRACT Clustering refers to the problem of partitioning data into several groups according to the predefined criterion. Graph-based method is one of main clustering approaches and has been shown impressive performance in many literatures. The core issue of graph-based clustering is how to construct a good adjacency graph. A large number of works employ the sparse representation of data as the similarity measure by ℓ_1 regularization. However, due to the flat nature of the ℓ_1 norm, such methods solve the sparse representation of each data point individually, which do not take into account the global structure of data. To exploit the global and essential structure in data, in contrast to existing methods, we propose to learn a graph with group sparsity. To incorporate more information into the graph, we also use the manifold regularization with adaptive similarity during the process of group sparse self-representation. The resulting model is handled by Alternating Direction Method of Multipliers (ADMM). Further, we employ Iterative Re-weighted Least Squares (IRLS) algorithm and threshold operator to solve the ADMM subproblems. Experimental results on real-world datasets demonstrate the superiority of our method compared to the competing clustering methods.

INDEX TERMS Clustering, graph, group sparsity.

I. INTRODUCTION

As a fundamental and important technique in machine learning and data mining, clustering aims to partition data points into several groups such that the data within the same group similar while data in different groups are dissimilar as much as possible [1]. It has been widely used in many domains, such as biological engineering, image processing, and social network [2]–[4]. Over the past decades, clustering algorithm has been well studied and a number of methods have been proposed so far [5]–[7]. The existing clustering algorithms can be roughly categorized into two families: density-based approaches, such as K-means and Expectation Maximization (EM) clustering [8], [9], and graph-based approaches, such as spectral clustering [10], normalized cut [11] and min-max cut [12].

Due to the good performance and simplicity, graph-based clustering methods [13]–[18] have gained considerable attention from a variety of communities, which usually adopt a two-step strategy. Specifically, a weighted undirected graph is first constructed, where the data points are the nodes and the affinities are the weights. The affinities between the data can be obtained via several similarity measures, such as RBF kernel function, binary function and dot-product function, representation coefficient obtained by sparse optimization problem, and so on. The data clustering is then accomplished by spectral or graph theoretic optimization procedures. During the process, it should be emphasized that how to build a good similarity matrix is the most crucial step.

Previous works [19], [20] get the affinity matrix by computing the Singular Value Decomposition (SVD) of the data matrix, which is sensitive to the noise and outlier in the data. Wang *et al.* [21] focused on learning distance measure by exploiting a graph structure of data samples, where an input similarity matrix can be improved through a propagation of graph random walk. Recently, with the development of sparse regularization, both theoretical and empirical studies have suggested that sparsity is one of the intrinsic

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

properties of real-world data. This motivates a large number of researchers to develop clustering models with sparse representation [14], [22]–[25]. Cheng *et al.* [23] employed the ℓ_1 norm to build the sparse graph weight matrix, which has been shown to be capable of finding data-adaptive neighborhood for the graph construction. Elhamifar and Vidal [22] have proposed sparse subspace clustering (SSC), which represents each data point as a linear or affine combination of the remaining data points and most combination coefficients are zero or close to zero. Essentially, SSC minimizes the ℓ_0 of representation coefficient, which denotes the number of non-zero elements in a vector. Based on SSC, several studies have been further developed to handle the noise [26] and outliers [27] in the data. Using the kernel trick, Patel and Vidal [28] extended SSC to non-linear manifolds, and shown that sparse representation obtained by non-linear mapping could obtain better performance than state-of-the-art methods. Also, Lerman *et al.* [29] proved that under certain conditions the multiple subspace structures can be exactly recovered via $\ell_p (p \leq 1)$ minimization.

It can be observed that most existing graph-based clustering methods use the ℓ_1 norm to promote the sparsity of coefficients, which have shown the promising performance. Moreover, the resulting minimization can be easily handled by a variety of techniques with global solutions. However, due to the flat nature of the ℓ_1 norm, the sparse representation of each data vector is found individually, which means that no global constraint are enforced on the solution. Therefore, this type of methods may be inaccurate at capturing global structure of data, or even are not able to exploiting the global structure. To address this issue, we propose to use the group sparse regularization to promote the sparsity of the coefficient matrix, which can exploit the global geometric structure and essential structure in data effectively and precisely. As a result, the performance can be improved. Besides, recent years have witnessed a great success of manifold learning in high-dimensional data analysis, which aims to find low-dimensional manifold embedding from original high-dimension data. Under the circumstance that if two points are similar, their low-dimensional embeddings are also similar, we integrate the manifold regularization into the group spare self-representation for learning a sparse graph.

In summary, we highlight our main contributions as follows,

- We propose a unified model for graph-based clustering. To exploit the global and essential structure, we propose to use group sparsity for self-representation. To well preserve the structure of the original data space, we utilize the manifold regularization.
- To minimize the proposed model, we derive the Alternating Direction Method of Multipliers (ADMM)-based optimization algorithm. Iterative Re-weighted Least Squares (IRLS) algorithm and thresholding operator are employed to obtain the solution of subproblems.
- We compare our method with the state-of-the-art clustering methods on several real-world datasets. The results demonstrate the effectiveness of our proposed method in terms of clustering accuracy, normalized mutual information and adjusted rand index.

The rest of this paper is organized as follows. In Section II, we provide some existing related works. In Section III, we present the proposed method. Section III-A introduces the group spare graph model for data clustering. In Section III-B, we investigate how to optimize the proposed model. The clustering details and convergence behavior are given in Section III-C and III-D, respectively. In Section IV, we compare the performance of the proposed method with the state-of-the-art on the four real-world datasets. Finally, Section V concludes the paper.

II. RELATED WORK

In this section, we give a brief overview from both Graph-based clustering approaches and group sparsity. Before we begin, we list notations to be used in this paper in Table 1.

A. GRAPH-BASED CLUSTERING APPROACHES

In recent years, spectral clustering has become one of the most popular modern clustering approaches with a huge number of variants being developed, whose main tools are the graph Laplacian matrices, including normalized and unnormalized graph Laplacian [1], [30]–[32]. To improve the performance of spectral clustering, two aspects have been considered by researchers. On one hand, one can construct a good or robust affinity matrix by using the standard spectral algorithms. Lee *et al.* [33] proposed an alternative approach to produce matrices with block-diagonal structures. On the other hand, many researchers focus on improving the clustering result when fixing the way of generating the data affinity matrix. Yan *et al.* [34] formulated spectral clustering as a semi-definite programming (SDP), which could find the closest doubly stochastic approximation to the affinity matrix more accurately.

As an extension of spectral clustering, sparse spectral clustering [1], [35] employs sparse regularization to enhance the robustness of spectral clustering. To promote the sparsity of representation coefficient, the ℓ_0 is more desired. However, such resulting optimization problems are in general non-convex and NP-hard. Yang *et al.* [24] proposed the ℓ_0 based clustering model solved by the Proximal Gradient Descent (PGD) method, which admitted a sub-optimal solution with theoretical guarantee. One of the well-known strategies is to replace the ℓ_0 by the convex ℓ_1 norm [14], [22], [25].

Based on Low-Rank Representation (LRR), Wang *et al.* [36], [38] and Wang and Wu [37] proposed spectral clustering models, where the ℓ_1 regularization was used to address the noise in the data. To better discover the latent group structure of data, Yin *et al.* [39] developed a pairwise sparse subspace representation model based on some prior information for clustering. More recently,

TABLE 1. The list of notations and their definitions in this paper.

Symbol	Description
x	a column vector in $\Re^n$
x_i	the i-th entry of vector x
$\|x\|_0$	the ℓ_0 of vector $x \in \Re^n$ which is equal to $\sum_{i=1}^n \|x_i\|^0$, where for a scalar x, $\|x\|^0 \triangleq \begin{cases} 1, & x_i \neq 0; \\ 0, & x_i = 0. \end{cases}$
$\|x\|_2$	the euclidean norm of vector $x \in \Re^n$ which is equal to $\sqrt{x^{\mathrm{T}}x}$
$\|x\|_p$	the ℓ_p norm of vector $x \in \Re^n$ which is equal to $\left(\sum_{i=1}^n \|x_i\|^p\right)^{1/p}$
X	a matrix in $\Re^{m\times n}$
X^{T}	the transpose of X
X_{ij}	the entry at the i-th row and the j-column of matrix X
$\|X\|_F$	the Frobenius norm of matrix $X \in \Re^{m\times n}$ which is equal to $\sqrt{\sum_{i=1}^m \sum_{j=1}^n X_{ij}^2}$
$\|X\|_{2,1}$	the $\ell_{2,1}$ norm of matrix $X \in \Re^{m\times n}$ which is equal to $\sum_{i=1}^m \sqrt{\sum_{j=1}^n X_{ij}^2}$
$\mathrm{Tr}(X)$	the trace of squire matrix $X \in \Re^{m\times m}$ which is equal to $\sum_{i=1}^m X_{ii}$,
$\langle X, Y\rangle$	the Euclidean inner product between two same scale matrix X and Y which is equal to $\sum_{i,j} X_{ij}Y_{ij} = \mathrm{Tr}\left(X^{\mathrm{T}}Y\right)$
I	the identity matrix of compatible size

Brbić and Kopriva [40] introduced the nonconvex generalized Minimax-Concave Penalty (MCP) and Schatten-0 quasi norm for low-rank sparse subspace clustering. Based on the convex hull of the fixed rank projection matrices, Lu *et al.* [1] proposed a novel convex relaxation to alleviate the nonconvex sparse spectral clustering model from the computational issue.

B. GROUP SPARSITY

Suppose that the features are independent and the structures of features are ignored completely, a variety of sparse regularizations have been proposed, including lasso (ℓ_1 norm), adaptive lasso, fused lasso, trace lasso and elastic net [41]. However, in practical applications, the features have some essential structures, such as disjoint groups [42], overlapping groups [43], and graphs [44]. Integrating some priori knowledge of the feature structures into model building can help identify the important features. Accordingly, the sparsity can be obtained by group lasso, overlapping group lasso, and graph lasso. As an extension of the group lasso, the sparse group lasso [45] combines both lasso and group lasso, which can produce a solution with simultaneous between- and within- group sparsity.

Among them, the $\ell_{2,1}$-norm is one of the most popular one, which is defined as the ℓ_1 norm of the vector containing of the ℓ_2 norm of the matrix rows on a matrix. It was first introduced in [46] as the rotational invariant of the ℓ_1 norm. By its definition, the $\ell_{2,1}$-norm encourages row sparsity, i.e., it enforces entire row of the matrix to have zero elements. It has been successfully used in feature selection [47], [48], dictionary learning [49], multi-task learning [50], [51], and multi-class classification [52], [53]. Nie *et al.* [47] developed a feature selection model via the $\ell_{2,1}$ norm joint minimization on the loss function and spare regularization. More feature selection works using the $\ell_{2,1}$ norm as sparse regularization can be found in [41]. Based on the $\ell_{2,1}$ norm, Cai *et al.* and Xiang *et al.* improved the Support Vector Machine (SVM) and least squares regression for multi-class classification. More recently, group sparsity has also been applied into deep neural works. Yoon and Hwang [54] combined group and exclusive sparsity as a regularization to enforce sparsity, by utilizing the sharing and competing relationships among various network weights.

III. THE PROPOSED METHODOLOGY

A. MODEL FORMULATION

1) GROUP SPARSE SELF-REPRESENTATION

Suppose that we have a collection of data points $\{x_1, x_2, \cdots, x_n\}$, where each sample x_i lies in $\Re^d$ Euclidean space. Let $X = [x_1, x_2, \cdots, x_n] \in \Re^{d\times n}$ be the data matrix. Self-representation learning aims to represent each sample as the linear combination of its most relevant samples. In addition, the sparsity of the representation coefficient is more desired, that is most combination coefficients are zero or close to zero. Mathematically, the sparse self-representation can be typically formulated as the following optimization problem,

$$\min_A \ loss(X, XA) + \alpha\Omega(A), \tag{1}$$

where the first term $loss(\cdot,\cdot)$ is the data fidelity term which encourages an accurate representation, the second term $\Omega(A)$ is the sparse regularization term which enforces the sparsity on the coefficient matrix $A = [a_1, a_2, \cdots, a_n] \in \Re^{n\times n}$, and α is a positive regularization parameter which balances these two terms in the formulation.

One commonly used loss function is the least squares loss, which takes advantage over several other loss functions due to its differentiability in the optimization of the resulting problem. However, it is sensitive to the outlier and noise in data points. To deal with this issue, in this paper, we use the

$\ell_{2,1}$ norm as the loss function. Then, (1) becomes,

$$\min_{A} \|X - XA\|_{2,1} + \alpha\Omega(A). \tag{2}$$

In (2), we expect that only few numbers of a_i are non-zeros, which implies that each sample can be represented by as few samples as possible. As a result, the samples corresponding to those non-zero rows are selected to regress the original data to its low-dimensional representation. To the end, different from existing works using the ℓ_1 minimization, we employ the group sparsity to exploit global and essential structure of the data. As stated earlier, the $\ell_{2,1}$ norm is the most common used and has shown the promising performance. Therefore, we use the $\ell_{2,1}$ norm to promote the group sparsity and obtain the $\ell_{2,1}$ norm regularized sparse self-representation formulation as follows,

$$\min_{A} \|X - XA\|_{2,1} + \alpha\|A\|_{2,1}. \tag{3}$$

It can be easily observed that the sparse coefficients represent the contribution of each data to the reconstruction of other data, which can be measured by the ℓ_2 norm. Due to the existence of the $\ell_{2,1}$ norm sparse regularization, the above formulation will lead to a small number of non-zero rows in representation matrix.

2) MANIFOLD REGULARIZATION WITH ADAPTIVE SIMILARITY

Let $S \in \Re^{n\times n}$ be a KNN adjacency matrix on the data, and $S_{ij} = 1$ if and only if either x_i is among the K-nearest neighbors of x_j. The manifold learning aims to find the low-dimensional embedding of each data in accordance with the manifold assumption by minimize the following regularization term,

$$\min_{a_i} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}\|a_i - a_j\|_2^2, \tag{4}$$

where a_i is the low-dimensional representation of data point x_i, for $i = 1, 2, \cdots, n$. With some simple linear algebra, we have the following,

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}\|a_i - a_j\|_2^2 \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}\|a_i\|_2^2 - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} 2S_{ij}\langle a_i, a_j\rangle \\
&\quad + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}\|a_j\|_2^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}\|\alpha_i\|_2^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}\langle a_i, a_j\rangle \\
&= \sum_{i=1}^{n} a_i^{\mathrm{T}} D a_i - \sum_{i=1}^{n} a_i^{\mathrm{T}} S a_i \\
&= \mathrm{Tr}(A^{\mathrm{T}} L_S A),
\end{aligned} \tag{5}$$

where $D = \sum_{j=1}^{n} S_{ij}$ is the degree matrix, and $L_S = D - S$ is the Laplacian matrix. Hence, (4) can be rewritten into,

$$\min_{A\neq 0} \mathrm{Tr}(A^{\mathrm{T}} L_S A). \tag{6}$$

The constraint $A \neq 0$ is to avoid the trivial solution. As we know, although the similarity matrix S constructed by KNN encourages the local smoothness of the low-dimensional representation in a neighborhood of each data point, it neglects the data that are far away from each other in the original data space. Inspired by the great success of the local smoothness of the low-dimensional embedding in clustering [14], we consider the following problem,

$$\begin{aligned}
&\min_{A} \mathrm{Tr}(A^{\mathrm{T}} L_W A) \\
&s.t.\ W = (|A| \circ S + |A^{\mathrm{T}}| \circ S^{\mathrm{T}})/2,
\end{aligned} \tag{7}$$

where the notation $|\cdot|$ stands for the absolute value of each element in matrix, and $\circ$ denotes the Hadamard product between two matrices with the same scale. The equality constraint would result in the low-dimensional varying smoothly along the geodesics of the data manifold through the graph Laplacian.

By combining the self-representation with the $\ell_{2,1}$ norm (3) and manifold regularization with adaptive similarity (7), our proposed model can be formulated as follows,

$$\begin{aligned}
&\min_{A,W} \|X - XA\|_{2,1} + \alpha\|A\|_{2,1} + \beta\mathrm{Tr}(A^{\mathrm{T}} L_W A) \\
&s.t.\ W = (|A| \circ S + |A^{\mathrm{T}}| \circ S^{\mathrm{T}})/2,
\end{aligned} \tag{8}$$

where α and β are two positive regularization parameters. The joint minimization will admit a sparse solution along the row, which can identify the most relevant features for each sample. Instead of the original data matrix, we perform the clustering on the sparse representation coefficient matrix.

B. OPTIMIZATION

In this subsection, we present the optimization algorithm of our proposed model. Before proceeding the process, we give another optimization problem, which is equivalent to (8) but can be solved more easily.

Lemma 1: Problem (8) is equivalent to the following constrained optimization problem,

$$\begin{aligned}
&\min_{W,A} \|X - XA\|_{2,1} + \alpha\|A\|_{2,1} + \beta\mathrm{Tr}(A^{\mathrm{T}} L_{S\circ|W|} A) \\
&s.t.\ W = A
\end{aligned} \tag{9}$$

Proof: To prove the equivalence, we make the following the calculation,

$$\begin{aligned}
\mathrm{Tr}(A^{\mathrm{T}} L_{S\circ|W|} A) &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}|W_{ij}|\|a_i - a_j\|_2^2 \\
&= \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}|W_{ij}|\|a_i - a_j\|_2^2 \\
&\quad + \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} S_{ji}|W_{ji}|\|a_j - a_i\|_2^2 \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{S_{ij}|W_{ij}| + S_{ji}|W_{ji}|}{2}\|a_i - a_j\|_2^2.
\end{aligned} \tag{10}$$

According to the equality constraint $W = A$ in (9) and the above equation, we arrive at that these two optimization problems are equivalent to each other. □

Clearly, the $|\cdot|$ operator and the transpose operator are removed from the equality constraint in (8), which leads to be more tractable than its preliminary form. We will focus on the optimization of problem (9), which is a convex optimization with global solutions. However, it involves the non-smooth term $\ell_{2,1}$-norm and has no closed-form solutions. In this paper, we propose to use ADMM, which achieves globally optimal solution for a class of convex problems. To this end, we first formulate the augmented Lagrangian of problem (9) as,

$$\mathcal{L}(A, W, \Sigma) = \|X - XA\|_{2,1} + \alpha\|A\|_{2,1} + \langle \Sigma, W - A\rangle + \beta \mathrm{Tr}(A^T L_{S\circ|W|}A) + \frac{\mu}{2}\|W - A\|_F^2, \quad (11)$$

where $\mu > 0$ is the penalty parameter. According to [55], ADMM consists of the following iterations,

$$A^{k+1} := \arg\min_A \mathcal{L}_\rho\left(A, W^k, \Sigma^k\right); \quad (12a)$$

$$W^{k+1} := \arg\min_W \mathcal{L}_\rho\left(A^{k+1}, W, \Sigma^k\right); \quad (12b)$$

$$\Sigma^{k+1} := \Sigma^k + \rho\left(A^{k+1} - W^{k+1}\right), \quad (12c)$$

where the superscript is the iteration counter. In what follows, we will describe the details of (12a) and (12b). For simplicity, we omit the superscript in (12a) and (12b).

1) UPDATING A

With W and Σ fixed in (12a), the ADMM subproblem with respect to A is reduced to,

$$\min_A \|X - XA\|_{2,1} + \alpha\|A\|_{2,1} + \beta \mathrm{Tr}(A^T L_{S\circ|W|}A) + \langle \Sigma, W - A\rangle + \frac{\mu}{2}\|W - A\|_F^2. \quad (13)$$

As seen from (13), although the objective function is convex which admits global solutions, it is non-smooth, making it difficult to be solved directly. Here, we utilize the Iterative Reweighted Least-Squares (IRLS) [56] algorithm to solve the ADMM subproblem, which is in an iterative way. To the end, we first construct two diagonal matrices $G_x^k = diag(G_x^1, G_x^2, \cdots, G_x^n)$ and $G_a^k = diag(G_a^1, G_a^2, \cdots, G_a^n)$ at the iteration k with,

$$G_x^i = \frac{1}{2\|x_i - Xa_i\|_2} \quad (14)$$

and

$$G_a^i = \frac{1}{2\|a_i\|_2}, \quad (15)$$

for $i = 1, 2, \cdots, n$, respectively. Then A^{k+1} can be updated by solving the following weighted least squares problem,

$$A^{k+1} := \arg\min \mathrm{Tr}((X - XA)^T G_x^k (X - XA)) + \alpha \mathrm{Tr}(A^T G_a^k A) + \beta \mathrm{Tr}(A^T L_{S\circ|W|}A) + \langle \Sigma, W - A\rangle + \frac{\mu}{2}\|W - A\|_F^2. \quad (16)$$

For the above unconstrained optimization problem, we can get the derivative and then set it to zero. Then, we have,

$$2X^T G_x^k XA - 2X^T G_x^k X + 2\alpha G_a^k A + 2\beta L_{S\circ|W|}A - \Sigma + \mu(A - W) = 0. \quad (17)$$

The closed-form solution can be given by

$$A^{k+1} = Y^{-1}(2X^T G_x^k X + \Sigma + \mu W), \quad (18)$$

where $Y = (2X^T G_x^k X + 2\alpha G_a^k + 2\beta L_{S\circ|W|} + \mu I)$. To obtain the next iteration point, we need to update G_k^x and G_k^a with Eqs.(14) and (15) based on A^{k+1}, respectively. The whole IRLS procedure is summarized in Algorithm 1.

Algorithm 1 IRLS for ADMM Subproblem (13)

Input: data matrix X and regularization parameter α

1: Initialize $k = 0$ and $A^0 = 0$
2: **repeat**
3: Calculate G_x^k and G_a^k with Eqs.(14) and (15), respectively.
4: Update A^{k+1} with Eq.(18)
5: $k := k + 1$
6: **until** The stopping criterion is satisfied

Output: The optimal solution A^*

2) UPDATING W

To update W, we fix variables A and Σ in (12b), and remove irrelevant terms that are irrelevant of W. Then the ADMM subproblem with respect to W becomes,

$$\min_W \beta \mathrm{Tr}(A^T L_{S\circ|W|}A) + \langle \Sigma, W - A\rangle + \frac{\mu}{2}\|W - A\|_F^2, \quad (19)$$

which is equivalent to,

$$\min_W \beta \mathrm{Tr}(A^T L_{S\circ|W|}A) + \frac{\mu}{2}\left\|W - \left(A - \frac{\Sigma}{\mu}\right)\right\|_F^2. \quad (20)$$

Recall that,

$$\mathrm{Tr}(A^T L_{S\circ|W|}A) = \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n S_{ij}|W_{ij}|\|a_i - a_j\|_2^2. \quad (21)$$

According to Eq.(21) and the definition of Frobenius norm, we can further reformulate problem (20) into the following form,

$$\min_{W_{ij}} \sum_{i=1}^n\sum_{j=1}^n \left(\frac{\beta}{2}S_{ij}|W_{ij}|\|a_i - a_j\|_2^2 + \frac{\mu}{2}(W_{ij} - \left(A_{ij} - \frac{\Sigma_{ij}}{\mu}\right))^2\right), \quad (22)$$

which is equivalent to solving n^2 element-wise subproblem simultaneously. The subproblem with respect to (w.r.t.) W_{ij}, for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, n$, is

$$\min_{W_{ij}} \frac{\mu}{2}\left(W_{ij} - \left(A_{ij} - \frac{\Sigma_{ij}}{\mu}\right)\right)^2 + \left(\frac{\beta}{2}S_{ij}\|a_i - a_j\|_2^2\right)|W_{ij}|. \quad (23)$$

Although the objective function of (23) is not differentiable, we can still easily compute a simple closed-form solution to this problem by using sub-differential calculus. The closed form solution to the above problem is given by,

$$W_{ij} = Shrinkage_t\left(\frac{\mu A_{ij} - \Sigma_{ij}}{\mu}\right), \tag{24}$$

where $t = \beta S_{ij}\|\alpha_i - \alpha_j\|_2^2/2\mu$ and the shrinkage operator is defined as,

$$Shrinkage_{\mathcal{K}}(a) = \begin{cases} a - \mathcal{K}, & \text{if } a > \mathcal{K}; \\ 0, & \text{if } |a| \le \mathcal{K}; \\ a + \mathcal{K}, & \text{if } a < -\mathcal{K}. \end{cases} \tag{25}$$

In detail, W_{ij} can be obtained by,

$$W_{ij} = \max\left(0, \frac{|\mu A_{ij} - \Sigma_{ij}|}{\mu} - t\right) sign(\mu A_{ij} - \Sigma_{ij}). \tag{26}$$

We summarize the whole algorithm of our proposed method in Algorithm 2.

Algorithm 2 ADMM for the Proposed Model (8)

Input: data matrix X, hyper-parameters α and β, penalty parameter ρ

1: Construct the normalized graph Laplacian matrix L and set $k = 0$
2: **repeat**
3: Update A^k by Algorithm 1
4: Update W^k by (26)
5: Update Σ^k by (12c)
6: $k := k + 1$
7: **until** ADMM stopping criterion is satisfied

Output: The optimal solution A^*

C. CLUSTERING BASED ON A

After obtaining the solution to the proposed model (8), we will conduct the data clustering task based on the sparse coefficients. Without loss of generality, assume that A^* is the optimal coefficient matrix. To construct a graph, we first symmetrize the coefficient matrix A^*, i.e., $\widetilde{A} = (A^* + A^{*\mathrm{T}})/2$. Then we construct the normalized graph Laplacian matrix $L = \widetilde{D}^{-1/2}(\widetilde{D} - \widetilde{A})\widetilde{D}^{-1/2}$, where $\widetilde{D}$ is a diagonal matrix with $\widetilde{D}_{ii} = \sum_{j=1}^{n}\widetilde{A}_{ij}$. Compute the eigenvectors $e_1, e_2, \cdots, e_K$ of L corresponding to the largest K eigenvalues, and form the matrix $E = [e_1, e_2, \cdots, e_K] \in \Re^{n\times k}$ by stacking the eigenvectors in columns. Take each row of E as a point and clustering them into clusters via the K-means method. Consequently, the clustering results are obtained.

D. CONVERGENCE ANALYSIS

As stated previously, the optimization algorithm is in the framework of ADMM. The convergence of ADMM has been well established in [55]. Since the ADMM subproblem w.r.t W is exactly solved, we only need to analyze the convergence of IRLS, which solves the ADMM subproblem w.r.t A in an alternative way. Let us recall a useful inequality, which was introduced in [47].

Lemma 2 ([47]): For any non-zero vectors a and b, the following inequality holds,

$$\|a\|_2 - \frac{\|a\|_2^2}{2\|b\|_2} \le \|b\|_2 - \frac{\|b\|_2^2}{2\|b\|_2} \tag{27}$$

The following theorem indicates that the objective function shown in Eq.(16) is non-increasing in each iteration.

Theorem 1: Algorithm 1 will monotonically decrease the objective function value in ADMM subproblem (13) in each iteration.

Proof: Let $C(A) = \beta\mathrm{Tr}(A^{\mathrm{T}}L_{S\circ|W|}A) + \langle\Sigma, W - A\rangle + \frac{\mu}{2}\|W - A\|_F^2$. According to (16), we have,

$$\begin{aligned} A^{k+1} = \arg\min_A \mathrm{Tr}((X - XA)^{\mathrm{T}}G_x^k(X - XA)) \\ + \alpha\mathrm{Tr}(A^{\mathrm{T}}G_a^kA) + C(A), \end{aligned} \tag{28}$$

which gives,

$$\begin{aligned} &\mathrm{Tr}((X - XA^{k+1})^{\mathrm{T}}G_x^k(X - XA^{k+1})) \\ &\quad + \alpha\mathrm{Tr}(A^{k+1^{\mathrm{T}}}G_a^kA^{k+1}) + C(A^{k+1}) \\ &\le \mathrm{Tr}((X - XA^k)^{\mathrm{T}}G_x^k(X - XA^k)) \\ &\quad + \alpha\mathrm{Tr}(A^{k^{\mathrm{T}}}G_a^kA^k) + C(A^k). \end{aligned} \tag{29}$$

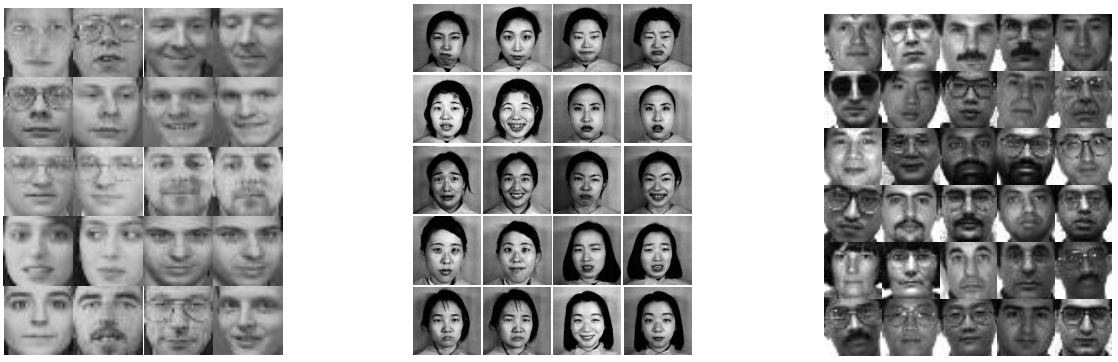
By the definition of the trace operator and the $\ell_{2,1}$ norm, and Eqs. (14) and (15), we have the following properties,

$$\begin{aligned} \mathrm{Tr}(A^{k+1^{\mathrm{T}}}G_a^kA^{k+1}) &= \sum_{i=1}^{n}\frac{\|\alpha_i^{k+1}\|_2^2}{2\|\alpha_i^k\|_2} \\ &= \|A^{k+1}\|_{2,1} \\ &\quad + \Big(\sum_{i=1}^{n}\frac{\|\alpha_i^{k+1}\|_2^2}{2\|\alpha_i^k\|_2} - \|\alpha_i^{k+1}\|_2\Big). \end{aligned} \tag{30}$$

Similarly, we also have,

$$\begin{aligned} \mathrm{Tr}(A^{k^{\mathrm{T}}}G_a^kA^k) &= \sum_{i=1}^{n}\frac{\|\alpha_i^k\|_2^2}{2\|\alpha_i^k\|_2} \\ &= \frac{1}{2}\|A^k\|_{2,1} \\ &= \|A^k\|_{2,1} \\ &\quad + \Big(\sum_{i=1}^{n}\frac{\|\alpha_i^k\|_2^2}{2\|\alpha_i^k\|_2} - \|\alpha_i^k\|_2\Big), \end{aligned} \tag{31}$$

$$\begin{aligned} &\mathrm{Tr}((X - XA^{k+1})^{\mathrm{T}}G_x^k(X - XA^{k+1}) \\ &= \sum_{i=1}^{n}\frac{\|x_i - X\alpha_i^{k+1}\|_2^2}{2\|x_i - X\alpha_i^k\|_2} \\ &= \|X - XA^{k+1}\|_{2,1} \\ &\quad + \sum_{i=1}^{n}\Big(\frac{\|x_i - X\alpha_i^{k+1}\|_2^2}{2\|x_i - X\alpha_i^k\|_2} - \|x_i - X\alpha_i^{k+1}\|_2\Big), \end{aligned} \tag{32}$$



(a)ORL (b)JAFFE (c)Yale

FIGURE 1. Sample images of ORL, JAFFE and Yale.

and

$$\begin{aligned}
&\mathrm{Tr}((X - XA^k)^{\mathrm{T}} G_x^k (X - XA^k)) \\
&= \sum_{i=1}^{n} \frac{\|x_i - X\alpha_i^k\|_2^2}{2\|x_i - X\alpha_i^k\|_2} \\
&= \frac{1}{2}\|X - XA^k\|_{2,1} \\
&= \|X - XA^k\|_{2,1} \\
&\quad + \sum_{i=1}^{n} \left(\frac{\|x_i - X\alpha_i^k\|_2^2}{2\|x_i - X\alpha_i^k\|_2} - \|x_i - X\alpha_i^k\|_2 \right). \qquad (33)
\end{aligned}$$

By substituting Eqs. (30), (31), (32), (33) into inequality (29), we have,

$$\begin{aligned}
&\|X - XA^{k+1}\|_{2,1} + \sum_{i=1}^{n} \left(\frac{\|x_i - X\alpha_i^{k+1}\|_2^2}{2\|x_i - X\alpha_i^k\|_2} - \|x_i - X\alpha_i^{k+1}\|_2 \right) \\
&\quad + \alpha\|A^{k+1}\|_{2,1} + \alpha \sum_{i=1}^{n} \left(\frac{\|\alpha_i^{k+1}\|_2^2}{2\|\alpha_i^k\|_2} - \|\alpha_i^{k+1}\|_2 \right) + C(A^{k+1}) \\
&\leq \|X - XA^k\|_{2,1} + \sum_{i=1}^{n} \left(\frac{\|x_i - X\alpha_i^k\|_2^2}{2\|x_i - X\alpha_i^k\|_2} - \|x_i - X\alpha_i^k\|_2 \right) \\
&\quad + \alpha\|A^k\|_{2,1} + \alpha \sum_{i=1}^{n} \left(\frac{\|\alpha_i^k\|_2^2}{2\|\alpha_t^i\|_2} - \|\alpha_t^i\|_2 \right) + C(A^k). \qquad (34)
\end{aligned}$$

Based on Lemma 2 and inequality (34), we can easily obtain the following inequality,

$$\begin{aligned}
&\|X - XA^{k+1}\|_{2,1} + \alpha\|A^{k+1}\|_{2,1} + C(A^{k+1}) \\
&\qquad \leq \|X - XA^k\|_{2,1} + \alpha\|A^k\|_{2,1} + C(A^k), \qquad (35)
\end{aligned}$$

which implies the result in the theorem. □

IV. EXPERIMENTS

In this section, to investigate the behavior of our proposed method, we carry out extensive experiments on the real-world databases. All of the experiments are implemented with Matlab R2018a on Windows 10 and the computer is deployed with CPU 3.60GHz and RAM 16GB.

TABLE 2. Datasets description.

Datasets	# of Features	# of Instances	# of Classes
Yale	1024	165	15
JAFFE	676	213	10
ORL	1024	400	40
oh15	3100	913	10
Tumors9	5726	60	9

A. DATASET

We use five real-world datasets in the experiments, including three image datasets (*i.e.*, ORL, Yale and JAFFE), one biomedical dataset (*i.e.*, Tumors9), and one text dataset (*i.e.*, oh15). Fig. 1 shows several sample images from the ORL, JAFFE and Yale database, respectively. We summarize the statistics of datasets in Table 2, and also provide the information of each dataset as follows.

- ORL[1]: The database contains 400 images of 40 distinct subjects. There are 10 images per subject. For some subjects, the images were taken at different times, varying the lighting, facial expressions and details. All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position.
- Yale[1]: The database consists of 165 grayscale images in GIF format of 15 individuals. Each subject has 11 different images. one per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink.
- JAFFE[1]: The database contains 213 images posed by 10 Japanese female models. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects. The size of each image is 26×26 pixels, with 256 gray levels per pixel.
- Tumors9 [57]: The dataset comes from a study of 9 human tumor types: NSCLC, colon, breast, ovary, leukemia, renal, melanoma, prostate, and CNS. There are in total 60 samples, each of which contains 5726 genes.

[1] http://featureselection.asu.edu/datasets.php

- oh15[2]: The text dataset contains 913 instances from 10 classes, including adenosine-diphosphate, aluminum, enzyme-activation, blood-coagulation-factors, blood-vessels, cell-movement, memory, staphylococcal-infections, leucine and uremia. Each sample is represented by 3100 feature words.

B. EVALUATION METRICS

To evaluate the performance, we consider the commonly used three metrics, i.e., clustering accuracy, normalized mutual information and adjusted rand index. The details about them are described as follows,

1) **Clustering Accuracy**(ACC): The clustering accuracy is defined as,

$$\mathrm{ACC} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(\mathrm{map}(l_i), y_i), \tag{36}$$

where l_i and y_i are the cluster label and ground truth label of x_i, respectively, n is the total number of data points. $\mathbb{I}(\cdot,\cdot)$ is the delta function, which indicates $\mathbb{I}(x, y) = 1$ if and only if $x = y$, and 0 otherwise. The permutation mapping function $\mathrm{map}(\cdot)$ maps each cluster label to the equivalent label. The best map can be obtained by the Kuhn-Munkres algorithm.

2) **Normalized Mutual Information** (NMI): The mutual information between X and Y is defined as,

$$\mathrm{MI}(X, Y) = \sum_{y\in Y}\sum_{x\in X} p(x, y)\log_2\left(\frac{p(x, y)}{p(x)p(y)}\right), \tag{37}$$

where $p(x)$ and $p(y)$ denote the marginal probability distribution functions of X and Y, respectively, and $p(x, y)$ is the joint probability distribution function of X and Y. Let $H(X)$ and $H(Y)$ be the entropies of $p(x)$ and $p(y)$, respectively. Then the normalized mutual information is given by,

$$\mathrm{NMI}(X, Y) = \frac{\mathrm{MI}(X, Y)}{\max(H(X), H(Y))} \tag{38}$$

3) **Adjusted Rand Index** (ARI): The adjusted rand index is defined as

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}]/\binom{n}{2}}{[\sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}]/2 - [\sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}]/\binom{n}{2}}, \tag{39}$$

where n_{ij} is the number of data points with true label i but they are assigned by j, n_i and n_j are the number of data points with label i and j, respectively.

C. EXPERIMENTAL RESULTS

We choose K-means and spectral clustering (SC) [10] methods as the baselines. Moreover, we consider non-negative matrix factorization (NMF) [58], [59] based method, which finds the cluster indicator by solving the NMF problem with the constraint of nonnegativity and orthogonality. To verify the effectiveness of $\ell_{2,1}$, we also compare our algorithm to ℓ_1-graph [23], which is based on ℓ_1 minimization. In addition, we compare our proposed with LDMGI [60], which used local discriminant models and global integration. To accord with the name (ℓ_1-graph) in [23], we name our proposed method as $\ell_{2,1}$-graph.

In order to randomize the experiments, we conduct the experiments using different cluster numbers. For each given cluster number, we run 20 times. The average over these 20 times are reported along with standard deviation. For the compared methods, we either use their existing parameter settings or tune them to obtain the best performances. There are some parameters to be set in advance. For SC, ℓ_1-graph and our proposed method, we set $k = 5$ for all the datasets to specify the size of neighborhoods. For fair evaluation, we tune the parameters in ℓ_1-graph, LDMGI and $\ell_{2,1}$-graph by the grid-search strategy from the range of $\{0.001, 0.01, 0.1, 1, 10\}$. According to the ground truth label, we consider the cluster number from the range of $\{5, 10, 20, 30, 40\}$, $\{3, 6, 9, 12, 15\}$, $\{2, 3, 5, 7, 9\}$, $\{2, 4, 6, 8, 10\}$ and $\{2, 4, 6, 8, 10\}$ for the ORL, Yale, Tumors9, JAFFE and oh15 dataset, respectively. The clustering results are listed in Tables 3, 4, 5, 6 and 7. Since LDMGI focuses on image clustering, we conduct the experiments on the image dataset when the number of clustering is set to the number of ground truth label. The results are listed in Table 8. To demonstrate the effectiveness of the manifold regularization with adaptive similarity, we use the regularizer proposed in [61] in our paper. We compare it with our model and present the results in Table 9.

From these tables, it can be observed that our proposed method $\ell_{2,1}$-graph achieves the best performance among all state-of-the art algorithms in many cases, i.e., 12 out of 15 for ORL (80%), 12 out of 15 for Yale (80%), 13 out of 15 for Tumors9 (86.67%), 11 out of 15 for JAFFE (73.33%) and 12 out of 15 for oh15 (80%). The advantage is more significant for the proposed method in some certain cases. For example, in the case that the cluster number is 9, the improvement of the proposed method $\ell_{2,1}$-graph over ℓ_1-graph(the best one among all the compared methods) is 8.15% in terms of ACC on the dataset Yale. It should be noted that, although in some cases, the performance of $\ell_{2,1}$-graph is not top-ranked, they are close to the best result.

Besides, we have also the following observations,

1) As can be seen, regardless of the datasets, as the cluster number increases, the corresponding performance, including ACC, NMI and ARI, of all the clustering methods always increases.
2) Graph-based clustering methods, both ℓ_1-graph and $\ell_{2,1}$-graph, often perform much better than other three compared methods, which indicates the effectiveness of graph in clustering. We can observe that ℓ_1-graph and $\ell_{2,1}$-graph improve the clustering performance considerably.
3) By comparing the performance of ℓ_1-graph and $\ell_{2,1}$-graph, we can clearly see the advantage of using

[2]http://tunedit.org/repo/Data/Text-wc/oh15.wc.arff

TABLE 3. Clustering results on the ORL dataset. The best results for these methods are highlighted in bold.

#Clusters	Measure	K-means	SC	NMF	ℓ_1-graph	$\ell_{2,1}$-graph
$c = 5$	ACC	11.91 ± 0.12	**12.50 ± 0.00**	10.75 ± 0.00	11.46 ± 0.36	**12.50 ± 0.00**
	NMI	33.46 ± 0.32	22.64 ± 0.00	29.50 ± 0.00	23.54 ± 1.30	**35.02 ± 0.26**
	ARI	8.200 ± 0.18	1.451 ± 0.00	6.160 ± 0.00	2.650 ± 0.63	**9.060 ± 0.18**
$c = 10$	ACC	22.14 ± 0.81	24.35 ± 0.28	21.93 ± 0.57	21.36 ± 0.61	**24.43 ± 0.66**
	NMI	45.98 ± 0.82	42.32 ± 2.59	44.21 ± 0.59	37.44 ± 0.75	**47.33 ± 1.07**
	ARI	15.13 ± 0.93	7.180 ± 1.96	13.87 ± 0.60	7.778 ± 0.77	**15.86 ± 1.04**
$c = 20$	ACC	37.20 ± 1.92	42.00 ± 2.13	35.23 ± 1.47	38.60 ± 1.54	**42.50 ± 1.71**
	NMI	59.46 ± 1.40	61.91 ± 2.10	57.43 ± 0.83	57.59 ± 1.10	**62.72 ± 1.66**
	ARI	23.88 ± 2.04	24.52 ± 3.40	23.40 ± 0.99	22.51 ± 1.54	**28.67 ± 2.50**
$c = 30$	ACC	47.49 ± 2.45	52.85 ± 2.41	45.24 ± 2.26	52.60 ± 2.37	**53.38 ± 2.64**
	NMI	68.20 ± 1.43	71.13 ± 1.79	66.28 ± 1.26	69.91 ± 1.43	**71.48 ± 2.06**
	ARI	32.74 ± 2.35	34.44 ± 3.89	30.31 ± 2.05	35.52 ± 2.09	**37.30 ± 3.65**
$c = 40$	ACC	53.68 ± 2.66	58.45 ± 2.68	50.69 ± 1.73	**60.93 ± 2.52**	57.88 ± 3.48
	NMI	73.22 ± 1.28	76.10 ± 1.41	70.81 ± 0.95	**77.40 ± 1.19**	75.84 ± 2.50
	ARI	38.51 ± 2.46	41.64 ± 3.59	35.12 ± 1.61	**46.52 ± 2.43**	44.11 ± 4.68

TABLE 4. Clustering results on the Yale dataset. The best results for these methods are highlighted in bold.

#Clusters	Measure	K-means	SC	NMF	ℓ_1-graph	$\ell_{2,1}$-graph
$c = 3$	ACC	17.48 ± 0.45	13.33 ± 0.00	17.58 ± 0.00	**19.33 ± 0.19**	17.55 ± 0.00
	NMI	19.94 ± 1.01	15.33 ± 0.00	20.14 ± 0.15	**25.53 ± 0.55**	24.03 ± 0.00
	ARI	5.16 ± 0.92	1.594 ± 0.00	4.791 ± 0.00	**7.017 ± 0.34**	5.901 ± 0.00
$c = 6$	ACC	22.58 ± 0.43	23.03 ± 0.00	23.97 ± 0.85	27.45 ± 1.87	**29.76 ± 0.65**
	NMI	25.07 ± 0.63	27.13 ± 0.00	27.66 ± 0.79	30.00 ± 2.30	**31.12 ± 0.71**
	ARI	7.60 ± 0.45	8.291 ± 0.00	9.480 ± 0.64	9.990 ± 1.61	**10.35 ± 0.84**
$c = 9$	ACC	32.94 ± 1.98	32.36 ± 0.36	32.82 ± 0.91	33.70 ± 2.36	**41.85 ± 1.22**
	NMI	37.08 ± 2.41	37.16 ± 0.54	36.01 ± 1.05	37.27 ± 1.72	**44.78 ± 1.14**
	ARI	14.76 ± 2.36	14.09 ± 0.55	14.25 ± 0.94	14.96 ± 1.74	**21.05 ± 1.11**
$c = 12$	ACC	37.73 ± 3.73	38.58 ± 0.79	37.64 ± 2.08	40.39 ± 3.18	**44.45 ± 2.18**
	NMI	43.76 ± 2.79	45.25 ± 0.82	42.63 ± 1.76	45.97 ± 2.57	**50.10 ± 3.10**
	ARI	18.33 ± 3.13	19.45 ± 1.14	17.88 ± 1.57	19.79 ± 1.69	**25.02 ± 2.52**
$c = 15$	ACC	41.94 ± 3.40	42.24 ± 1.82	33.91 ± 2.08	46.67 ± 3.80	**48.09 ± 4.09**
	NMI	47.56 ± 2.62	49.35 ± 1.03	38.94 ± 1.52	51.21 ± 2.95	**51.22 ± 2.93**
	ARI	21.49 ± 2.80	22.60 ± 1.30	12.30 ± 1.33	23.89 ± 3.48	**26.51 ± 3.78**

TABLE 5. Clustering results on the JAFFE dataset. The best results for these methods are highlighted in bold.

#Clusters	Measure	K-means	SC	NMF	ℓ_1-graph	$\ell_{2,1}$-graph
$c = 2$	ACC	20.66 ± 0.01	21.13 ± 0.00	21.33 ± 0.00	20.19 ± 0.00	**21.47 ± 0.00**
	NMI	24.22 ± 0.00	**31.87 ± 0.00**	29.96 ± 0.00	16.86 ± 0.00	30.31 ± 0.00
	ARI	13.20 ± 0.00	**19.01 ± 0.00**	17.38 ± 0.00	4.531 ± 0.00	17.13 ± 0.00
$c = 4$	ACC	39.18 ± 0.24	38.03 ± 0.00	39.84 ± 0.24	40.12 ± 0.36	**40.59 ± 0.17**
	NMI	45.19 ± 1.02	40.35 ± 0.00	46.08 ± 0.55	47.34 ± 1.38	**51.63 ± 0.72**
	ARI	26.43 ± 1.15	16.29 ± 0.00	29.13 ± 0.70	31.19 ± 1.92	**32.06 ± 0.54**
$c = 6$	ACC	57.77 ± 0.10	57.28 ± 0.00	54.44 ± 0.36	58.69 ± 0.82	**60.56 ± 0.00**
	NMI	66.28 ± 1.25	62.24 ± 0.00	58.72 ± 0.67	69.54 ± 0.64	**70.88 ± 1.04**
	ARI	51.59 ± 2.07	36.13 ± 0.00	41.75 ± 0.74	52.15 ± 2.31	**57.72 ± 1.52**
$c = 8$	ACC	74.08 ± 3.12	76.06 ± 0.00	76.67 ± 0.68	78.94 ± 2.60	**78.99 ± 1.50**
	NMI	77.55 ± 2.27	80.86 ± 0.00	80.16 ± 1.39	83.48 ± 2.46	**84.91 ± 1.98**
	ARI	64.03 ± 3.22	67.21 ± 0.00	69.75 ± 2.65	72.68 ± 6.22	**74.14 ± 1.58**
$c = 10$	ACC	84.32 ± 5.98	96.24 ± 0.00	82.84 ± 4.79	**96.46 ± 5.37**	95.99 ± 2.42
	NMI	86.49 ± 4.00	95.41 ± 0.00	84.21 ± 2.34	**97.16 ± 2.41**	95.44 ± 2.17
	ARI	75.84 ± 7.24	92.27 ± 0.00	73.25 ± 4.05	92.44 ± 4.39	**95.37 ± 5.38**

the group sparsity. This verifies that it is beneficial to adopt our proposed group sparse adaptive graph for clustering.

4) It can be observed that $\ell_{2,1}$-graph outperforms LRGA on the most cases (11 out of 15, i.e., 73.33%), which indicated the superiority of the manifold regularization with adaptive similarity.

D. PARAMETER SENSITIVITY

In order to study the influence of the regularization parameters α and β, we test the parameter sensitivity of our proposed method in terms of three evaluation metrics on all the datasets. On these datasets, the regularization parameters α and β vary in the same ranges as provided in Section IV-C. We investigate the case that the cluster number is set to be the

TABLE 6. Clustering results on the Tumors9 dataset. The best results for these methods are highlighted in bold.

#Clusters	Measure	K-means	SC	NMF	ℓ_1-graph	$\ell_{2,1}$-graph
$c=2$	ACC	24.08 ± 0.85	25.00 ± 0.00	25.00 ± 0.00	23.75 ± 0.74	$\mathbf{25.11 \pm 0.00}$
	NMI	24.16 ± 0.74	27.70 ± 0.00	19.61 ± 0.00	26.52 ± 0.00	$\mathbf{28.68 \pm 0.00}$
	ARI	7.479 ± 0.91	$\mathbf{10.71 \pm 0.00}$	3.341 ± 0.00	10.18 ± 0.00	$\mathbf{10.71 \pm 0.00}$
$c=3$	ACC	32.92 ± 2.75	35.00 ± 0.00	31.76 ± 0.00	35.00 ± 0.75	$\mathbf{36.67 \pm 0.00}$
	NMI	32.11 ± 3.60	34.12 ± 0.00	27.33 ± 0.00	36.23 ± 1.52	$\mathbf{37.99 \pm 0.00}$
	ARI	12.87 ± 3.30	12.68 ± 0.00	8.040 ± 0.00	$\mathbf{17.62 \pm 0.00}$	16.97 ± 0.00
$c=5$	ACC	39.25 ± 3.40	36.67 ± 0.00	38.50 ± 2.16	43.58 ± 2.25	$\mathbf{48.33 \pm 0.00}$
	NMI	37.54 ± 3.69	37.92 ± 0.00	36.69 ± 1.16	39.59 ± 2.66	$\mathbf{46.49 \pm 0.00}$
	ARI	14.23 ± 3.84	12.44 ± 0.00	12.75 ± 1.09	17.10 ± 3.04	$\mathbf{23.71 \pm 0.00}$
$c=7$	ACC	40.83 ± 3.57	39.92 ± 0.85	45.67 ± 1.74	44.83 ± 3.62	$\mathbf{47.92 \pm 2.01}$
	NMI	40.74 ± 2.58	43.59 ± 1.55	44.39 ± 1.26	45.70 ± 2.79	$\mathbf{47.42 \pm 2.37}$
	ARI	14.60 ± 3.54	14.36 ± 1.88	17.34 ± 1.18	18.74 ± 4.14	$\mathbf{20.64 \pm 2.97}$
$c=9$	ACC	41.58 ± 3.17	41.50 ± 2.35	45.75 ± 3.57	42.50 ± 2.73	$\mathbf{49.75 \pm 3.72}$
	NMI	39.99 ± 3.12	40.95 ± 1.77	43.08 ± 2.83	44.17 ± 2.79	$\mathbf{46.79 \pm 2.76}$
	ARI	14.23 ± 3.37	10.16 ± 1.95	19.42 ± 2.27	15.32 ± 3.22	$\mathbf{20.96 \pm 3.16}$

TABLE 7. Clustering results on the oh15 dataset. The best results for these algorithms are highlighted in bold.

#Clusters	Measure	K-means	SC	NMF	ℓ_1-graph	$\ell_{2,1}$-graph
$c=2$	ACC	23.66 ± 0.00	20.92 ± 0.00	23.55 ± 0.00	33.08 ± 0.00	$\mathbf{33.52 \pm 0.00}$
	NMI	5.298 ± 0.00	3.545 ± 0.00	5.266 ± 0.00	15.42 ± 0.00	$\mathbf{17.83 \pm 0.00}$
	ARI	0.670 ± 0.00	0.170 ± 0.00	0.643 ± 0.00	14.45 ± 0.00	$\mathbf{15.99 \pm 0.00}$
$c=4$	ACC	26.60 ± 1.60	23.77 ± 0.00	27.65 ± 0.16	40.85 ± 0.00	$\mathbf{45.67 \pm 0.00}$
	NMI	14.07 ± 1.29	7.367 ± 0.00	14.80 ± 0.10	$\mathbf{30.35 \pm 0.00}$	28.31 ± 0.00
	ARI	5.455 ± 1.76	0.879 ± 0.00	6.343 ± 0.15	24.76 ± 0.00	$\mathbf{26.18 \pm 0.00}$
$c=6$	ACC	29.18 ± 2.40	30.12 ± 0.00	29.27 ± 0.29	50.06 ± 0.00	$\mathbf{54.88 \pm 0.00}$
	NMI	16.74 ± 2.42	16.16 ± 0.00	16.56 ± 0.28	36.31 ± 0.22	$\mathbf{39.52 \pm 0.10}$
	ARI	6.441 ± 1.86	4.068 ± 0.00	8.536 ± 0.22	$\mathbf{36.00 \pm 0.20}$	33.90 ± 0.00
$c=8$	ACC	30.33 ± 2.83	35.43 ± 0.19	29.40 ± 0.14	53.64 ± 2.64	$\mathbf{55.29 \pm 2.32}$
	NMI	19.70 ± 2.98	23.75 ± 0.19	18.63 ± 0.31	41.39 ± 0.79	$\mathbf{44.25 \pm 0.54}$
	ARI	8.470 ± 2.36	6.654 ± 0.16	8.845 ± 0.10	$\mathbf{37.00 \pm 2.01}$	35.38 ± 0.98
$c=10$	ACC	31.19 ± 3.57	34.46 ± 0.88	32.88 ± 0.56	60.24 ± 1.05	$\mathbf{62.87 \pm 0.23}$
	NMI	21.05 ± 2.99	23.57 ± 0.65	24.01 ± 0.66	46.94 ± 0.86	$\mathbf{48.69 \pm 0.32}$
	ARI	8.139 ± 3.02	5.866 ± 0.58	8.620 ± 1.14	41.04 ± 1.15	$\mathbf{43.62 \pm 0.83}$

TABLE 8. Clustering results of different methods on the image datasets. The best results for these algorithms are highlighted in bold.

	Measure	ORL	Yale	JAFFE
LDMGI	ACC	$\mathbf{63.50 \pm 2.52}$	36.36 ± 3.80	91.55 ± 5.37
	NMI	74.43 ± 1.19	41.38 ± 2.95	92.64 ± 2.17
	ARI	$\mathbf{49.66 \pm 2.43}$	16.48 ± 3.48	85.87 ± 4.39
$\ell_{2,1}$-graph	ACC	57.88 ± 3.48	$\mathbf{48.09 \pm 4.09}$	$\mathbf{95.99 \pm 2.42}$
	NMI	$\mathbf{75.84 \pm 2.50}$	$\mathbf{51.22 \pm 2.93}$	$\mathbf{95.44 \pm 2.17}$
	ARI	44.11 ± 4.68	$\mathbf{26.51 \pm 3.78}$	$\mathbf{95.37 \pm 5.38}$

number of ground truth labels for all the datasets. The results are visualized in Figs. 2 and 3. From these figures, we can observe that our proposed method often gives reasonable results in a wide range of parameters. However, compared with β, the performance is much sensitive to α on the most datasets. Thus, the parameter α should be tuned carefully.

E. CONVERGENCE STUDY

To solve the proposed model $\ell_{2,1}$-graph, we employ the ADMM framework, where the subproblems can be effectively handled by shrinkage operator and IRLS algorithm, respectively. As stated previously, we have presented the convergence of IRLS and analyzed the computational complexity. In this subsection, we will investigate the trend of the objective function values during the iteration process.

TABLE 9. Clustering results comparisons with different regularizations.

Regularization	Measures	ORL	Yale	JAFFE	oh15	Tumors9
LRGA	ACC	$\mathbf{60.64 \pm 2.29}$	46.33 ± 3.04	$\mathbf{97.19 \pm 0.22}$	61.38 ± 0.25	45.42 ± 3.15
	NMI	73.19 ± 1.64	50.97 ± 2.43	$\mathbf{98.59 \pm 0.39}$	46.40 ± 0.40	45.50 ± 2.11
	ARI	42.02 ± 4.23	24.81 ± 2.43	$\mathbf{96.44 \pm 0.55}$	41.05 ± 0.35	19.32 ± 3.51
Ours	ACC	57.88 ± 3.48	$\mathbf{48.09 \pm 4.09}$	95.99 ± 2.42	$\mathbf{62.87 \pm 0.23}$	$\mathbf{49.75 \pm 3.72}$
	NMI	$\mathbf{75.84 \pm 2.50}$	$\mathbf{51.22 \pm 2.93}$	95.44 ± 2.17	$\mathbf{48.69 \pm 0.32}$	$\mathbf{46.79 \pm 2.76}$
	ARI	$\mathbf{44.11 \pm 4.68}$	$\mathbf{26.51 \pm 3.78}$	95.37 ± 5.38	$\mathbf{43.62 \pm 0.83}$	$\mathbf{20.96 \pm 3.16}$

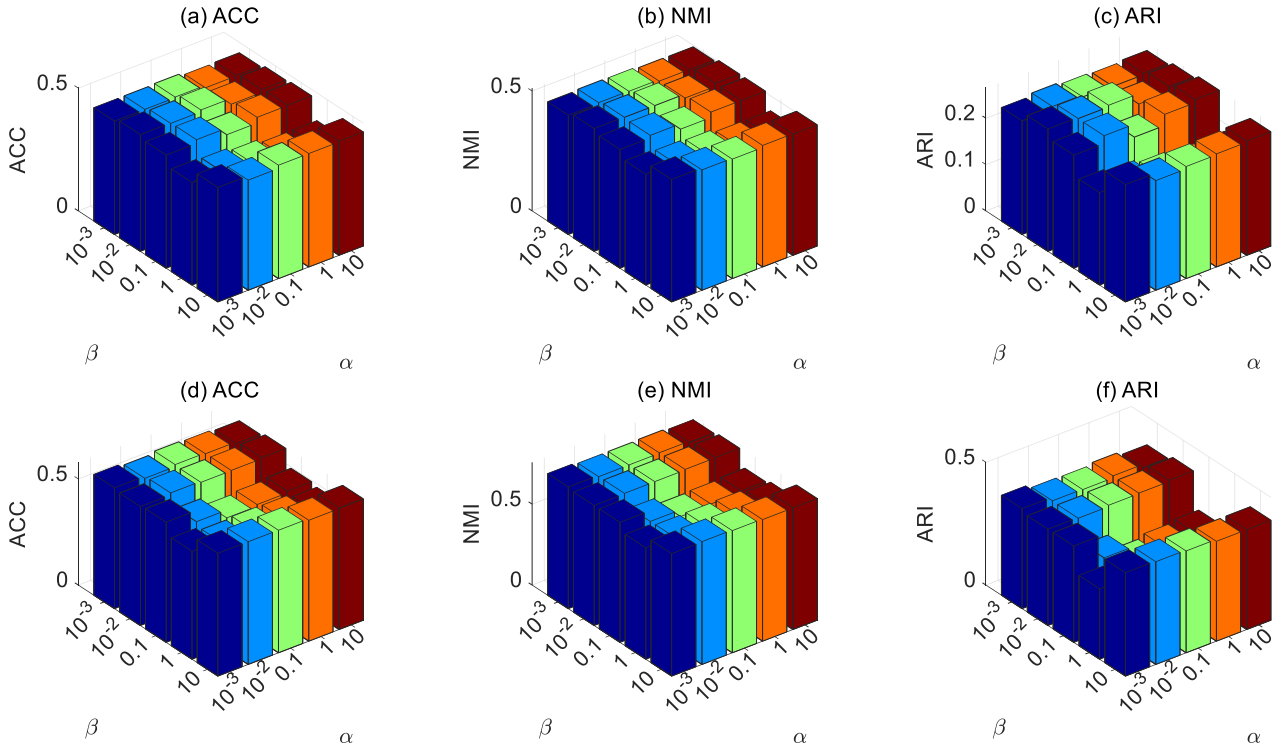


FIGURE 2. The clustering results of our proposed method w.r.t. α and β. Top: Yale dataset; Bottom: ORL dataset.

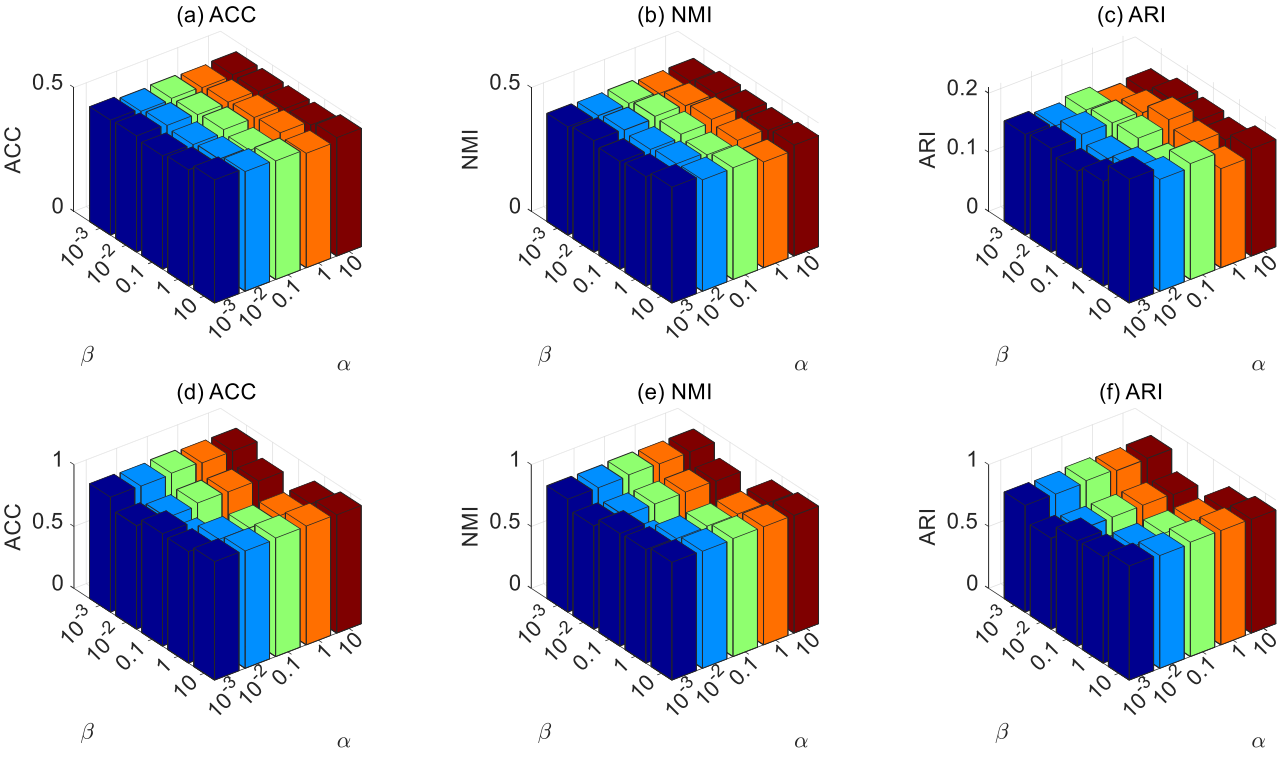


FIGURE 3. The clustering results of our proposed method w.r.t. α and β. Top: Tumors9 dataset; Bottom: JAFFE dataset.

In the algorithm, the stopping criterion is set to $(obj(k-1) - obj(k))/obj(k-1) < 1e-5$, where the $obj(k)$ denotes the objective function value at the k-th step. The variations of the objective function values on the ORL, Yale and JAFFE dataset are shown in Fig. 4. It can be observed that the objective function value decreases rapidly and converges after 5 iterations, which shows the effectiveness and efficiency of our algorithm.

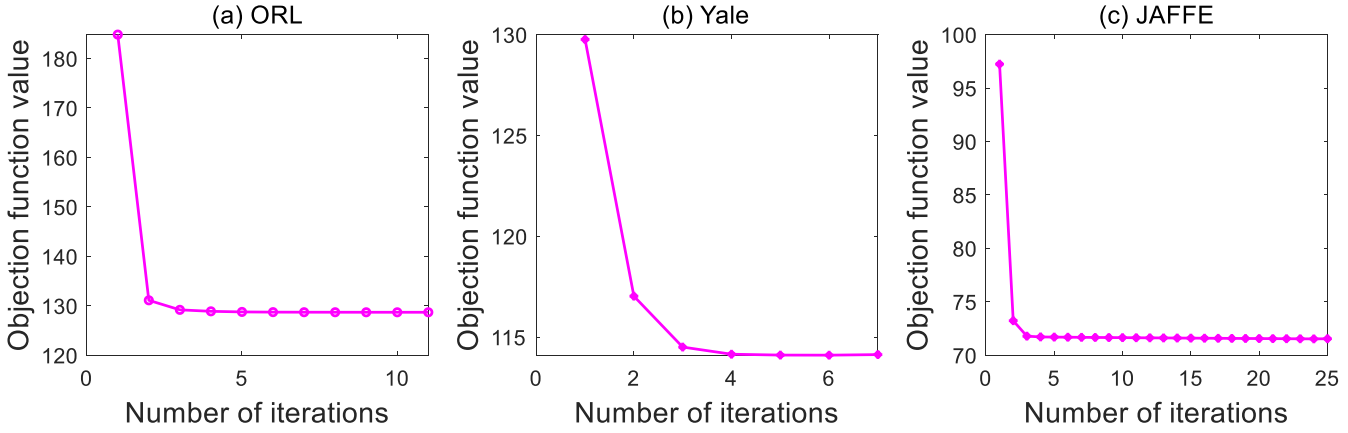


FIGURE 4. Variation of the objective function values over ORL, Yale and JAFFE datasets.

V. CONCLUSION

In this paper, we propose a novel graph-based method for data clustering, which resorts to the $\ell_{2,1}$ norm to exploit the global structure and manifold regularization to preserve the original data structure. Compared to the existing methods, the learned graph is more discriminative and informative. The resulting optimization problem can be handled by the ADMM, where the solutions to ADMM subproblems are obtained by the IRLS and shrinkage operator, respectively. We further provide the convergence analysis for IRLS. Extensive experimental results on various real-world datasets demonstrate the effectiveness and superiority of $\ell_{2,1}$-graph over other competing clustering methods.

REFERENCES

[1] C. Lu, S. Yan, and Z. Lin, "Convex sparse spectral clustering: Single-view to multi-view," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2833–2843, Jun. 2016.
[2] A. Staiano, R. Tagliaferri, and W. Pedrycz, "Improving RBF networks performance in regression tasks by means of a supervised fuzzy clustering," *Neurocomputing*, vol. 69, nos. 13–15, pp. 1570–1581, 2006.
[3] Y. Han and P. Shi, "An improved ant colony algorithm for fuzzy clustering in image segmentation," *Neurocomputing*, vol. 70, nos. 4–6, pp. 665–671, 2007.
[4] M. A. de Luis Balaguer and C. M. Williams, "Hierarchical modularization of biochemical pathways using fuzzy-C means clustering," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1473–1484, Aug. 2014.
[5] Y. Yang, Y. Yang, H. T. Shen, Y. Zhang, X. Du, and X. Zhou, "Discriminative nonnegative spectral clustering with out-of-sample extension," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 8, pp. 1760–1771, Aug. 2013.
[6] Y. Yang, F. Shen, Z. Huang, and H. T. Shen, "A unified framework for discrete spectral clustering," in *Proc. IJCAI*, 2016, pp. 2273–2279.
[7] P. Zhou, Y.-D. Shen, L. Du, F. Ye, and X. Li, "Incremental multi-view spectral clustering," *Knowl.-Based Syst.*, vol. 174, pp. 73–86, Jun. 2019.
[8] L. Lucchese and S. K. Mitra, "Unsupervised segmentation of color images based on k-means clustering in the chromaticity plane," in *Proc. IEEE Workshop Content-Based Access Image Video Libraries (CBAIVL)*, Jun. 1999, pp. 74–78.
[9] A. W. Moore, "Very fast EM-based mixture model clustering using multiresolution kd-trees," in *Proc. Adv. Neural Inf. Process. Syst.*, 1999, pp. 543–549.
[10] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 849–856.
[11] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.
[12] C. H. Q. Ding, X. He, H. Zha, M. Gu, and H. D. Simon, "A min-max cut algorithm for graph partitioning and data clustering," in *Proc. IEEE Int. Conf. Data Mining*, Nov./Dec. 2001, pp. 107–114.
[13] Y. Yang, Z. Ma, Y. Yang, F. Nie, and H. T. Shen, "Multitask spectral clustering by exploring intertask correlation," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 1083–1094, May 2015.
[14] Y. Yang, Z. Wang, J. Yang, J. Han, and T. S. Huang, "Regularized ℓ^1-graph for data clustering," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–11.
[15] F. Nie, X. Wang, M. I. Jordan, and H. Huang, "The constrained Laplacian rank algorithm for graph-based clustering," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 1969–1976.
[16] Y. Yan, G. Liu, S. Wang, J. Zhang, and K. Zheng, "Graph-based clustering and ranking for diversified image search," *Multimedia Syst.*, vol. 23, no. 1, pp. 41–52, 2017.
[17] P. Das and A. K. Das, "Graph-based clustering of extracted paraphrases for labelling crime reports," *Knowl.-Based Syst.*, vol. 179, pp. 55–76, Sep. 2019.
[18] H. Wang, Y. Yang, B. Liu, and H. Fujita, "A study of graph-based system for multi-view clustering," *Knowl.-Based Syst.*, vol. 163, pp. 1009–1019, Jan. 2019.
[19] T. E. Boult and L. G. Brown, "Factorization-based segmentation of motions," in *Proc. IEEE Workshop Vis. Motion*, Oct. 1991, pp. 179–186.
[20] J. P. Costeira and T. Kanade, "A multibody factorization method for independently moving objects," *Int. J. Comput. Vis.*, vol. 29, no. 3, pp. 159–179, 1998.
[21] Y. Wang, W. Zhang, L. Wu, X. Lin, and X. Zhao, "Unsupervised metric fusion over multiview data by graph random walk-based cross-view diffusion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 1, pp. 57–70, Jan. 2017.
[22] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2790–2797.
[23] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. S. Huang, "Learning with ℓ^1-graph for image analysis," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 858–866, Apr. 2010.
[24] Y. Yang, J. Feng, N. Jojic, J. Yang, and T. S. Huang, "ℓ^0-sparse subspace clustering," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 731–747.
[25] Y. Sui, G. Wang, and L. Zhang, "Sparse subspace clustering via low-rank structure propagation," *Pattern Recognit.*, vol. 95, pp. 261–271, Nov. 2019.
[26] Y.-X. Wang and H. Xu, "Noisy sparse subspace clustering," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 320–360, 2016.
[27] M. Soltanolkotabi and E. J. Candés, "A geometric analysis of subspace clustering with outliers," *Ann. Statist.*, vol. 40, no. 4, pp. 2195–2238, 2012.
[28] V. M. Patel and R. Vidal, "Kernel sparse subspace clustering," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 2849–2853.
[29] G. Lerman and T. Zhang, "Robust recovery of multiple subspaces by geometric l_p minimization," *Ann. Statist.*, vol. 39, no. 5, pp. 2686–2715, 2011.
[30] S. Wu, X. Feng, and W. Zhou, "Spectral clustering of high-dimensional data exploiting sparse representation vectors," *Neurocomputing*, vol. 135, pp. 229–239, Jul. 2014.
[31] A. Adler, M. Elad, and Y. Hel-Or, "Linear-time subspace clustering via bipartite graph modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2234–2246, Oct. 2015.
[32] Y. Wei, C. Niu, Y. Wang, H. Wang, and D. Liu, "The fast spectral clustering based on spatial information for large scale hyperspectral image," *IEEE Access*, vol. 7, pp. 141045–141054, 2019.

[33] M. Lee, J. Lee, H. Lee, and N. Kwak, "Membership representation for detecting block-diagonal structure in low-rank or sparse subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1648–1656.
[34] Y. Yan, C. Shen, and H. Wang, "Efficient semidefinite spectral clustering via Lagrange duality," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3522–3534, Aug. 2014.
[35] H. Lu, Z. Fu, and X. Shu, "Non-negative and sparse spectral clustering," *Pattern Recognit.*, vol. 47, no. 1, pp. 418–426, 2014.
[36] 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. (IJCAI)*, 2016, pp. 2153–2159.
[37] 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, Jul. 2018.
[38] Y. Wang, L. Wu, X. Lin, and J. Gao, "Multiview spectral clustering via structured low-rank matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4833–4843, Oct. 2018.
[39] Q. Yin, S. Wu, R. He, and L. Wang, "Multi-view clustering via pairwise sparse subspace representation," *Neurocomputing*, vol. 156, pp. 12–21, May 2015.
[40] M. Brbić and I. Kopriva, "ℓ_0-motivated low-rank sparse subspace clustering," *IEEE Trans. Cybern.*, to be published.
[41] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan, "Feature selection based on structured sparsity: A comprehensive study," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1490–1507, Jul. 2017.
[42] S. Zhang, J. Huang, H. Li, and D. N. Metaxas, "Automatic image annotation and retrieval using group sparsity," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 3, pp. 838–849, Jun. 2012.
[43] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *J. Mach. Learn. Res.*, vol. 12, pp. 2777–2824, Feb. 2011.
[44] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," *J. Mach. Learn. Res.*, vol. 12, pp. 3371–3412, Jan. 2011.
[45] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1095–1103.
[46] C. Ding, D. Zhou, X. He, and H. Zha, "R_1-PCA: Rotational invariant L_1-norm principal component analysis for robust subspace factorization," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 281–288.
[47] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.
[48] Y. Shi, J. Miao, Z. Wang, P. Zhang, and L. Niu, "Feature selection with $\ell_{2,1-2}$ regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4967–4982, Jan. 2018.
[49] J. Miao, H. Cao, X.-B. Jin, R. Ma, X. Fei, and L. Niu, "Joint sparse regularization for dictionary learning," *Cogn. Comput.*, vol. 11, no. 5, pp. 697–710, 2019.
[50] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 41–48.
[51] J. Zhang, J. Miao, K. Zhao, and Y. Tian, "Multi-task feature selection with sparse regularization to extract common and task-specific features," *Neurocomputing*, vol. 340, pp. 76–89, May 2019.
[52] X. Cai, F. Nie, H. Huang, and C. Ding, "Multi-class $\ell_{2,1}$-norm support vector machine," in *Proc. IEEE 11th Int. Conf. Data Mining*, Dec. 2011, pp. 91–100.
[53] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 11, pp. 1738–1754, Nov. 2012.
[54] J. Yoon and S. J. Hwang, "Combined group and exclusive sparsity for deep neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3958–3966.
[55] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.
[56] 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.
[57] J. E. Staunton, D. K. Slonim, H. A. Coller, P. Tamayo, M. J. Angelo, J. Park, U. Scherf, J. K. Lee, W. O. Reinhold, J. N. Weinstein, J. P. Mesirov, E. S. Lander, and T. R. Golub, "Chemosensitivity prediction by transcriptional profiling," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 19, pp. 10787–10792, 2001.
[58] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2003, pp. 267–273.
[59] T. Li and C. Ding, "The relationships among various nonnegative matrix factorization methods for clustering," in *Proc. 6th Int. Conf. Data Mining (ICDM)*, Dec. 2006, pp. 362–371.
[60] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang, "Image clustering using local discriminant models and global integration," *IEEE Trans. Image Process.*, vol. 19, no. 10, pp. 2761–2773, Oct. 2010.
[61] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

JIANYU MIAO received the M.S. degree in applied mathematics from Zhengzhou University, Zhengzhou, China, in 2014, and the Ph.D. degree in operational research and cybernetics from the University of Chinese Academy of Sciences, Beijing, China, in 2018. Since 2018, he has been on the faculty of the College of Information Science and Engineering, Henan University of Technology, China, where he is currently a Lecturer. His current research interests include optimization and machine learning.

TIEJUN YANG was born in Shangcheng, Henan, China, in 1975. He received the B.S. degree in applied electronic technology from the Xi'an University of Technology, Shanxi, China, in 1999, and the M.S. and Ph.D. degrees in communication and information system from the University of Electronics Science and Technology of China (UESTC), Chengdu, China, in 1999 and 2003, respectively. Since 2011, he has been a Professor with the College of Information Science and Engineering, Henan University of Technology, China. His research interests include signal and information processing, image processing, and wireless communication.

JUNWEI JIN received the B.S. degree from Ningxia University, Ningxia, China, in 2013, and the M.S. and Ph.D. degrees from the University of Macau, Macau, in 2015 and 2019, respectively. Since May 2019, he has been an Assistant Professor with the School of Information Science and Engineering, Henan University of Technology. His current research interests include machine learning, computer vision, and neural networks.

LINGFENG NIU received the B.S. degree in mathematics from Xi'an Jiaotong University, Xi'an, China, in 2004, and the Ph.D. degree in mathematics from the Chinese Academy of Sciences, Beijing, China, in 2009. She has been an Associate Professor with the Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, since 2009. Her current research interests include optimization and machine learning.