JGSED: An End-to-End Spectral Clustering Model for Joint Graph Construction, Spectral Embedding and Discretization

Yong Peng, *Member, IEEE*, Wenna Huang, Wanzeng Kong, *Member, IEEE*, Feiping Nie, *Senior Member, IEEE*, and Bao-Liang Lu, *Fellow, IEEE*

***Abstract*—Most of the existing graph-based clustering models performed clustering by adopting a two-stage strategy which first completes the spectral embedding from a given fixed graph and then resorts to other clustering methods such as $k$means to achieve discrete cluster results. On one hand, such a discretization operation easily causes that the obtained results deviate far from the true solution. On the other hand, clustering performance heavily relies on the quality of graph; therefore, the fixed graph is usually not optimal enough. In addition, clustering by separated steps inevitably breaks the underlying connections among the graph construction, spectral embedding and discretization. To address these drawbacks, in this paper, we propose a new spectral clustering model termed JGSED which integrates the graph construction, spectral embedding and spectral rotation together into a unified objective. JGSED is an end-to-end framework to directly take data as input and output the final binary cluster indicator matrix. An efficient algorithm is proposed to optimize the model variables in JGSED, which can be co-evolved towards the optimum. Extensive experiments are conducted on both synthetic and real data sets and the results demonstrate that JGSED outperforms the other state-of-the-art spectral clustering models, indicating the effectiveness of joint optimization.**

***Index Terms*—Graph construction, joint optimization, spectral clustering, spectral embedding, spectral rotation.**

I. Introduction

AS ONE of the most popular unsupervised learning paradigms, clustering has been widely applied in many practical scenes such as image processing [1], bioinformatics [2], remote sensing [3] and change detection [4], which has always been a research hotspot in diverse fields such as machine learning, data mining and computational intelligence. Generally, clustering aims to group data points into different clusters based on their similarities [5]. Recently a lot of methods such as neural networks and fuzzy systems were proposed to perform data clustering [6], [7]. Current clustering models mainly include $k$means clustering, hierarchical clustering [8], [9], density-based clustering [10], subspace clustering [11], [12], spectral clustering [13], [14], multi-view clustering [15], [16] and some others [17], [18], [19]. Among the existing clustering models, spectral-based ones mainly focus on modeling the spectral structure of data by a similarity graph first and then partitioning the graph vertices into respective clusters, which have drawn increasing attention recently due to its effectiveness. The essence of spectral clustering is to convert clustering into a graph cutting problem.

Generally, a complete spectral clustering process consists of three consecutive stages, *i.e.*, first constructing a similarity to depict the structure information of data as accurately as possible, then calculating the spectral embedding on the graph Laplacian matrix to get the scaled cluster indicator matrix with continuous values, and finally performing post-processing on the scaled indicator matrix to extract the corresponding discrete indicator matrix by an additional post-processing step such as $k$means or spectral rotation [20]. However, performing spectral embedding and post-processing separately may easily result in far deviation from the true discrete indicator matrix, leading to sub-optimality as well as degenerated clustering performance. In [21], a rank-constrained spectral clustering with flexible embedding framework was proposed by integrating the three components of adaptive probabilistic neighborhood learning, flexible embedding and the rank constraint on graph Laplacian matrix. In [22], a robust graph affinity matrix was adaptively learned from multiple features. Accordingly, a set of projection matrices are learned to determine the optimal subspaces of the different types of features. In [23], Yang et al. proposed a unified framework for discrete spectral clustering (UFDSC) which can directly output final indicator matrix with predefined graph. Further, Pang et al. proposed to jointly perform spectral embedding and spectral rotation (JSESR) [24], aiming at addressing the

Manuscript received 9 April 2022; revised 3 November 2022 and 27 November 2022; accepted 2 February 2023. Date of publication 22 February 2023; date of current version 23 November 2023. This work was supported in part by the Natural Science Foundation of Zhejiang Province under Grant LY21F030005, in part by the National Natural Science Foundation of China under Grants 61971173 and U20B2074, in part by the Key Research and Development Project of Zhejiang Province under Grants 2023C03026, 2021C03001, and 2021C03003, and in part by the China Postdoctoral Science Foundation under Grant 2017M620470. *(Corresponding author: Wanzeng Kong.)*

Yong Peng, Wenna Huang, and Wanzeng Kong are with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China, and also with the Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: yongpeng@hdu.edu.cn; hwn@hdu.edu.cn; kongwanzeng@hdu.edu.cn).

Feiping Nie is with the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: feipingnie@gmail.com).

Bao-Liang Lu is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: bllu@sjtu.edu.cn).

Digital Object Identifier 10.1109/TETCI.2023.3243945

unbalance limitation in UFDSC. In [25], an variant $k$means was proposed by jointly performing spectral embedding and rotation on a specific data similarity graph.

In spectral clustering, the quality of input graph plays a crucial role in determining whether clustering models can achieve promising performance. Traditionally, graph was constructed according to some fixed rules such as the '0-1' and 'Heatkernel'-based weighting schemes [26], [27], which have limited capacities in characterizing the inner connection between sample pairs. For example, in [28], an adaptive semi-supervised feature selection model was proposed for cross-modal retrieval, in which the semantic regression was utilized to strengthen the neighboring relationship between the data with the same semantic. Meanwhile, a graph-constraint based on the 'Heatkernel' function was adopted to perform label estimation of the unlabeled data. Therefore, a lot of efforts were devoted to adaptively learn graph from data and consequently many high-quality graphs were proposed such as the sparse representation-based graph [29], low-rank representation-based graph [30], [31], maximum entropy graph [32], [33], doubly stochastic graph [14] and weight-adaptive graphs [34]. In [35], Nie et al. proposed a clustering with adaptive neighbors (CAN) model in which the target graph was enforced to satisfy triple desirable properties of non-negativity, row normalization and constrained rank. Besides, CAN can adaptively determine the neighborhood size. Its projected version PCAN can jointly perform graph construction and subspace exploration. Yang et al. proposed a joint spectral embedding and clustering method with structured graph optimization, in which the similarity matrix is optimized with embedded low-dimensional data [36]. In the LAPIN model, an optimal bipartite graph is learned to extract the duality relationship between samples and featues for co-clustering [37]. To directly achieve the discrete clustering results, spectral rotation is integrated into the graph learning to form a discrete optimal graph clustering (DOGC) model [38]. In [39], CAN was extended to self-weighted CAN (SWCAN) by additionally exploring the contributions of different feature dimensions in graph construction. This is completed by introducing a self-weighted variable in characterizing the connection between data pairs. SWCAN can output a better graph and result in better clustering performance than CAN. However, its learned scaled indicator matrix is still in real-valued form. Therefore, for all these mentioned graph construction models, a post-processing step is always necessary to discretize the obtained scaled indicator matrix to finally indicate the cluster assignments of data points [20].

Motivated by that joint optimization of the graph construction, spectral embedding and discretization post-processing can effectively capture the underlying connections among them, in this paper, we propose an end-to-end spectral clustering model termed JGSED to unify them together into a single objective function, which allows the dynamic updates of the variables corresponding to these three operations. For the graph construction, a variant SWCAN is adopted by simultaneously taking the rank constraint, adaptive neighbors, and self-weighted feature importance into account. For the discretization, improved spectral rotation is adopted to minimize the discrepancy between scaled cluster indicator and real clustering solution [40]. Below we summarize the main contributions of this work.

- A complete spectral clustering model JGSED is proposed to perform end-to-end data clustering, which unifies all the three stages of graph construction, spectral embedding and discretization together into a single objective. JGSED directly takes data as input and outputs the final clustering results as well as the constructed graph, which effectively avoids the sub-optimality caused by executing these stages separately and sequentially.
- An efficient optimization algorithm is developed to solve the JGSED objective function. The involved model variables in JGSED are iteratively updated in the form of alternate optimization. Especially, the utilization of the generalized power iteration (GPI) method in optimizing the scaled indicator matrix greatly decreases the model computational complexity in comparison with the traditional eigen-decomposition operation on the Laplacian matrix.
- Extensive experiments are conducted on both synthetic and real data sets and the obtained results show that JGSED outperforms the other state-of-the-art models in data clustering. Besides, the model robustness, convergence and parameter sensitivity analysis JGSED are provided.

The remainder of this paper is organized as follows. Section II briefly reviews some related models to current work. In Section III, we introduce the proposed JGSED model in detail including its objective function, optimization algorithm, complexity and convergence analysis. The evaluations of JGSED performance on synthetic and benchmark data clustering are respectively provided in Section IV and Section V. Section VI concludes the whole paper.

Notations: Throughout this paper, matrices and vectors are respectively denoted by boldface uppercase and lowercase letters. For example, for matrix $\mathbf{M}$, we use m_{ij} to represent its (i,j)-th element, $\mathbf{m}_i$ and $\mathbf{m}^j$ to represent its i-th column and j-th row, respectively. The boldfaced $\mathbf{1}$ is an all-one column vector whose length is determined by the context. To facilitate the understanding of the derivations in the following sections, we summarize the main notations in Table I.

TABLE I
SUMMARY OF THE MAIN NOTATIONS

Notation	Definition
n	number of data points
d	feature dimensionality
c	number of clusters
$\mathbb{R}$	real domain
$\mathbb{B}$	binary domain
$\mathbf{x}_i \in \mathbb{R}^d$	the i-th sample
$\mathbf{X} \in \mathbb{R}^{d\times n}$	the data matrix
$\mathbf{A} \in \mathbb{R}^{n\times n}$	graph similarity matrix
$\mathbf{D} \in \mathbb{R}^{n\times n}$	degree matrix, $d_{ii} = \sum_j \frac{a_{ij}+a_{ji}}{2}$
$\mathbf{L} \in \mathbb{R}^{n\times n}$	Laplacian matrix, $\mathbf{L} = \frac{\mathbf{A}+\mathbf{A}^T}{2}$
$\boldsymbol{\theta} \in \mathbb{R}^d$	the feature weight vector
$\boldsymbol{\Theta} \in \mathbb{R}^{d\times d}$	diagonal matrix, $\boldsymbol{\Theta} = diag(\boldsymbol{\theta})$
$\mathbf{I}$	identity matrix
$\mathbf{Y}$	binary cluster indicator matrix

II. Related Works

In this section, we introduce some related works including the conventional spectral clustering, self-weighted CAN and JSESR.

A. Spectral Clustering Revisit

Given a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ where d is the dimensionality and n is the number of data points, the clustering task aims to partition $\mathbf{X}$ into c groups and usually c is a predefined number of clusters. Denote $\mathbf{Y}$ as the cluster indicator matrix (*i.e.*, $\mathbf{Y} \in$ Ind). $\mathbf{Y} = [\mathbf{y}^1; \mathbf{y}^2; \cdots; \mathbf{y}^n] \in \mathbb{B}^{n\times c}$ and the unique non-zero element 1 in its i-th row $\mathbf{y}^i \in \mathbb{B}^{1\times c}$ indicates the corresponding cluster membership of the sample $\mathbf{x}_i$. We denote the similarity matrix by $\mathbf{A} \in \mathbb{R}^{n\times n}$ in which element a_{ij} represents the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. Then the Laplacian matrix $\mathbf{L} \in \mathbb{R}^{n\times n}$ can be calculated by $\mathbf{L} = \mathbf{D} - (\mathbf{A} + \mathbf{A}^T)/2$, where $\mathbf{D} \in \mathbb{R}^{n\times n}$ is a diagonal degree matrix with its i-th element defined as $d_{ii} = \sum_{j=1}^{n}(a_{ij} + a_{ij})/2$. By introducing a scaled cluster indicator matrix $\mathbf{F} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}} \in \mathbb{R}^{n\times c}$ with its j-th column defined as

$$\mathbf{f}_j = [\underbrace{0,\ldots,0}_{\sum_{i=1}^{j-1} n_i}, \underbrace{1,\ldots,1}_{n_j}, \underbrace{0,\ldots,0}_{\sum_{i=j+1}^{c} n_i}]^T/\sqrt{n_j}, \tag{1}$$

where n_j is the number of samples in the j-th cluster, the objective function of spectral clustering is formulated as

$$\min_{\mathbf{F}} \mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}), \ \text{s.t. } \mathbf{F} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}. \tag{2}$$

Since $\mathbf{F}^T\mathbf{F} = (\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}} = \mathbf{I}$, the optimal solution $\mathbf{F}^*$ to problem (2) can be obtained by stacking the eigenvectors of $\mathbf{L}$ corresponding to its c smallest eigenvalues. The process of computing $\mathbf{F}^*$ is called spectral embedding.

B. Self-Weighted CAN

For graph-based clustering models, their performance highly depends on the quality of the input graph. To let the two processes of graph construction and spectral embedding better match each other, Nie et al. proposed a CAN model to adaptively learn the similarity matrix by exploring its local connectivity. Meanwhile, the rank constraint is imposed to the Laplacian matrix with the expectation that the number of the connected graph components is equal to the number of clusters. The CAN model objective function is

$$\min_{\mathbf{A}} \sum_{i,j=1}^{n} \left(\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 a_{ij} + \gamma a_{ij}^2\right),$$
$$\text{s.t. } \forall i, \ \mathbf{a}^i\mathbf{1} = 1, \mathbf{a}^i \geq \mathbf{0}, \mathrm{rank}(\mathbf{L}) = n - c, \tag{3}$$

where $\mathbf{a}^i$ is the i-th row of $\mathbf{A}$, γ is a tradeoff parameter.

Supposing that $\sigma_i(\mathbf{L})$ is the i-th eigenvalue of $\mathbf{L}$, $\sigma_i(\mathbf{L})$ is non-negative since $\mathbf{L}$ is positive semi-definite. By introducing a large enough parameter λ, the rank constraint can be absorbed in the objective function as

$$\min_{\mathbf{A}} \sum_{i,j=1}^{n} \left(\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 a_{ij} + \gamma a_{ij}^2\right) + 2\lambda \sum_{i=1}^{c} \sigma_i(\mathbf{L}),$$
$$\text{s.t. } \forall i, \ \mathbf{a}^i\mathbf{1} = 1, \mathbf{a}^i \geq \mathbf{0}. \tag{4}$$

According to the Ky Fan's Theorem [41], we have

$$\sum_{i=1}^{c} \sigma_i(\mathbf{L}) = \min_{\mathbf{F}\in\mathbb{R}^{n\times c}, \mathbf{F}^T\mathbf{F}=\mathbf{I}} \mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}), \tag{5}$$

and then the final objective of CAN is formulated as

$$\min_{\mathbf{A},\mathbf{F}} \sum_{i=1}^{n}\sum_{j=1}^{n} \left(\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 a_{ij} + \gamma a_{ij}^2\right) + 2\lambda \mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}),$$
$$\text{s.t. } \forall i, \ \mathbf{a}^i\mathbf{1} = 1, \mathbf{a}^i \geq \mathbf{0}, \mathbf{F} \in \mathbb{R}^{n\times c}, \mathbf{F}^T\mathbf{F} = \mathbf{I}. \tag{6}$$

However, it is not reasonable to treat all feature dimensions equally; that is, they hold the same importance in determining the cluster assignments of samples. To this end, Nie et al. proposed to adaptively learn the feature weights by introducing a self-weighted variable $\boldsymbol{\theta}$ which satisfies $\boldsymbol{\theta} \geq \mathbf{0}$ and $\sum_{t=1}^{d} \theta_t = 1$ [42], [43], [44]. The objective function of SWCAN is

$$\min_{\mathbf{A},\boldsymbol{\Theta},\mathbf{F}} \sum_{i,j=1}^{n} \left(\|\boldsymbol{\Theta}\mathbf{x}_i - \boldsymbol{\Theta}\mathbf{x}_j\|_2^2 a_{ij} + \gamma a_{ij}^2\right) + 2\lambda \mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}),$$
$$\text{s.t. } \forall i, \ \mathbf{a}^i\mathbf{1} = 1, \mathbf{a}^i \geq \mathbf{0}, \boldsymbol{\theta} \geq \mathbf{0}, \boldsymbol{\theta}^T\mathbf{1} = 1,$$
$$\boldsymbol{\Theta} = \mathrm{diag}(\boldsymbol{\theta}), \mathbf{F} \in \mathbb{R}^{n\times c}, \mathbf{F}^T\mathbf{F} = \mathbf{I}, \tag{7}$$

where θ_t characterizes the importance of the t-th feature. Intuitively, θ_t should be large if the t-th feature is discriminative while it should be small if it is noisy or redundant. $\boldsymbol{\Theta} \in \mathbb{R}^{d\times d}$ is a diagonal matrix reshaped from vector $\boldsymbol{\theta}$.

It should be noted that (5) is consistent with the spectral embedding objective function (2) to some extent. From this perspective, we can decompose the second term in objective function (7) into two subitems which respectively correspond to the rank constraint and spectral embedding since parameter λ is large enough. Therefore, SWCAN can be viewed as a joint model of graph construction and spectral embedding.

C. JSESR

Given a graph, conventional spectral clustering models perform spectral embedding first and then discretize the scaled cluster indicator matrix by $k$means or spectral rotation. As pointed by [20], [40], spectral rotation usually outperforms $k$means, which therefore becomes a better choice for post-processing. In practice, such processing strategy cannot achieve the optimum and both stages should be unified together. In [24], Pang et al. proposed to jointly perform spectral embedding and improved spectral rotation (JSESR). Based on k-way normalized cut (NCut), the objective function of JSESR was formulated as

$$\min_{\mathbf{F}_1,\mathbf{R},\mathbf{Y}} \mathrm{Tr}(\mathbf{F}_1^T\tilde{\mathbf{L}}\mathbf{F}_1) + \alpha \left\|\mathbf{F}_1\mathbf{R} - \mathbf{D}^{\frac{1}{2}}\mathbf{Y}(\mathbf{Y}^T\mathbf{D}\mathbf{Y})^{-\frac{1}{2}}\right\|_2^2,$$
$$\text{s.t. } \mathbf{F}_1^T\mathbf{F}_1 = \mathbf{I}, \mathbf{R}^T\mathbf{R} = \mathbf{I}, \mathbf{Y} \in \text{Ind}, \tag{8}$$

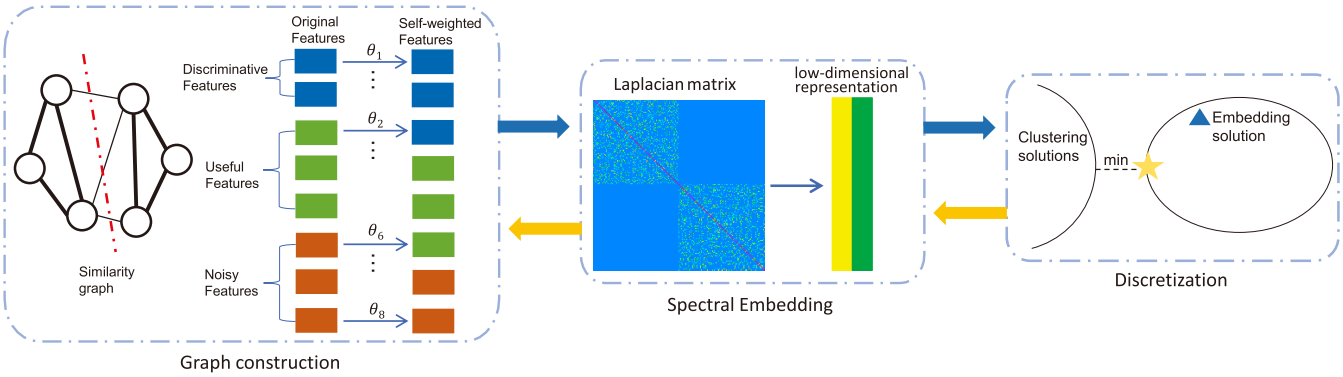


Fig. 1. The overall framework of the proposed JGSED model.

where α is a regularization parameter to balance the effects of the two items. $\tilde{\mathbf{L}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$ is the normalized Laplacian matrix. The first item corresponds to the spectral embedding and the second implements the spectral rotation. The NCut objective referred in JSESR is

$$\min_{\mathbf{H}} \mathrm{Tr}(\mathbf{H}^T\mathbf{L}\mathbf{H}),\ \text{s.t. } \mathbf{H}^T\mathbf{D}\mathbf{H} = \mathbf{I}, \tag{9}$$

where $\mathbf{H}$ is the scaled indicator matrix and defined as $\mathbf{H} = \mathbf{Y}(\mathbf{Y}^T\mathbf{D}\mathbf{Y})^{-\frac{1}{2}}$. By defining $\mathbf{F}_1 = \mathbf{D}^{\frac{1}{2}}\mathbf{H} = \mathbf{D}^{\frac{1}{2}}\mathbf{Y}(\mathbf{Y}^T\mathbf{D}\mathbf{Y})^{-\frac{1}{2}}$, we have $\mathbf{F}_1^T\mathbf{F}_1 = (\mathbf{Y}^T\mathbf{D}\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{D}^{\frac{1}{2}} \cdot \mathbf{D}^{\frac{1}{2}}\mathbf{Y}(\mathbf{Y}^T\mathbf{D}\mathbf{Y})^{-\frac{1}{2}} = \mathbf{I}$. Then objective function (9) can be rewritten as

$$\min_{\mathbf{F}_1} \mathrm{Tr}(\mathbf{F}_1^T\tilde{\mathbf{L}}\mathbf{F}_1),\ \text{s.t. } \mathbf{F}_1 = \mathbf{D}^{\frac{1}{2}}\mathbf{Y}(\mathbf{Y}^T\mathbf{D}\mathbf{Y})^{-\frac{1}{2}}. \tag{10}$$

Since $\mathbf{F}_1$ is a real-valued matrix, JSESR utilized the improved spectral rotation as the postprocessing method to get final clustering results. The objective function of the postprocessing process is

$$\min_{\mathbf{R}^T\mathbf{R}=\mathbf{I},\mathbf{Y}\in\mathrm{Ind}} \left\|\mathbf{F}_1\mathbf{R} - \mathbf{D}^{\frac{1}{2}}\mathbf{Y}(\mathbf{Y}^T\mathbf{D}\mathbf{Y})^{-\frac{1}{2}}\right\|_2^2, \tag{11}$$

where $\mathbf{R} \in \mathbb{R}^{c\times c}$ is an orthonormal matrix such that $\mathbf{F}_1\mathbf{R}$ is closest to the true solution $\mathbf{D}^{\frac{1}{2}}\mathbf{Y}(\mathbf{Y}^T\mathbf{D}\mathbf{Y})^{-\frac{1}{2}}$ for $\mathbf{F}_1$.

III. The Proposed JGSED Model

In this section, we present a unified framework of graph construction, spectral embedding and discretization for spectral clustering (JGSED) including its formulation and optimization. Besides, its computational complexity and convergence property are analyzed in detail. The overall framework of the proposed JGSED model is provided in Fig. 1.

A. Model Formulation

According to the analysis of SWCAN and JSESR in Section II, we draw a conclusion that SWCAN model can be understood as the joint framework of graph construction and spectral embedding, while the JSESR is a unified model that integrates the procedure of spectral embedding and discretization.

For the general graph-based spectral clustering objective function (2), inspired by JSESR, we adopt the improved spectral rotation as the post-processing method to discretize the obtained scaled indicator matrix $\mathbf{F}$. That is

$$\min_{\mathbf{Y},\mathbf{R}} \left\|\mathbf{F}\mathbf{R} - \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\right\|_2^2, \tag{12}$$

where $\mathbf{R} \in \mathbb{R}^{c\times c}$ is an arbitrary orthonormal matrix and $\mathbf{Y} \in$ Ind is the binary cluster indicator matrix.

Based on SWCAN, we add an extra setting to $\mathbf{F}$ as $\mathbf{F} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}$, which also satisfies the orthonormal constraint. Then, an variant SWCAN model can be represented as

$$\begin{aligned}&\min_{\mathbf{A},\mathbf{\Theta},\mathbf{F}} \sum_{i,j=1}^{n}\left(\|\mathbf{\Theta}\mathbf{x}_i - \mathbf{\Theta}\mathbf{x}_j\|_2^2 a_{ij} + \gamma a_{ij}^2\right) + 2\lambda\mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}),\\ &\text{s.t. } \forall i,\ \mathbf{a}^i\mathbf{1} = 1, \mathbf{a}^i \geq \mathbf{0}, \mathbf{\Theta} = \mathrm{diag}(\boldsymbol{\theta}), \boldsymbol{\theta}^T\mathbf{1} = 1, \boldsymbol{\theta} \geq \mathbf{0},\\ &\mathbf{F} \in \mathbb{R}^{n\times c}, \mathbf{F} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}, \mathbf{F}^T\mathbf{F} = \mathbf{I}.\end{aligned} \tag{13}$$

Though the above variant SWCAN model optimizes the graph similarity matrix $\mathbf{A}$ and the scaled indicator matrix $\mathbf{F}$ in one step, we still need to discretize $\mathbf{F}$ to get the final cluster assignments. In this paper, we propose to jointly optimize the objective functions of (12) and (13), and then get the objective function of the proposed JGSED model as

$$\begin{aligned}&\min_{\mathbf{A},\mathbf{\Theta},\mathbf{F},\mathbf{R},\mathbf{Y}} \sum_{i,j=1}^{n}(\|\mathbf{\Theta}\mathbf{x}_i - \mathbf{\Theta}\mathbf{x}_j\|_2^2 a_{ij} + \gamma a_{ij}^2)\\ &+ 2\lambda\mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) + \beta\left\|\mathbf{F}\mathbf{R} - \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\right\|_2^2,\\ &\text{s.t. } \forall i,\ \mathbf{a}^i\mathbf{1} = 1, \mathbf{a}^i \geq \mathbf{0}, \boldsymbol{\theta}^T\mathbf{1} = 1, \boldsymbol{\theta} \geq \mathbf{0}, \mathbf{\Theta} = \mathrm{diag}(\boldsymbol{\theta}),\\ &\mathbf{F} \in \mathbb{R}^{n\times c}, \mathbf{F}^T\mathbf{F} = \mathbf{I}, \mathbf{R} \in \mathbb{R}^{c\times c}, \mathbf{R}^T\mathbf{R} = \mathbf{I}, \mathbf{Y} \in \mathrm{Ind}.\end{aligned} \tag{14}$$

where γ, λ, and β are regularization parameters.

It is obvious that there are three terms in the JGSED objective function (14). Here it should be noted that the second term $\mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F})$ has multiple meanings. First, it corresponds to the rank constraint (*i.e.*, $\mathrm{rank}(\mathbf{L}) = n - c$) for structured graph learning. Second, this term together with the orthonormal constraint $\mathbf{F}^T\mathbf{F} = \mathbf{I}$ performs the spectral embedding, which is

similar to the one in objective function (2). Therefore, it can be regarded as a compound term, based on which the former two terms in (14) not only form the objective function of the SWCAN model but also fuse the two processes of graph construction and spectral embedding together. Similarly, the combination of the latter two terms jointly performs spectral embedding and spectral rotation. As a summary, our JGSED model fully integrates the three processes in spectral clustering, *i.e.*, graph construction, spectral embedding and spectral rotation-based discretization, which is an end-to-end spectral clustering model to directly operate on data and output cluster assignments.

What is more, it should be illustrated that there are two reasons for adopting SWCAN as the basic method for the graph construction stage of JGSED. First, SWCAN simultaneously takes the non-negativity, row normalization, rank constraint, adaptive neighbors and self-weighted feature importance into account, which is quite advanced among existing graph-construction methods. The first three properties aim to learn a structured graph which has exact c connected components corresponding to the clusters [45]. Adaptive neighborhood is also one of the desirable properties to construct an informative graph [35], [46]. The adaptive feature importance learning can improve the discriminative ability of the learned graph based on the consensus that different features have different contributions in characterizing the semantic information of data. Second, the primary focus of this paper is to build an end-to-end spectral clustering model to avoid the limitations caused by separately performing graph construction, spectral embedding and discretization, rather than proposing a new graph construction method. In this sense, our proposed JGSED is a general framework whose building blocks corresponding to the three operations are replaceable by others on condition that the mathematics are tractable.

B. Model Optimization

There are total five variables, *i.e.*, $\mathbf{A}$, $\mathbf{\Theta}$, $\mathbf{F}$, $\mathbf{R}$ and $\mathbf{Y}$, in the JGSED objective function (14). In this section, we apply the alternative optimization approach to update these variables; that is, updating one with the others fixed. Below are the detailed derivations.

1) Update $\mathbf{A}$*:* The sub-objective associated with $\mathbf{A}$ is

$$\min_{\mathbf{A}} \sum_{i,j=1}^{n} (\|\mathbf{\Theta}\mathbf{x}_i - \mathbf{\Theta}\mathbf{x}_j\|_2^2 a_{ij} + \gamma a_{ij}^2) + 2\lambda \mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}),$$
$$\text{s.t. } \forall i,\ \mathbf{a}^i\mathbf{1} = 1, \mathbf{a}^i \geq \mathbf{0}. \tag{15}$$

Since $\mathbf{f}^i \in \mathbb{R}^{1\times c}$ is the i-th row of $\mathbf{F}$, it can be verified that

$$2\mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \left\|\mathbf{f}^i - \mathbf{f}^j\right\|_2^2 a_{ij}. \tag{16}$$

Then objective function (15) can be rewritten as

$$\min_{\mathbf{A}} \sum_{i,j=1}^{n} \left[(\|\mathbf{\Theta}\mathbf{x}_i - \mathbf{\Theta}\mathbf{x}_j\|_2^2 + \lambda\|\mathbf{f}^i - \mathbf{f}^j\|_2^2)a_{ij} + \gamma a_{ij}^2\right],$$
$$\text{s.t. } \forall i,\ \mathbf{a}^i\mathbf{1} = 1, \mathbf{a}^i \geq \mathbf{0}. \tag{17}$$

Denote that $\mathbf{d}^i \in \mathbb{R}^{1\times n}$ is an vector whose j-th element is $d_{ij} = \|\mathbf{\Theta}\mathbf{x}_i - \mathbf{\Theta}\mathbf{x}_j\|_2^2 + \lambda\|\mathbf{f}^i - \mathbf{f}^j\|_2^2$. Since different row vectors of $\mathbf{A}$ are independent, problem (15) can be decoupled to the following optimization problem

$$\begin{aligned}
&\min_{\mathbf{a}^i\mathbf{1}=1,\mathbf{a}^i\geq\mathbf{0}} \sum_{j=1}^{n}(d_{ij}a_{ij} + \gamma a_{ij}^2) \\
\Leftrightarrow\ &\min_{\mathbf{a}^i\mathbf{1}=1,\mathbf{a}^i\geq\mathbf{0}} \gamma\sum_{j=1}^{n}\left(a_{ij}^2 + \frac{1}{\gamma}d_{ij}a_{ij} + \frac{1}{4\gamma^2}d_{ij}^2\right) \\
\Leftrightarrow\ &\min_{\mathbf{a}^i\mathbf{1}=1,\mathbf{a}^i\geq\mathbf{0}} \left\|\mathbf{a}^i + \frac{1}{2\gamma}\mathbf{d}^i\right\|_2^2,
\end{aligned} \tag{18}$$

which defines a squared Euclidean distance on a simplex constraint. This can be efficiently solved by the method proposed in [47]. Usually, the parameter γ can be tuned by grid search and below we provide a more elegant approach to determine it.

Inspired by [35], it is preferred to learn a sparse $\mathbf{a}^i$, meaning that $\mathbf{x}_i$ is only associated with its k nearest neighbors. Then $\mathbf{a}^i$ has exact k nonzero elements. Supposing that the elements of $\mathbf{d}^i$ (denoted as $d_{i1}, d_{i2}, \ldots, d_{in}$) were sorted in ascending order, the overall γ can be set as

$$\gamma = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{k}{2}d_{i,k+1} - \frac{1}{2}\sum_{j=1}^{k}d_{ij}\right). \tag{19}$$

Therefore, tuning the regularization parameter γ can be completed by tuning the neighborhood size parameter k, which is much easier since k is an integer with explicit meaning. The optimal solution to $\mathbf{a}^i$ is

$$a_{ij} = \left[\frac{d_{i,k+1} - d_{ij}}{kd_{i,k+1} - \sum_{j=1}^{k} d_{ij}}\right]_+, \tag{20}$$

which indicates that the similarity graph $\mathbf{A}$ can be initialized by

$$a_{ij} = \left[\frac{d_{i,k+1}^x - d_{ij}^x}{kd_{i,k+1}^x - \sum_{j=1}^{k} d_{ij}^x}\right]_+. \tag{21}$$

Here $d_{ij}^x = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$.

2) Update $\mathbf{\Theta}$*:* When the other variables are fixed, problem (14) becomes

$$\min_{\mathbf{\Theta}} \sum_{i,j=1}^{n} \|\mathbf{\Theta}\mathbf{x}_i - \mathbf{\Theta}\mathbf{x}_j\|_2^2 a_{ij},$$
$$\text{s.t. } \boldsymbol{\theta}^T\mathbf{1} = 1, \boldsymbol{\theta} \geq \mathbf{0}, \mathbf{\Theta} = \mathrm{diag}(\boldsymbol{\theta}). \tag{22}$$

Based on (16), by setting $\mathbf{q}_i = \mathbf{\Theta}\mathbf{x}_i \in \mathbb{R}^{d\times 1}$ and $\mathbf{Q} = \mathbf{X}^T\mathbf{\Theta} \in \mathbb{R}^{n\times d}$, we have

$$\begin{aligned}
&\min_{\mathbf{\Theta}} \sum_{i=1}^{n}\sum_{j=1}^{n} \|\mathbf{q}_i - \mathbf{q}_j\|_2^2 a_{ij} = \min_{\mathbf{Q}} 2\mathrm{Tr}(\mathbf{Q}^T\mathbf{L}\mathbf{Q}) \\
\Leftrightarrow\ &\min_{\mathbf{\Theta}} \mathrm{Tr}(\mathbf{\Theta}\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{\Theta}).
\end{aligned} \tag{23}$$

Denote $\mathbf{M} = \mathbf{X}\mathbf{L}\mathbf{X}^T \in \mathbb{R}^{d\times d}$ and m_i^* is the i-th diagonal element of $\mathbf{M}$, we have

$$\min_{\theta_i} \sum_{i=1}^{d} \theta_i^2 m_i^*, \text{ s.t. } \boldsymbol{\theta}^T\mathbf{1} = 1, \ \boldsymbol{\theta} \geq \mathbf{0}. \tag{24}$$

Obviously, even if $\theta_i < 0, \theta_i^2$ is still positive and therefore we can discard the non-negative constraint $\boldsymbol{\theta} \geq \mathbf{0}$. Then, the Lagrange function of (24) is

$$\mathcal{L}(\theta_i, z) = \sum_{i=1}^{d} \theta_i^2 m_i^* + z\left(\sum_{i=1}^{d} \theta_i - 1\right), \tag{25}$$

where z is a Lagrange multiplier. By taking the derivative with respect to θ_i and setting it to zero, we have

$$\frac{\partial \mathcal{L}(\theta_i, z)}{\partial \theta_i} = 2\theta_i m_i^* + z = 0 \Rightarrow \theta_i = \frac{-z}{2m_i^*}. \tag{26}$$

Considering the constraint $\boldsymbol{\theta}^T\mathbf{1} = 1$, we have

$$z = \frac{-2}{\sum_{i=1}^{d} \frac{1}{m_i^*}}, \tag{27}$$

and get the final solution to θ_i as

$$\theta_i = \frac{1}{m_i^* \sum_{i=1}^{d} \frac{1}{m_i^*}}. \tag{28}$$

We denote $\mathbf{x}_{.i} \in \mathbb{R}^{n\times 1}$ as the i-th row of $\mathbf{X}$. Since m_i^* is real-valued and can be represented as $m_i^* = \mathbf{x}_{.i}^T\mathbf{L}\mathbf{x}_{.i}$, we get

$$\mathbf{x}_{.i}^T\mathbf{L}\mathbf{x}_{.i} = \text{Tr}(\mathbf{x}_{.i}^T\mathbf{L}\mathbf{x}_{.i}) \Leftrightarrow \sum_{j=1}^{n}\sum_{p=1}^{n} \|x_{.i,j} - x_{.i,p}\|_2^2 \, a_{jp}. \tag{29}$$

According to our prior knowledge, both $\|x_{.i,j} - x_{.i,p}\|_2^2$ and a_{jp} are non-negative, which implies that $m_i^* \geq 0$. Hence, $\theta_i \geq 0$ and we can come to the conclusion that $\boldsymbol{\theta} \geq \mathbf{0}$.

3) Update $\mathbf{F}$*:* The objective function in terms of variable $\mathbf{F}$ is

$$\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}} \text{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) + \frac{\beta}{2\lambda}\left\|\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}} - \mathbf{F}\mathbf{R}\right\|_2^2. \tag{30}$$

Denoting that $\mathbf{M}_Y = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}$ and $v = \frac{\beta}{2\lambda}$, since both $\mathbf{R}$ and $\mathbf{F}$ are orthonormal matrices, the above problem is equivalent to

$$\begin{aligned}
&\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}} \text{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) + v\left\|\mathbf{M}_Y - \mathbf{F}\mathbf{R}\right\|_2^2 \\
\Leftrightarrow &\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}} \text{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) + v\text{Tr}(\mathbf{F}\mathbf{R}\mathbf{R}^T\mathbf{F}^T) - 2v\text{Tr}(\mathbf{F}^T\mathbf{M}_Y\mathbf{R}^T) \\
\Leftrightarrow &\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}} \text{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) - 2v\text{Tr}(\mathbf{F}^T\mathbf{M}_Y\mathbf{R}^T).
\end{aligned} \tag{31}$$

This is the typical form of the quadratic problem on the Stiefel manifold (QPSM) [48] and can be further relaxed to

$$\max_{\mathbf{F}^T\mathbf{F}=\mathbf{I}} \text{Tr}(\mathbf{F}^T\bar{\mathbf{L}}\mathbf{F}) + 2v\text{Tr}(\mathbf{F}^T\mathbf{B}), \tag{32}$$

where $\bar{\mathbf{L}} = \eta\mathbf{I} - \mathbf{L} \in \mathbb{R}^{n\times n}$ and $\mathbf{B} = \mathbf{M}_Y\mathbf{R}^T \in \mathbb{R}^{n\times c}$. The relaxation parameter η can be easily set as the largest eigenvalue of $\mathbf{L}$.

Algorithm 1: The GPI-based method to solve problem (32).

Input: The Laplacian matrix $\mathbf{L} \in \mathbb{R}^{n\times n}$, matrix $\mathbf{B} \in \mathbb{R}^{n\times c}$, and parameters η, v;
Output: The scaled cluster indicator matrix $\mathbf{F} \in \mathbb{R}^{n\times c}$.

```
Initialize F which satisfies F^T F = I;
Compute L̄ = ηI − L;
while not converged do
    Update M_F = 2L̄F + 2vB;
    Calculate the compact SVD of M_F as
    M_F = UΣV^T, where U ∈ R^{n×c}, Σ ∈ R^{c×c},
    V ∈ R^{c×c};
    Update F = UV^T;
end while
```

According to the generalized power iteration method (GPI), we summarize the optimization method to $\mathbf{F}$ in Algorithm 1.

4) Update $\mathbf{R}$*:* When considering the variable $\mathbf{R}$ only, problem (14) degenerates to

$$\min_{\mathbf{R}^T\mathbf{R}=\mathbf{I}} \left\|\mathbf{F}\mathbf{R} - \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\right\|_2^2. \tag{33}$$

As previously denoted that $\mathbf{M}_Y = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}$, the above objective function can be simplified as

$$\begin{aligned}
&\min_{\mathbf{R}^T\mathbf{R}=\mathbf{I}} \|\mathbf{F}\mathbf{R} - \mathbf{M}_Y\|_2^2 \\
\Leftrightarrow &\min_{\mathbf{R}^T\mathbf{R}=\mathbf{I}} \text{Tr}(\mathbf{F}\mathbf{R}\mathbf{R}^T\mathbf{F}^T) - 2\text{Tr}(\mathbf{F}^T\mathbf{M}_Y\mathbf{R}^T) \\
\Leftrightarrow &\max_{\mathbf{R}^T\mathbf{R}=\mathbf{I}} \text{Tr}(\mathbf{M}_Y^T\mathbf{F}\mathbf{R}),
\end{aligned} \tag{34}$$

which can be further represented as

$$\max_{\mathbf{R}^T\mathbf{R}=\mathbf{I}} \text{Tr}(\mathbf{M}_R\mathbf{R}). \tag{35}$$

Here $\mathbf{M}_R = \mathbf{M}_Y^T\mathbf{F} = (\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{F}$. Let the SVD of matrix $\mathbf{M}_R$ be $\mathbf{M}_R = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T$, and then we have $\text{Tr}(\mathbf{M}_R\mathbf{R}) = \text{Tr}(\mathbf{R}\mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T) = \text{Tr}(\mathbf{\Sigma}_r\mathbf{E}) = \sum_{i=1}^{c} \lambda_{ii}e_{ii}$, where $\mathbf{E} = \mathbf{V}_r^T\mathbf{R}\mathbf{U}_r$, λ_{ii} and e_{ii} are (i,i)-th element of matrix $\mathbf{\Sigma}_r$ and $\mathbf{E}$ respectively.

Because $\mathbf{E}^T\mathbf{E} = \mathbf{U}_r^T\mathbf{R}^T\mathbf{V}_r\mathbf{V}_r^T\mathbf{R}\mathbf{U}_r = \mathbf{I}$ meaning that $\sum_{j=1}^{c} e_{ji}^2 = 1$, we have $e_{ii} \leq 1$ $(1 \leq i \leq c)$. Meanwhile, λ_{ii} is the i-th singular value which should be non-negative. That is $\text{Tr}(\mathbf{M}_R\mathbf{R}) = \sum_{i=1}^{c} \lambda_{ii}e_{ii} \leq \sum_{i=1}^{c} \lambda_{ii}$. Hence, we can infer that the maximum of $\text{Tr}(\mathbf{M}_R\mathbf{R})$ is obtained when $\mathbf{E} = \mathbf{I} = \mathbf{V}_r^T\mathbf{R}\mathbf{U}_r$. Then, the optimal solution to $\mathbf{R}$ is

$$\mathbf{R} = \mathbf{V}_r\mathbf{U}_r^T. \tag{36}$$

5) Update $\mathbf{Y}$*:* We optimize $\mathbf{Y}$ by addressing the following problem

$$\begin{aligned}
&\min_{\mathbf{Y}\in\text{Ind}} \left\|\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}} - \mathbf{F}\mathbf{R}\right\|_2^2 \\
\Leftrightarrow &\max_{\mathbf{Y}\in\text{Ind}} \text{Tr}(\mathbf{M}_Y^T\mathbf{F}\mathbf{R}).
\end{aligned} \tag{37}$$

Denote $\mathbf{G} = \mathbf{FR}$, then objective function (37) can be rewritten as

$$\max_{\mathbf{Y}\in\text{Ind}} \text{Tr}((\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{G}). \tag{38}$$

Let $\mathbf{W} = \mathbf{Y}^T\mathbf{Y}$, we have $w_{ij} = \sum_{l=1}^{n} y_{li}y_{lj}$. Since only one element in $\mathbf{y}^i$ is one and the others are zeros, we can derive that $y_{li}y_{lj} = 0$ and $w_{ij} = 0$ if $i \neq j$. Therefore, $\mathbf{W}$ should be diagonal and the (j, j)-th diagonal element of $\mathbf{W}^{-\frac{1}{2}}$ is $1/\sqrt{\mathbf{y}_j^T\mathbf{y}_j}$. Then, problem (38) is equivalent to

$$\max_{\mathbf{Y}\in\text{Ind}} \sum_{j=1}^{c} \frac{\mathbf{y}_j^T\mathbf{g}_j}{\sqrt{\mathbf{y}_j^T\mathbf{y}_j}}. \tag{39}$$

Since samples are independent to each other, we can solve $\mathbf{Y}$ in row-wise manner, which updates one row of $\mathbf{Y}$ by fixing the others. When solving $\mathbf{y}^i$, we only consider the increment of the objective function (39) from $y_{ij} = 0$ to $y_{ij} = 1$. Assume that the optimal solution is $\bar{\mathbf{Y}}$, and then we can calculate $\bar{\mathbf{y}}_j^T\bar{\mathbf{y}}_j$ and $\bar{\mathbf{y}}_j^T\mathbf{g}_j$ in advance before updating $\mathbf{y}^i$. The increment can be calculated as

$$s_{ij} = \frac{\bar{\mathbf{y}}_j^T\mathbf{g}_j + g_{ij}(1-\bar{y}_{ij})}{\sqrt{\bar{\mathbf{y}}_j^T\mathbf{y}_j + (1-\bar{y}_{ij})}} - \frac{\bar{\mathbf{y}}_j^T\mathbf{g}_j - \bar{y}_{ij}g_{ij}}{\sqrt{\bar{\mathbf{y}}_j^T\mathbf{y}_j - \bar{y}_{ij}}}. \tag{40}$$

Finally, the optimal solution to $\mathbf{y}^i$ is

$$y_{ij} = \langle j = \arg\max_{j'\in[1,c]} s_{ij'}\rangle, \tag{41}$$

where $\langle\cdot\rangle$ is one if the argument is true or zero otherwise.

We summarize the detailed optimization procedure to problem (14) in Algorithm 2.

Algorithm 2: The optimization procedure to problem (14).

Input: Data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$, cluster number c, parameter β, and neighborhood parameter k;
Output: The binary cluster indicator matrix $\mathbf{Y}$.

1: Initialize a sparse graph $\mathbf{A} \in \mathbb{R}^{n\times n}$ by (21);
2: Initialize $\gamma = \lambda = \frac{1}{n}\sum_{i=1}^{n}(\frac{k}{2}d_{i,k+1}^x - \frac{1}{2}\sum_{j=1}^{k} d_{ij}^x)$ where $d_{ij}^x = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$;
3: Compute $\mathbf{L} = \mathbf{D} - (\mathbf{A} + \mathbf{A}^T)/2$ where $\mathbf{D}$ is the diagonal degree matrix with i-th diagonal element $d_{ii} = \sum_{j=1}^{n} a_{ij}$;
4: Compute $\mathbf{F}$ formed by the c eigenvectors of $\mathbf{L}$ corresponding to the c smallest eigenvalues;
5: Initialize $\mathbf{Y} \in \mathbb{B}^{n\times c}$ according to $\mathbf{Y}^* = \text{diag}(\mathbf{FF}^T)^{-\frac{1}{2}}\mathbf{F}$ and $y_{ij} = \langle j = \arg\max_{j'\in[1,c]} y_{ij'}^*\rangle$;
6: **while** not converged **do**
7: Update $\mathbf{\Theta} \in \mathbb{R}^{d\times d}$ with its i-th diagonal element as $\theta_i = 1/(m_i^*\sum_{i=1}^{d}\frac{1}{m_i^*})$ where m_i^* is the i-th diagonal element of $\mathbf{M} = \mathbf{XLX}^T$;
8: Update $\mathbf{R} = \mathbf{U}_r\mathbf{V}_r^T$ where the compact SVD of $\mathbf{M}_R = (\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{F}$ is $\mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T$;
9: Construct $\bar{\mathbf{L}} = \eta\mathbf{I} - \mathbf{L}$ where η is the domain eigenvalue of $\mathbf{L}$ and $\mathbf{B} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{R}^T$, calculate $v = \frac{\beta}{2\lambda}$;
10: Update $\mathbf{F}$ according to Algorithm 1;
11: Calculate $\mathbf{G} = \mathbf{FR}$ and update $\mathbf{Y}$ according to (41);
12: Update $\mathbf{A}$ row by row, the i-th row of $\mathbf{A}$ is updated by solving (20);
13: Recompute $\mathbf{L} = \mathbf{D} - (\mathbf{A} + \mathbf{A}^T)/2$;
14: **end while**

C. Analysis of Computational Complexity and Convergence

Generally, spectral clustering methods consist of three separated steps, graph construction, spectral embedding and post-processing. The state-of-art spectral clustering methods such as the constrained Laplacian rank (CLR) [45], CAN [35] and SWCAN [39] jointly perform graph construction and spectral embedding, which share the same computational complexity $\mathcal{O}(n^3)$ because they directly perform eigen-decomposition operation on an $n \times n$ graph Laplacian matrix. Besides, if spectral rotation is used as the post-processing method, the total complexity will be $\mathcal{O}(c^3 + tnc^2)$ where t is its average number of iterations. As a result, the time complexity of popular spectral clustering methods is $\mathcal{O}(n^3)$.

For JGSED, we adopted an alternative framework to optimize the objective function. Variables $\mathbf{A}$ and $\mathbf{Y}$ are updated row-wisely and the complexities of updating each row of them are $\mathcal{O}(n)$ and $\mathcal{O}(c)$, respectively. Therefore, the time complexities of updating variables $\mathbf{A}$ and $\mathbf{Y}$ are $\mathcal{O}(n^2)$ and $\mathcal{O}(nc)$. When updating $\mathbf{R}$, it costs $\mathcal{O}(c^3)$ because we performed singular value decomposition on $\mathbf{M}_R \in \mathbb{R}^{c\times c}$. According to the GPI method, the complexity of updating $\mathbf{F}$ is $\mathcal{O}(n^2c)$ [48]. When updating $\mathbf{\Theta}$, the complexity mainly comes from the calculation of $\mathbf{M} = \mathbf{XLX}^T$, which has the complexity of $\mathcal{O}(nd^2 + n^2d)$. Assuming that T is the maximum number of iterations, t_1, t_2 and t_3 are the average numbers of iterations to update $\mathbf{A}$, $\mathbf{F}$ and $\mathbf{Y}$ respectively. Therefore, the time complexity of the proposed JGSED is $\mathcal{O}(T(t_1n^2 + nd^2 + n^2d + t_2n^2c + t_3nc + c^3))$. On the convergence property of the JGSED model, we learn from the model optimization that variables $\mathbf{R}$ and $\mathbf{\Theta}$ have analytical solutions in each iteration. When updating $\mathbf{F}$, we adopted the GPI method to solve the corresponding QPSM objective function. Inspired by [48], it converges to the global optimum which accordingly guarantees the convergence of Algorithm 1. When updating $\mathbf{Y}$, it is optimized row-wisely by utilizing the independence among samples. According to the updating rule (41), each row of $\mathbf{Y}$ has exact one non-zero element and the optimal one should be one of the c feasible possibilities. For the variable $\mathbf{A}$, problem (18) can be solved with a closed form solution [35]. In summary, the optimization procedure of JGSED objective function shown in Algorithm 2 is expected to be convergent.

IV. Experiments on Synthetic Data

In this section, we evaluate the performance of JGSED on some synthetic data sets and verify the rationality of the feature importance measure $\boldsymbol{\theta}$.

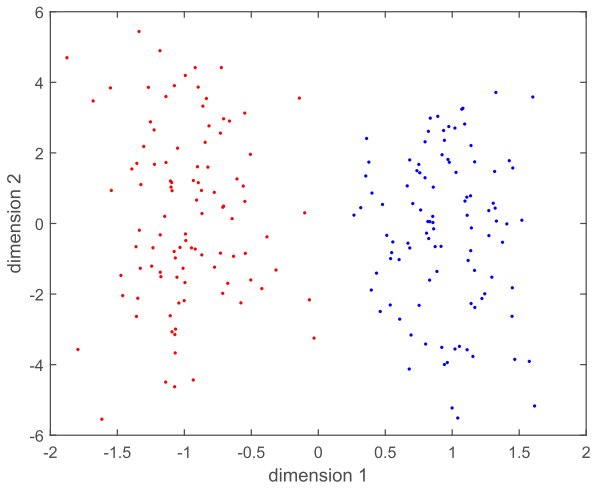


Fig. 2. The 2D visualization of the two-gaussian data. The learned weights for the five dimensions by JGSED are 0.9820, 0.0047, 0.0044, 0.0044, and 0.0045, respectively.

A. Rationality Verification of θ on Synthetic Data

A synthetic two-gaussian data set is used in this experiment. It has two clusters and each has 100 samples. The dimensionality is five. The first two dimensions are distributed in a sphere while the latter three dimensions are deliberately fabricated noise. The first two dimensions of this two-gaussian data set are visualized in Fig. 2. It is obvious that the first dimension is discriminative and the second one is not. By training JGSED on this data set, we get the weights for the first dimension, the second dimension and all the others are respectively 0.9820, 0.0047 and 0.0133. This result coincides with our intuitive understanding that more discriminative dimension should be assigned larger weight.

B. Clustering Performance Evaluation

The two synthetic data sets used in this subsection are described below.

- *Two-gaussian Data Set*. We synthesized this data set using three clusters with 166 samples per cluster. In the total four dimensions, a three-sphere shape is presented by the former two dimensions, while the latter two dimensions are useless.
- *Multi-cluster Data Set*. This data set has 64 clusters each of which has 15 samples. The dimensionality is five, among which the distribution of the first two dimensions has a spherical form and the third one is noisy.

We evaluate the clustering performance of JGSED on these two data sets by considering SWCAN as a baseline model. For the three-gaussian data set, we additionally included two noisy dimensions to the original data. In Fig. 3(a), we respectively show the original three-gaussian synthetic data set, clustering results of SWCAN and JGSED. Generally, we find that both SWCAN and JGSED are competent for this clustering task by handling data with noisy dimensions. In Fig. 3(b), some data points which are correctly clustered by JGSED but wrongly clustered by SWCAN were marked with red circles, demonstrating that our JGSED is more competitive in dealing with data points in boundary regions.

For the multi-cluster data set, we respectively show the original data points, clustering results of SWCAN and JGSED in Fig. 3(d), (e), and (f). The clustering accuracy of JGSED is slightly higher than that of SWCAN because the five marked data points in Fig. 3(f) were correctly clustered. It can be inferred from these two toy experiments that JGSED has more powerful ability in handling hard samples, leading to improved performance in comparison with SWCAN.

C. Robustness Evaluation

The two synthetic data sets used in this subsection are described below.

- *Two-moon Data Set*. The data set is constructed with two crescent-shaped clusters on 2D plane and two additional noise dimensions. Each cluster has 250 samples.
- *Three-ring Data Set*. We designed this data set with three cocentric circles and each corresponds to a cluster. The numbers of the three circles are 100, 300, 600 respectively. The latter two of the total four feature dimensions are noisy.

In this section, we presented experiments on these two synthetic data sets by respectively adding small and large noisy dimensions to investigate the robustness of our proposed JGSED model. We compared JGSED with $k$means, CAN, and SWCAN.

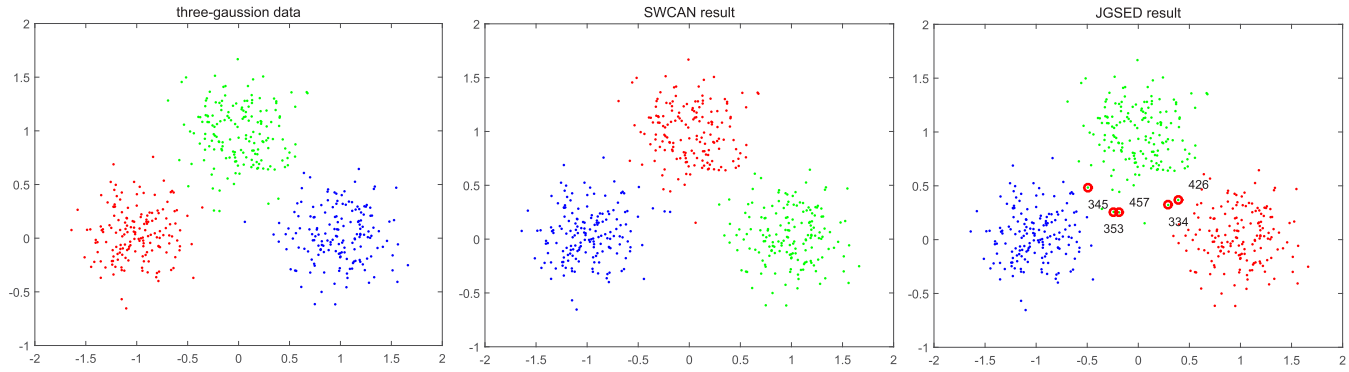
The first one is the two-moon data set which has two noisy dimensions. In Fig. 4(a), we respectively show the results of $k$means, CAN, SWCAN, and JGSED in clustering this data set with small noisy dimensions. Fig. 4(b) corresponds to the clustering results of these four models under the large noise setting. It is observed that $k$means fails to handle this data set and CAN is only suitable for clustering this data with small noisy dimensions. SWCAN and JGSED obtained promising results in both settings, indicating their desirable robustness. Similarly, Fig. 4(c) and (d) show the clustering results of these four models on the three-ring data set with small and large noisy dimensions, respectively. These results indicate that on this data set, all the CAN, SWCAN, and JGSED models can handle both small and large noise settings except for $k$means. Overall, we can simply draw a conclusion that JGSED can effectively process data with noisy dimensions whose adverse effects could be suppressed by the self-weighted variable.

V. Experiments on Benchmark Data

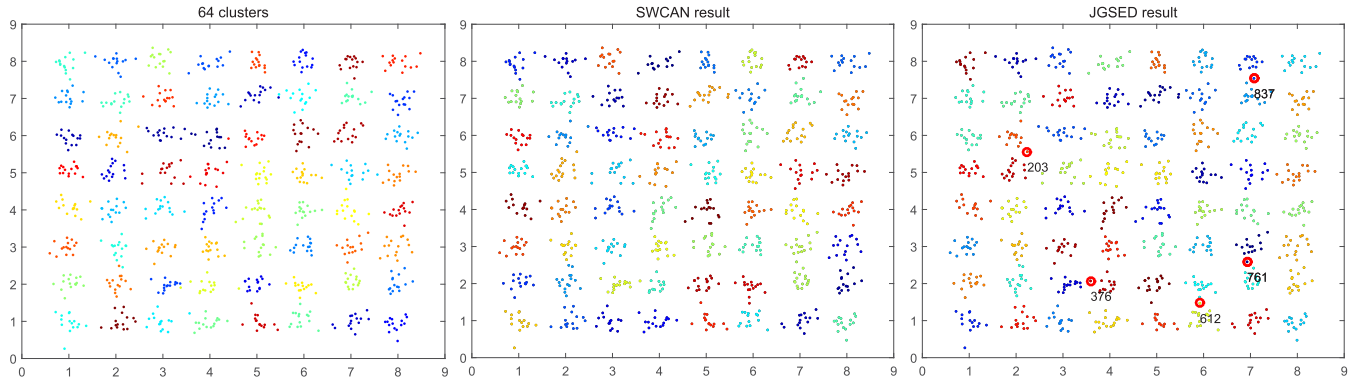
In this section, several benchmark data sets are used to show the learned feature self-weighted variable by JGSED and its clustering performance.

A. Rationality Verification of θ on AR

We randomly selected 150 images without scarf and 150 images with scarves from the AR face image data set, forming a subset with two clusters. The size of each face image is 165×120; therefore, the dimensionality of each sample is 19800. From Fig. 5(a), We can easily find that the features within the scarf block are more discriminative, meaning that these feature



(a) Clustering results of SWCAN and JGSED on the three-gaussian data set.



(b) Clustering results of SWCAN and JGSED on a multi-cluster data set.

Fig. 3. Clustering performance evaluation of SWCAN and JGSED on the three-gaussian and multi-cluster synthetic data sets.

dimensions should be assigned larger weights. Similar to the pipeline in [49], we trained JGSED on this customized data set and reshaped the learned $\boldsymbol{\theta}$ in 2D representation by the left-most one in Fig. 5(d) where brighter pixels correspond to larger assigned weights. Because the scarf block occupies approximately 40% of the total area of each face image, we calculated the sum of feature weights within this block. By repeating the experiment 20 times, we obtained that the average value of the summed feature weights within this block is 0.7640.

Moreover, we performed experiments on another two binary-cluster data sets cropped from AR. As shown in Fig. 5(b), we aim to group these face images into two clusters according to the gender information. Fig. 5(c) shows the male face images with sunglasses and the female face images without sunglasses, and our task is to perform clustering according to the sunglass information but ignoring the gender information. The learned weights corresponding to these two cropped data sets are respectively shown as the middle and the right-most two figures in Fig. 5(d). These experimental results effectively verified that the weights obtained by JGSED are reasonable.

B. Benchmark Data Sets

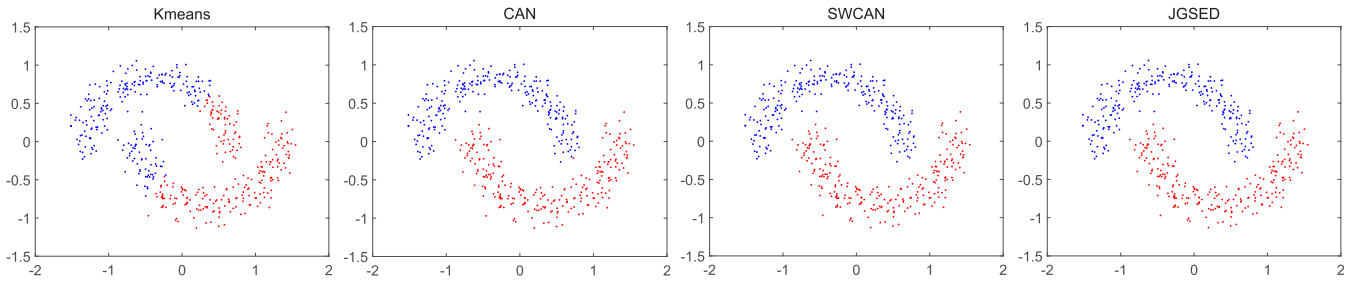
Ten benchmark data sets were used in the following experiments including the glass, vehicle, jaffe, umist, Yale, YaleB, Binalpha, AT&T, COIL20, and MSRA25. We summarized the main characteristics of these data sets in Table II.

TABLE II
MAIN CHARACTERISTICS OF THE BENCHMARK DATA SETS

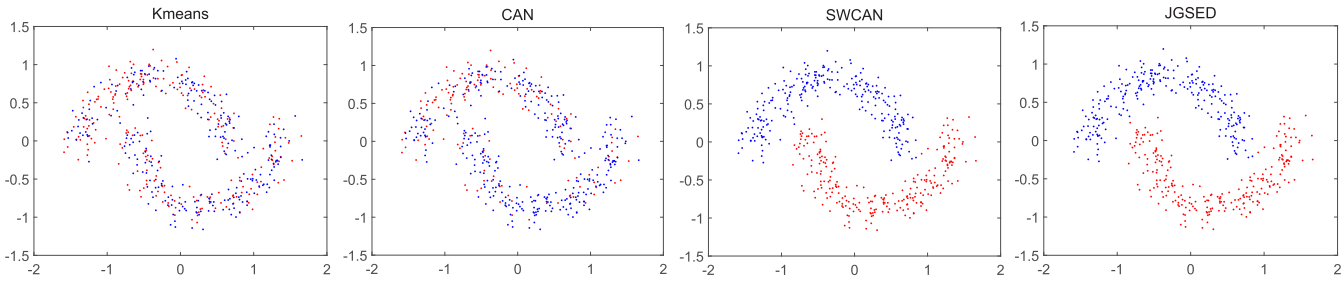
Data sets	Number of Samples	Dimensions	Clusters
glass	214	9	6
vehicle	864	18	4
jaffe	212	177	7
umist	575	644	20
YaleB	2414	1024	38
Yale	165	105	15
Binalpha	1404	320	36
AT&T	400	1024	36
COIL20	1440	1024	20
MSRA25	1799	256	12

C. Experimental Settings

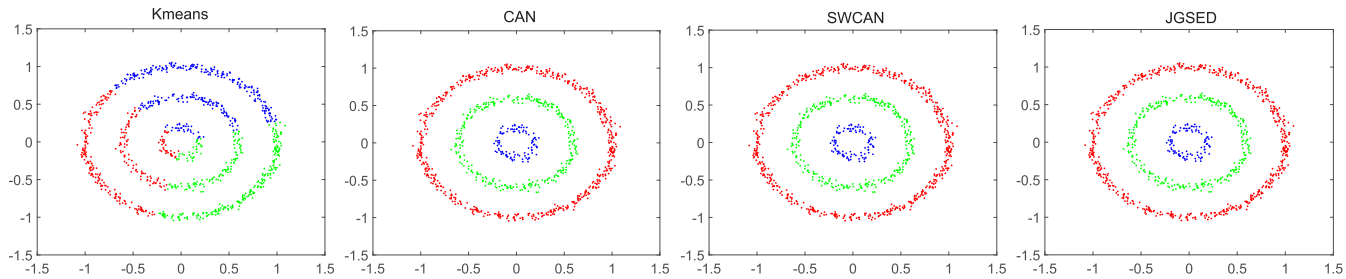
Besides the $k$means, we compared JGSED with two classical spectral clustering models including NCut and ratio cut (RCut), two graph learning models including the sparse subspace clustering (SSC) based on sparse representation and the graph based on low-rank representation (LRR), one joint model to simultaneously perform spectral embedding and spectral rotation-based post-processing (*i.e.*, JSESR), five joint graph construction and spectral embedding models (*i.e.*, CAN, PCAN, SWCAN, DOGC and LAPIN). $k$means was adopted as the discretization method for NCut and RCut. For $k$means, we repeated it 50 times with



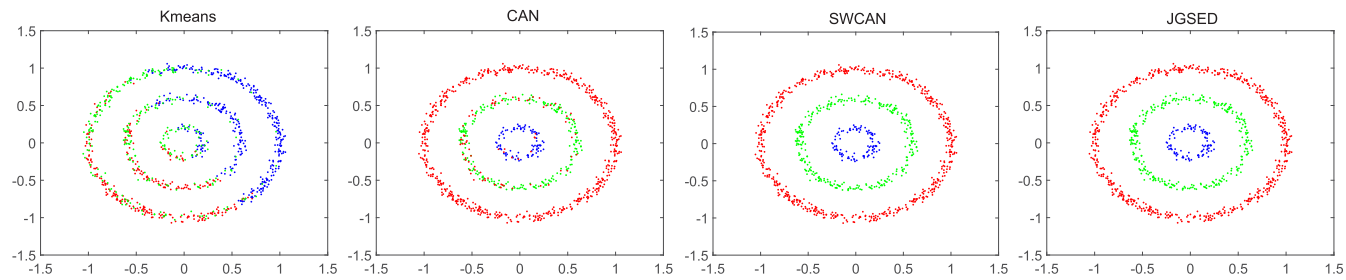
(a) Clustering results on the two-moon data set with small noisy dimensions.



(b) Clustering results on the two-moon data set with large noisy dimensions.



(c) Clustering results on the three-ring data set with small noisy dimensions.



(d) Clustering results on the three-ring data set with large noisy dimensions.

Fig. 4. Model robustness evaluation of $k$means, CAN, SWCAN, and JGSED on the two-moon and three-ring synthetic data sets.

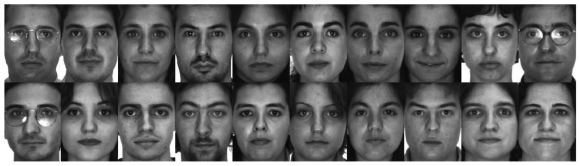
different initializations and the best result in terms of its objective function values was recorded.

For the graph learning models, LRR and SSC, the regularization parameter α in SSC was tuned from $\{1, 2, \cdots, 500\}$, the threshold and maximal iteration number are set as 1.0×10^{-4} and 100, respectively. As for SSC, we tuned the regularization parameter λ from $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$.

For the graph-based clustering models, NCut, RCut and JSESR, we constructed the initial similarity graph by using the 'Heatkernel' function in which the number of the nearest neighbors was set as five and the bandwidth parameter was set as one. All these three models were repeated 20 times. Since there is a free regularization parameter α in JSESR, we tuned it from candidate values $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$ to let JSESR achieve the best performance.

For the joint graph construction and spectral embedding models, CAN, PCAN, and SWCAN, they were executed once since their results are stable. Inspired by [39], λ in SWCAN is not a tradeoff parameter since it should be a large enough value. The parameter λ is initialized the same value as γ and adjusted by investigating the number of connected components in the learned graph $\mathbf{A}$. When the number of connected components is greater than c, λ is decreased by $\lambda = \lambda/2$; otherwise, it is increased by $\lambda = \lambda \times 2$. By the way, γ can be precomputed by (19) which relies on the neighborhood size parameter k. That is, SWCAN has only one parameter k to adjust. The neighborhood

(a) Sample face images with/without scarves.



(b) Male and female face images without occlusion.

(c) Male face images with sunglasses and female face images without sunglasses.

(d) Learned $\boldsymbol{\theta}$s corresponding to samples in Fig. 5a)-c)

Fig. 5. Sample face images from cropped AR data sets and the corresponding learned feature weights.

size parameter k in CAN, PCAN, and SWCAN were tuned from 5 to 100 with a step size 5. For both DOGC and LAPIN, we use the source codes provided by the authors and the related model parameters are set as described by the original papers.

For our proposed JGSED model, the three parameters γ, λ and β can be reduced to two tradeoff parameters k and β with the same technique in SWCAN. The neighborhood size parameter k was tuned from 5 to 100 and the regularization parameter β was tuned from $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. The maximal number of iterations T for optimizing the JGSED was set as 30 and we ran it only once with the initial graph described by (21).

For all the compared models, all the features were standardized to [0, 1] before experiments. The number of clusters were set equal to the ground-truth value. The clustering performance was evaluated in terms of the three widely used metrics, *Acc*, *NMI* and *Purity* [47], and the larger values of them indicate better clustering performance. The average clustering results and standard deviations were reported.

D. Experimental Results and Analysis

The clustering results of all the compared models on these benchmark data sets are reported in Table III in which the best results are highlighted in boldface. It is observed that JGSED significantly outperforms the other models in most cases especially on the data sets of glass, vehicle, jaffe, YaleB and MSRA25. Specifically, we have the following three findings.

- It can be observed that JGSED has superior performance to both SSC and LRR on most of the data sets. From our point of view, the reasons accounting for such phenomenon are two folds. One is that the graph similarity matrices based on the self-representation coefficient matrices cannot capture the underlying connections of data points effectively. SSC enforces the element-level sparsity of the coefficient matrix while LRR enforces the singular values of the representation matrix to be sparse. The other is that both SSC and LRR pay only attention to the graph learning whilst the subsequent steps of spectral embedding and discretization are separately performed.
- By comparing the results respectively obtained by JSESR and JGSED, the latter achieved better performance on nine of the total ten benchmark data sets. The difference between them is that the latter involves the graph learning process into the whole model training process. Therefore, the experimental results of JGSED demonstrate that jointly performing graph construction, spectral embedding and rotation can effectively avoid the information loss in graph construction. Meanwhile, the graph similarity matrix can be continuously optimized in iterations, which in turn further improves the learning of the resultant cluster indicator matrix.
- On the difference between SWCAN and JGSED, we know that the spectral rotation-based discretization operation has been seamlessly integrated in the JGSED model objective function. After it is fitted by data, we directly obtain the binary cluster indicator matrix $\mathbf{Y}$. However, in SWCAN, though the graph similarity matrix is expected to have block diagonal structures corresponding to the clusters, it is not always the case in dealing with difficult data sets. In addition, $k$means is used to judge the cluster assignment of each data point according to the learned continuous indicator matrix $\mathbf{F}$, which has been proven to have inferior performance to spectral rotation. As a result, JGSED obtained better performance than SWCAN on the whole.
- Generally, the proposed JGSED mdoel is superior to both DOGC and LAPIN in data clustering. Though DOGC is also a unified framework to learn the graph similarity matrix and discrete cluster results, there are some obvious differences between JGSED and DOGC. First, the data similarity in DOGC is directly modeled in the original feature space rather than the projected subspace. Second, the feature importance is not considered. Third, the traditional spectral rotation is utilized in DOGC to learn an orthogonal rotation matrix which is not necessarily an orthonormal matrix [24]. LAPIN is an extension of the constrained Laplacian rank method [45] to construct an optimal bipartite graph, which can extract the duality relationship between samples and features to achieve co-clustering. Similar to DOGC, LAPIN takes the non-negative, row-normalization and rank constraint properties into account in graph construction, which neglects measuring the importance of different feature dimensions. In addition, its

TABLE III
CLUSTERING PERFORMANCE (%) OF JGSED AND THE COMPARED MODELS ON THE BENCHMARK DATA SETS

	*k*means	NCut	RCut	JSESR	LRR	SSC	CAN	PCAN	SWCAN	LAPIN	DOGC	JGSED
Acc												
glass	43.47±3.33	36.92±0.00	38.64±1.85	42.40±0.96	49.07	50.68	45.79	51.40	51.40	**54.67**	**54.67**	**54.67**
vehicle	37.29±1.11	39.95±0.02	40.79±0.82	42.74±2.70	42.43	47.28	41.25	39.72	45.15	45.50	44.61	**47.52**
jaffe	29.55±2.61	23.65±0.17	24.75±1.41	27.83±0.00	29.72	29.72	28.77	28.30	30.19	30.68	31.79	**32.08**
umist	38.55±2.16	58.19±2.30	55.27±2.98	61.03±1.54	56.17	65.74	64.87	57.74	**71.30**	**71.30**	62.17	**71.30**
YaleB	29.57±0.79	34.61±0.66	33.67±1.26	37.81±0.37	34.80	39.03	36.58	36.21	37.32	39.25	42.09	**46.73**
Yale	18.61±2.39	27.28±1.48	27.41±1.16	30.18±0.27	26.67	24.48	24.24	25.45	26.67	30.45	30.93	**32.12**
Binalpha	41.27±2.04	43.64±0.90	44.56±1.43	46.59±1.00	44.59	45.87	42.24	43.59	46.08	46.83	38.53	**47.86**
AT&T	50.97±3.04	63.59±1.56	60.75±2.17	66.15±0.80	**70.50**	63.18	56.75	57.75	60.00	66.50	62.50	65.00
COIL20	55.41±6.14	78.70±2.41	73.12±5.09	81.37±0.04	73.40	76.18	**83.54**	83.33	79.03	81.96	80.90	82.64
MSRA25	53.46±4.84	51.59±2.40	47.30±3.14	50.33±2.34	64.54	57.87	57.87	60.87	57.87	57.87	65.20	**66.76**
NMI												
glass	30.77±3.83	27.41±0.00	26.58±2.93	30.91±3.10	34.68	33.85	27.67	32.75	34.90	36.71	35.38	**37.97**
vehicle	11.73±2.42	15.14±0.02	14.89±1.74	16.58±2.37	16.88	17.96	15.04	14.46	19.24	20.66	**25.83**	20.53
jaffe	14.05±3.14	5.87±0.08	7.02±1.55	9.97±0.00	16.62	14.92	13.33	12.88	15.33	15.91	**18.06**	17.59
umist	57.64±1.74	74.30±0.95	72.53±1.63	75.18±0.72	70.22	77.68	78.21	70.05	**82.95**	**82.95**	78.55	**82.95**
YaleB	13.30±1.20	37.75±0.70	39.24±0.66	40.81±0.31	46.27	50.16	37.11	35.91	42.77	45.70	46.66	**54.64**
Yale	15.22±2.61	29.69±1.56	28.86±1.12	33.48±0.31	27.53	25.69	20.71	23.54	24.05	32.98	33.42	**33.78**
Binalpha	57.19±1.04	57.89±0.69	59.70±0.73	60.83±0.60	57.93	**62.50**	54.24	49.55	57.98	59.04	49.82	60.64
AT&T	71.76±1.60	79.18±0.64	78.39±0.74	80.37±0.39	**83.20**	79.26	76.45	73.16	76.51	82.74	78.09	81.80
COIL20	70.53±2.88	86.81±1.11	84.86±2.31	89.25±0.06	82.90	90.29	91.09	89.08	89.19	90.03	88.96	**91.17**
MSRA25	62.41±3.98	60.18±1.72	59.36±2.48	64.49±2.36	72.21	69.17	67.56	72.58	68.36	69.32	72.04	**74.32**
Purity												
glass	53.34±3.54	58.41±0.00	52.04±2.28	61.07±2.24	55.61	54.18	54.21	53.27	61.68	65.12	63.08	**65.42**
vehicle	40.31±0.64	40.66±0.02	42.12±1.80	44.39±3.65	43.74	47.28	44.21	42.91	45.15	**50.71**	44.96	49.65
jaffe	31.03±2.57	24.06±0.00	24.83±1.42	27.83±0.00	32.55	30.19	28.77	28.30	31.13	32.40	**36.79**	34.91
umist	45.10±2.22	67.65±1.66	64.97±2.25	69.91±1.01	67.65	69.91	71.48	65.57	**76.35**	**76.35**	70.87	**76.35**
YaleB	10.23±0.81	35.28±0.59	34.78±1.20	38.28±0.38	36.00	40.77	38.03	36.58	41.18	44.21	42.71	**47.93**
Yale	21.09±2.24	28.87±1.49	28.65±1.11	31.03±0.27	30.30	27.73	26.67	28.48	29.70	30.87	31.18	**33.94**
Binalpha	44.20±1.73	46.90±0.78	47.75±1.00	49.48±1.02	47.58	49.29	44.66	45.23	48.15	49.96	39.89	**50.07**
AT&T	56.11±2.55	66.80±1.05	65.15±1.85	69.20±0.41	47.58	70.35	65.00	65.25	67.25	70.03	64.75	**70.50**
COIL20	59.90±5.16	81.48±1.83	76.98±4.62	85.65±0.04	78.96	82.43	**87.36**	85.14	83.54	84.92	82.53	86.46
MSRA25	57.23±4.06	56.42±2.26	52.30±3.27	59.02±2.02	68.10	61.42	61.42	64.42	61.48	63.44	67.76	**68.54**

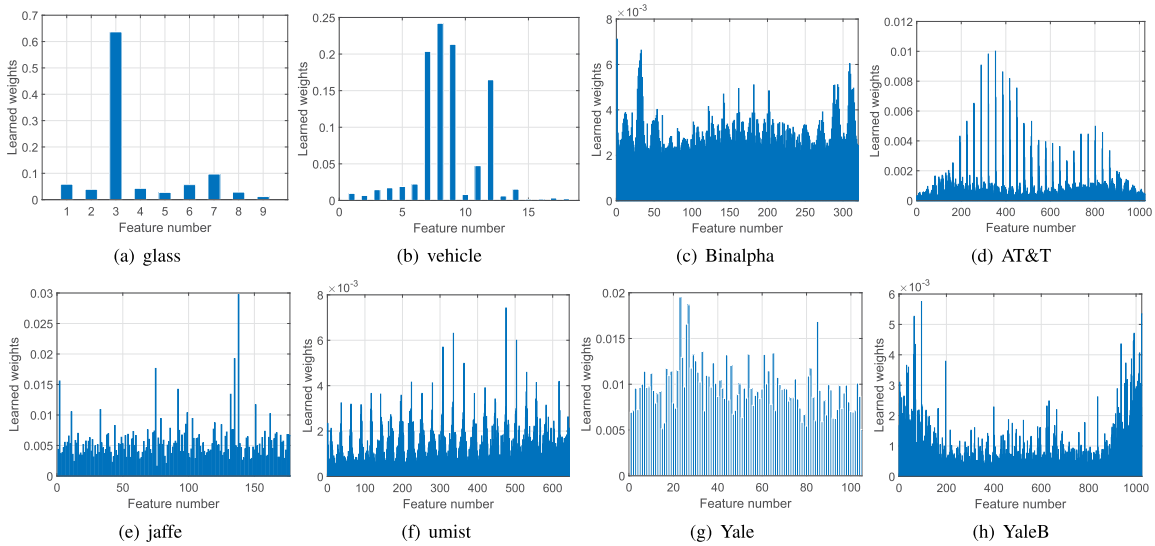


Fig. 6. The learned feature importance measure $\boldsymbol{\theta}$ on the eight data sets.

learned real-valued cluster indicator matrix still needs to be discretized to achieve the final cluster assignments of samples.

Besides the above clustering metrics, we randomly selected eight data sets and show the learned feature weights in Fig. 6 when JGSED achieved the best clustering performance. It is obvious that the learned feature weights are clearly distinguishable on data sets including the glass, vehicle, AT&T, jaffe, umist, and YaleB, indicating that these data sets may contain both discriminative and redundant features. On the Binalpha and

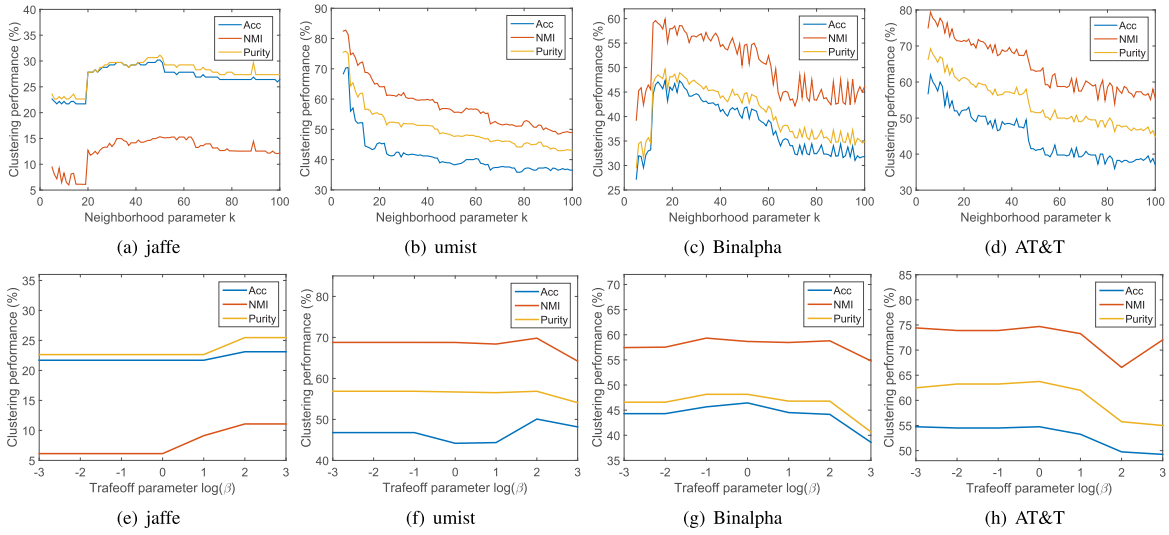


Fig. 7. Clustering performance of JGSED varies with neighborhood size k (the first row) and regularization parameter β (the second row) on the four data sets.

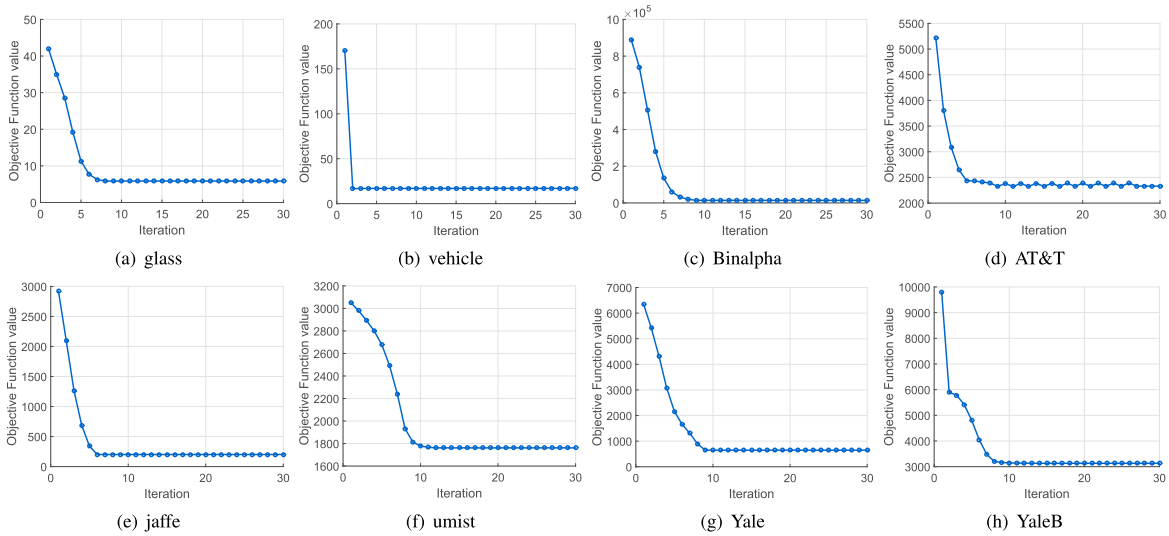


Fig. 8. Convergence curves of JGSED on the eight data sets.

Yale data sets, the learned feature weights are almost uniformly distributed, meaning that features in both data sets share similar contributions in determining the cluster assignment of data points.

We also studied the influence of different parameter settings on the clustering performance of JGSED. Taking the data sets of jaffe, umist, Binalpha, and AT&T as examples, we performed experiments to evaluate the clustering performance of JGSED in terms of different neighborhood sizes (*i.e.*, different ks) and regularization parameters (*i.e.*, different βs). In the first row of Fig. 7, we show the clustering performance of JGSED varies with different neighborhood sizes by fixing parameter β as one. Similarly, the sensitivity of JGSED on β by fixing k as 15 is provided in the second row of Fig. 7. Based on these results, we conclude that JGSED relies more on the neighborhood size parameter k but insensitive to the regularization parameter β.

In addition to the theoretical convergence analysis of JGSED in Section III-C, in Fig. 8 we experimentally show its convergence curves on some data sets. We observe that the objective function values of JGSED monotonically decrease as the number of iterations increases. Moreover, JGSED often converges within a few iterations, indicating its fast convergence.

VI. Conclusion

The typical three steps in spectral clustering are graph construction, spectral embedding, and postprocessing (usually performed by $k$means or spectral rotation to discretize the continuous cluster indicator). Existing studies performed them in a sequential manner or just unified the former or latter two steps, which inevitably cause the sub-optimality problem. In this paper, we proposed a complete spectral clustering model

termed JGSED to unify the three stages together to form a single objective function. That is, JGSED is an end-to-end spectral clustering model by directly taking data as input and outputting the clustering result. In JGSED, the sub-objectives respectively corresponding to the three operations can co-evolve to the optimum and the sub-optimality limitation in existing spectral clustering models can be effectively avoid. Extensive experiments on both synthetic and benchmark data sets demonstrate the validity of the proposed joint optimization mode, and JGSED outperforms some state-of-the-art models in data clustering.

REFERENCES

[1] T. Lei, X. Jia, Y. Zhang, S. Liu, H. Meng, and A. K. Nandi, "Superpixel-based fast fuzzy C-means clustering for color image segmentation," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 9, pp. 1753–1766, Sep. 2019.

[2] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, "Sequence clustering in bioinformatics: An empirical study," *Brief. Bioinf.*, vol. 21, no. 1, pp. 1–10, 2020.

[3] H. Xie et al., "Unsupervised hyperspectral remote sensing image clustering based on adaptive density," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 632–636, Apr. 2018.

[4] N. Gupta, S. Ari, and N. Panigrahi, "Change detection in landsat images using unsupervised learning and RBF-based clustering," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 2, pp. 284–297, Apr. 2021.

[5] A. Saxena et al., "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[6] R. Xu and D. Wunsch, "Computational intelligence in clustering algorithms, with applications," in *Algorithms for Approximation*. Berlin, Germany: Springer, 2007, pp. 31–50.

[7] P. H. Thong and L. H. Son, "Picture fuzzy clustering: A new computational intelligence method," *Soft Comput.*, vol. 20, no. 9, pp. 3549–3562, 2016.

[8] W. Zhu, F. Nie, and X. Li, "Fast spectral clustering with efficient large graph construction," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 2492–2496.

[9] W. Xie, Y. Lee, C. Wang, D. Chen, and T. Zhou, "Hierarchical clustering supported by reciprocal nearest neighbors," *Inf. Sci.*, vol. 527, pp. 279–292, 2020.

[10] Y. Chen et al., "KNN-BLOCK DBSCAN: Fast clustering for large-scale data," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 51, no. 6, pp. 3939–3953, Jun. 2021.

[11] G. Liu, Z. Zhang, Q. Liu, and H. Xiong, "Robust subspace clustering with compressed data," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 5161–5170, Oct. 2019.

[12] K. Chen, Y. Tang, L. Wei, P. Wang, Y. Liu, and Z. Jin, "Sparse subspace clustering for stream data," *IEEE Access*, vol. 9, pp. 57271–57279, 2021.

[13] X. Chen, W. Hong, F. Nie, D. He, M. Yang, and J. Z. Huang, "Spectral clustering of large-scale data by directly solving normalized cut," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2018, pp. 1206–1215.

[14] Y. Peng, X. Zhu, F. Nie, W. Kong, and Y. Ge, "Fuzzy graph clustering," *Inf. Sci.*, vol. 571, pp. 38–49, 2021.

[15] F. Nie, J. Li, and X. Li, "Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification," in *Proc. Int. J. Conf. Artif. Intell.*, 2016, pp. 1881–1887.

[16] J. Wen, Z. Zhang, Z. Zhang, L. Fei, and M. Wang, "Generalized incomplete multiview clustering with flexible locality structure diffusion," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 101–114, Jan. 2021.

[17] R. Liu, M. Chen, Q. Wang, and X. Li, "Robust rank constrained sparse learning: A graph-based method for clustering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 4217–4221.

[18] Z. Yang, Y. Zhang, Y. Xiang, W. Yan, and S. Xie, "Non-negative matrix factorization with dual constraints for image clustering," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 50, no. 7, pp. 2524–2533, Jul. 2020.

[19] S. Sun, S. Wang, Y. Wei, and G. Zhang, "A clustering-based nonlinear ensemble approach for exchange rates forecastinn," *IEEE Trans. Syst. Man Cybern. Systs*, vol. 50, no. 6, pp. 2284–2292, Jun. 2020.

[20] J. Huang, F. Nie, and H. Huang, "Spectral rotation versus k-means in spectral clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2013, pp. 431–437.

[21] Z. Li, F. Nie, X. Chang, L. Nie, H. Zhang, and Y. Yang, "Rank-constrained spectral clustering with flexible embedding," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6073–6082, Dec. 2018.

[22] Z. Li, F. Nie, X. Chang, Y. Yang, C. Zhang, and N. Sebe, "Dynamic affinity graph construction for spectral clustering using multiple features," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6323–6332, Dec. 2018.

[23] Y. Yang, F. Shen, Z. Huang, and H. Shen, "A unified framework for discrete spectral clustering," in *Proc. Int. J. Conf. Artif. Intell.*, 2016, pp. 2273–2279.

[24] Y. Pang, J. Xie, F. Nie, and X. Li, "Spectral clustering by joint spectral embedding and spectral rotation," *IEEE Trans. Cybern.*, vol. 50, no. 1, pp. 247–258, Jan. 2020.

[25] W. Huang, Y. Peng, Y. Ge, and W. Kong, "A new Kmeans clustering model and its generalization achieved by joint spectral embedding and rotation," *PeerJ Comput. Sci.*, vol. 7, 2021, Art. no. e450.

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

[27] M. Wu, S. Pan, and X. Zhu, "Attraction and repulsion: Unsupervised domain adaptive graph contrastive learning network," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 5, pp. 1079–1091, Oct. 2022.

[28] E. Yu, J. Sun, J. Li, X. Chang, X.-H. Han, and A. G. Hauptmann, "Adaptive semi-supervised feature selection for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1276–1288, May 2019.

[29] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

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

[31] Y. Peng, B.-L. Lu, and S. Wang, "Enhanced low-rank representation via sparse manifold adaption for semi-supervised learning," *Neural Netw.*, vol. 65, pp. 1–17, 2015.

[32] X. Li, H. Zhang, R. Zhang, Y. Liu, and F. Nie, "Generalized uncorrelated regression with adaptive graph for unsupervised feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1587–1595, May 2019.

[33] Z. Wang, F. Nie, R. Wang, H. Yang, and X. Li, "Local structured feature learning with dynamic maximum entropy graph," *Pattern Recognit.*, vol. 111, 2021, Art. no. 107673.

[34] D. Cheng, F. Nie, J. Sun, and Y. Gong, "A weight-adaptive Laplacian embedding for graph-based clustering," *Neural Comput.*, vol. 29, no. 7, pp. 1902–1918, 2017.

[35] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2014, pp. 977–986.

[36] X. Yang, S. Li, K. Liang, F. Nie, and L. Lin, "Structured graph optimization for joint spectral embedding and clustering," *Neurocomputing*, vol. 503, pp. 62–72, 2022.

[37] F. Nie, W. Chang, R. Wang, and X. Li, "Learning an optimal bipartite graph for subspace clustering via constrained Laplacian rank," *IEEE Trans. Cybern.*, vol. 53, no. 2, pp. 1235–1247, Feb. 2023.

[38] Y. Han, L. Zhu, Z. Cheng, J. Li, and X. Liu, "Discrete optimal graph clustering," *IEEE Trans. Cybern.*, vol. 50, no. 4, pp. 1697–1710, Apr. 2020.

[39] F. Nie, D. Wu, R. Wang, and X. Li, "Self-weighted clustering with adaptive neighbors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3428–3441, Sep. 2020.

[40] X. Chen, F. Nie, J. Z. Huang, and M. Yang, "Scalable normalized cut with improved spectral rotation," in *Proc. Int. J. Conf. Artif. Intell.*, 2017, pp. 1518–1524.

[41] K. Fan, "On a theorem of Weyl concerning eigenvalues of linear transformations I," *Proc. Nat. Acad. Sci.*, vol. 35, no. 11, pp. 652–655, 1949.

[42] X. Chen, G. Yuan, F. Nie, and J. Z. Huang, "Semi-supervised Feature Selection via Rescaled Linear Regression," in *Proc. Int. J. Conf. Artif. Intell.*, 2017, pp. 1525–1531.

[43] F. Nie, S. Shi, and X. Li, "Semi-supervised learning with auto-weighting feature and adaptive graph," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 6, pp. 1167–1178, Jun. 2020.

[44] X. Chen, G. Yuan, F. Nie, and Z. Ming, "Semi-supervised feature selection via sparse rescaled linear square regression," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 1, pp. 165–176, Jan. 2020.

[45] F. Nie, X. Wang, M. I. Jordan, and H. Huang, "The constrained Laplacian rank algorithm for graph-based clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 1969–1976.
[46] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu, "Non-negative low rank and sparse graph for semi-supervised learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2328–2335.
[47] J. Huang, F. Nie, and H. Huang, "A new simplex sparse learning model to measure data similarity for clustering," in *Proc. Int. J. Conf. Artif. Intell.*, 2015, pp. 3569–3575.
[48] F. Nie, R. Zhang, and X. Li, "A generalized power iteration method for solving quadratic problem on the Stiefel manifold," *Sci. China–Inf. Sci.*, vol. 60, no. 11, pp. 1–10, 2017.
[49] Y. Zhang, Y. Peng, H. Bian, Y. Ge, F. Qin, and W. Kong, "Auto-weighted concept factorization for joint feature map and data representation learning," *J. Intell. Fuzzy Syst.*, vol. 41, no. 1, pp. 69–81, 2021.

Yong Peng (Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2015. He is currently a Full Professor with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. His main research interests include machine learning, pattern recognition, and EEG-based brain-computer interfaces. Dr. Peng was the recipient of the President Prize from the Chinese Academy of Sciences in 2009 and Third Prize from the Chinese Institute of Electronics in 2018.

Wenna Huang received the B.S. degree from the Ningbo University of Technology, Ningbo, China, in 2018, and the M.S. degree from the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China, in 2022. Her research interests include machine leaning and data clustering.

Wanzeng Kong (Member, IEEE) received the B.S. degree in automation and the Ph.D. degree in control theory and control engineering from Zhejiang University, Hangzhou, China, in 2003 and 2008, respectively. From 2012 to 2013, he was a Visiting Research Associate with the Department of Biomedical Engineering, University of Minnesota Twin Cities, Minneapolis, MN, USA. He is currently a Full Professor with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, and the Dean with the Zhejiang Key Laboratory of Brain-Machine Collaborative Intelligence. His research interests include biomedical signal processing, brain-computer interface, cognitive computing, and pattern recognition.

Feiping Nie (Senior Member, IEEE) received the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2009. He is currently a Full Professor with Northwestern Polytechnical University, Xi'an, China. He has authored more than 100 articles in top journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *International Journal of Computer Vision*, *Bioinformatics*, International Conference on Machine Learning, Conference on Neural Information Processing Systems, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, International Joint Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, International Conference on Computer Vision, Conference on Computer Vision and Pattern Recognition, and ACM Multimedia. His articles have been cited more than 10,000 times. His research interests include machine learning and its applications, such as pattern recognition, data mining, computer vision, image processing, and information retrieval. Dr. Nie is also an Associate Editor or a PC Member for several prestigious journals and conferences in the related fields.

Bao-Liang Lu (Fellow, IEEE) received the B.S. degree in instrument and control engineering from Qingdao University of Science and Technology, Qingdao, China, in 1982, the M.S. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 1989, and the Dr.Eng. degree in electrical engineering from Kyoto University, Kyoto, Japan, in 1994. From 1982 to 1986, he was with the Qingdao University of Science and Technology. From 1994 to 1999, he was a Frontier Researcher with the Bio-Mimetic Control Research Center, Institute of Physical and Chemical Research (RIKEN), Nagoya, Japan, and from 1999 to 2002, a Research Scientist with RIKEN Brain Science Institute, Wako, Japan. Since 2002, he has been a Full Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include brain-like computing, neural networks, machine learning, brain-computer interaction, and affective computing. Prof. Lu was the recipient of the IEEE Transactions on Autonomous Mental Development Outstanding Paper Award in 2018. He is an Associate Editor for IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS, and *Journal of Neural Engineering*.