Category-Oriented Self-Learning Graph Embedding for Efficient Image Compact Representation

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

16-08-2021 / 26-08-2021

CITATION

Hu, Liangchen (2021): Category-Oriented Self-Learning Graph Embedding for Efficient Image Compact Representation. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.15173058.v1>

DOI

[10.36227/techrxiv.15173058.v1](https://doi.org/10.36227/techrxiv.15173058.v1)

Category-Oriented Self-Learning Graph Embedding for Efficient Image Compact Representation

Liangchen Hu, Zhenlei Dai, Lei Tian, and Wensheng Zhang

Abstract—As one of the ways to acquire efficient image compact representation, graph embedding (GE) based manifold learning has been widely developed over the last two decades. Good graph embedding depends on the construction of graphs concerning intra-class compactness and inter-class separability, which are crucial indicators of the effectiveness of a model in generating discriminative features. Unsupervised approaches are intended to reveal the data structure information from a local or global perspective, but the resulting compact representation often has poorly inter-class margins due to the lack of label information. Moreover, supervised techniques only consider enhancing the adjacency affinity within the class but excluding the affinity of different classes, which results in the inability to fully capture the marginal structure between distributions of different classes. To overcome these issues, we propose a learning framework that implements Category-Oriented Self-Learning Graph Embedding (COSLGE), in which we achieve a flexible low-dimensional compact representation by imposing an adaptive graph learning process across the entire data while examining the inter-class separability of low-dimensional embedding by jointly learning a linear classifier. Besides, our framework can easily be extended to the semi-supervised situation. Extensive experiments on several widely-used benchmark databases demonstrate the effectiveness of the proposed method comparing with some state-of-the-art approaches.

Index Terms—Adaptive graph learning, separability examination, $L_{2,p}$ -norm sparsity, compact representation.

I. INTRODUCTION

CAPTURING an efficient image compact representation by taking full advantage of valid information about the given data is essential for the applications of machine learning as well as data mining. Here, the compact representation of data we refer to contains two levels: one is to embed samples from the high-dimensional observation space into the low-dimensional compact space, namely, by means of dimensionality reduction for feature compression efficiently; the other is to aggregate samples with high affinity or similarity to each other into a compact domain. Both help to reveal

the latent data distribution structure and facilitate excellent classification and generalization performance.

Serving the above two purposes, we consider a data-driven graph $G(\mathbf{Y}, \mathbf{S})$ on a r -dimensional manifold embedded in a d -dimensional space ($r < d$) [1], where $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n] \in \mathbb{R}^{r \times n}$ refers to the data matrix composed of n embedded data samples and $\mathbf{S} \in \mathbb{R}^{n \times n}$ represents the affinity matrix of interconnected sample pairs. Now, by embedding the graph into a r -dimensional Euclidean space for locality preserving, we arrive at the following problem

$$\min_{\mathbf{Y}_i^T \mathbf{H} \mathbf{Y}_i = m} \|\mathbf{Y}_i - \mathbf{Y}_j\|_2^2 S_{ij} \quad (1)$$

which is known as the Laplacian Eigenmaps (LE) [2] for providing an efficient approach to nonlinear dimensionality reduction, where $\mathbf{H}$ is a constraint matrix and m is a constant. Nevertheless, LE faces a major problem of out-of-sample extension. Behind many out-of-sample attempts [3], a series of improved algorithms based on linear assumption $\mathbf{Y} = \mathbf{P}^T \mathbf{X}$ is the focus of research, among which Locality Preserving Projections (LPP) [4], [5] are the pioneering work, as follows

$$\min_{\mathbf{P} \mathbf{X}_i^T \mathbf{H} \mathbf{X}_i \mathbf{P}^T = m} \|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2^2 S_{ij}. \quad (2)$$

The developed LPP-based algorithms fall into three categories of unsupervised, supervised, and semi-supervised. In unsupervised manners [6], [24], [18], [30], [21], the lack of label information leads to the insignificant inter-class margins in the low-dimensional space, thus affecting subsequent clustering and classification tasks. By contrast, the ways of supervising or semi-supervising [7], [8], [9], [10], [11] can strikingly yield the sharper margins between different classes. Unfortunately, as indicated in Fig. 1 (top), these methods only preserve the adjacency information among samples with high similarity by constructing the intra-class graphs but ignore the adjacency relationships between inter-class boundaries, which is not conducive to the generalization of the model. Thus, it is critical to induce large inter-class margins by guiding label information while constructing the graph Laplacians over the entire data distribution. A feasible framework is illustrated in Fig. 1 (bottom).

In these LPP-based approaches, the success of achieving a superior compact representation depends on the construction of affinity graphs, including partial joins and full joins, where partial joins are skilled in mining the structural information of data distribution. However, the thorny issue is that the conventional ways of constructing the affinity graph rely on pre-definition in the original ambient space, including label based 0-1 affinity [8], local geodesic distance [12], sparse

Manuscript received April 19, 2005; revised August 26, 2015. This work was supported in part by the National Key R&D Program of China under Grant 2017YFC0806500, and in part by the National Natural Science Foundation of China under Grant U1636220, Grant 61876183 and Grant 61772525. (Corresponding author: Wensheng Zhang.)

Liangchen Hu is with the School of Computer and Information, Anhui Normal University, Wuhu 241003, China (e-mail: hlc_clear@foxmail.com).

L. Dai is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: zldai720@163.com).

L. Tian, and W. Zhang are with the Research Center of Precision Sensing and Control, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: tianlei2017@ia.ac.cn; zhangwenshengia@hotmail.com).

Fig. 1. The general discriminant criterion (top) vs. the proposed learning framework (bottom).

representation [14], [15], and a few other attempts [16], [17], [10]. Since the affinity of samples in the ambient space cannot be strictly inherited by means of linear dimensionality reduction [19], learning the affinity of samples jointly in the process of projection learning is beneficial to reveal the latent structure of the data manifold. To address this problem, a variety of more or less successful methods have been devised over the past few years, which can be roughly classified into four categories, namely regularization penalty [18], [19], [20], [21], power weighting [25], [26], [27], norm-induced self-weighting [28], and non-negative self-representation [30], [32], [31]. These graph learning strategies are revisited in detail later in the related work.

As mentioned earlier, all of the supervised algorithms discussed above can be based on the general discriminant criterion of imposing intra-class compactness and inter-class separability, as illustrated in Fig. 1 (top). Moreover, not considering the marginal adjacency structure is not conducive to the generalization of learning subspace. As illustrated in Fig. 1 (bottom), we propose a framework that covers global graph learning with marginal information preserving and category-oriented separability examination, thus solving the above issue. In what follows, we summarize several favorable and attractive contributions of our framework.

1) We propose a novel supervised graph embedding framework which implements the joint learning of global adaptive graph learning and inter-class margin examination. Instead of focusing on intra-class compactness as the previous supervised strategies, we are more inclined to consider local compactness over the entire data domain, thus boosting the generalization performance of the learned subspace along with classifying by a joint classifier.

2) Employing relative entropy as regularization, we develop a graph learning approach that encourages each other with subspace learning, which has several excellent properties, such as affinity retentivity, flexible nonlinearity, numerical stability, and approximate sparsity, compared with existing graph learning techniques.

3) We provide an effective iterative optimization algorithm in which we optimize all available variables except the data matrix, which ensures the flexibility of learning low-dimensional compact representation. Besides, the algorithm can be easily extended to semi-supervised scenarios.

The rest of the paper is organized as follows. Section II reviews the related work of dynamic graph structure learning. In Section III, details of the framework we presented are provided and an effective iterative algorithm is then derived for this purpose. Meanwhile, computational complexity and some necessary discussions are given in this section. The results of our experiments are described in Section V. Section VI concludes the paper with some additional summary and future work.

II. RELATED WORK

A. Notations

In our writing, all vectors and matrices are in bold italics, and all scalars are in simple italics. Given the data matrix $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n] \in \mathbb{R}^{d \times n}$ composed of n samples belonging to c classes, where n_i is the number of samples in class i , X_{ij} is the i -th row and j -th column entry of $\mathbf{X}$, and $\mathbf{X}_i \in \mathbb{R}^{d \times 1}$ refer to a d -dimensional sample. And $l_i \in \{1, 2, \dots, c\}$ is the class label of sample $\mathbf{X}_i$. Moreover, $\mathbf{X}^T$, $\mathbf{X}^{-1}$ and $Tr(\mathbf{X})$ represent $\mathbf{X}$'s transpose, inverse and trace, respectively. The Frobenius norm, L_2 norm, and $L_{2,p}$ norm are defined as $\|\mathbf{X}\|_F = \sqrt{\sum_{i,j} X_{ij}^2}$, $\|\mathbf{X}_{i,:}\|_2 = \sqrt{\sum_j X_{ij}^2}$, and $\|\mathbf{X}\|_{2,p} = (\sum_i \|\mathbf{X}_{i,:}\|_2^p)^{\frac{1}{p}}$ respectively, where $\mathbf{X}_{i,:}$ means the i -th row vector of the matrix $\mathbf{X}$.

B. Dynamic Graph Structure Learning

As mentioned earlier, the key of achieving a compact low-dimensional representation via solving problem (2) lies in the construction of a suitable affinity graph, including fully connected and locally connected. Unfortunately, among the numerous spectral graph based approaches [9], [8], [7], [12],

[10], [11], affinity graphs are often constructed independently of the learning process. Instead of preconstructing the affinity graph, some adaptive weighted approaches have emerged in recent years for graph learning. Here, we present a brief review of four representative strategies.

1) *Regularization Penalty*: Regulationization is a commonly-used technique for machine learning, which can effectively avoid trivial solutions. The graph learning with regularization penalty can be written as follows

$$\min_{\mathbf{S}, \mathbf{1}=1, \mathbf{S} \geq 0} \sum_{i,j} \|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2^2 S_{ij} + \eta \sum_{i,j} \Psi(S_{ij}) \quad (3)$$

where $\mathbf{S}_i$ is the row vector whose the j -th entry is S_{ij} , $\mathbf{1}$ is the column vector whose entries are all 1, and η controls the evenness of the entries in $\mathbf{S}_i$ and the smaller η causes $\mathbf{S}_i$ to be more uneven. Taking the case of $\sum_j \Psi(S_{ij}) = \|\mathbf{S}_i\|_2^2$ [18], [19] or $\sum_j \Psi(S_{ij}) = \|\mathbf{S}_i - \mathbf{A}_i\|_2^2$ where $\mathbf{A}_i$ is the i -th row of the predefined affinity matrix $\mathbf{A}$ [20], [21], the equivalent subproblem can be extracted from problem (3) as

$$\min_{\mathbf{S}_i, \mathbf{1}=1, \mathbf{S}_i \geq 0} \|\mathbf{S}_i - \mathbf{R}_i\|_2^2 \quad (4)$$

where $R_{ij} = \begin{cases} -\frac{\|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2^2}{2\eta}, & \|\mathbf{S}_i\|_2^2 \\ A_{ij} - \frac{\|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2^2}{2\eta}, & \|\mathbf{S}_i - \mathbf{A}_i\|_2^2 \end{cases}$ is the j -th entry in row vector $\mathbf{R}_i$. And then, problem (4) can be effectively solved by referring to [22], [18], [23]. Generally speaking, the results obtained by solving (4) are not sparse, which is not conducive to revealing the local manifold structure of data.

2) *Power Weighting*: An alternative approach is to transform the linear problem concerning affinity S_{ij} into the following power nonlinear problem

$$\min_{\mathbf{S}_i, \mathbf{1}=1, \mathbf{S}_i \geq 0} \sum_{i,j} \|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2^2 S_{ij}^v \quad (5)$$

where $v = 2$ taken in [25], [26] and it is generalized in [27] as any case where $v > 1$. When $\mathbf{P}$ is fixed, the general solution of (5) can be derived as follows

$$S_{ij} = \frac{\|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2^{\frac{2}{1-v}}}{\sum_j \|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2^{\frac{2}{1-v}}}. \quad (6)$$

The power weighting strategy requires no additional parameters and has a closed-form solution. But when $v > 1$, the $\|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2^2$ that approximates to 0 falls in the denominator, thus resulting in the absolute dominance of S_{ij} in $\mathbf{S}_i$, which may lead to the construction of morbidity graph.

3) *Norm-Induced Self-Weighting*: Besides, we can directly employ the non-squared L_2 norm to measure the distance between two projected samples [28], as follows

$$\min_{\|\mathbf{S}_i\|_0=K} \sum_{i,j} \|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2^2 S_{ij} \quad (7)$$

where $\|\mathbf{S}_i\|_0 = K$ [29] means that each sample has only K nearest neighbors and $S_{ij} = \frac{1}{\|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2}$ [28] means that larger weights can be assigned to pairs of samples with smaller distances. Solving problem (7) is equivalent to calculating the

optimal KNN graph under the non-squared L_2 norm, which is constructed as follows

$$\min_{\|\mathbf{V}_i\|_0=K, \mathbf{V}_{ij} \in \{0,1\}} \sum_{i,j} \|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2 V_{ij}. \quad (8)$$

However, the same hidden danger of S_{ij} that occurs in (6) may affect the performance of the model.

4) *Non-Negative Self-Representation*: Besides the above methods of relying solely on distance, the affinity on the graph can also be derived from the non-negative contribution of other samples to the reconstruction of a sample, also known as non-negative self-representation of data. Here we mention two more representative approaches [30], [31]. In [30], fixing other unrelated variables, we arrive at

$$\min_{\mathbf{S}} \sum_{i,j} \|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2^2 \frac{S_{ij} + S_{ji}}{2} + \alpha \|\mathbf{S}\|_* \quad (9)$$

$s.t. \mathbf{X} = \mathbf{B}\mathbf{S} + \mathbf{E}.$

where $\mathbf{E}$ is a sparse additive error matrix. Problem (9) strives to achieve a symmetric graph in which the affinity of a sample pair is $\frac{S_{ij} + S_{ji}}{2}$ derived from these low-rank representation coefficients of the original data under a dictionary $\mathbf{B}$. Another approach proposed in [31] is to express the sample relationships on the graph through the non-negative self-representation on the projected data, namely

$$\min_{\mathbf{S}} \|\mathbf{P}^T (\mathbf{X} - \mathbf{X}\mathbf{S})\|_{2,1} + \alpha \|\mathbf{E} \odot \mathbf{S}\|_1 + \beta \sum_{i,j} \|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2 S_{ij} \quad s.t. \mathbf{S} \geq 0. \quad (10)$$

where $\odot$ is the entry-wise multiplication and $\mathbf{E}$ here refers to the locality adaptor matrix. Similar work can also refer to [32]. The above two methods still construct $\mathbf{S}$ in a linear way and require more parameters to be turned, but the performance of mining latent subspace would be improved by preserving the structural representations between samples.

Remark 1. From the above several ways of constructing graphs, we summarize several properties that a high-quality graph should have.

- 1) **Retentivity**: The structure information in high dimensional space should be preserved to a certain extent in low dimensional embedding;
- 2) **Nonlinearity**: The nonlinear manifold distribution of data forces us to measure the relationships between samples in a nonlinear way;
- 3) **Stability**: The affinity between samples cannot trend to be infinity as the distance between samples varies;
- 4) **Sparsity**: The affinity of sample pairs in the graph should be sparse to better reveal the local manifold structure.

III. METHODOLOGY

In this section, we formulate our motivations step by step and derive an efficient iterative optimization algorithm. Then, the computational complexity and some bright spots of the proposed algorithm are discussed.

A. Framework

Assisted by the separability examination of a classifier, we are eager to reveal the low-dimensional manifold structure of the data, in which the marginal distribution information is preserved simultaneously with the intra-class structure information. Here, we present the overall framework to express our motives

$$\min_{\mathbf{P}, \mathbf{S}, f} \Xi(\mathbf{X}, \mathbf{P}, \mathbf{S}) + \delta \Omega(\mathbf{P}) + \gamma (\Pi(f(\mathbf{Y}), \mathbf{T}) + \alpha \Gamma(f)) \quad (11)$$

where f is a classifier and $\mathbf{T}$ is a target matrix containing label information. $\Xi(\mathbf{X}, \mathbf{P}, \mathbf{S})$ is employed to perform self-learning graph embedding across the entire data domain, and $\Omega(\mathbf{P})$ is a regularization function for $\mathbf{P}$ with δ as a regularization parameter. $\Pi(f(\mathbf{Y}), \mathbf{T})$ checks the separability of projected samples of different classes in the learning process of graph embedding. $\Gamma(f)$ is an item that prevents the classifier f from overfitting. γ is a tradeoff coefficient and α is a regularization parameter.

B. Self-Learning Graph Embedding

As indicated in Section II-B, a variety of dynamic graph learning strategies exist for diverse assumptions. For flexibility, we are more inclined to the graph learning strategy of "Regularization Penalty". Innovatively, we provide a way to regularize graph learning in terms of relative entropy below

$$\begin{aligned} \Xi(\mathbf{X}, \mathbf{P}, \mathbf{S}) &= \sum_{i,j=1}^n \|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2^2 S_{ij} + \sigma \sum_{i=1}^n D_{KL}(\mathbf{S}_i \| \mathbf{A}_i) \quad (12) \\ \text{s.t. } \|\mathbf{S}_i\|_0 &= K, \mathbf{S}\mathbf{1} = \mathbf{1}, \mathbf{S} \geq 0. \end{aligned}$$

where S_{ij} refers to the probability of the j -th sample being the nearest neighbor to the i -th sample, $\|\mathbf{S}_i\|_0 = K$ is the scale employed to constrain local graph construction. The relative entropy $D_{KL}(\mathbf{S}_i \| \mathbf{A}_i)$ in (12), also known as Kullback-Leibler divergence [33], is defined as follows

$$D_{KL}(\mathbf{S}_i \| \mathbf{A}_i) = \sum_{j=1}^n S_{ij} \ln \frac{S_{ij}}{A_{ij}} \quad (13)$$

which is an asymmetric measure of the difference between the two probability distributions $\mathbf{S}_i$ and $\mathbf{A}_i$. Specially, we define $0 \ln 0 = 0$.

Besides, regularization of $\mathbf{P}$ can ensure the mining of resultful latent manifold structure, and there are many strategies available. However, the samples collected in high dimensional space are often doped with noise and redundancy. In view of this, $L_{2,0}$ norm should be the best choice, but the difficulty in solving such problems makes us prefer $L_{2,p}$ norm with $0 < p \leq 1$ [36], namely,

$$\Omega(\mathbf{P}) = \|\mathbf{P}\|_{2,p}^p. \quad (14)$$

C. Inter-class Separability Examination

Although solving problem (12) with the regularization term (14) can capture the overall structural information, the resulting $\mathbf{P}$ lacks discriminant power due to the unutilization of

label information. As a result, we import the obtained low-dimensional representation $\mathbf{Y}$ into a classifier f to examine its classification performance. Then, we arrive at

$$\Pi(f(\mathbf{Y}), \mathbf{T}) = \sum_{i=1}^n \psi_i(f(\mathbf{Y}_i), \mathbf{T}_i) \quad (15)$$

where Π is the total classification loss and ψ_i is the i -th loss measure for sample $\mathbf{Y}_i$.

Since it is of no practical significance to generate soft labels for training samples, the classifier allows continuous variable prediction. Here, for simplicity we take the linear predictor as $f(\mathbf{Y}_i) = \mathbf{W}^T \mathbf{Y}_i$ where $\mathbf{W}$ is the learned combination coefficient matrix. And, we expect the predicted results to correspond to the targets of those samples, namely

$$\begin{bmatrix} T_{i1} \\ T_{i2} \\ \vdots \\ T_{ic} \end{bmatrix} \approx \begin{bmatrix} W_{11}Y_{1i} + W_{21}Y_{2i} + \cdots + W_{r1}Y_{ri} \\ W_{12}Y_{1i} + W_{22}Y_{2i} + \cdots + W_{r2}Y_{ri} \\ \vdots \\ W_{1c}Y_{1i} + W_{2c}Y_{2i} + \cdots + W_{rc}Y_{ri} \end{bmatrix}. \quad (16)$$

where $\mathbf{T}$ is flexible and satisfies the marginal constraint $T_{il_i} - \max_{j \neq l_i} T_{ij} \geq 1$ [34].

Taking the squared L_2 norm for measuring the approximation in (16), we arrive at the result below

$$\begin{aligned} \psi_i(f(\mathbf{Y}_i), \mathbf{T}_i) &= \|\mathbf{Y}_i^T \mathbf{W} - \mathbf{T}_i\|_2^2 = \|\mathbf{X}_i^T \mathbf{P} \mathbf{W} - \mathbf{T}_i\|_2^2 \\ \text{s.t. } T_{il_i} - \max_{j \neq l_i} T_{ij} &\geq 1. \end{aligned} \quad (17)$$

Simply taking $\Gamma(\mathbf{W}) = \|\mathbf{W}\|_F^2$ and integrating it with (12), (14), (15) and (17) as a joint learning framework, our model arrives at

$$\begin{aligned} \min_{\mathbf{P}, \mathbf{S}, \mathbf{W}, \mathbf{T}} \sum_{i,j=1}^n \|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2^2 S_{ij} + \sigma \sum_{i=1}^n D_{KL}(\mathbf{S}_i \| \mathbf{A}_i) \\ + \delta \|\mathbf{P}\|_{2,p}^p + \gamma \left(\sum_{i=1}^n \|\mathbf{X}_i^T \mathbf{P} \mathbf{W} - \mathbf{T}_i\|_2^2 + \alpha \|\mathbf{W}\|_F^2 \right) \quad (18) \\ \text{s.t. } \|\mathbf{S}_i\|_0 = K, \mathbf{S}\mathbf{1} = \mathbf{1}, \mathbf{S} \geq 0, T_{il_i} - \max_{j \neq l_i} T_{ij} \geq 1. \end{aligned}$$

With some mathematical tricks, problem (18) can be further written in matrix form

$$\begin{aligned} \min_{\mathbf{P}, \mathbf{S}, \mathbf{W}, \mathbf{T}} \text{Tr}(\mathbf{P}^T \mathbf{X} \mathbf{L}_s \mathbf{X}^T \mathbf{P}) + \sigma D_{KL}(\mathbf{S} \| \mathbf{A}) \\ + \delta \|\mathbf{P}\|_{2,p}^p + \gamma (\|\mathbf{X}^T \mathbf{P} \mathbf{W} - \mathbf{T}\|_F^2 + \alpha \|\mathbf{W}\|_F^2) \quad (19) \\ \text{s.t. } \|\mathbf{S}_i\|_0 = K, \mathbf{S}\mathbf{1} = \mathbf{1}, \mathbf{S} \geq 0, T_{il_i} - \max_{j \neq l_i} T_{ij} \geq 1 \end{aligned}$$

in which the Laplacian matrix $\mathbf{L}_s$ is defined as $\mathbf{L}_s = \mathbf{D} - \frac{\mathbf{S}^T + \mathbf{S}}{2}$ where $\mathbf{D} \in \mathbb{R}^{n \times n}$ is the diagonal matrix whose ii -th element is $\sum_j (S_{ij} + S_{ji})/2$, and $D_{KL}(\mathbf{S} \| \mathbf{A}) = \sum_{i=1}^n D_{KL}(\mathbf{S}_i \| \mathbf{A}_i)$.

D. Optimization

The optimization problem established in (19) involves several variables, which interact with each other, making it difficult to directly derive analytic results. As an executable strategy, we utilize the alternating direction multipliers method (ADMM) [41] for updating one variable alternately while fixing the other variables each time.

To make (19) separable, we introduce the auxiliary variables $\mathbf{Z}$ and $\mathbf{U}$ with two additional constraints $\mathbf{Z} = \mathbf{U}\mathbf{W}$, $\mathbf{U} = \mathbf{P}$. And then problem (19) can be expressed loosely as follows

$$\begin{aligned} \min_{\mathbf{Z}, \mathbf{T}, \mathbf{P}, \mathbf{S}} & Tr(\mathbf{P}^T \mathbf{X} \mathbf{L}_s \mathbf{X}^T \mathbf{P}) + \sigma D_{KL}(\mathbf{S} \| \mathbf{A}) \\ & + \delta \|\mathbf{P}\|_{2,p}^p + \gamma (\|\mathbf{X}^T \mathbf{Z} - \mathbf{T}\|_F^2 + \alpha \|\mathbf{W}\|_F^2) \\ \text{s.t. } & \|\mathbf{S}_i\|_0 = K, \mathbf{S} \mathbf{1} = \mathbf{1}, \mathbf{S} \geq 0, T_{il_i} - \max_{j \neq l_i} T_{ij} \geq 1, \\ & \mathbf{Z} = \mathbf{U}\mathbf{W}, \mathbf{U} = \mathbf{P}. \end{aligned} \quad (20)$$

Consequently, we derive the augmented Lagrangian function of problem (20) below

$$\begin{aligned} \mathcal{L}(\mathbf{Z}, \mathbf{T}, \mathbf{P}, \mathbf{S}, \mathbf{W}, \mathbf{U}) &= \frac{1}{2} Tr(\mathbf{P}^T \mathbf{X} \mathbf{L}_s \mathbf{X}^T \mathbf{P}) + \frac{\sigma}{2} D_{KL}(\mathbf{S} \| \mathbf{A}) \\ &+ \frac{\delta}{2} \|\mathbf{P}\|_{2,p}^p + \frac{\gamma}{2} (\|\mathbf{X}^T \mathbf{Z} - \mathbf{T}\|_F^2 + \alpha \|\mathbf{W}\|_F^2) \\ &+ \langle \mathbf{C}_1, \mathbf{Z} - \mathbf{U}\mathbf{W} \rangle + \langle \mathbf{C}_2, \mathbf{U} - \mathbf{P} \rangle \\ &+ \frac{\mu}{2} (\|\mathbf{Z} - \mathbf{U}\mathbf{W}\|_F^2 + \|\mathbf{U} - \mathbf{P}\|_F^2) \end{aligned} \quad (21)$$

where $\mathbf{C}_1$ and $\mathbf{C}_2$ are Lagrange multipliers, μ is a positive penalty parameter. With some simple mathematical reasoning, Eq. (21) can be reformulated as

$$\begin{aligned} \mathcal{L}(\mathbf{Z}, \mathbf{T}, \mathbf{P}, \mathbf{S}, \mathbf{W}, \mathbf{U}) &= \frac{1}{2} Tr(\mathbf{P}^T \mathbf{X} \mathbf{L}_s \mathbf{X}^T \mathbf{P}) + \frac{\sigma}{2} D_{KL}(\mathbf{S} \| \mathbf{A}) \\ &+ \frac{\delta}{2} \|\mathbf{P}\|_{2,p}^p + \frac{\gamma}{2} (\|\mathbf{X}^T \mathbf{Z} - \mathbf{T}\|_F^2 + \alpha \|\mathbf{W}\|_F^2) \\ &- \frac{1}{2\mu} \|\mathbf{C}_1\|_F^2 - \frac{1}{2\mu} \|\mathbf{C}_2\|_F^2 \\ &+ \frac{\mu}{2} (\|\mathbf{Z} - \mathbf{U}\mathbf{W} + \frac{\mathbf{C}_1}{\mu}\|_F^2 + \|\mathbf{U} - \mathbf{P} + \frac{\mathbf{C}_2}{\mu}\|_F^2). \end{aligned} \quad (22)$$

By alternately optimizing each variable of (22), we ultimately achieve the optimal solution for all variables. The detailed steps are shown below.

Update $\mathbf{Z}$: Since the variables $\mathbf{P}$, $\mathbf{S}$, $\mathbf{W}$, $\mathbf{T}$ and $\mathbf{U}$ are fixed, (22) can be reduced to the following problem without considering other irrelevant terms

$$\mathcal{L}(\mathbf{Z}) = \frac{\gamma}{2} \|\mathbf{X}^T \mathbf{Z} - \mathbf{T}\|_F^2 + \frac{\mu}{2} \|\mathbf{Z} - \mathbf{U}\mathbf{W} + \frac{\mathbf{C}_1}{\mu}\|_F^2. \quad (23)$$

Taking the derivative of Eq. (23) w.r.t. $\mathbf{Z}$ and setting it to zero, we have

$$\mathbf{Z} = (\gamma \mathbf{X} \mathbf{X}^T + \mu \mathbf{I})^{-1} (\gamma \mathbf{X} \mathbf{T} + \mu \mathbf{U}\mathbf{W} - \mathbf{C}_1). \quad (24)$$

Update $\mathbf{T}$: By fixing $\mathbf{P}$ and $\mathbf{W}$, and simply denoting the regression result $\mathbf{X}^T \mathbf{P}\mathbf{W}$ as $\mathbf{V} \in \mathbb{R}^{n \times c}$, Eq. (22) degenerates into a retargeting problem

$$\mathcal{L}(\mathbf{T}) = \|\mathbf{V} - \mathbf{T}\|_F^2 \quad \text{s.t. } T_{il_i} - \max_{j \neq l_i} T_{ij} \geq 1. \quad (25)$$

Then, n subproblems with an independent constraint on each other can be extracted from problem (25), and the i -th subproblem is described below

$$\begin{aligned} \mathcal{L}(T_i) &= \|\mathbf{V}_i - \mathbf{T}_i\|_2^2 = \sum_{j=1}^c (V_{ij} - T_{ij})^2 \\ \text{s.t. } & T_{il_i} - \max_{j \neq l_i} T_{ij} \geq 1. \end{aligned} \quad (26)$$

Obviously, $\mathbf{T}_i$ depends only on $\mathbf{V}_i$. As in [34], we redefine the i -th target vector $\mathbf{T}_i$ below

$$T_{ij} = \begin{cases} V_{ij} + \Delta_i, & j = l_i \\ V_{ij} + \min(\Delta_i - \tau_j, 0), & j \neq l_i \end{cases} \quad (27)$$

where $\Delta_i = T_{il_i} - V_{il_i}$ is a step parameter, $\tau_j = V_{ij} + 1 - V_{il_i}$ is an indicator and $\tau_j \leq 0$ means that class i and class l_i satisfy the margin constraint. Using the new formulation in Eq. (27), we can rewrite optimization problem (26) as

$$f(\Delta_i) = \Delta_i^2 + \sum_{j \neq l_i} (\min(\Delta_i - \tau_j, 0))^2. \quad (28)$$

Taking the derivative of Eq. (28) w.r.t. Δ_i , we have

$$f'(\Delta_i) = 2(\Delta_i + \sum_{j \neq l_i} \min(\Delta_i - \tau_j, 0)). \quad (29)$$

And, when $\Delta_i + \sum_{j \neq l_i} \min(\Delta_i - \tau_j, 0) = 0$, problem (28) minimizes, and the optimal Δ_i is calculated as

$$\Delta_i = \frac{\sum_{j \neq l_i} \tau_j \Theta(\tau_j)}{1 + \sum_{j \neq l_i} \Theta(\tau_j)} \quad (30)$$

where $\Theta(\tau_j) = \begin{cases} 1, & f'(\tau_j) > 0 \\ 0, & \text{other} \end{cases}$. Then, the optimal target matrix $\mathbf{T}$ can be derived from Eq. (27).

Update $\mathbf{W}$: Given $\mathbf{Z}$, $\mathbf{U}$ and other variables, the subproblem for updating $\mathbf{W}$ is

$$\mathcal{L}(\mathbf{W}) = \frac{\mu}{2} \|\mathbf{Z} - \mathbf{U}\mathbf{W} + \frac{\mathbf{C}_1}{\mu}\|_F^2 + \frac{\gamma\alpha}{2} \|\mathbf{W}\|_F^2. \quad (31)$$

Similarly, taking the derivative of Eq. (31) w.r.t. $\mathbf{W}$ and setting it to zero, we have

$$\mathbf{W} = (\mu \mathbf{U}^T \mathbf{U} + \gamma\alpha \mathbf{I})^{-1} (\mathbf{U}^T (\mu \mathbf{Z} + \mathbf{C}_1)). \quad (32)$$

Update $\mathbf{U}$: Analogously, $\mathbf{U}$ can be solved by minimizing

$$\mathcal{L}(\mathbf{U}) = \|\mathbf{Z} - \mathbf{U}\mathbf{W} + \frac{\mathbf{C}_1}{\mu}\|_F^2 + \|\mathbf{U} - \mathbf{P} + \frac{\mathbf{C}_2}{\mu}\|_F^2. \quad (33)$$

Taking $\partial \mathcal{L}(\mathbf{U}) / \partial \mathbf{U} = \mathbf{0}$, we have closed-form solution

$$\mathbf{U} = ((\mathbf{Z} + \frac{\mathbf{C}_1}{\mu}) \mathbf{W}^T + \mathbf{P} - \frac{\mathbf{C}_2}{\mu}) (\mathbf{W} \mathbf{W}^T + \mathbf{I})^{-1}. \quad (34)$$

Update $\mathbf{P}$: When the variables $\mathbf{Z}$, $\mathbf{T}$, $\mathbf{S}$, $\mathbf{W}$ and $\mathbf{U}$ are fixed, the subproblem w.r.t. $\mathbf{P}$ can be rewritten as

$$\begin{aligned} \mathcal{L}(\mathbf{P}) &= \frac{1}{2} Tr(\mathbf{P}^T \mathbf{X} \mathbf{L}_s \mathbf{X}^T \mathbf{P}) + \frac{\delta}{2} \|\mathbf{P}\|_{2,p}^p \\ &+ \frac{\mu}{2} \|\mathbf{U} - \mathbf{P} + \frac{\mathbf{C}_2}{\mu}\|_F^2. \end{aligned} \quad (35)$$

Then, by setting $\partial \mathcal{L}(\mathbf{P}) / \partial \mathbf{P} = \mathbf{0}$, we have

$$\mathbf{X} \mathbf{L}_s \mathbf{X}^T \mathbf{P} + \delta \tilde{\mathbf{D}} \mathbf{P} + \mu \mathbf{P} - \mu \mathbf{U} - \mathbf{C}_2 = \mathbf{0} \quad (36)$$

where $\tilde{\mathbf{D}} \in \mathbb{R}^{d \times d}$ is a diagonal matrix with diagonal elements $\tilde{D}_{ii} = \frac{p}{2} \|\mathbf{P}_i\|_2^{p-2}$ in which $\mathbf{P}_i$ is the i -th row vector of $\mathbf{P}$. And then, we express the optimal $\mathbf{P}$ as

$$\mathbf{P} = (\mathbf{X} \mathbf{L}_s \mathbf{X}^T + \delta \tilde{\mathbf{D}} + \mu \mathbf{I})^{-1} (\mu \mathbf{U} + \mathbf{C}_2). \quad (37)$$

Incontestably, $\tilde{\mathbf{D}}$ depends on $\mathbf{P}$, making the optimal $\mathbf{P}$ hard to achieve, but here a suboptimal choice can be made through the strategy of iterative solution.

Lemma 1. [36], [37] When $0 < p \leq 1$, the following inequality holds for any positive real number a and b ,

$$\sqrt{a^p} - \frac{p}{2}a\sqrt{b^{p-2}} \leq \sqrt{b^p} - \frac{p}{2}b\sqrt{b^{p-2}}. \quad (38)$$

Theorem 1. Assuming that $\mathbf{P}^t$ and $\tilde{\mathbf{D}}^t$ exist in the t -th iteration, and $\mathbf{P}^{t+1}$ is further solved by Eq. (37), then the inequality $\mathcal{L}(\mathbf{P}^t) \geq \mathcal{L}(\mathbf{P}^{t+1})$ holds.

Proof. Equivalently, (35) can be reformulated as

$$\begin{aligned} \mathcal{L}(\mathbf{P}, \tilde{\mathbf{D}}) = & \frac{1}{2} \text{Tr}(\mathbf{P}^T \mathbf{X} \mathbf{L}_s \mathbf{X}^T \mathbf{P}) + \frac{\delta}{2} \text{Tr}(\mathbf{P}^T \tilde{\mathbf{D}} \mathbf{P}) \\ & + \frac{\mu}{2} \|\mathbf{U} - \mathbf{P} + \frac{\mathbf{C}_2}{\mu}\|_F^2. \end{aligned} \quad (39)$$

And obviously, by Eq. (37), the following inequality holds

$$\mathcal{L}(\mathbf{P}^t, \tilde{\mathbf{D}}^t) \geq \mathcal{L}(\mathbf{P}^{t+1}, \tilde{\mathbf{D}}^t). \quad (40)$$

Denoting $\Phi(\mathbf{P}) = \text{Tr}(\mathbf{P}^T \mathbf{X} \mathbf{L}_s \mathbf{X}^T \mathbf{P})$, $\Psi(\mathbf{P}) = \|\mathbf{U} - \mathbf{P} + \frac{\mathbf{C}_2}{\mu}\|_F^2$ and expanding (40), we have

$$\begin{aligned} & \Phi(\mathbf{P}^{t+1}) + \delta \sum_{i=1}^d \frac{p}{2} \|\mathbf{P}_i^{t+1}\|_2^2 \|\mathbf{P}_i^t\|_2^{p-2} + \mu \Psi(\mathbf{P}^{t+1}) \\ & \leq \Phi(\mathbf{P}^t) + \delta \sum_{i=1}^d \frac{p}{2} \|\mathbf{P}_i^t\|_2^2 \|\mathbf{P}_i^t\|_2^{p-2} + \mu \Psi(\mathbf{P}^t). \end{aligned} \quad (41)$$

From Lemma 1, we know

$$\begin{aligned} & \sum_{i=1}^d \|\mathbf{P}_i^{t+1}\|_2^p - \delta \sum_{i=1}^d \frac{p}{2} \|\mathbf{P}_i^{t+1}\|_2^2 \|\mathbf{P}_i^t\|_2^{p-2} \\ & \leq \sum_{i=1}^d \|\mathbf{P}_i^t\|_2^p - \delta \sum_{i=1}^d \frac{p}{2} \|\mathbf{P}_i^t\|_2^2 \|\mathbf{P}_i^t\|_2^{p-2}. \end{aligned} \quad (42)$$

Combining inequalities (41) and (42), we arrive at

$$\begin{aligned} & \Phi(\mathbf{P}^{t+1}) + \delta \|\mathbf{P}^{t+1}\|_{2,p}^p + \mu \Psi(\mathbf{P}^{t+1}) \\ & \leq \Phi(\mathbf{P}^t) + \delta \|\mathbf{P}^t\|_{2,p}^p + \mu \Psi(\mathbf{P}^t) \end{aligned} \quad (43)$$

which completes the proof. $\square$

Thus, after receiving the $\mathbf{P}^t$ generated by the last iteration and calculating $\tilde{\mathbf{D}}^t$, the suboptimal solution of (35) can be $\mathbf{P}^{t+1}$ obtained through (37).

Update S: Similarly, by removing all terms that do not involve the variable $\mathbf{S}$, the objective function is reduced to

$$\begin{aligned} \mathcal{L}(\mathbf{S}) = & \sum_{i,j=1}^n \|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2^2 S_{ij} + \sigma \sum_{i,j=1}^n S_{ij} \ln \frac{S_{ij}}{A_{ij}} \quad (44) \\ \text{s.t. } & \|\mathbf{S}_i\|_0 = K, \mathbf{S} \mathbf{1} = \mathbf{1}, \mathbf{S} \geq 0. \end{aligned}$$

We decompose problem (44) into n independent subproblems, each of which is equivalent to solving the following problem

$$\begin{aligned} \mathcal{L}(\mathbf{S}_i) = & \sum_{j=1}^n Q_{ij} S_{ij} + \sigma \sum_{j=1}^n S_{ij} \ln \frac{S_{ij}}{A_{ij}} \quad (45) \\ \text{s.t. } & \|\mathbf{S}_i\|_0 = K, \mathbf{S}_i \mathbf{1} = 1, \mathbf{S}_i \geq 0. \end{aligned}$$

where Q_i is the i -th row of $\mathbf{Q} \in \mathbb{R}^{n \times n}$ whose ij -th entry is $\|\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j\|_2^2$. Obviously, the constraint $\|\mathbf{S}_i\|_0 = K$

can be implemented by constraining K minimum non-zero elements in $\mathbf{Q}_i$. Denoting the subscripts of these K non-zero elements as $\{\rho_1, \rho_2, \dots, \rho_K\}$, (45) can be reexpressed as

$$\min_{\sum_{j=1}^K S_{i\rho_j} = 1, S_i \geq 0} \sum_{j=1}^K (Q_{i\rho_j} S_{i\rho_j} + \sigma S_{i\rho_j} \ln \frac{S_{i\rho_j}}{A_{i\rho_j}}). \quad (46)$$

And, the Lagrangian function of problem (46) is

$$L(\mathbf{S}_i, \mu) = \sum_{j=1}^K (Q_{i\rho_j} S_{i\rho_j} + \sigma S_{i\rho_j} \ln \frac{S_{i\rho_j}}{A_{i\rho_j}}) - \mu (\sum_{j=1}^K S_{i\rho_j} - 1) \quad (47)$$

where μ is the Lagrangian multiplier. Taking the derivative $\nabla_{S_{i\rho_j}} L = Q_{i\rho_j} + \sigma (\ln \frac{S_{i\rho_j}}{A_{i\rho_j}} + 1) - \mu = 0$, we have

$$S_{i\rho_j} = A_{i\rho_j} \exp(-\frac{Q_{i\rho_j}}{\sigma}) \exp(\frac{\mu}{\sigma} - 1). \quad (48)$$

Substituting (48) into $\sum_{j=1}^K S_{i\rho_j} = 1$, we can easily strip out

$$\exp(\frac{\mu}{\sigma} - 1) = \frac{1}{\sum_{j=1}^K A_{i\rho_j} \exp(-Q_{i\rho_j} / \sigma)}. \quad (49)$$

Then, we substitute (49) into (48) to produce the final result

$$S_{ik} = \begin{cases} \frac{A_{i\rho_j} \exp(-Q_{i\rho_j} / \sigma)}{\sum_{j=1}^K A_{i\rho_j} \exp(-Q_{i\rho_j} / \sigma)}, & k = \rho_j \\ 0, & \text{otherwise} \end{cases} \quad (50)$$

Update $\mathbf{C}_1$, $\mathbf{C}_2$ and μ : Lagrange multipliers $\mathbf{C}_1$ and $\mathbf{C}_2$, and penalty parameter μ are respectively updated as follows

$$\mathbf{C}_1 = \mathbf{C}_1 + \mu(\mathbf{Z} - \mathbf{U}\mathbf{W}), \quad (51)$$

$$\mathbf{C}_2 = \mathbf{C}_2 + \mu(\mathbf{U} - \mathbf{P}), \quad (52)$$

$$\mu = \min(\mu_{max}, \rho\mu) \quad (53)$$

where μ_{max} and ρ are both constants.

The detailed procedure for solving problem (19) is summarized in Algorithm 1.

E. Computational complexity

Here, we analyze each step of Algorithm 1 and generalize the total computational complexity. Initializing the probability matrix $\mathbf{S}^0$ costs $O(n^2d)$. In Step 6, Computing $\mathbf{Z}$ requires $O(nd^2 + ndc + d^3 + d^2c + drc)$ operations. In Step 7, the calculation of $\mathbf{T}$ with less time consuming is approximately $O(nc)$ operations. In Steps 8 and 9, update $\mathbf{W}$ and $\mathbf{U}$ are both of order $O(dr^2 + r^3 + r^2c + drc)$. Computing $\mathbf{P}$ in Step 10 requires $O(n^2d + nd^2 + d^2r)$ operations. In step 11, the probabilistic matrix $\mathbf{S}$ needs a time complexity of order $O(n^2r)$. By contrast, the complexity in Step 12 is negligible. Since $\{r, c\} \leq d$ and the size of n is uncertain, the total computational complexity is at most $O(t(nd^2 + d^3 + n^2d))$, where t is the number of iterations before algorithm convergence.

Algorithm 1 Our COSLGE algorithm of solving problem (19)

Input: Training matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$, label matrix $\mathbf{T} \in \mathbb{R}^{n \times c}$, predefined $\mathbf{A}$, dimension r , parameters α, γ, δ and σ .

- 1: Initialize $\mathbf{P}^0 \in \mathbb{R}^{d \times r}$ and $\mathbf{W}^0 \in \mathbb{R}^{r \times c}$ randomly.
- 2: Initialize $\mathbf{S}^0$ by solving

$$\min_{\|\mathbf{S}_i\|_0=K, \mathbf{S}_i \mathbf{1}=1, \mathbf{S} \geq 0} \sum_{i,j=1}^n \|\mathbf{X}_i - \mathbf{X}_j\|_2^2 S_{ij} + \sigma \sum_{i,j=1}^n S_{ij} \ln \frac{S_{ij}}{A_{ij}}.$$

- 3: Initialize $\mathbf{T}^0$ as a strict zero-one target matrix in which only the l_i -th entry of $\mathbf{T}_i^0$ is 1 while the others are all 0.
- 4: Initialize $\mathbf{U}^0 = \mathbf{P}^0$, $\mathbf{C}_1^0 = \mathbf{0}$, $\mathbf{C}_2^0 = \mathbf{0}$, $\mu^0 = 10^{-1}$, $\rho = 1.05$, $\mu_{max} = 10^5$.
- 5: **repeat**
- 6:   Update $\mathbf{Z}^{t+1}$ with (24).
- 7:   Update $\mathbf{T}^{t+1}$ with (27) and (30).
- 8:   Update $\mathbf{W}^{t+1}$ with (32).
- 9:   Update $\mathbf{U}^{t+1}$ with (34).
- 10:   Update $\mathbf{P}^{t+1}$ with (37).
- 11:   Update $\mathbf{S}^{t+1}$ with (50).
- 12:   Update $\mathbf{C}_1^{t+1}$, $\mathbf{C}_2^{t+1}$ and μ^{t+1} with (51), (52) and (53).
- 13: **until** $(\mathcal{L}^t - \mathcal{L}^{t+1}) \leq 10^{-3}$, $\|\mathbf{Z}^{t+1} - \mathbf{U}^{t+1}\mathbf{W}^{t+1}\|_\infty \leq 10^{-3}$ and $\|\mathbf{U}^{t+1} - \mathbf{P}^{t+1}\|_\infty \leq 10^{-3}$.

Output: $\mathbf{P}$, $\mathbf{S}$, $\mathbf{W}$, $\mathbf{T}$.

Fig. 2. Illustration of affinity sparsity. Assuming $\mathbf{A}_i$ is a unit vector and $\mathbf{Q}_i = [1, 1.1, 1.2, 1.3, 1.4, 1.5, 2, 3, 4, 5, 10, 20, 30, 40, 50]$ with $K=15$, we calculate $\mathbf{S}_i$ based on (50) with $\sigma = 1$.

F. Discussion

Here we discuss the following interesting points concerning our proposal and its theoretical derivations.

1) In Eq. (50), the predefined affinity $A_{i\rho_j}$ is performed as the penalty weight, thus establishing a "Retentivity" from the high dimensional space to the low dimensional space. Then, $\exp(-\frac{Q_{i\rho_j}}{\sigma})$ in (50) makes the distribution of affinity to yield flexible "Nonlinearity". And $\exp(-\frac{Q_{i\rho_j}}{\sigma}) \rightarrow 1$ versus $Q_{i\rho_j} \rightarrow 0$ ensures numerical "Stability". As illustrated in Fig. 2, Q_{i,ρ_j} , which is not on the same order of magnitude as Q_{i,ρ_1} , may make S_{i,ρ_j} trend to 0, which highlights a certain "Sparsity".

2) Essentially, our proposal yields two projection matrices $\mathbf{P} \in \mathbb{R}^{d \times r}$ and $\mathbf{PW} \in \mathbb{R}^{d \times c}$, both of which can be used for subsequent tasks. $\mathbf{P}$ preserves the adjacency structure information of the entire data distribution and can project data to any dimensional subspace, while $\mathbf{PW}$ can only project data to a c -dimensional subspace. Besides, for $\mathbf{PW}$, assuming $r < c$, the resulting $\mathbf{PW}$ has a discriminative low-rank property [35].

3) The sparsity of $L_{2,p}$ norm can realize the joint learning of feature selection and feature extraction, making the input features of joint classification learning as pure as possible.

Moreover, in the process of the joint classification learning, the targets can adaptively respond to the changes of low-dimensional embedding and learn more discriminative features.

IV. SEMI-SUPERVISED EXTENSION OF COSLGE

Assuming that the data matrix $\mathbf{X} = [\hat{\mathbf{X}}, \tilde{\mathbf{X}}]$, where $\hat{\mathbf{X}} \in \mathbb{R}^{d \times l}$ is labeled and $\tilde{\mathbf{X}} \in \mathbb{R}^{d \times u}$ is unlabeled ($l + u = n$), then we can easily extend our model to a semi-supervised version

$$\min_{\mathbf{P}, \mathbf{S}, \mathbf{W}, \mathbf{T}} Tr(\mathbf{P}^T \mathbf{X} \mathbf{L}_s \mathbf{X}^T \mathbf{P}) + \sigma D_{KL}(\mathbf{S} \| \mathbf{A}) + \delta \|\mathbf{P}\|_{2,p}^p + \gamma (\|\hat{\mathbf{X}}^T \mathbf{P} \mathbf{W} - \mathbf{T}\|_F^2 + \alpha \|\mathbf{W}\|_F^2) \quad (54)$$

s.t. $\|\mathbf{S}_i\|_0 = K, \mathbf{S} \mathbf{1} = \mathbf{1}, \mathbf{S} \geq 0, T_{li} - \max_{j \neq l_i} T_{ij} \geq 1$.

In (54), local adaptive graph learning is consistent with (19) across the entire data field, while the separability examination is implemented only on those labeled samples. Labeled data $\hat{\mathbf{X}}$ can be projected into a desired subspace by means of the separability examination, while unlabeled data $\tilde{\mathbf{X}}$ can be drawn closer to labeled data $\hat{\mathbf{X}}$ with the help of its underlying structure, distribution patterns, and adjacent-labeled data. Problem (54) can also be solved using Algorithm 1, but $\mathbf{W}$ is only learned on the labeled projected data $\mathbf{P}^T \hat{\mathbf{X}}$.

V. EXPERIMENTS

In this section, we conduct a comprehensive analysis of the proposed method experimentally, including comparisons with other methods on several public benchmark databases, validation of the effectiveness of each component in our model using ablation experiments, and study of parameter sensitivity and algorithm convergence.

A. Databases

Fig. 3. Typical images sampled from six databases. From top to bottom are GTdb, UMIST, YaleB, PIE, USPS, and COIL20.

Here, six distinctive and widely-used databases are employed to demonstrate the validity of our algorithm, including GTdb, UMIST, YaleB, PIE, USPS, and COIL20, four of which are to evaluate face recognition performance, and two for handwritten and object recognition performance, respectively.

Some typical sample images are illustrated in Fig. 3. Besides, more details are described below.

GTdb¹ - The Georgia Tech face database contains images of 50 people, each with 15 sets of images. The pictures show frontal and/or tilted faces with different facial expressions, lighting conditions and scale. We resize each cropped image to 60×40 pixels and directly train the grayscale pixel features.

UMIST [43] - The UMIST face database consists of 575 images of 20 subjects, each changing a series of views from the side to the front. Each subject has no less than 20 images and a maximum of 38 images, each of which is resized to 32×32 pixels in our experiments.

YaleB [44] - The cropped Extended Yale B face database consists of 2414 frontal facial images of 38 subjects, each photographed under about 64 different lighting conditions. We resize all the images to 32×32 pixels, with 256 gray levels per pixel.

PIE [45] - The CMU PIE database includes 68 persons and a total of 41368 face images taken by multiple synchronous cameras and flashes in different poses, lighting and expressions. Our study focuses on five near-frontal image sets (C05, C07, C09, C27, C29), with approximately 170 images per subject for a total of 11554 images. All images are grayscale and cropped to 32×32 pixels.

USPS [46] - The USPS handwritten digital database we refer to here is a popular subset of the original database that contains 9298 images of 10 Arabic numerals, each resized to 16×16 pixels, meaning that each sample is a 256 dimensional vector.

COIL20 [47] - The Colombian Object Image Library database is composed of 20 objects with a total of 1440 images. As objects rotate on the turntable, images of each object are taken at five-degree intervals, with 72 images for each object. Each image is adjusted to 32×32 pixels.

B. Competitors

Training a Deep Neural Network (DNN) requires a large number of data points, which is resource-based and not suitable for small scale datasets [48]. As a result, we focus mainly on several state-of-the-art graph-based methods for comparisons, including supervised LPP (SLPP) [5], LPP-based local Fisher discriminant analysis (LFDA) [7], marginal Fisher analysis (MFA) [8], nonnegative sparse graph (NNSG) based linear regression [42], robust latent subspace learning (RLatSL) [32], joint graph optimization and projection learning (JGOPL) [31], simultaneously learning neighborhood and projection (SLNP) [19], adaptive local linear discriminant analysis (ALLDA) [27]. Except that JGOPL and SLNP are our duplicate codes based on the published papers, the codes of other methods are released by the original authors. To be fair, for unsupervised JGOPL, we extend it to supervised scenarios

below

$$\begin{aligned} \min_{P, S} & \|P^T(X - XS)\|_{2,1} + \alpha \|E \odot S\|_1 \\ & + \beta \sum_{k=1}^c \sum_{i,j=1}^{n_k} \|P^T X_i^k - P^T X_j^k\|_2 S_{ij}^k \quad (55) \\ \text{s.t. } & S \geq 0, S_{ij} = 0 \text{ if } l_i \neq l_j. \end{aligned}$$

where X_i^k is the i -th sample belonging to the k -th class, and S_{ij}^k is the affinity between X_i^k and X_j^k .

Besides, some information about these methods is described in Table I. Among them, there are 3 predefined graph methods, 3 dynamic intra-class graph learning methods and 2 global dynamic graph learning methods. NNSG builds a non-negative self-representation graph for the learned targets, while ours builds a global graph of inter-sample relationships in the projected subspace.

TABLE I
ATTRIBUTES OF ALL METHODS INVOLVED IN COMPARISONS.

Method	Dynamic graph	Neighborhood	Hyper-parameter
SLPP	No	20%	None
LFDA	No	20%	None
MFA	No	20%	None
NNSG	Global	Adaptive	4
RLatSL	—	—	3
JGOPL	Intra-class	Adaptive	2
SLNP	Intra-class	20%	None
ALLDA	Intra-class	20%	None
COSLGE	Global	20%	3

C. Experimental Setup

All the experimental images are expanded into vector samples and normalized. n_s samples of each category in each database are randomly selected for training, and the remaining samples form a test set. To be specific, in the following experimental results, the value of n_s is denoted as “# number”. Considering that some methods involve inverse problems, PCA algorithm is performed to reserve 98% energy for each data in all our experiments. Since the regression-based algorithms can only reduce to c dimensions, we set the reduced dimensions of each algorithm to c for simplicity and convenience.

In Table I, there are some parameters that need to be set up in advance for the methods being compared. As for the neighborhood size of each method constructing the graph, we uniformly set it at 20% of each class of samples. Self-representation graph learning adaptively determines the neighborhood size. Then, we simply accept the suggested hyper-parameter settings in the original papers. After extracting compact features by various methods, we simply classify them by the nearest neighbor classifier (1-NN) and repeat 10 trials randomly to evaluate their performance with average recognition accuracies and standard deviations.

D. Experimental Results

1) *Classification Performance*: In this section, we investigate the superiority of our method over other methods

¹http://www.anefian.com/research/face_reco.htm

in classification performance. These databases involve face recognition, object recognition and handwritten recognition, which is helpful to reflect the universality of the proposed method. In the supervised version, our COSLGE can learn two different projections simultaneously, namely $P \in \mathbb{R}^{d \times r}$ and $PW \in \mathbb{R}^{d \times c}$. Thus, to compare the classification performance of these two projections, we denote their corresponding methods as COSLGE_P and COSLGE_PW. Instead of comparing our semi-supervised version S-COSLGE separately, the training data of COSLGE is taken as the labeled samples of S-COSLGE and the test data is taken as the unlabeled samples for training. The final experimental results are shown in Table II. And then, we discuss the following concerns.

1) Our approach surpasses other approaches at almost every split in all of these databases, especially for GTdb databases with more complex distributions. The two projections learned are all valid and have similar classification performance. Because of adjacency graph learning across training sets and test sets, the results of S-COSLGE are usually better than supervised COSLGE.

2) Some graph optimization algorithms can not even exceed the predefined graph algorithms, the potential reasons may be caused by two aspects: one is the algorithm's ability to extract discriminative information is insufficient, and the other is the rule assumption of graph optimization is unreasonable. Both LFDA and MFA attempt to construct local penalty graph, which is helpful to enhance the ability of model to extract discriminant information.

3) JGOPL, SLNP, and ALLDA all require the inverse of the global scatter matrix S_t , thus it is necessary to remove the null space of the matrix S_t in advance or to impose regularization perturbation on the matrix S_t . Our approach, by contrast, does not require these concerns and does special preprocessing of the data. Moreover, RLatSL and NNSG can only project samples into c -dimensional space, and our method COSLGE_P can project samples into any dimensional space.

4) A stable and efficient model needs to yield stronger compactness in the sample domain with high similarity and induce certain separability among different classes. Thus, we prefer to construct graph globally rather than within classes. Moreover, the tendency to preserve marginal adjacency information helps to enhance the generalization ability of the model. Then in our model, the joint classifier promotes the generation of separability.

2) *Visualization Results:* In this section, we present some visual results to support the validity of our proposed COSLGE. Due to space limitations, we cannot illustrate the results on all databases, but only on the PIE and YaleB databases.

With the increase of dimension, the inter-sample affinity based on Euclidean metric gradually fails.² Measuring the affinity between similar samples in the original space becomes extremely hard, as illustrated on the left in Fig. 4. An alternative strategy is to project these samples into a low-dimensional space so that similar samples are clustered in a compact domain, which is what our model craves. By contrast,

all the non-zero affinities learned on the right side of Fig. 4 have a distinct meaning and are almost all concentrated in the same class. The meaning of "almost" here is that some marginal information remains during the learning process. On the right side of Fig. 4, we illustrate two sample pairs {1630, 121} and {1290, 2239} that are not in the same class and have the non-zero affinities, where the number refers to the order in the training set. The two sample pairs are highly similar in terms of facial features, face orientation, and lighting conditions, which confirms the importance of preserving marginal information illustrated in the framework of Fig. 1.

Fig. 4. Two affinity matrices generated from class 5 to 10 on the PIE database of 40 training samples per class. Initial affinity matrix in original space (left). Final affinity matrix in projected space (right).

Besides, to investigate the validity of the separability examination in our model more intuitively, we specially made t-SNE visualization for the original sample data and the projected sample data. In Fig. 5, it is obvious that the data samples belonging to different classes after projection are significantly separated from each other, while the samples in the original space are mixed together.

Fig. 5. t-SNE visualization of all original samples (left) and all embedded samples (right) in YaleB database.

As for the classification targets, we also visualized them in Fig. 6, and they differ significantly from the 0-1 targets. The following ablation experiments will verify their superiority over the strict 0-1 targets.

Fig. 6. Flexible targets learned through COSLGE on YaleB database.

3) *Ablation Study:* In this section, we investigate the performance of our model after removing the relevant modules. Above all, we denote our model without optimizing

²<https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/>

TABLE II
MEAN CLASSIFICATION ACCURACIES $\pm$ STANDARD DEVIATIONS (%) OF DIFFERENT METHODS ON VARIOUS SPLITS OF SIX DATABASES.

Method	GTdb				UMIST				YaleB			
	# 6	# 7	# 8	# 9	# 6	# 8	# 10	# 12	# 15	# 20	# 25	# 30
SLPP	54.38 ± 3.51	54.38 ± 2.22	55.43 ± 2.52	55.90 ± 3.34	89.12 ± 2.88	91.83 ± 2.79	93.87 ± 2.23	95.79 ± 1.26	91.80 ± 0.79	93.98 ± 0.93	95.13 ± 0.62	96.23 ± 0.37
LFDA	63.42 ± 1.38	66.70 ± 2.27	70.20 ± 2.12	73.13 ± 1.83	92.66 ± 2.49	95.13 ± 1.83	97.07 ± 1.25	98.36 ± 0.97	92.81 ± 0.90	95.28 ± 0.81	96.37 ± 0.58	97.42 ± 0.49
MFA	66.09 ± 1.49	67.73 ± 1.92	70.00 ± 1.85	71.33 ± 2.48	91.67 ± 2.76	95.18 ± 1.52	96.67 ± 1.39	97.67 ± 0.94	90.34 ± 0.97	92.08 ± 0.79	93.31 ± 0.89	94.06 ± 0.74
NNSG	66.58 ± 1.52	67.55 ± 2.08	69.60 ± 1.66	71.97 ± 2.05	87.76 ± 1.77	91.73 ± 2.23	94.69 ± 1.73	96.00 ± 1.29	87.94 ± 0.86	90.27 ± 0.93	91.40 ± 0.84	91.92 ± 0.64
RLatSL	63.42 ± 2.06	64.40 ± 1.76	65.54 ± 2.00	67.07 ± 2.53	89.89 ± 2.46	93.23 ± 1.87	95.01 ± 1.92	96.06 ± 1.35	93.11 ± 0.95	95.37 ± 0.66	96.69 ± 0.45	97.56 ± 0.57
JGOPL	45.89 ± 3.85	46.83 ± 3.18	47.31 ± 1.98	48.83 ± 3.10	89.14 ± 1.82	93.13 ± 1.57	95.68 ± 1.99	97.01 ± 1.30	88.99 ± 0.93	92.30 ± 0.75	93.35 ± 0.47	94.36 ± 0.46
SLNP	48.40 ± 3.10	49.38 ± 2.00	51.94 ± 2.89	54.03 ± 2.70	90.02 ± 1.99	93.78 ± 1.63	96.08 ± 1.78	97.94 ± 1.19	91.98 ± 1.04	94.37 ± 0.93	95.70 ± 0.56	96.61 ± 0.38
ALLDA	46.67 ± 1.87	46.45 ± 2.06	47.29 ± 2.43	48.50 ± 2.93	86.99 ± 2.24	90.96 ± 2.06	93.23 ± 1.77	95.28 ± 1.53	89.77 ± 0.92	92.62 ± 0.87	94.31 ± 0.77	95.67 ± 0.28
COSLGE_P	77.49 ± 1.38	80.75 ± 1.34	83.09 ± 1.82	84.57 ± 1.49	92.99 ± 2.33	95.57 ± 1.11	97.41 ± 1.46	98.24 ± 0.91	94.92 ± 0.77	96.93 ± 0.63	98.00 ± 0.53	98.81 ± 0.28
COSLGE_PW	77.44 ± 1.35	79.60 ± 1.49	81.71 ± 1.73	83.97 ± 1.59	93.21 ± 2.34	95.78 ± 1.10	97.49 ± 1.38	98.27 ± 0.99	94.96 ± 0.76	96.95 ± 0.59	98.00 ± 0.47	98.81 ± 0.33
S-COSLGE	78.07 ± 1.45	81.25 ± 1.77	83.54 ± 2.04	85.00 ± 1.81	93.19 ± 2.55	95.73 ± 1.22	97.47 ± 1.55	98.33 ± 0.88	96.69 ± 0.81	98.11 ± 0.47	98.70 ± 0.31	99.24 ± 0.29
Method	PIE				USPS				COIL20			
	# 10	# 20	# 30	# 40	# 10	# 20	# 30	# 40	# 15	# 20	# 25	# 30
SLPP	77.02 ± 0.79	89.47 ± 0.48	93.62 ± 0.29	95.16 ± 0.16	71.94 ± 2.74	79.83 ± 1.44	84.54 ± 1.35	87.03 ± 1.07	89.76 ± 1.05	92.79 ± 1.13	94.83 ± 0.77	95.80 ± 0.69
LFDA	85.10 ± 0.71	93.38 ± 0.47	95.45 ± 0.35	96.49 ± 0.16	82.42 ± 1.84	86.29 ± 0.88	88.14 ± 0.76	89.40 ± 0.60	94.32 ± 0.84	96.58 ± 0.80	97.51 ± 0.73	98.19 ± 0.58
MFA	80.04 ± 0.85	89.29 ± 0.67	92.36 ± 0.54	93.89 ± 0.25	72.53 ± 2.80	72.23 ± 2.05	69.98 ± 2.02	69.78 ± 2.39	93.33 ± 1.29	95.11 ± 1.26	95.95 ± 1.10	96.77 ± 0.61
NNSG	82.74 ± 0.97	90.52 ± 0.48	93.05 ± 0.51	94.39 ± 0.27	79.43 ± 1.84	82.35 ± 1.22	83.72 ± 1.23	84.97 ± 0.50	90.71 ± 1.36	92.21 ± 1.17	93.72 ± 1.24	94.60 ± 0.92
RLatSL	84.66 ± 0.89	91.48 ± 0.39	93.91 ± 0.16	94.86 ± 0.12	78.68 ± 2.57	83.00 ± 1.21	85.34 ± 0.93	86.60 ± 0.85	89.03 ± 1.06	91.43 ± 1.36	92.79 ± 1.29	94.14 ± 1.22
JGOPL	78.11 ± 0.70	89.62 ± 0.65	93.30 ± 0.54	94.81 ± 0.31	78.31 ± 2.36	82.92 ± 1.81	85.38 ± 1.31	86.29 ± 1.23	93.59 ± 0.81	96.16 ± 0.75	97.64 ± 0.76	98.62 ± 0.35
SLNP	81.97 ± 0.70	92.33 ± 0.48	94.97 ± 0.41	96.12 ± 0.22	75.51 ± 2.34	83.74 ± 1.24	85.93 ± 1.80	87.48 ± 1.26	93.03 ± 0.73	96.18 ± 0.80	97.43 ± 0.65	98.23 ± 0.66
ALLDA	76.47 ± 0.86	87.90 ± 0.50	92.23 ± 0.42	93.86 ± 0.31	74.01 ± 2.66	79.65 ± 1.18	83.20 ± 2.00	85.43 ± 1.66	89.71 ± 1.17	93.30 ± 0.94	95.69 ± 1.04	97.05 ± 0.45
COSLGE_P	88.64 ± 0.77	94.46 ± 0.33	96.07 ± 0.14	96.78 ± 0.14	85.52 ± 1.21	88.70 ± 0.82	90.29 ± 0.62	91.10 ± 0.72	96.75 ± 0.70	98.37 ± 0.47	98.97 ± 0.48	99.31 ± 0.48
COSLGE_PW	89.71 ± 0.67	94.85 ± 0.33	96.31 ± 0.24	97.03 ± 0.19	85.44 ± 1.20	88.55 ± 0.77	90.05 ± 0.63	90.87 ± 0.63	96.75 ± 0.83	98.47 ± 0.39	99.03 ± 0.44	99.32 ± 0.56
S-COSLGE	87.52 0.57	94.26 0.39	96.02 0.16	96.73 0.11	84.78 ± 1.54	88.83 ± 1.25	90.47 ± 0.87	91.40 ± 0.90	96.78 ± 0.87	98.44 ± 0.45	99.13 ± 0.50	99.36 ± 0.50

"# number" means the number of training samples per class, and bold indicates the best and the second best.

TABLE III
THE PERFORMANCE EVALUATION OF OUR MODEL AFTER REMOVING THE RELEVANT MODULES.

Database	Split	COSLGE_P	COSLGE_noT	COSLGE_noA	COSLGE_noS
GTdb	# 6	77.49± 1.38	76.40 ± 1.10	77.33 ± 1.24	74.67 ± 1.15
	# 9	84.57 ± 1.49	83.60 ± 1.46	84.73± 1.57	81.83 ± 1.71
UMIST	# 6	92.99 ± 2.33	93.19 ± 2.49	93.23± 2.22	92.24 ± 2.56
	# 12	98.24± 0.91	98.00 ± 0.98	98.18 ± 1.00	98.09 ± 0.92
YaleB	# 15	94.92± 0.77	94.12 ± 0.71	94.47 ± 0.99	93.15 ± 1.02
	# 30	98.81± 0.28	98.19 ± 0.39	98.71 ± 0.35	98.25 ± 0.36
PIE	# 10	88.64± 0.77	86.74 ± 1.00	88.42 ± 0.77	83.81 ± 0.79
	# 40	96.78± 0.14	95.53 ± 0.10	96.65 ± 0.12	96.22 ± 0.14

Bold indicates the best.

target matrix T as COSLGE_noT, our model without A as COSLGE_noA, and our model optimizing S only in original space as COSLGE_noS. The average experimental results for the two splits across all databases are shown in Table III. Although COSLGE_noA outperforms the others in GTdb and UMIST, the worst performing COSLGE_noS shows that S is more efficient at capturing nearest-neighbor relationships in projected subspace. Besides that, COSLGE_P performs best in all other situations, which also indicates the indispensability of the three modules.

4) *Sensitivity to Reduced Dimensions*: In the above experimental setup, for fairness, we assume that all methods are projected onto a c -dimensional subspace. In this section, we specially set up a set of experiments to observe the effect of the proposed method with respect to dimension changes. At 10-dimension intervals between 10 and 100 dimensions, we repeat 10 randomized trials across all databases, and the average results are illustrated in Fig. 7. Except for a slight underperformance in 10 dimensions, there is almost no significant difference in the remaining dimensions, indicating that our algorithm is relatively stable to any dimension.

Fig. 7. Classification performance (%) versus reduced dimensions on six different databases with 40, 30, 9, 12, 30 and 40 training samples per class, respectively. The numbers in the legend are the reduced dimensions.

5) *Parameter Analysis*: In our model, three hyper-parameters α , δ and γ need to be tuned before use, where α and δ are regularization parameters. We investigate the parameter sensitivity of the proposed method by observing the classification accuracy under different parameter combinations. Given the properties of the regularization parameter, we recommend that α and δ be set to the same value. Then, we recommend that two regularization parameters $[\alpha, \delta]$ can be selected from set $\{a, b, c, d\} = \{[0.5, 0.5], [0.1, 0.1], [0.05, 0.05], [0.01, 0.01]\}$, and the trade-off coefficient γ can be selected from set $\{0.5, 1, 5, 10\}$. The classification representation is shown in Fig. 8. In general, our model has strong stability for parameter combinations except for GTdb and UMIST databases with smaller sample sizes. The performance on GTdb reflects that the model is more affected by regularization.

6) *Convergence Analysis*: Here, we demonstrate the convergence of the proposed COSLGE. In Fig. 9, we plot the changes of the three convergence conditions of our algorithm with the number of iterations on YaleB and PIE databases. In the early stage of convergence, the convergence curve may fluctuate under the influence of iteration step size but eventually converges. Of the three, $\|Z - UW\|_\infty$ and $\|U - PW\|_\infty$ converge the fastest within 20 iterations. The convergence of our method to other databases is not covered due to the limitation of space.

Fig. 8. Classification performance (%) of the proposed COSLGE versus hyper-parameters α & δ and γ on six different databases.

Fig. 9. Convergence curves satisfying the termination conditions in Algorithm 1.

VI. CONCLUSION

In this paper, we propose a compact self-learning representation framework that combines adjacency graph learning with joint separability examination. Different from the existing supervised self-learning graph embedding methods, we tend to preserve both intra-class and inter-class adjacency information for better generalization performance. And the obtained low-dimensional embedded samples are input into the joint classifier to test their inter-class separability, which is beneficial to the subsequent classification tasks. Extensive experiments on six widely used databases demonstrate that the proposed method has strong generalization, stability and superiority. Besides, our approach can easily be extended to semi-supervised version and does well. In future work, we further consider improving the robustness of our model and the processing of large data sets.

REFERENCES

- [1] H. S. Seung and D. D. Lee, "The manifold ways of perception," *Science*, vol. 290, no. 5500, pp. 2268-2269, 2000.
- [2] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: *Advances in Neural Information Processing Systems (NIPS)*, 2002, pp. 585-591.
- [3] E. Vural and C. Guillemot, "Out-of-Sample Generalizations for Supervised Manifold Learning for Classification," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1410-1424, 2016.
- [4] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems*, 2004, vol. 16, no. December, pp. 153-160.

- [5] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328-340, 2005.
- [6] J. Yang, D. Zhang, J. Y. Yang, and B. Niu, "Globally maximizing, locally minimizing: Unsupervised discriminant projection with applications to face and palm biometrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 650-664, 2007.
- [7] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1027-1061, 2007.
- [8] S. Yan, D. Xu, B. Zhang, H. J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40-51, 2007.
- [9] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis," in *IJCAI International Joint Conference on Artificial Intelligence*, 2007, pp. 708-713.
- [10] Q. Gao, J. Ma, H. Zhang, X. Gao, and Y. Liu, "Stable orthogonal local discriminant embedding for linear dimensionality reduction," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2521-2531, 2013.
- [11] L. Hu and W. Zhang, "Orthogonal neighborhood preserving discriminant analysis with patch embedding for face recognition," *Pattern Recognit.*, vol. 106, Oct. 2020.
- [12] R. Wang and X. Chen, "Manifold discriminant analysis," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, pp. 429-436.
- [13] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugen.*, vol. 7, no. 2, pp. 179-188, 1936.
- [14] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognit.*, vol. 43, no. 1, pp. 331-341, 2010.
- [15] 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, 2013.
- [16] H. Zhao, S. Sun, Z. Jing, and J. Yang, "Local structure based supervised feature extraction," *Pattern Recognit.*, vol. 39, no. 8, pp. 1546-1550, 2006.
- [17] W. K. Wong and H. T. Zhao, "Supervised optimal locality preserving projection," *Pattern Recognit.*, vol. 45, no. 1, pp. 186-197, 2012.
- [18] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 977-986.
- [19] Y. Pang, B. Zhou, and F. Nie, "Simultaneously Learning Neighborhood and Projection Matrix for Supervised Dimensionality Reduction," *IEEE Trans. neural networks Learn. Syst.*, vol. 30, no. 9, pp. 2779-2793, 2019.
- [20] L. Qiao, L. Zhang, and S. Chen, "Dimensionality reduction with adaptive graph," *Front. Comput. Sci.*, vol. 7, no. 5, pp. 745-75, 2013.
- [21] W. Wang, Y. Yan, F. Nie, S. Yan, and N. Sebe, "Flexible Manifold Learning with Optimal Graph for Image and Video Representation," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2664-2675, 2018.
- [22] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the ℓ_1 -ball for learning in high dimensions," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 272-279.
- [23] J. Huang, F. Nie, and H. Huang, "A new simplex sparse learning model to measure data similarity for clustering," in *IJCAI International Joint Conference on Artificial Intelligence*, 2015, pp. 3569-3575.
- [24] L. Zhang, L. Qiao, and S. Chen, "Graph-optimized locality preserving projections," *Pattern Recognit.*, vol. 43, no. 6, pp. 1993-2002, 2010.
- [25] X. Li, M. Chen, F. Nie, and Q. Wang, "Locality adaptive discriminant analysis," in *IJCAI International Joint Conference on Artificial Intelligence*, 2017, pp. 2201-2207.
- [26] J. Wen, Z. Zhong, Z. Zhang, L. Fei, Z. Lai, and R. Chen, "Adaptive Locality Preserving Regression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 75-88, 2020.
- [27] F. Nie, Z. Wang, R. Wang, Z. Wang, and X. Li, "Adaptive Local Linear Discriminant Analysis," *ACM Trans. Knowl. Discov. Data*, vol. 14, no. 1, pp. 1-19, 2020.
- [28] L. Hu, W. Zhang, and Z. Dai, "Joint Sparse Locality-Aware Regression for Robust Discriminative Learning," *IEEE Trans. Cybern.*, pp. 1-14, 2021.
- [29] F. Nie, Z. Wang, R. Wang, and X. Li, "Submanifold-Preserving Discriminant Analysis with an Auto-Optimized Graph," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3682-3695, 2020.
- [30] Y. Lu, Z. Lai, Y. Xu, X. Li, D. Zhang, and C. Yuan, "Low-Rank Preserving Projections," *IEEE Trans. Cybern.*, vol. 46, no. 8, pp. 1900-1913, 2016.
- [31] Y. Yi, J. Wang, W. Zhou, Y. Fang, J. Kong, and Y. Lu, "Joint graph optimization and projection learning for dimensionality reduction," *Pattern Recognit.*, vol. 92, pp. 258-273, 2019.
- [32] X. Fang, Y. Xu, X. Li, Z. Lai, S. Teng, and L. Fei, "Orthogonal self-guided similarity preserving projection for classification and clustering," *Neural Networks*, vol. 88, pp. 1-8, 2017.
- [33] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79-86, 1951.
- [34] X. Y. Zhang, L. Wang, S. Xiang, and C. L. Liu, "Retargeted Least Squares Regression Algorithm," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 26, no. 9, pp. 2206-2213, 2015.
- [35] X. Cai, C. Ding, F. Nie, and H. Huang, "On the equivalent of low-rank regressions and linear discriminant analysis based regressions," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 1124-1132.
- [36] H. Tao, C. Hou, F. Nie, Y. Jiao, and D. Yi, "Effective Discriminative Feature Selection With Nontrivial Solution," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 27, no. 4, pp. 796-808, 2016.
- [37] F. Nie, W. Zhu, and X. Li, "Structured Graph Optimization for Unsupervised Feature Selection," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 3, pp. 1210-1222, 2021.
- [38] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge university press, 2004.
- [39] S. Xiang, F. Nie, G. Meng, C. Pan and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," *IEEE Trans. Neural New. Learn. Syst.*, vol. 23, no. 11, pp. 1738-1754, Nov. 2012.
- [40] L. Wang and C. Pan, "Groupwise Retargeted Least-Squares Regression," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 4, pp. 1352-1358, 2018.
- [41] 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, 2010.
- [42] X. Fang, Y. Xu, X. Li, Z. Lai, and W. K. Wong, "Learning a nonnegative sparse graph for linear regression," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2760-2771, 2015.
- [43] D. B. Graham and N. M. Allinson, "Characterising Virtual Eigensignatures for General Purpose Face Recognition," in *Face Recognition*, 1998, pp. 446-456.
- [44] A. S. Georgiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643-660, 2001.
- [45] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression Database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615-1618, 2003.
- [46] J. J. Hull, "A Database for Handwritten Text Recognition Research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550-554, 1994.
- [47] S. Nene, S. Nayar, and H. Murase, "Columbia Object Image Library (COIL-20)," *Tech. Rep.*, vol. 95, pp. 223-303, 1996.
- [48] F. Nie, Z. Wang, R. Wang, Z. Wang, and X. Li, "Towards Robust Discriminative Projections Learning via Non-Greedy $\ell_{2,1}$ -Norm MinMax," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2086-2100, 2021.