Double low-rank representation with projection distance penalty for clustering

Zhiqiang Fu[1,2], Yao Zhao[1,2]*, Dongxia Chang[1,2], Xingxing Zhang[3], Yiming Wang[1,2]

[1]Institute of Information Science, Beijing Jiaotong University

[2]Beijing Key Laboratory of Advanced Information Science and Network Technology

[3]Department of Computer Science and Technology, Tsinghua University

{zhiqiangfu, yzhao, dxchang}@bjtu.edu.cn, xxzhang2020@mail.tsinghua.edu.cn, wangym@bjtu.edu.cn

Abstract

This paper presents a novel, simple yet robust self-representation method, i.e., Double Low-Rank Representation with Projection Distance penalty (DLRRPD) for clustering. With the learned optimal projected representations, DLRRPD is capable of obtaining an effective similarity graph to capture the multi-subspace structure. Besides the global low-rank constraint, the local geometrical structure is additionally exploited via a projection distance penalty in our DLRRPD, thus facilitating a more favorable graph. Moreover, to improve the robustness of DLRRPD to noises, we introduce a Laplacian rank constraint, which can further encourage the learned graph to be more discriminative for clustering tasks. Meanwhile, Frobenius norm (instead of the popularly used nuclear norm) is employed to enforce the graph to be more block-diagonal with lower complexity. Extensive experiments have been conducted on synthetic, real, and noisy data to show that the proposed method outperforms currently available alternatives by a margin of 1.0%∼10.1%.

1. Introduction

Clustering is one of the most fundamental unsupervised problems, aiming to group samples into categories such that samples in the same category are similar in some sense and differentiate from those of other categories in the same sense. It has been widely used in many areas (e.g., image processing [31], image segmentation [31], camera source identification [15], and data mining [29, 30, 34, 35]). Many clustering methods (e.g., kmeans based methods [1, 24], density based methods [4, 3], and graph based methods [2]) have been proposed. Among these methods, graph based clustering methods (e.g., Ncut [22]), which classify the samples according to a similarity graph, have attracted lots of attention because of their good performance and solid mathematical foundation. Therefore, constructing a good similarity graph is important for clustering algorithm to obtain good clustering results.

Self-representation models are effective to construct the similarity graph because they are proposed to exploit the subspace structure of data. These methods base on the assumption that a database with k clusters is drawn from a union of k independent low-dimensional subspaces. Furthermore, self-representation theory shows that a sample in a subspace can be linearly represented by the other points in the same subspace [5]. Hence, self-representation methods use the database as the dictionary and learn a similarity graph by capturing the subspace structure.

Sparse subspace clustering (SSC) [6] is the first self-representation method proposed to represent each data point with a few neighbors, hence it can capture the local neighbor relationship. In fact, SSC can't learn the global structure, and the learned graph may be too sparse for clustering [14]. In order to capture the global structure, low-rank representation [17] was proposed to learn the low-dimensional subspace structure with a global low-rank constraint. Compared with SSC, LRR can learn a more denser similarity graph and capture more global information [28]. However, some elements of the graphs obtained by SSC and LRR are negative, while the similarity should be nonnegative. To overcome this problem, Zhuang et al. proposed a non-negative low-rank learning method which adopted both low-rank and sparse constraint and could show the similarity among samples directly [37]. Recently, some evidence showed that using nuclear norm to capture the subspace structure could lose the local intrinsic structure [10, 12]. Motivated by this, some Laplacian regularized LRR methods [11, 19, 33] were proposed to preserve the local information by learning the manifold structure embedded in the data space. Futhermore, these LRR methods use the original features which may contain some redundancy and noise. To address this issue, Wen et al. proposed an adaptive weighted nonnegative low-rank representation that used a sparse weighted matrix to reduce the bad influence of noise and redundancy information [28].

*Corresponding author

The LRR methods mentioned above use the observed data as the dictionary. When the observed data is insufficient or corrupted by noise, the performance of these methods may deteriorate [18]. Therefore, the latent low-rank representation (LatLRR) [18] was proposed to represent the data using both observed and unobserved data. Although LatLRR performs better than LRR in matrix recovering, it still ignores the structure of the feature. In order to learn the relationship among samples and features sufficiently, the double low-rank representation was proposed [32], which performs better than LRR and LatLRR in face recognition. In fact, DLRR uses a global low-rank constraint to learn the structure among the samples and ignores the local structure; in addition, DLRR may not learn the optimal projection since it doesn't take the class information into account. To address these issues, a novel double low-rank representation model, i.e., double low-rank representation with projection distance penalty (DLRRPD), is proposed in this paper, and the major highlights of our method are as follows,

- We develop DLRRPD: a Double Low-Rank Representation with Projection Distance penalty, to improve the discrimination and robustness of similarity graph for clustering tasks.
- An additional projection distance penalty is introduced to capture both the global and local geometrical structures, thus facilitating a sparse and discriminative graph.
- With a Laplacian rank constraint, the robustness of DLRRPD to noises is guaranteed, and meanwhile, the effectiveness of the learned graph is further enhanced.
- Frobenius norm (instead of the widely used nuclear norm) is employed to enforce the graph to be more block-diagonal with a lower complexity, so as to improve the clustering performance.

2. Related work

In this section, we will review some of the most related works in detail and the symbols used are shown in Table.1.

Table 1. Description of the symbols used in the paper

Symbol	Description
X	the original feature of the database
x_i	the i-th column vector of X
x^i	the i-th row vector of X
$x_{i,j}$	the element on the i-th row and j-th column of X
X^T	the transpose of X
tr(X)	the trace of X
X^{-1}	the inverse of X
$\|X\|_1$	the L_1-norm of X
$\|X\|_F$	the Frobenius norm of X
$\|X\|_{2,1}$	the $L_{2,1}$-norm of X
$\|X\|_*$	the nuclear norm of X
$\|x_i\|_2$	the L_2-norm of x_i
$\mathbf{1}$	the vector in which elements are 1
$\odot$	the element-wise multiplication
rank(X)	the rank of X
I	the identity matrix

Given a dataset $X = [x_1, x_2, x_3, ..., x_n] \in \mathcal{R}^{d\times n}$, where $x_i \in \mathcal{R}^{d\times 1}$ is i-th sample with d dimension and n is the number of samples. LRR aims to learn the linear representation with the lowest rank and can be formulated as

$$\min_{Z,E} \|Z\|_* + \lambda\|E\|_{2,1}, s.t. X = XZ + E \qquad (1)$$

where Z is a representation matrix that is also called similarity graph, E is the reconstruction error matrix, and λ is a parameter to balance the effect of different terms. In fact, only the given data is used as the dictionary by LRR but there still exit many unobserved samples. Based on LRR, LatLRR uses both the observed and unobserved data as the dictionary. And the objective function of LatLRR model is

$$\min_{Z,P,E} \|Z\|_* + \|P\|_* + \lambda\|E\|_1, s.t. X = XZ + PX + E \quad (2)$$

where XZ is the principal features, and PX can learn the salient features. Although LatLRR can capture the principal features and the salient features simultaneously, some relationships among feature dimensions are missed. Therefore, DLRR aims to capture structure among samples and features as follows,

$$\min_{Z,P,E} \|Z\|_* + \|P\|_* + \lambda\|E\|_1, s.t. X = PXZ + E \quad (3)$$

where P is a projection matrix that can learn the structure in features (column space), and the similarity graph Z can capture the structure among the samples (row space). However, DLRR uses a global low-rank constraint to capture the global structure and ignores the local structure. Besides, DLRR doesn't use class information to guide the projection, and thus the learned projection maybe not the most suitable one. To address these problems, a novel DLRR method, i.e., DLRRPD, is proposed in this paper.

3. Double low-rank representation with projection distance penalty

As previously analyzed, constructing a good similarity graph is an effective way to improve clustering performance. To improve the quality of the similarity graph, three strategies have been introduced to make the graph more discriminative and robust.

3.1. DLRRPD: Formulation

The formulation of DLRRPD is introduced in this section. Since the nearby samples have a high possibility from the same cluster, the similarity graph should capture this neighbor relationship structure (local structure). Motivated by this, a projection distance penalty is used to capture more local information. Then, a Laplacian rank constraint is adopted to make use of the class information. Therefore,

the initial model of DLRRPD with nuclear norm can be formulated as

$$\begin{aligned}\min_{Z,P,E} \underbrace{\sum_{i,j} \|Px_i - Px_j\|_2^2 z_{i,j}}_{\text{projection distance penalty}} + \frac{\lambda_1}{2}(\|Z\|_* + \|P\|_*) \\ +\lambda_2\|E\|_1, \quad s.t. X = PXZ + E, \\ \underbrace{\text{rank}(L_Z) = n-k}_{\text{Laplacian rank constraint}}, \quad \underbrace{Z \geq 0, \quad z_{i,i} = 0, \quad z^i\mathbf{1} = 1}_{\text{other constraints}}\end{aligned} \tag{4}$$

where Px_i denotes the x_i in the projection space and L_Z is the Laplacian matrix of Z obtained by $L_Z = D_Z - W_Z$, $D_Z = \text{diag}(\text{sum}(W_Z))$ and $W_Z = (Z + Z^T)/2$. $Z \geq 0$ can make sure that each element is positive and satisfies the physical meaning of similarity. $z_{i,i} = 0$ is used to avoid the influence of self-representation. $z^i\mathbf{1} = 1$ can avoid the extreme case that elements of any row of Z are all zeros. By jointly adopting the two constraints, model (4) holds the following good properties:

- Introducing the projection distance penalty has several good properties: 1) as the graph $Z \geq 0$, this penalty can be regarded as a weighted sparse regularization[1], which can ensure the sparsity and locality; 2) model (4) can simultaneously learn the local and global structure to obtain a more discriminative graph; 3) the leaned graph can guide the projection learning.
- The Laplacian rank constraint can ensure that the graph Z consists of k connected components corresponding to k clusters, which is an optimal clustering structure.
- By combining these two terms, the projection learning can be guided with clusters. Consequently, the samples in the same cluster are nearby in the projection space with high similarity. This can further alleviate the adverse effects of noises.

Model (4) learns the subspace structure by minimizing nuclear norm. However, some theoretical analyses and experimental evidence have pointed out that Frobenius norm is another convex surrogate of low-rank constraint [21]. Furthermore, theoretical states that the Frobenius norm satisfies the enforced block diagonal conditions [20], which improves the performance by making the graph more block-diagonal. By taking advantages of Frobenius norm, model (4) can be reformulated as the following problem

$$\begin{aligned}\min_{Z,P,E} \sum_{i,j} \|Px_i - Px_j\|_2^2 z_{i,j} + \underbrace{\frac{\lambda_1}{2}(\|Z\|_F^2 + \|P\|_F^2)}_{\text{Frobenius norm}} \\ +\lambda_2\|E\|_1, \quad s.t. X = PXZ + E, \\ \text{rank}(L_Z) = n-k, Z \geq 0, z_{i,i} = 0, z^i\mathbf{1} = 1\end{aligned} \tag{5}$$

[1]If we define $d_{i,j} = \|Px_i - Px_j\|_2^2$, it is obvious that $\sum_{i,j} \|Px_i - Px_j\|_2^2 z_{i,j} = \|D \odot Z\|_1$

Using Frobenius norm can bring two additional benefits. First, using Frobenius norm can make the graph coefficients of correlated samples be approximately equal to avoid a too sparse graph. Second, while nuclear norm should be solved by SVD which needs lots of computational cost, using Frobenius norm can reduce the computational cost because it can be solved by derivation.

For convenience of calculations, we rewrite model (5) as

$$\begin{aligned}\min_{Z,P,E} 2\text{tr}(PXL_ZX^TP^T) + \frac{\lambda_1}{2}(\|Z\|_F^2 + \|P\|_F^2) \\ +\lambda_2\|E\|_1, \quad s.t. X = PXZ + E, \\ \text{rank}(L_Z) = n-k, Z \geq 0, z_{i,i} = 0, z^i\mathbf{1} = 1\end{aligned} \tag{6}$$

Next, we provide the optimization procedures of DLRRPD.

3.2. DLRRPD: Algorithm

In this section, the proposed model is solved using the alternating direction method of multipliers (ADMM). Since it is difficult to solve problem (6) directly, problem (6) can be relaxed according to [9] as

$$\begin{aligned}\min_{Z,P,E,F} 2\text{tr}(PXL_ZX^TP^T) + \frac{\lambda_1}{2}(\|Z\|_F^2 + \|P\|_F^2) \\ +\lambda_2\|E\|_1 + 2\lambda_3\text{tr}(F^TL_ZF), s.t. X = PXZ + E, \\ Z \geq 0, F^TF = I, z_{i,i} = 0, z^i\mathbf{1} = 1\end{aligned} \tag{7}$$

where $F \in R^{n\times k}$. A variable S is introduced to separate (7) as

$$\begin{aligned}\min_{Z,P,E,S,F} 2\text{tr}(PXL_SX^TP^T) + \frac{\lambda_1}{2}(\|Z\|_F^2 + \\ \|P\|_F^2) + \lambda_2\|E\|_1 + 2\lambda_3\text{tr}(F^TL_SF), \\ s.t. X = PXZ + E, S \geq 0, F^TF = I, s_{i,i} = 0, \\ s^i\mathbf{1} = 1, Z = S\end{aligned} \tag{8}$$

Then the corresponding augmented Lagrangian function of Eq.(8) is

$$\begin{aligned}\min_{Z,P,E,S,F} 2\text{tr}(PXL_SX^TP^T) + \frac{\lambda_1}{2}(\|Z\|_F^2 \\ +\|P\|_F^2) + \lambda_2\|E\|_1 + 2\lambda_3\text{tr}(F^TL_SF) + \\ \frac{\mu}{2}(\|X - PXZ - E + \frac{C_1}{\mu}\|_F^2 + \|Z - S + \frac{C_2}{\mu}\|_F^2)\end{aligned} \tag{9}$$

where C_1 and C_2 are Lagrange multipliers, and μ is a positive penalty parameter. Using the alternative update strategy, the objective function (9) can be divided into the following subproblems:

Update Z: Fixing S, P, E and F, Z can be updated by solving the following problem:

$$\begin{aligned}\min_Z \frac{\lambda_1}{2}\|Z\|_F^2 + \frac{\mu}{2}(\|X - PXZ - E + \frac{C_1}{\mu}\|_F^2 + \\ \|Z - S + \frac{C_2}{\mu}\|_F^2)\end{aligned} \tag{10}$$

By setting the derivative of Eq.(10) to zero, Z can be updated as

$$Z = (\lambda_1 I + \mu I + \mu(PX)^T PX)^{-1}\mu(L_1 + L_2) \quad (11)$$

where $L_1 = (PX)^T(X - E + C_1/\mu)$ and $L_2 = S - C_2/\mu$.

Update S: When Z, P, E and F are fixed, S can be computed by minimizing the following formula,

$$\begin{aligned}\min_{S\geq 0, s_{i,i}=0, s^i\mathbf{1}=1} \sum_{i,j} \|Px_i - Px_j\|_2^2 s_{i,j} + \\ 2\lambda_3 \text{tr}(F^T L_S F) + \frac{\mu}{2}\|Z - S + \frac{C_2}{\mu}\|_F^2\end{aligned} \quad (12)$$

Formula (12) can be rewritten as

$$\begin{aligned}\min_{S\geq 0, s_{i,i}=0, s^i\mathbf{1}=1} \sum_{i,j} \|Px_i - Px_j\|_2^2 s_{i,j} + \\ \lambda_3 \sum_{i,j} \|f_i - f_j\|_2^2 s_{i,j} + \frac{\mu}{2}\|Z - S + \frac{C_2}{\mu}\|_F^2\end{aligned} \quad (13)$$

In order to simplify the calculation, we define that $q_{i,j} = \|Px_i - Px_j\|_2^2$ and $h_{i,j} = \|f_i - f_j\|_2^2$, and then this subproblem can be rewritten as

$$\begin{aligned}\min_{S\geq 0, s_{i,i}=0, s^i\mathbf{1}=1} \lambda_2 \text{tr}(Q^T S) + \lambda_3 \text{tr}(H^T S) \\ + \frac{\mu}{2}\|Z - S + \frac{C_2}{\mu}\|_F^2\end{aligned} \quad (14)$$

To improve the efficiency, problem (14) can be solved by two steps. Firstly, a latent solution $\bar{S}$ can be obtained by minimizing following problem,

$$\begin{aligned}\min_S \lambda_2 \text{tr}(Q^T S) + \lambda_3 \text{tr}(H^T S) \\ + \frac{\mu}{2}\|Z - S + \frac{C_2}{\mu}\|_F^2\end{aligned} \quad (15)$$

This formula has a closed solution as

$$\bar{S} = Z + \frac{C_2 - \lambda_2 Q - \lambda_3 H}{\mu} \quad (16)$$

Then we can obtain S by solving following problem,

$$\min_{S\geq 0, s_{i,i}=0, s^i\mathbf{1}=1} \|S - \bar{S}\|_F^2 \quad (17)$$

This problem can be regarded as n independent subproblems, and it can be calculated as

$$s^i = \max(\sigma^i \hat{\mathbf{1}}_i + \bar{s}^i, 0) \quad (18)$$

where $\hat{\mathbf{1}}_i$ is a vector that the i-th element is 0, and the other elements are 1. σ is the Lagrangian multiplier which is defined as

$$\sigma^i = \left(1 + \bar{s}^i \mathbf{1}\right)/(n-1) \quad (19)$$

Update F: F can be obtained by solving the following problem with the other variables fixed.

$$\min_F \text{tr}(F^T L_S F), s.t. F^T F = I \quad (20)$$

This problem has a close solution which is the set of k eigenvectors corresponding to the first k smallest eigenvalues of L_S.

Update P: When the other variables are fixed, P can be updated by solving the following sub-problem,

$$\begin{aligned}\min_P 2\text{tr}(PXL_S X^T P^T) + \frac{\lambda_1}{2}\|P\|_F^2 + \\ \frac{\mu}{2}\|X - PXZ - E + \frac{\mu}{2}\|_F^2\end{aligned} \quad (21)$$

This problem can be directly solved as

$$P = \mu L_3 Z^T X^T L_4^{-1} \quad (22)$$

where $L_3 = X - E + C_1/\mu$ and $L_4 = \lambda_1 I + 4XL_S X^T + \mu XZZ^T X^T$.

Update E: E can be obtained with the other variables fixed as

$$\min_E \lambda_2\|E\|_1 + \frac{\mu}{2}\|X - PXZ - E + \frac{C_1}{\mu}\|_F^2 \quad (23)$$

This problem can solved directly by

$$E = \Omega_{\lambda_2/\mu}(X - PXZ + C_1/\mu) \quad (24)$$

where Ω is the shrinkage operator mentioned in [16].

Update the other parameters: Penalty parameter μ, lagrange multipliers C_1 and C_2 can be updated as follows,

$$\mu = \min(\rho\mu, \mu_{\max}) \quad (25)$$

$$C_1 = C_1 + \mu(X - PXZ - E) \quad (26)$$

$$C_2 = C_2 + \mu(Z - S) \quad (27)$$

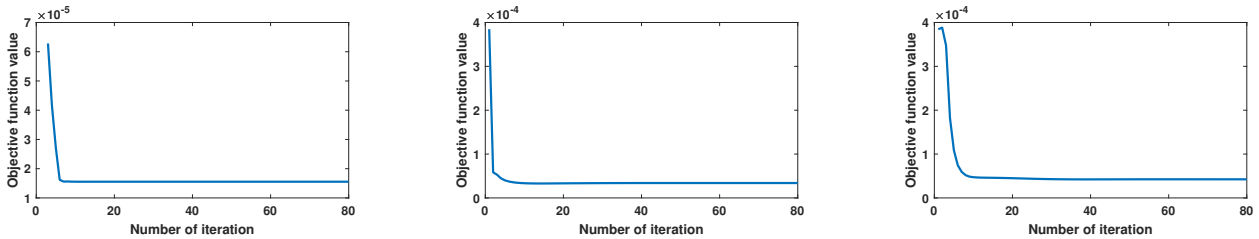
where ρ and μ_{max} are two constants. The proposed solution of model (9) is summarized as Algorithm 1.

4. Analysis of our method

In this section, we further analyze the computation complexity, convergence, and connections to other methods.

4.1. Complexity and convergence analysis

DLRRPD is solved as Algorithm 1 that contains five main steps, i.e., step 3-7. Step 3, 4 and 6 use inverse operation, so their computational complexities are $\mathcal{O}(n^3)$, $\mathcal{O}(n^3)$ and $\mathcal{O}(d^3)$, respectively. Step 5 is updated by eigendecomposition whose computational complexity is $\mathcal{O}(kn^2)$. Since step 7 is solved by singular value thresholding, its computational complexity is $\mathcal{O}(n^3)$. Then we can know that the



(a) Auto (b) Cars (c) Glass

Figure 1. Convergence curve of our DLRRPD on Auto, Cars and Glass, in which all classes of each database are selected.

Algorithm 1: Solving DLRRPD

Input: Data matrix X and parameters λ_1, λ_2, λ_3
Output: Z,P,S,E,F

Initialization: Initializing Z by constructing the k-nearest neighbor graph, initializing F by Eq.(20), $S = Z$,$P = I$,$E = X - PXZ$, $C_1 = 0$, $C_2 = 0$ $\mu = 0.01$, $\rho = 1.1$, $\mu_{max} = 10^8$;
while *not converged* **do**
  Update Z by Eq. (11);
  Update S by Eq. (18);
  Update F by Eq. (20);
  Update P by Eq. (22);
  Update E by Eq. (24);
  Update μ, C_1 and C_2 by Eq.(25)(26)(27);

computation complexity of our method is $\mathcal{O}(\tau(d^3 + 3n^3 + kn^2))$, where τ is the number of iterations.

Our DLRRPD uses ADMM methods to get the solution, and it's a five-block ADMM problem. The strong convex of two-block ADMM has been proved in [13, 8], but as far as we know, it is still unrealistic to prove the five-block ADMM is convex. Hence, we will prove the convergence of our method empirically. In order to show the convergence, the objective function value with respect to the number of iterations is shown in Fig.1. The objective function value can be obtained by Obj $= (2\text{tr}(PXL_ZX^TP^T) + \frac{\lambda_1}{2}(\|Z\|_F^2 + \|P\|_F^2) + \lambda_2\|E\|_1 + 2\lambda_3\text{tr}(F^TL_ZF))/\|X\|_F^2$. As shown in Fig.1, the objective function value of DLRRPD monotonically decreases until the local optimal point, and it is obvious that the proposed method can converge fastly.

4.2. Connections to other methods

In this section, the connections among the proposed method and two most related methods (i.e, DLRR, and RSEC) are analyzed.

Connections to DLRR: The model of DLRR is shown as Eq.(28).

$$\min_{Z,P,E} \|Z\|_* + \|P\|_* + \lambda\|E\|_1, s.t. X = PXZ + E \quad (28)$$

Compared with DLRR, there are lots of improvements in DLRRPD. Firstly, a projection distance penalty is introduced to DLRRPD to capture more local structure, which makes the graph more discriminative. Then a rank constraint is adopted to DLRRPD to make sure that the similarity graph contains k connected component. Moreover, Frobenius norm is used to learn a better graph with lower computational complexity. Hence, DLRRPD performs much better than DLRR.

Connections to RSEC: The model of RSEC is shown as Eq.(29).

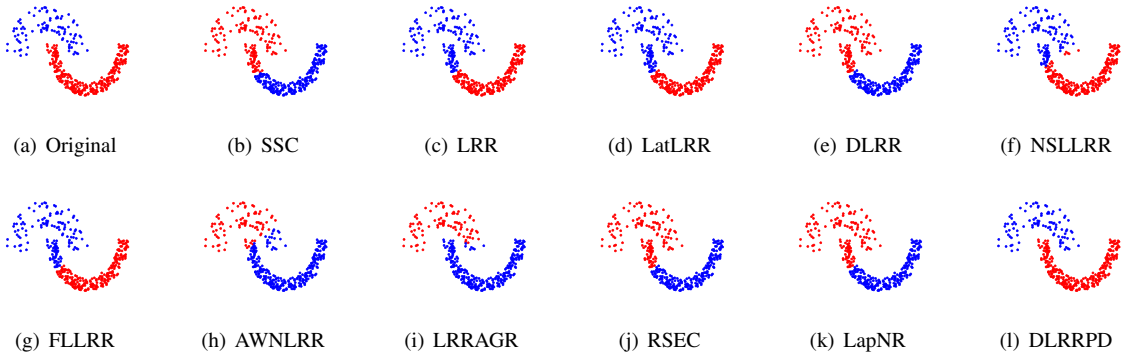
$$\begin{aligned}\min_{Z,F,E} \text{tr}(F^TL_ZF) + \lambda_1\|Z\|_* + \lambda_2\|E\|_{2,1}, \\ s.t. X = XZ + E, F^TF = I\end{aligned} \quad (29)$$

As shown in Eq.(29), RSEC can be regarded as a special case of DLRRPD. If we set the $P = I$ and remove the distance penalty, then DLRRPD will degrade to RSEC with Frobenius norm. Compared with RSEC, DLRRPD introduces a distance penalty to preserve more intrinsic structure and adopts projection learning to learn a better feature. Moreover, DLRRPD uses $z_{i,i} = 0$ and $z^i\mathbf{1} = 1$ to avoid the trivial solution. Thus, DLRRPD can achieve better performance.

5. Experiments and analysis

In order to show the effectiveness of our method, some experiments are conducted on synthetic, real and noisy databases. Here, the performances of DLRRPD and several related algorithms, i.e., Ncut [23], SSC [7], LRR [17], LatLRR [18], DLRR [32], NSLLRR [33], FLLRR [25], AWNLRR [28], LRRAGR [27], RSEC [26] and LapNR [36] are compared through the experiments.

Ncut is a classical clustering method that is always used as the baseline of clustering. Moreover, Ncut is also used to handle the similarity graph obtained by other methods. SSC, LRR, LatLRR, DLRR, and FLLRR are five basic self-representation methods. NSLLRR, AWNLRR, LRRAGR, RSEC, and LapNR are five improved methods that can achieve better performance. To make it fair, each method's parameters are varied in a wide range to find the best performance. Moreover, all experiments are conducted on a



(a) Original (b) SSC (c) LRR (d) LatLRR (e) DLRR (f) NSLLRR

(g) FLLRR (h) AWNLRR (i) LRRAGR (j) RSEC (k) LapNR (l) DLRRPD

Figure 2. Experimental results on the two-moon database.

PC with Intel Core I7-10700 CPU @ 4.6GHz 32G.

5.1. Experiments on two-moon database

In this section, a synthetic database, i.e., the two-moon database shown in Fig.2(a), is used to evaluate the methods. Here different colors of dots represent different clusters. It can be seen that some samples in different clusters are close. As shown in Fig.2(b) - Fig.2(l), while the comparison methods are misled by the nearby sample in the different clusters, our DLRRPD can obtain the ground truth because it can preserve more intrinsic structure by the projection distance penalty.

Table 2. Description of the databases

Type	Database	Samples	Dim	Classes
UCI	Auto	205	25	6
	Cars	392	8	3
	Contral	600	60	6
	Glass	214	9	6
	Isolet	1560	617	2
Handwritten	Dig	1797	64	10
	USPS	1000	256	10
	MNIST	1000	256	10
Face	Jaffe	213	676	10
	MSRA	1799	256	12
	Umist	575	645	20

5.2. Experiments on real databases

In this section, eleven real databases are used to evaluate the performance of all the methods mentioned above. These databases include five UCI databases: Auto, Cars, Control, Glass, and Isolet, three handwritten databases: Dig, USPS, and MNIST, three face databases: Jaffe, MSRA, and UMIST. The details of these databases are shown in Table 2. For comparison purposes, three typical performance metrics are used: accuracy (ACC), normalized mutual information (NMI) and F1-score.

The clustering performance is shown in Table 3, and we can conclude as:

- Overall, DLRRPD achieves very competitive and stable performance compared to most compared methods. Taking the image databases, handwritten databases and Isolet with large dimension feature for example, DLRRPD can significantly outperform the other methods. For the Umist database, the proposed DLRRPD achieves more than 10% scores of ACC in comparison with LRRAGR (the second-best method). Moreover, the proposed method can also obtain higher ACC than other methods on the remaining databases. This indicates that projection learning can capture the important features and reduce the redundant information of the high-dimensional data.
- From the comparison between LatLRR and FLLRR, we can find that FLLRR performs better in most cases. Since FLLRR improves LatLRR by using Frobenius norm instead of nuclear norm, this phenomenon proves that Frobenius norm is more efficient than nuclear norm in clustering.
- With respect to NSLLRR, LRRAGR, AWNLRR and LapNR, our proposed DLRRPD often shows better performance. This fully demonstrates our DLRRPD captures the actual structure among the samples using the projection distance penalty.
- In addition, DLRRPD consistently performs better than RSEC on almost the database. RSEC just introduce a rank constraint to LRR to make sure the learned graph contains k connected components. This can show the effectiveness of using class information. DLRRPD uses the class information to jointly guide the projection learning and graph learning, hence a more discriminative graph can be obtained leading to better performance.

In summary, these observations validate the efficacy of our projection distance penalty, the Laplacian rank constraint and Frobenius norm. With the integration of the

Table 3. Clustering results on real databases

Database	Metric	Ncut	SSC	LRR	LatLRR	DLRR	FLLRR	NSLLRR	AWNLRR	LRRAGR	RSEC	LapNR	DLRRPD
Auto	ACC	41.95	40.00	40.98	41.46	41.46	41.46	41.95	40.98	39.02	44.39	41.46	**46.83**
	NMI	19.22	16.30	16.61	17.22	17.27	17.22	18.39	18.57	16.67	17.86	17.82	**20.53**
	F1	32.53	33.32	34.05	34.62	32.47	34.62	33.48	32.46	34.20	35.16	32.36	**36.70**
Cars	ACC	48.72	61.99	62.76	62.76	62.76	61.99	63.52	66.33	62.76	63.01	57.14	**68.37**
	NMI	22.01	1.33	4.55	4.55	4.55	1.33	6.83	20.97	20.09	22.53	23.99	**24.07**
	F1	48.60	62.58	63.17	63.17	63.17	62.58	63.67	66.04	59.25	58.92	50.54	**66.18**
Control	ACC	51.50	54.83	48.17	47.33	74.00	40.83	65.00	53.17	56.83	54.33	37.83	**76.33**
	NMI	67.11	69.59	63.37	61.11	61.51	58.09	61.37	61.13	71.94	62.42	67.81	**74.78**
	F1	58.35	64.39	57.25	55.77	62.03	52.58	62.11	53.47	68.10	57.61	54.28	**68.85**
Glass	ACC	54.21	48.60	53.27	51.87	47.66	51.40	57.48	54.67	55.61	54.67	53.74	**58.48**
	NMI	39.58	35.26	33.13	39.22	25.05	38.28	39.87	43.75	**45.90**	38.87	38.99	39.91
	F1	44.08	42.18	40.22	41.65	43.05	41.16	48.70	49.17	**51.05**	48.57	42.63	47.09
Isolet	ACC	55.58	54.29	56.00	55.64	57.12	54.10	59.36	58.40	54.49	62.95	58.65	**67.37**
	NMI	0.90	0.60	1.03	0.93	1.48	0.49	2.60	2.07	0.59	4.93	2.28	**9.00**
	F1	50.62	52.16	50.60	50.77	51.11	50.31	51.75	51.53	50.51	53.43	52.24	**56.15**
Dig	ACC	76.85	14.08	79.13	79.12	60.77	78.95	67.78	79.86	59.32	79.19	76.02	**88.81**
	NMI	71.51	1.38	77.11	74.02	48.05	74.52	71.85	84.27	70.47	76.83	78.93	**88.81**
	F1	67.71	10.15	72.85	71.12	43.35	71.63	63.23	76.90	45.25	72.32	71.70	**83.22**
MNIST	ACC	56.30	56.70	55.40	48.80	33.80	55.80	56.00	61.60	55.70	55.70	64.00	**68.20**
	NMI	47.72	58.32	50.95	47.21	26.40	50.44	54.74	59.70	59.61	51.91	61.33	**66.15**
	F1	42.36	49.15	43.96	40.61	24.98	43.47	46.66	50.78	45.16	44.54	53.52	**59.79**
USPS	ACC	49.50	52.50	53.30	50.60	33.70	49.10	54.20	55.00	40.90	53.80	57.20	**59.90**
	NMI	43.78	52.46	49.97	46.77	26.57	46.97	48.76	55.75	50.15	50.66	55.55	**55.58**
	F1	36.96	42.67	43.82	40.76	24.66	39.82	42.35	46.81	38.90	44.33	47.24	**47.82**
Jaffe	ACC	90.00	96.71	99.53	**100**	75.12	**100**	99.53	98.59	98.59	**100**	98.12	**100**
	NMI	87.57	95.99	99.18	**100**	71.66	**100**	99.17	97.52	98.16	**100**	97.36	**100**
	F1	82.45	93.62	99.05	**100**	62.63	**100**	99.03	97.11	97.10	**100**	96.32	**100**
MSRA	ACC	52.64	60.92	65.87	63.76	39.02	65.81	69.65	55.98	55.98	69.71	61.42	**72.98**
	NMI	57.95	73.48	73.58	69.10	44.05	72.50	**74.99**	62.63	71.08	70.73	74.79	73.71
	F1	42.75	52.06	55.47	51.65	33.91	55.27	61.49	47.41	46.20	55.06	55.13	**63.31**
Umist	ACC	47.83	62.43	45.57	39.83	29.39	43.83	55.48	65.39	69.91	45.74	53.04	**80.17**
	NMI	62.62	77.50	61.30	61.90	44.07	59.91	72.50	80.85	82.95	65.26	70.51	**89.57**
	F1	38.25	53.35	35.36	30.91	21.32	34.11	42.12	58.90	59.35	39.76	45.49	**72.73**

$*$The variances of experiments are all 0.

above factors, the proposed method achieves better performance than the other methods.

5.3. Image clustering against corruptions

In this section, the robustness property of our DLRRPD is explored. Here, MSRA and UMIST databases are used to evaluate the robustness. For computational efficiency, we select the first 10 samples of each class to construct two sub-databases. In the experiments, salt & pepper noise with a fixed percentage is added to the image, which may break the distance relationship among samples. The clustering results of noisy data are shown in Fig.3 in which the noise percentage is set to $[0, 10, 20, 30, 40, 50]$, and some noisy images are also shown. We can find that: 1) the ACC decreases monotonically when the noise level is increased; 2) our DLRRPD achieves higher accuracies than other methods under different noise levels. Specifically, DLRRPD degrades slower than other methods with the percentage increasing, which means that our DLRRPD method is more robust than other methods for salt & pepper noise. Moreover, Fig.4 shows some original faces, noised faces and recovered faces.[2] We can see that DLRR and DLRRPD can recover images accurately because of projection learning. In particular, by utilizing class information, DLRRPD obtains the best recovery, proving the robustness of DLRRPD.

5.4. Parameter sensitivity and selection

As shown in model (9), the proposed DLRRPD contains three parameters, i.e., λ_1, λ_2 and λ_3, which balance the low-rank constraint, error and Laplacian rank constraint, respectively. In this section, we test the sensitivity of these three parameters by performing the proposed method with different combinations of three param-

[2]More results are shown in supplementary materials.

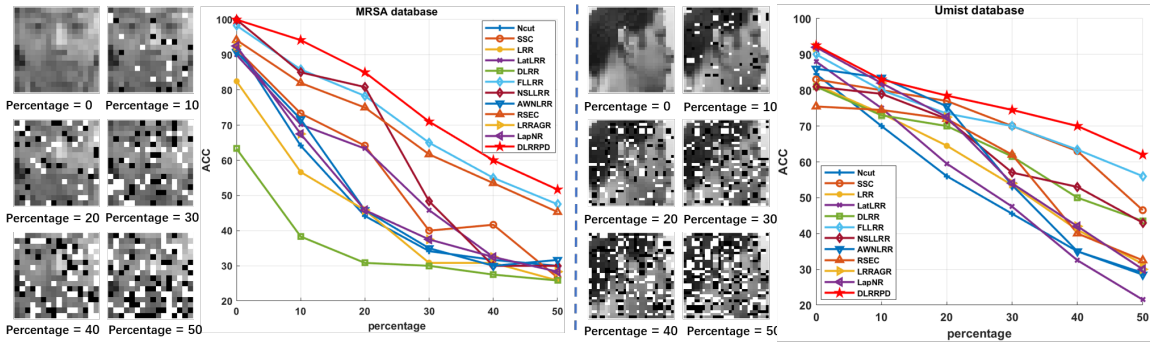


Figure 3. Clustering performance vs. varying percentage on MSRA (left) and UMIST (right) databases.

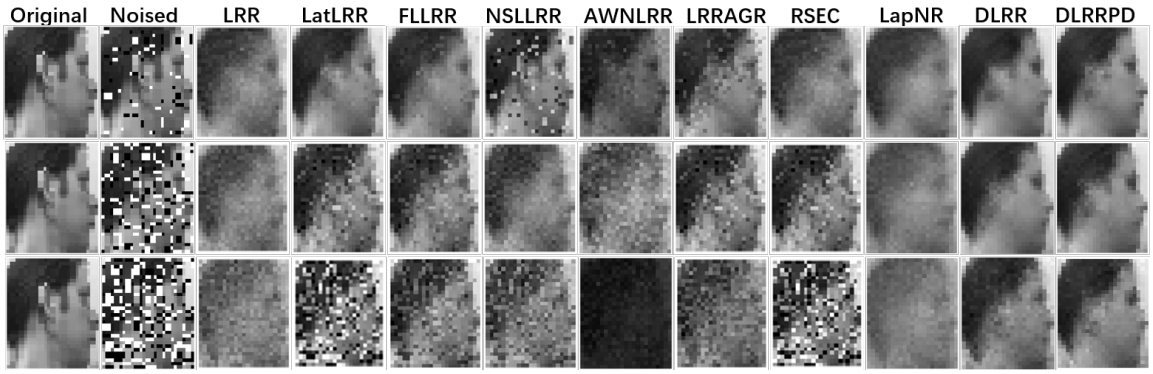


Figure 4. Results about recovering the face from noised images. The resulted images of each row are recovered from noised images with 10%, 30% and 50% salt & pepper noise, respectively.

eters, and each parameter is varied in a wide range, i.e., $[10^{-5}, 10^{-4}, 10^{-3}, ..., 10^{4}, 10^{5}]$. First, we fix $\lambda_3 = 10^{-1}$ and tune λ_1 and λ_2, thus the sensitivity of λ_2 and λ_3 as Fig 5(a). Then, λ_1 and λ_2 are fixed, and the influence of λ_3 is showed by performing the proposed method with different λ_3 on the Jaffe database. As shown in Fig 5(b), we can find that DLRRDP can deliver good results with $\lambda_3 \leq 10^{0}$. We can find that DLRRDP can deliver good results with $\lambda_1 \leq 10^{-2}$ and $\lambda_2 \leq 10^{-2}$. However, finding a suitable combination of parameters is still an open problem, and we just confirm that the most suitable parameters in our method can be found in a small range, i.e., $[10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}]$.

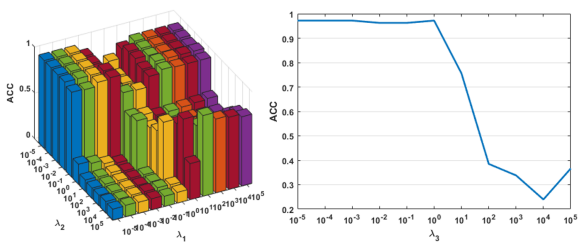


Figure 5. Parameter sensitivity analysis of DLRRPD on the Jaffe, where (a) fix λ_3 to tune $\lambda_1$and λ_2; (b) fix λ_1 and λ_2 to tune λ_3

6. Conclusion and future work

A novel self-representation learning model, i.e., Double Low-Rank Representation with Projection Distance penalty (DLRRPD), is proposed in this paper. It adopts a projection distance penalty to exploit more intrinsic structure, thus making the model preserve both the global and local structures. And then, a Laplacian rank constraint is employed to simultaneously guide the projection learning and graph learning, thus facilitating a more discriminative and robust graph. Moreover, using Frobenius norm instead of the widely used nuclear norm, we can obtain a more block-diagonal graph with lower complexity.

The effectiveness of our DLRRPD has been evaluated on several benchmark databases for data clustering. The clustering of the data with salt & pepper noise can also show the robustness of our method. In the future, we will try to extend this model to semi-supervised and weak supervised cases. Since labeled samples contain more prior information, this model is promising to handle some more complex real tasks.

Acknowledgments This work was supported by the National Key Research and Development of China (No. 2018AAA0102100), and the National Natural Science Foundation of China (No. U1936212).

References

[1] D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of Annual Symposium on Discrete Algorithms*, pages 1027–1035, 2007.

[2] X. Chen, W. Hong, F. Nie, J. Z. Huang, and L. Shen. Enhanced balanced min cut. *Int. J. Comput. Vis.*, 128(7):1982–1995, 2020.

[3] Y. Chen, L. Zhou, N. Bouguila, C. Wang, Y. Chen, and J. Du. BLOCK-DBSCAN: fast clustering for large scale data. *Pattern Recognition*, 109:107624, 2021.

[4] H. Ding, F. Yang, and M. Wang. On metric DBSCAN with low doubling dimension. In *IJCAI*, pages 3080–3086, 2020.

[5] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2790–2797, 2009.

[6] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2765–2781, 2013.

[7] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2765–2781, 2013.

[8] E. Esser. Applications of lagrangian-based alternating direction methods and connections to split bregman. *CAM report*, 9:31, 2009.

[9] K. Fan. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences of the United States of America*, 35(11):652, 1949.

[10] L. Fei, Y. Xu, X. Fang, and J. Yang. Low rank representation with adaptive distance penalty for semi-supervised subspace classification. *Pattern Recognition*, 67:252–262, 2017.

[11] J. Feng, Z. Lin, H. Xu, and S. Yan. Robust subspace segmentation with block-diagonal prior. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3818–3825, 2014.

[12] Z. Fu, Y. Zhao, D. Chang, and Y. Wang. A hierarchical weighted low-rank representation for image clustering and classification. *Pattern Recognition*, 112:107736, 2021.

[13] R. Glowinski and P. Le Tallec. *Augmented Lagrangian and operator-splitting methods in nonlinear mechanics*. SIAM, 1989.

[14] P. Ji, M. Salzmann, and H. Li. Efficient dense subspace clustering. In *IEEE Winter Conference on Applications of Computer Vision*, pages 461–468, 2014.

[15] X. Jiang, S. Wei, T. Liu, R. Zhao, Y. Zhao, and H. Huang. Blind image clustering for camera source identification via row-sparsity optimization. *IEEE Trans. Multimedia*, pages 1–1, 2020.

[16] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Adv. Neural Inform. Process. Syst.*, pages 612–620, 2011.

[17] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *Int. Conf. Mach. Learn.*, pages 663–670, 2010.

[18] G. Liu and S. Yan. Latent low-rank representation for subspace segmentation and feature extraction. In *Int. Conf. Comput. Vis.*, pages 1615–1622, 2011.

[19] J. Liu, Y. Chen, J. Zhang, and Z. Xu. Enhancing low-rank subspace clustering by manifold regularization. *IEEE Trans. Image Process.*, 23(9):4022–4030, 2014.

[20] C. Lu, H. Min, Z. Zhao, L. Zhu, D. Huang, and S. Yan. Robust and efficient subspace segmentation via least squares regression. In *Eur. Conf. Comput. Vis.*, pages 347–360, 2012.

[21] X. Peng, C. Lu, Z. Yi, and H. Tang. Connections between nuclear-norm and frobenius-norm-based representations. *IEEE Trans. Neural Networks Learn. Syst.*, 29(1):218–224, 2018.

[22] J. Shi and J. Malik. Normalized cuts and image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 731–737, 1997.

[23] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[24] K. Song, X. Yao, F. Nie, X. Li, and M. Xu. Weighted bilateral K-means algorithm for fast co-clustering and fast spectral clustering. *Pattern Recognition*, 109:107560, 2021.

[25] Y. Song and Y. Wu. Subspace clustering based on latent low rank representation with frobenius norm minimization. *Neurocomputing*, 275:2479–2489, 2018.

[26] Z. Tao, H. Liu, S. Li, Z. Ding, and Y. Fu. Robust spectral ensemble clustering via rank minimization. *ACM Trans. Knowl. Discov. Data*, 13(1):4:1–4:25, 2019.

[27] J. Wen, X. Fang, Y. Xu, C. Tian, and L. Fei. Low-rank representation with adaptive graph regularization. *Neural Networks*, 108:83–96, 2018.

[28] J. Wen, B. Zhang, Y. Xu, J. Yang, and N. Han. Adaptive weighted nonnegative low-rank representation. *Pattern Recognition*, 81:326–340, 2018.

[29] J. Wen, Z. Zhang, Y. Xu, B. Zhang, L. Fei, and G. Xie. Cdimc-net: Cognitive deep incomplete multi-view clustering network. In *IJCAI*, pages 3230–3236, 2020.

[30] J. Wen, Z. Zhang, Z. Zhang, Z. Wu, L. Fei, Y. Xu, and B. Zhang. Dimc-net: Deep incomplete multi-view clustering network. In *ACM Int. Conf. Multimedia*, pages 3753–3761, 2020.

[31] C. Wu and Y. Chen. Adaptive entropy weighted picture fuzzy clustering algorithm with spatial information for image segmentation. *Appl. Soft Comput.*, 86, 2020.

[32] M. Yin, S. Cai, and J. Gao. Robust face recognition via double low-rank matrix recovery for feature extraction. In *IEEE Int. Conf. Image Process.*, pages 3770–3774, 2013.

[33] M. Yin, J. Gao, and Z. Lin. Laplacian regularized low-rank representation and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(3):504–517, 2016.

[34] X. Zhang, Z. Zhu, Y. Zhao, and D. Chang. Learning a general assignment model for video analytics. *IEEE Trans. Circuit Syst. Video Technol.*, 28(10):3066–3076, 2018.

[35] X. Zhang, Z. Zhu, Y. Zhao, and D. Kong. Self-supervised deep low-rank assignment model for prototype selection. In *IJCAI*, pages 3141–3147, 2018.

[36] Y. Zhao, L. Chen, and C. L. P. Chen. Laplacian regularized nonnegative representation for clustering and dimensionality reduction. *IEEE Trans. Circuit Syst. Video Technol.*, pages 1–1, 2020.

[37] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2328–2335, 2012.