SSC-l_0: Sparse Subspace Clustering with the l_0 Inequality Constraint

Yangbo Wang[1,2], Jie Zhou[1,2,3(✉)], Qingshui Lin[4], Jianglin Lu[5], and Can Gao[2,3]

[1] National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, China
2070276074@email.szu.edu.cn, jie_jpu@163.com
[2] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
[3] SZU Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518060, China
[4] Basic Teaching Department, Liaoning Technical University, Huludao 125105, China
[5] College of Engineering, Northeastern University, Boston, MA 02115, USA
lu.jiang@northeastern.edu

Abstract. Self-expression learning methods often obtain a coefficient matrix to measure the similarity between pairs of samples. However, directly using all points to represent a fixed sample in a class under the self-expression framework may not be ideal, as points from other classes participate in the representing process. To alleviate this issue, this study attempts to achieve representation learning between points only coming from the same class. In practice, it is easier for data points from the same class to represent each other than that from different classes. So, when reconstructing a point, if the number of non-zero elements in the coefficient vector is limited, a model is more likely to select data points from the class where the reconstructed point lies to complete the reconstruction work. Based on this idea, we propose Sparse Subspace Clustering with the l_0 inequality constraint (SSC-l_0). In SSC-l_0, the l_0 inequality constraint determines the maximum number of non-zero elements in the coefficient vector, which helps SSC-l_0 to conduct representation learning among the points in the same class. After introducing the simplex constraint to ensure the translation invariance of the model, an optimization method concerning l_0 inequality constraint is formed to solve the proposed SSC-l_0, and its convergence is theoretically analyzed. Extensive experiments on well-known datasets demonstrate the superiority of SSC-l_0 compared to several state-of-the-art methods.

Keywords: Sparse subspace clustering · l_0 inequality constraint · simplex constraint · self-expression learning · graph learning

This work was supported by the National Natural Science Foundation of China (No. 62076164), Shenzhen Science and Technology Program (No. JCYJ20210324094601005), and Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515011861).

H. Lu et al. (Eds.): ACPR 2023, LNCS 14408, pp. 136–149, 2023.
https://doi.org/10.1007/978-3-031-47665-5_12

1 Introduction

As an important branch of the field of pattern recognition, clustering has been extensively developed in the past few decades. Commonly used clustering methods include prototype-based clustering [1], matrix factorization-based clustering [2], and graph-based clustering [3], etc. Among them, graph-based clustering has attracted much attention because it can exploit the geometrical structure information of the data [4,5].

As a typical graph-based clustering method, Spectral Clustering (SC) [6] often achieves superior clustering performance when handling datasets with high dimensions. In the spectral clustering method, a low-dimensional representation of data is first constructed by utilizing the predetermined similarity matrix. Subsequently, SC yields the discrete clustering result by calling the spectral rotation method [7] or using k-Means methods [8] based on the relaxed spectral solutions. So, the similarity matrix plays a critical role in spectral clustering. Similarity matrix construction methods often include two types, point-pairwise distance-based methods and self-expression learning methods-based. The typical point-pairwise distance-based methods include the Gaussian kernel method, adaptive neighbors [9,10], typicality-aware adaptive graph [11], and so on. These methods construct sparse similarity matrices by learning sample neighborhood structure information.

The self-expression learning methods, including the Sparse Subspace Clustering (SSC) [12], Low-Rank Representation (LRR) [13], Block-Diagonal Representation (BDR) [14], and their extensions [15,16], have been reported and achieved desired performance. Among them, SSC [12] aims to group data drawn from a union of multiple linear subspaces. When the subspaces from which data are drawn are independent, SSC can obtain desired performance by learning a sparse affinity matrix. LRR [13] minimizes the rank of the coefficient matrix to recover the row space of the data. BDR [14] proposes a novel regularizer for directly pursuing a coefficient matrix with the block diagonal property. Iteratively Reweighted Least Squares (IRLS) [15] solves a joint low-rank and sparse minimization problem. Least Squares Regression (LSR) [16] minimizes the F-norm of the coefficient matrix to achieve that the correlated samples have similar coefficient vectors. Simplex Sparse Representation (SSR) [17] introduces the simplex constraint to ensure the translation invariance of the model.

The above self-expression learning-based methods seek sparse or low-rank representation coefficient matrices to measure the similarity between pairs of samples. However, in these methods, the way that directly uses all sample points to represent a fixed sample may not be ideal since points from other classes also participate in the representing process. Therefore, the generated similarity matrix is not reliable and affects the performance of downstream tasks.

To alleviate this issue, we propose Sparse Subspace Clustering with the l_0 inequality constraint (SSC-l_0). SSC-l_0 aims to achieve a representation learning mechanism in which only the points coming from the same class are expected to be used for representing each sample in this class.

The main contributions of this work are listed as follows:

1. A new self-expression learning method, named Sparse Subspace Clustering with the l_0 inequality constraint (SSC-l_0), is proposed. In SSC-l_0, the l_0 inequality constraint constrains the number of non-zero elements in the coefficient vector, which helps SSC-l_0 to conduct representation learning among the points in the same class.
2. By introducing the simplex constraint to ensure the translation invariance of the model, an optimization method concerning l_0 inequality constraint is presented to solve the proposed SSC-l_0, and its convergence is theoretically analyzed. Since the l_0 inequality constraint problem is difficult to be solved, the proposed optimization method has the potential to be widely used sparse learning models.
3. Extensive experiments on benchmark datasets demonstrate the superiority of SSC-l_0.

The rest of this paper is organized as follows. In Sect. 2, we introduce some notations and the related methods. In Sect. 3, we elaborate on the proposed SSC-l_0 and its corresponding optimization algorithm. Experimental results are reported in Sect. 4. Finally, Sect. 5 concludes the paper.

2 Related Works

2.1 Notations

In this study, the bold uppercase letters and lowercase italic letters are used to stand for the matrices and scalars, respectively, such as $\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; \cdots; \mathbf{x}_n] \in \mathbb{R}^{n \times d}$ represents a data matrix, where d and n denote the dimensions and the number of samples, respectively. The bold lowercase letters stand for vectors, such as $\mathbf{x}_i \in \mathbb{R}^{1 \times d}$ is a sample point. $Tr(\mathbf{A})$ and $\mathbf{A}^T$ stand for the trace and the transpose of matrix $\mathbf{A}$, respectively. $\mathbf{A} \geq \mathbf{0}$ means that each element $a_{ij} \geq 0$ in $\mathbf{A}$.

2.2 Sparse Subspace Clustering

Let $\left\{\mathbf{x}_j \in \mathbb{R}^{1 \times d}\right\}_{j=1}^{n}$ be a set of data points drawn from a union of c independent linear subspaces $\{\mathcal{S}_i\}_{i=1}^{c}$. Let n_i is the number of data points drawn from the subspace $\mathcal{S}_i$, d_i is dimension of $\mathcal{S}_i$, and $\mathbf{Y}_i$ a data matrix corresponding to subspace $\mathcal{S}_i$. If $n_i \geq d_i$, the points that are drawn from the subspace $\mathcal{S}_i$ is *self-expressive* [12]. This means that if $\mathbf{x}$ is a new data point in $\mathcal{S}_i$, then it can be represented as a linear combination of d_i points in the same subspace. Denoting $\mathbf{z}_i \in \mathbb{R}^{1 \times n_i}$ is a fragment of the coefficient vector, and it corresponds to the data matrix $\mathbf{Y}_i$. If we let $\mathbf{X}$ be a data matrix with proper permutation, i.e., $\mathbf{X} = [\mathbf{Y}_1, \mathbf{Y}_1, \cdots, \mathbf{Y}_c]$, for any data point $\mathbf{x}$, there exists a vector $\mathbf{s} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_c] \in \mathbb{R}^{1 \times n}$ such that $\mathbf{x} = \mathbf{s}\mathbf{X}$, where $\mathbf{z}_i \neq \mathbf{0}$ and $\mathbf{z}_j = \mathbf{0}$ for all $j \neq i$ ($\mathbf{z}_i \in \mathbb{R}^{1 \times n_i}$ is the fragment corresponding to points in the same subspace as $\mathbf{x}$). And such a $\mathbf{s}$ can be sought by solving the following problem:

$$\min_{\mathbf{s}} ||\mathbf{s}||_0, \quad \text{s.t.} \quad \mathbf{x} = \mathbf{s}\mathbf{X}, \tag{1}$$

where $||\mathbf{s}||_0$ represents the number of non-zero elements in row vector $\mathbf{s}$. This can be proved by the following Theorem 1.

Theorem 1. *Denoting* $\mathbf{s}^*$ *is the optimal solution to the problem* (1). *Let* $\mathbf{X}$ *be a data matrix with proper permutation and* $\mathbf{x} \in \mathcal{S}_i$, *then* $\mathbf{s}^* = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_c] \in \mathbb{R}^{1\times n}$, *where* $\mathbf{z}_i \neq \mathbf{0}$ *and* $\mathbf{z}_j = \mathbf{0}$ *for all* $j \neq i$.

Proof. Let $\mathbf{s}^* = \mathbf{s}^*_\uparrow + \mathbf{s}^*_\downarrow$, where $\mathbf{s}^*_\uparrow = [\mathbf{0}, \mathbf{0}, \cdots, \mathbf{z}_i, \cdots \mathbf{0}] \in \mathbb{R}^{1\times n}$ and $\mathbf{s}^*_\downarrow = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_{i-1}, \mathbf{0}, \mathbf{z}_{i+1}, \cdots, \mathbf{z}_c] \in \mathbb{R}^{1\times n}$. It has Theorem 1 holds when we show that $\mathbf{s}^*_\downarrow = \mathbf{0}$.

Let $\mathbf{s}^*_\downarrow \neq \mathbf{0}$, since $\mathbf{s}^* = \mathbf{s}^*_\uparrow + \mathbf{s}^*_\downarrow$, it has $\mathbf{x} = \mathbf{s}^*\mathbf{X} = (\mathbf{s}^*_\uparrow + \mathbf{s}^*_\downarrow)\mathbf{X}$. According to the independence assumption [12], it has $\mathbf{x} \in \mathcal{S}_i$, $\mathbf{s}^*_\uparrow\mathbf{X} \in \mathcal{S}_i$, and $\mathbf{s}^*_\downarrow\mathbf{X} \notin \mathcal{S}_i$. Thus, we have $\mathbf{s}^*_\downarrow\mathbf{X} = \mathbf{0}$. This implies that

$$\mathbf{x} = \mathbf{s}^*\mathbf{X} = \mathbf{s}^*_\uparrow\mathbf{X} \tag{2}$$

which means that $\mathbf{s}^*_\uparrow$ is also a solution of Eq. (1). And we have $||\mathbf{s}^*_\uparrow||_0 < ||\mathbf{s}^*_\uparrow + \mathbf{s}^*_\downarrow||_0 = ||\mathbf{s}^*||_0$, which conflicts with $\mathbf{s}^*$ being the optimal solution to Eq. (1). So we have $\mathbf{s}^*_\downarrow = \mathbf{0}$. The Theorem 1 holds. ■

The matrix format of Eq. (1) is as follows:

$$\min_{\mathbf{S}} \sum_{i=1}^{n} ||\mathbf{s}_i||_0, \quad \text{s.t.} \quad \mathbf{X} = \mathbf{S}\mathbf{X}, \tag{3}$$

where $\mathbf{S} = [\mathbf{s}_1; \mathbf{s}_2; \cdots; \mathbf{s}_n] \in \mathbb{R}^{n\times n}$ is a coefficient matrix, and $\mathbf{s}_i \in \mathbb{R}^{1\times n}$ is a coefficient vector corresponding to the point $\mathbf{x}_i$.

Theorem 1 demonstrates that when the subspaces are independent, any sample $\mathbf{x}_i$ can be represented by the samples coming from the same class. And $s_{ij} = 0$ if samples $\mathbf{x}_j$ and $\mathbf{x}_i$ come from different classes.

According to Theorem 1, when $\mathbf{X}$ is properly permuted and subspaces are independent, the coefficient matrix $\mathbf{S}$ has a block diagonal structure, which can be used to improve clustering performance.

3 Sparse Subspace Clustering with the l_0 Inequality Constraint

3.1 Motivation of SSC-l_0

Most self-expression learning methods seek a sparse or low-rank representation coefficient matrix and minimize the difference between the original samples and their reconstructed estimations. The objective functions of SSC, LRR, and their extensions [14–16] can be generalized as follows:

$$\min_{\mathbf{S}} ||\mathbf{X} - \mathbf{S}\mathbf{X}||_F^2 + \gamma\mathcal{R}(\mathbf{S}), \tag{4}$$

where $\mathbf{S}$ is a reconstruction coefficient matrix, $||\mathbf{X} - \mathbf{SX}||_F^2$ indicates the self-expression reconstruction loss. $\mathcal{R}(\mathbf{S})$ is a regularization term, which guarantees that $\mathbf{S}$ is sparse or low-rank, or has a strict block diagonal structure, etc.

Numerous studies have shown that self-expression learning methods achieve impressive performance in clustering tasks [13]. However, in Eq. (4), directly using all points to represent a fixed sample in a class may degrade the quality of the learned similarity matrix, as points from other classes also participate in the representing process. To alleviate this issue, we attempt to achieve representation learning between points of the same class. However, this work is not easy because the ground truth labels are not provided beforehand. According to Theorem 1, the permuted coefficient matrix has a block diagonal structure when the subspaces from which the data are drawn are independent. Thus, such block diagonal structures can be used to guide the similarity measurements among samples. However, real data often do not qualify the subspace independence assumption [18].

In practice, it is easier for data points from the same class to represent each other than that from different classes. Thus, when reconstructing a point, if the number of non-zero elements in the coefficient vector is limited, a model is more likely to select data points from the class where the reconstructed point lies to complete the reconstruction process. This means that introducing the l_0 inequality constraint may improve the performance of the self-expression learning model.

3.2 Objective Function of SSC-l_0

According to the discussion in Sect. 3.1, introducing the l_0 inequality constraint may improve the performance of the self-expression learning model. The proposed objective function is as follows:

$$\begin{aligned} \min_{\mathbf{S}} \quad & ||\mathbf{X} - \mathbf{SX}||_F^2 + \gamma ||\mathbf{S}||_F^2, \\ \text{s.t.} \quad & \mathbf{S} \geq 0, \mathbf{S}\mathbf{1}^T = \mathbf{1}^T, ||\mathbf{s}_i||_0 \leq k \ (i = 1, 2, \cdots, n), \end{aligned} \tag{5}$$

where $\mathbf{S} = [\mathbf{s}_1; \mathbf{s}_2; \cdots; \mathbf{s}_n] \in \mathbb{R}^{n \times n}$ is a reconstruction coefficient matrix, γ is a non-negative balance parameter, and k is a constant. The constraint $||\mathbf{s}_i||_0 \leq k$ means that the max number of non-zero elements in the coefficient vector $\mathbf{s}_i$ is k.

By introducing the l_0 inequality constraint $||\mathbf{s}_i||_0 \leq k$, in the first term of Eq. (5), each sample is represented as a linear combination of the samples that are more likely from the class that the represented sample belongs to. Since the coefficient matrix is non-negative, the elements in the coefficient matrix can reflect the similarity between the sample pairs. The l_2 norm $||\mathbf{s}_i||_2^2$ can be utilized to avoid overfitting on $\mathbf{s}_i$. The constraint $\mathbf{s}_i \mathbf{1}^T = 1$ can be used to ensure the translation invariance [17].

3.3 Optimization of SSC-l_0

First, because of the difficulty of solving model Eq. (5), we turn to optimize the following model:

$$\begin{aligned} \min_{\mathbf{S},\mathbf{M}} \ & ||\mathbf{X}-\mathbf{M}\mathbf{X}||_F^2+\alpha||\mathbf{M}-\mathbf{S}||_F^2+\gamma||\mathbf{S}||_F^2, \\ \text{s.t.} \ & \mathbf{S}\geq\mathbf{0},\mathbf{S}\mathbf{1}^T=\mathbf{1}^T,||\mathbf{s}_i||_0\leq k \ (i=1,2,\cdots,n), \end{aligned} \tag{6}$$

where α is a non-negative parameter. Problem (6) can be transformed into the problem (5) when α is large enough. And the problem (6) can be minimized using the following iterative algorithm:

Updating M. When $\mathbf{S}$ is fixed, problem (6) becomes:

$$\min_{\mathbf{M}} \ ||\mathbf{X}-\mathbf{M}\mathbf{X}||_F^2+\alpha||\mathbf{M}-\mathbf{S}||_F^2. \tag{7}$$

Setting the derivative of Eq. (7) with respect to $\mathbf{M}$ to zero, it has:

$$\mathbf{M}=\left(\mathbf{X}\mathbf{X}^T+\alpha\mathbf{S}\right)\left(\mathbf{X}\mathbf{X}^T+\alpha\mathbf{I}\right)^{-1}. \tag{8}$$

Updating S. When $\mathbf{M}$ is fixed, problem (6) becomes:

$$\begin{aligned} \min_{\mathbf{S}} \ & \alpha||\mathbf{M}-\mathbf{S}||_F^2+\gamma||\mathbf{S}||_F^2, \\ \text{s.t.} \ & \mathbf{S}\geq\mathbf{0},\mathbf{S}\mathbf{1}^T=\mathbf{1}^T,||\mathbf{s}_i||_0\leq k \ (i=1,2,\cdots,n). \end{aligned} \tag{9}$$

Since problem (9) is independent of each $\mathbf{s}_i$, problem (9) can be converted into subproblems:

$$\begin{aligned} & \min_{\mathbf{s}_i} \ \alpha||\mathbf{m}_i-\mathbf{s}_i||_2^2+\gamma||\mathbf{s}_i||_2^2, \\ & \text{s.t.} \ \ \mathbf{s}_i\geq\mathbf{0},\mathbf{s}_i\mathbf{1}^T=1,||\mathbf{s}_i||_0\leq k, \\ \Leftrightarrow & \min_{\mathbf{s}_i} \ \mathcal{O}(\mathbf{s}_i)=||\mathbf{s}_i-\mathbf{u}_i||_2^2, \\ & \text{s.t.} \ \ \mathbf{s}_i\geq\mathbf{0},\mathbf{s}_i\mathbf{1}^T=1,||\mathbf{s}_i||_0\leq k, \end{aligned} \tag{10}$$

where $\mathbf{u}_i=\frac{\alpha}{\alpha+\gamma}\mathbf{m}_i$.

Obviously, Eq. (10) seems to be an NP-hard problem. Some methods sparsify the vector $\mathbf{s}_i$ obtained from all samples [19]. However, the convergence of this solving method cannot be guaranteed theoretically. Here, we attempt to identify an equivalent problem to Eq. (10) that is readily solvable.

Denoting $\check{\Gamma}$ is a set that includes the indices corresponding to the first k largest elements in $\mathbf{u}_i$. Then, the optimal solution to the following problem (11) is also an optimal solution to the problem (10). To prove this, we propose further the following definitions, lemmas, and theorems.

$$\min_{\mathbf{s}_i\geq\mathbf{0},\mathbf{s}_i\mathbf{1}^T=1,s_{ij}=0 \ if j\notin\check{\Gamma}(i)} \left\{\mathcal{O}_1=\sum_{j\in\check{\Gamma}(i)}\left(s_{ij}-u_{ij}\right)^2\right\}. \tag{11}$$

Definition 1. *Let* $\mathbf{s}_i$ *satisfy the constraints of Eq.* (10)*. If there is at least one non-zero element* s_{ij} $(s_{ij} > 0)$ *in* $\mathbf{s}_i$ *that satisfies* $j \notin \check{\Gamma}(i)$*, then* $\mathbf{s}_i$ *is called a* $\check{\zeta}$ *solution of Eq.* (10)*.*

Definition 2. $\Pi\left(\mathbf{s}_i, \check{\Gamma}(i)\right) = \sum_{j=1}^{n} \mathcal{A}(s_{ij}, j)$*, where* $\mathcal{A}(s_{ij}, j) = 1$ *if* $s_{ij} > 0$ *and* $j \in \check{\Gamma}(i)$*; otherwise* $\mathcal{A}(s_{ij}, j) = 0$*.*

Lemma 1. *For any* $\check{\zeta}$ *solution* $\ddot{\mathbf{s}}_i$ *of Eq.* (10)*, there exists another solution* $\widetilde{\mathbf{s}}_i$ *of Eq.* (10) *that satisfies* $\Pi\left(\widetilde{\mathbf{s}}_i, \check{\Gamma}(i)\right) = \Pi\left(\ddot{\mathbf{s}}_i, \check{\Gamma}(i)\right) + 1$ *such that* $\mathcal{O}(\widetilde{\mathbf{s}}_i) \leq \mathcal{O}(\ddot{\mathbf{s}}_i)$*.*

Proof. Let $\tau \notin \check{\Gamma}(i)$ and $\epsilon \in \check{\Gamma}(i)$ be indices that satisfy $\ddot{\mathbf{s}}_i = [\cdots, \ddot{s}_{i\tau}, \cdots, \ddot{s}_{i\epsilon}, \cdots]$ and $\widetilde{\mathbf{s}}_i = [\cdots, \widetilde{s}_{i\tau}, \cdots, \widetilde{s}_{i\epsilon}, \cdots]$, where $u_{i\tau} \leq u_{i\epsilon}$, $\ddot{s}_{i\epsilon} = 0$, $\widetilde{s}_{i\tau} = 0$, $\ddot{s}_{i\tau} = \widetilde{s}_{i\epsilon} > 0$, and $\ddot{s}_{ij} = \widetilde{s}_{ij}$ for any $j \neq \tau$ and $j \neq \epsilon$.

Thus, one has:

$$\left(\ddot{s}_{i\tau} - u_{i\tau}\right)^2 + \left(\ddot{s}_{i\epsilon} - u_{i\epsilon}\right)^2 \geq \left(\widetilde{s}_{i\tau} - u_{i\tau}\right)^2 + \left(\widetilde{s}_{i\epsilon} - u_{i\epsilon}\right)^2, \tag{12}$$

which means $\mathcal{O}(\widetilde{\mathbf{s}}_i) \leq \mathcal{O}(\ddot{\mathbf{s}}_i)$. Thus, Lemma 1 holds. ■

Lemma 2. *For any* $\check{\zeta}$ *solution* $\ddot{\mathbf{s}}_i$ *of Eq.* (10)*, there exists another non-*$\check{\zeta}$ *solution* $\widetilde{\mathbf{s}}_i$ *of Eq.* (10) *such that* $\mathcal{O}(\widetilde{\mathbf{s}}_i) \leq \mathcal{O}(\ddot{\mathbf{s}}_i)$*.*

Proof. Lemma 2 holds when Lemma 1 holds. ■

Lemma 3. *Let* $\mathbf{s}_i^*$ *be the optimal solution of Eq.* (10)*, there exists a non-*$\check{\zeta}$ *solution* $\dddot{\mathbf{s}}_i^*$ *that satisfy* $\mathcal{O}(\dddot{\mathbf{s}}_i^*) = \mathcal{O}(\mathbf{s}_i^*)$*.*

Proof. Obviously, Lemma 3 holds when Lemma 2 holds. ■

Lemma 3 demonstrates that if $\mathbf{s}_i^*$ is an optimal solution of Eq. (10) and it is a $\check{\zeta}$ solution, then there exists a non-$\check{\zeta}$ solution $\dddot{\mathbf{s}}_i^*$ which is also an optimal solution to the problem (10).

Theorem 2. *Let* $\bar{\mathbf{s}}_i^*$ *be the optimal solution to the problem* (11)*, then* $\bar{\mathbf{s}}_i^*$ *is also an optimal solution to the problem* (10)*.*

Proof. Let $\mathbf{s}_i^*$ be the optimal solution of Eq. (10), according to Lemma 3, there exists a non-$\check{\zeta}$ solution $\dddot{\mathbf{s}}_i^*$ that satisfy $\mathcal{O}(\dddot{\mathbf{s}}_i^*) = \mathcal{O}(\mathbf{s}_i^*)$. And it has:

$$\mathcal{O}(\dddot{\mathbf{s}}_i^*) = \check{\sigma} + \mathcal{O}_1(\dddot{\mathbf{s}}_i^*), \tag{13}$$

where $\check{\sigma} = \sum_{j=1, j\notin\check{\Gamma}(i)}^{n} (0 - u_{ij})^2$, and $\dddot{\mathbf{s}}_i^* \geq \mathbf{0}$, $\dddot{\mathbf{s}}_i^* \mathbf{1}^T = 1$, $||\dddot{\mathbf{s}}_i^*||_0 \leq k$. Considering $\check{\Gamma}(i)$ is fixed, $\check{\sigma}$ is a constant. And we have $\mathcal{O}(\bar{\mathbf{s}}_i^*) = \check{\sigma} + \mathcal{O}_1(\bar{\mathbf{s}}_i^*)$. Since $\bar{\mathbf{s}}_i^*$ is the optimal solution to the problem (11), we have $\mathcal{O}_1(\bar{\mathbf{s}}_i^*) \leq \mathcal{O}_1(\dddot{\mathbf{s}}_i^*)$. So, $\mathcal{O}(\bar{\mathbf{s}}_i^*) \leq \mathcal{O}(\dddot{\mathbf{s}}_i^*) = \mathcal{O}(\mathbf{s}_i^*)$. Thus, Theorem 2 holds. ■

The problem (11) can be transformed into a vector format. We can solve it easily by utilizing the KKT conditions and the Newton method [17].

Algorithm 1 Sparse Subspace Clustering with the l_0 inequality constraint (SSC-l_0)

Input: Data matrix $\mathbf{X} \in \mathbf{R}^{n \times d}$
Parameter: α, γ, and k.
Output: $\mathbf{S} \in \mathbf{R}^{n \times n}$.

1: Initialize $\mathbf{S}$.
2: Initialize $\mathbf{M}$ with formula (8).
3: **while** Iterations $t \leq T$ and not converge **do**
4: Update $\check{\Gamma}(i)$ for any i;
5: Update $\mathbf{S}$ by solving the problem (11);
6: Update $\mathbf{M}$ with formula (8).
7: **end while**
8: **return** $\mathbf{S}$.

The algorithm of SSC-l_0 is summarized in Algorithm 1. Assume that Algorithm 1 iterates at most T times. The time cost of updating $\mathbf{M}$ is $O(T\left(dn^2 + n^3\right))$, where n is the number of samples and d is the number of features. The time cost of solving the problem (11) is $O(T\,(n\ k\ log\ k))$ [20], where k is the maximum number of non-zero elements in the coefficient vector. Given that $k \ll n$, the overall time complexity of SSC-l_0 is $O(n^3)$.

After a non-negative coefficient matrix $\mathbf{S}$ is learned, spectral clustering can be executed based on the produced $\mathbf{S}$.

In Algorithm 1, the optimal solutions for $\mathbf{M}$ and $\mathbf{S}$ can be obtained by solving problems (8) and (11). Then,

$$J_{SSC-l_0}\left(\mathbf{M}^{(t)}, \mathbf{S}^{(t)}\right) \leq J_{SSC-l_0}\left(\mathbf{M}^{(t-1)}, \mathbf{S}^{(t)}\right) \leq J_{SSC-l_0}\left(\mathbf{M}^{(t-1)}, \mathbf{S}^{(t-1)}\right), \tag{14}$$

where $t-1$ and t represent the $(t-1)$th and tth iteration, respectively. Inequality Eq. (14) indicates that the objective function values of SSC-l_0 decrease monotonically, i.e., SSC-l_0 is convergent.

4 Experiments

In this study, all experiments are conducted on a personal computer with i5-9500 CPU @3.00 GHz and 8 GB RAM. The codes are implemented in MATLAB R2021a 64 bit.

4.1 Datasets and Comparison Methods

Ten image datasets are utilized for experiments, including five face image datasets (Yale, Jaffe [21], ORL1024 [22], PIE [23] and FERET [24]), and four

handwritten digit image datasets (Binary[1], Semeion[2], Digit[3], and Minist[4]), and one palm image dataset (Palm) [9]. The details of these datasets are shown in Table 1.

Table 1. The Benchmark Datasets.

Datasets	Samples	Size	Clusters	Datasets	Samples	Size	Clusters
Yale	165	32 * 32	15	Binary	1404	20 * 16	36
Jaffe	213	26 * 26	10	Semeion	1593	16 * 16	10
ORL1024	400	32 * 32	40	Digit	1797	8 * 8	10
FERET	1400	40 * 40	200	Minist	4000	28 * 28	10
PIE	1632	32 * 32	68	Palm	2000	16 * 16	100

Seven clustering methods are selected for comparison, including Least Squares Regression (LSR1) [16], Sparse Subspace Clustering (SSC) [12], Low-Rank Representation (LRR) [13], Iteratively Reweighted Least Squares (IRLS) [15], Block-Diagonal Representation (BDR) [14], Clustering with Typicality-aware Adaptive Graph (CTAG) [11], and the Simplex Sparse Representation (SSR) [17]. Please see Sect. 1 for more details on these methods.

SSC-l_0 and seven other methods vary parameters in the range of $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$. Spectral clustering is applied to coefficient matrices, followed by post-processing with k-means. To reduce sensitivity to initialization, k-means is repeated 50 times, and the reported result is the average of the 20 trials with the lowest loss.

4.2 Clustering Analysis

Clustering Performance. Table 2 shows the ACC values obtained by the different methods on the selected benchmark datasets. In Table 2, the best results are exhibited in bold and the second-best results are marked in brackets. The average results of each method over all selected datasets are listed as the last row in Table 2. From Table 2, we have the following observations:

1. LRR and IRLS rank second and third, respectively, in terms of average ACC values among the eight methods. This can be attributed to their ability to restore the row space of the data by learning a coefficient matrix with a low-rank structure. This characteristic aids in enhancing their robustness and improving their clustering ability.
2. SSR exhibits the lowest average ACC value, which can be attributed to the absence of an effective regularization term.

[1] https://cs.nyu.edu/~roweis/data/.
[2] https://archive.ics.uci.edu/ml/index.php.
[3] http://www.escience.cn/people/fpnie/index.html.
[4] http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html.

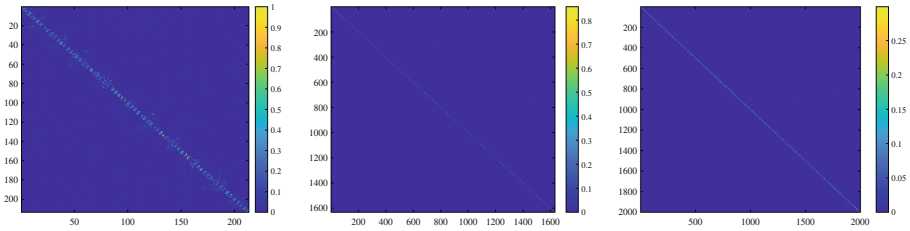
Table 2. ACC Values of the SSC-l_0 and the Selected Methods on Benchmark Datasets. (%)

Datasets	LSR1	SSC	LLR	IRLS	BDR	SSR	CTAG	SSC-l_0
Yale	49.33	46.79	(50.61)	47.06	48.55	48.09	49.67	**54.97**
Jaffe	96.74	94.37	95.02	93.57	**100.00**	95.96	(97.18)	**100.00**
ORL	64.71	63.35	(65.53)	65.13	59.66	57.20	62.94	**70.64**
FERET	32.75	38.30	33.54	34.19	(38.62)	34.83	29.60	**39.50**
PIE	70.64	48.92	**71.24**	69.61	56.87	50.89	31.95	(70.94)
Binary	36.05	25.77	35.68	35.08	6.12	5.46	(43.54)	**45.17**
Semeion	57.29	56.57	58.26	(58.46)	57.61	57.74	54.11	**60.18**
Digit	47.74	**70.60**	64.15	59.55	58.13	60.52	61.62	(69.79)
Minist	37.49	(39.41)	37.04	38.30	37.58	31.01	38.34	**44.10**
Palm	80.32	67.35	(82.03)	80.41	66.98	66.08	82.01	**91.13**
Avg.	57.31	55.14	(59.31)	58.14	53.01	50.78	55.10	**64.64**

3. SSC-l_0 achieves the best ACC values across all selected datasets except PIE and Digit. The reason is that SSC-l_0 tends to represent each sample as a linear combination of the samples from the same class under the l_0 inequality constraint. However, the other seven methods lack this ability. In general, data points coming from the same class are easier to represent each other. When reconstructing a point, if the number of non-zero elements in the coefficient vector is limited, it is more likely to select data points from the same class to complete the corresponding reconstruction work. Therefore, the l_0 inequality constraint can improve the performance of the SSC-l_0.

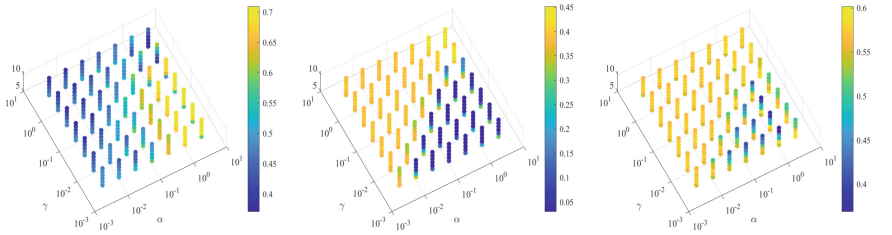
Visualization of the Similarity Matrix. Figure 1 visualizes the coefficient matrix obtained by SSC-l_0 on datasets including Jaffe, PIE, and Palm. In Fig. 1, the coefficient matrix exhibits a block-diagonal structure, indicating that similar samples in SSC-l_0 tend to belong to the same class. This characteristic enhances the clustering performance of SSC-l_0. The reason for showing the block-diagonal structure is that if the number of non-zero elements in the coefficient vector is limited, when reconstructing a point, it is more likely to select a point of the class to which the reconstruction point belongs.

Parameter Sensitivity. SSC-l_0 has three parameters: α, γ, and k. Figure 2 illustrates the impact of these parameters on SSC-l_0 performance across different datasets. The results indicate that SSC-l_0 is sensitive to the values of α, γ, and k. To obtain optimal performance, a grid search method is recommended for parameter selection in SSC-l_0.



(a) (b) (c)

Fig. 1. The coefficient matrix obtained by SSC-l_0 on the different datasets. (a) Jaffe. (b) PIE. (c) Palm.



(a) (b) (c)

Fig. 2. ACC values of SSC-l_0 on different datasets when varying parameters α, γ, and k. (a) PIE. (b) Binary. (c) Semeion. The color in the figure reflects the ACC values.

Convergence Analysis. Figure 3 shows the convergence curves of SSC-l_0 on datasets Yale, ORL1024, and FERET. From Fig. 3, the objective function values of SSC-l_0 decrease monotonically and SSC-l_0 converges after 100 iterations.

4.3 Robustness Analysis

To compare the robustness of the eight methods, the following experiments are designed:

1. In the ORL2116 dataset, 20% of the samples are randomly selected and added salt and pepper noise. The noisy densities are 0.1, 0.2, 0.3, 0.4, and 0.5. Figure 4(a) illustrates the original ORL2116 dataset and noisy ORL2116 datasets. Figure 5(a) presents the performances of eight methods on these noisy datasets.
2. In the ORL1024 dataset, 20% of the samples are randomly selected and added random noisy blocks. The sizes of the noisy blocks were 5×5, 10×10, 15×15, 20×20, and 25×25 pixels. Figure 4(b) displays the original ORL1024 dataset and the corresponding noisy ORL1024 datasets. The performances of different methods on these noisy datasets are presented in Fig. 5(b).
3. In the Jaffe dataset, 20% of the samples are randomly selected and added random noisy blocks. The sizes of the noisy blocks were $4 * 4$, $8 * 8$, $12 * 12$,

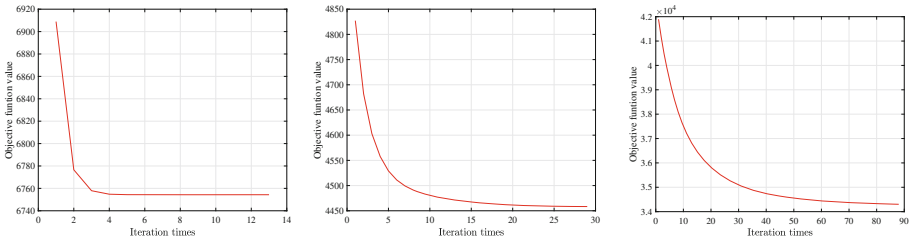


Fig. 3. Convergence curves of SSC-l_0 on different datasets. (a) Yale.(b) ORL1024. (c) FERET.

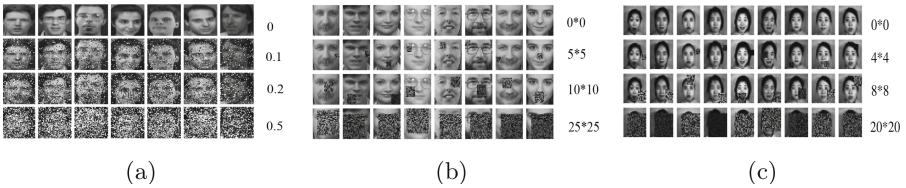
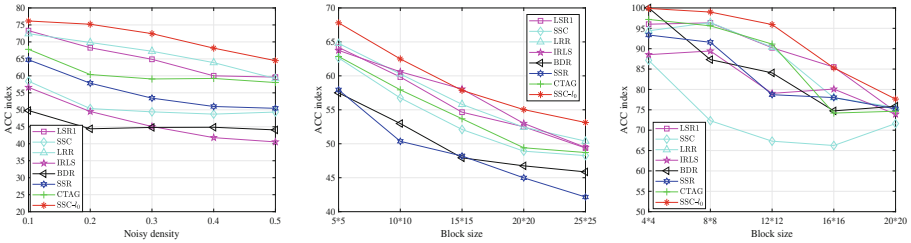


Fig. 4. Original dataset and noisy datasets. (a) ORL2116. (b) ORL1024. (c) Jaffe.

16 * 16 and 20 * 20 pixels. Figure 4(c) displays the original Jaafe dataset and the corresponding noisy Jaffe datasets. The performances of different methods on these noisy datasets are presented in Fig. 5(c).

Form Fig. 5, although the performance of the eight methods degrades as increasing the noise levels, SSC-l_0 achieve almost the best results no matter which noise levels are involved, which demonstrates further that SSC-l_0 has high robustness. The reason is as follows. Clean points have better representation ability compared to noisy points. When the number of non-zero elements in the coefficient vector is limited, clean points are more likely to be used for reconstruction. Thus, constraining the number of non-zero elements can improve model robustness. SSC and SSR use l_1 regularization and l_1 equation constraint to induce sparsity in the coefficient matrix, but they do not directly constrain the number of non-zero elements in the coefficient vector. In contrast, SSC-l_0 can precisely control the number of non-zero elements, leading to better performance.



(a) (b) (c)

Fig. 5. Performance of different clustering methods on the noisy datasets with different noise levels. (a) Noisy ORL2116 datasets with different noise densities. (b) Noisy ORL1024 datasets with different sizes noisy blocks. (c) Noisy Jaffe datasets with different sizes noisy blocks.

5 Conclusions

Considering data points coming from the same class are easier to represent each other and when the number of non-zero elements in the coefficient vector is limited, a model is more likely to select data points from the same class to complete corresponding reconstruction work, we propose Sparse Subspace Clustering with the l_0 inequality constraint (SSC-l_0) to conduct representation learning among the points in the same class. By introducing the simplex constraint, an optimization method concerning l_0 inequality constraint is proposed, and its convergence is also theoretically analyzed. Since the l_0 inequality constraint problem is difficult to be solved, the proposed optimization method can be widely used in lots of sparse learning models. Extensive experiments demonstrate the superiority of SSC-l_0. Establishing a one-step clustering method based on SSC-l_0 may further improve the performance of the proposed clustering model, which is our next work.

References

1. Zhou, J., Pedrycz, W., Wan, J., et al.: Low-rank linear embedding for robust clustering. IEEE Trans. Knowl. Data Eng. **35**(5), 5060–5075 (2022)
2. Li, X., Chen, M., Wang, Q.: Discrimination-aware projected matrix factorization. IEEE Trans. Knowl. Data Eng. **32**, 809–814 (2019)
3. Lu, J., Wang, H., Zhou, J., et al.: Low-rank adaptive graph embedding for unsupervised feature extraction. Pattern Recogn. **113**, 107758 (2021)
4. Wang, Y., Gao, C., Zhou, J.: Geometrical structure preservation joint with self-expression maintenance for adaptive graph learning. Neurocomputing **501**, 436–450 (2022)
5. Lu, J., Lai, Z., Wang, H., et al.: Generalized embedding regression: a framework for supervised feature extraction. IEEE Trans. Neural Netw. Learn. Syst. **33**, 185–199 (2022)

6. Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: analysis and an algorithm. In: Advances in Neural Information Processing Systems, Vancouver, pp. 849–856. MIT (2002)
7. Stella, X., Shi, J.: Multiclass spectral clustering. In: Proceedings of the 9th IEEE International Conference on Computer Vision, Nice, pp. 313–319. IEEE Computer Society (2003)
8. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297 (1967)
9. Nie, F., Wang, X., Huang, H.: Clustering and projected clustering with adaptive neighbors. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, pp. 977–986. ACM (2014)
10. Lu, J., Lin, J., Lai, Z., et al.: Target redirected regression with dynamic neighborhood structure. Inf. Sci. **544**, 564–584 (2021)
11. Zhou, J., Gao, C., Wang, X., et al.: Typicality-aware adaptive similarity matrix for unsupervised learning. IEEE Trans. Neural Netw. Learn. Syst. (2023). https://doi.org/10.1109/TNNLS.2023.3243914
12. Elhamifar, E., Vidal, R.: Sparse subspace clustering: algorithm, theory, and applications. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 2765–2781 (2013)
13. Liu, G., Lin, Z., Yan, S., et al.: Robust recovery of subspace structures by low-rank representation. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 171–184 (2013)
14. Lu, C., Feng, J., Lin, Z., et al.: Subspace clustering by block diagonal representation. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 487–501 (2019)
15. Lu, C., Lin, Z., Yan, S.: Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization. IEEE Trans. Image Process. **24**, 646–654 (2015)
16. Lu, C.-Y., Min, H., Zhao, Z.-Q., Zhu, L., Huang, D.-S., Yan, S.: Robust and efficient subspace segmentation via least squares regression. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7578, pp. 347–360. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33786-4_26
17. Huang, J., Nie, F., Huang, H.: A new simplex sparse learning model to measure data similarity for clustering. In: Proceedings of the 24th International Conference on Artificial Intelligence, Buenos Aires, pp. 3569–3575. AAAI (2015)
18. Maggu, J., Majumdar, A., Chouzenoux, E.: Transformed subspace clustering. IEEE Trans. Knowl. Data Eng. **33**, 1796–1801 (2020)
19. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering, in: Advances in Neural Information Processing Systems, Montreal, pp. 1601–1608 (2004)
20. Condat, L.: Fast projection onto the simplex and the l-1 ball. Math. Program. **158**, 575–585 (2016)
21. Lyons, M., Budynek, J., Akamatsu, S.: Automatic classification of single facial images. IEEE Trans. Pattern Anal. Mach. Intell. **21**, 1357–1362 (1999)
22. Samaria, F., Harter, A.: Parameterisation of a stochastic model for human face identification. In: Proceedings of 1994 IEEE Workshop on Applications of Computer Vision, Sarasota, pp. 138–142. IEEE (1994)
23. Sim, T., Baker, S., Bsat, M.: The CMU Pose, illumination and expression database of human faces. Carnegie Mellon University Technical Report CMU-RI-TR-OI-02 (2001)
24. Phillips, P., Moon, H., Rizvi, S., et al.: The FERET evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1090–1104 (2000)