mathematics

MDPI

Review

A Survey on High-Dimensional Subspace Clustering

Wentao Qu [1], Xianchao Xiu [2], Huangyue Chen [3] and Lingchen Kong [1,*]

[1] School of Mathematics and Statistics, Beijing Jiaotong University, Beijing 100044, China
[2] School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China
[3] Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
* Correspondence: lchkong@bjtu.edu.cn

Abstract: With the rapid development of science and technology, high-dimensional data have been widely used in various fields. Due to the complex characteristics of high-dimensional data, it is usually distributed in the union of several low-dimensional subspaces. In the past several decades, subspace clustering (SC) methods have been widely studied as they can restore the underlying subspace of high-dimensional data and perform fast clustering with the help of the data self-expressiveness property. The SC methods aim to construct an affinity matrix by the self-representation coefficient of high-dimensional data and then obtain the clustering results using the spectral clustering method. The key is how to design a self-expressiveness model that can reveal the real subspace structure of data. In this survey, we focus on the development of SC methods in the past two decades and present a new classification criterion to divide them into three categories based on the purpose of clustering, i.e., low-rank sparse SC, local structure preserving SC, and kernel SC. We further divide them into subcategories according to the strategy of constructing the representation coefficient. In addition, the applications of SC methods in face recognition, motion segmentation, handwritten digits recognition, and speech emotion recognition are introduced. Finally, we have discussed several interesting and meaningful future research directions.

Keywords: high-dimensional data; kernel learning; machine learning; sparse optimization; subspace clustering

MSC: 62H30; 65K05; 90C30

check for updates

Citation: Qu, W.; Xiu, X.; Chen H.; Kong, L. A Survey on High-Dimensional Subspace Clustering. *Mathematics* **2023**, *11*, 436. https://doi.org/10.3390/math11020436

Academic Editor: Alfonso Mateos Caballero

Received: 13 December 2022
Revised: 10 January 2023
Accepted: 11 January 2023
Published: 13 January 2023

1. Introduction

Clustering has been one of the most important research topics in both statistical machine learning and data mining, which aims to segment unlabeled samples into several disjoint groups (clusters) by maximizing the difference among different groups and minimizing the difference in the same group. Clustering originated from the anthropological research by Driver and Kroeber in 1932 [1] and was then introduced into psychology [2,3]. Over the last few decades, various clustering methods have been proposed, such as *k*-means clustering [4–7], spectral clustering [8–11], model-based clustering [12–14], hierarchical clustering [15–17], and online clustering [18,19].

With the rapid development of information technology and the advent of the big data era, data have shown explosive growth and presents complex characteristics such as high dimensionality, nonlinearity, and incompleteness. Although traditional clustering methods have achieved excellent results in low-dimensional data mining tasks, they often encounter serious bottlenecks in high-dimensional data mining tasks, which cannot meet the sparsity of high-dimensional data and avoid the impact of the "curse of dimensionality" [20,21]. Since the inherent characteristics of high-dimensional data, it usually lies in a union of low-dimensional structures instead of being uniformly distributed across the whole space. For instance, Figure 1 shows a set of sample points in $\mathbb{R}^3$ from a two-dimensional subspace (the plane $\mathcal{S}_1$) and two one-dimensional subspaces (the lines $\mathcal{S}_2$ and $\mathcal{S}_3$). It is worth pointing out that it is difficult to analyze the data as a whole, while the natural properties of this data can be better reflected in its underlying low-dimensional subspaces (plane or line). In fact, in many practical problems, the data in each class can be well represented by low-dimensional subspaces. For example, the face images of a person under different lighting conditions can be distributed in a nine-dimensional subspace [22,23], and the motion trajectories of rigidly moving objects in a video belong to different subspaces whose dimensions do not exceed three [22]. To explore high-dimensional data in a low-dimensional space, subspace clustering arises at the opportune time [24]. The subspace clustering aims to search for the underlying subspaces in dimension space or data space to obtain corresponding subspaces and clustering results. Over the past several decades, subspace clustering has been widely used in face recognition [25–27], motion segmentation [28–30], image processing [31–33], speech emotion recognition [34,35], social network [36,37], and other fields.

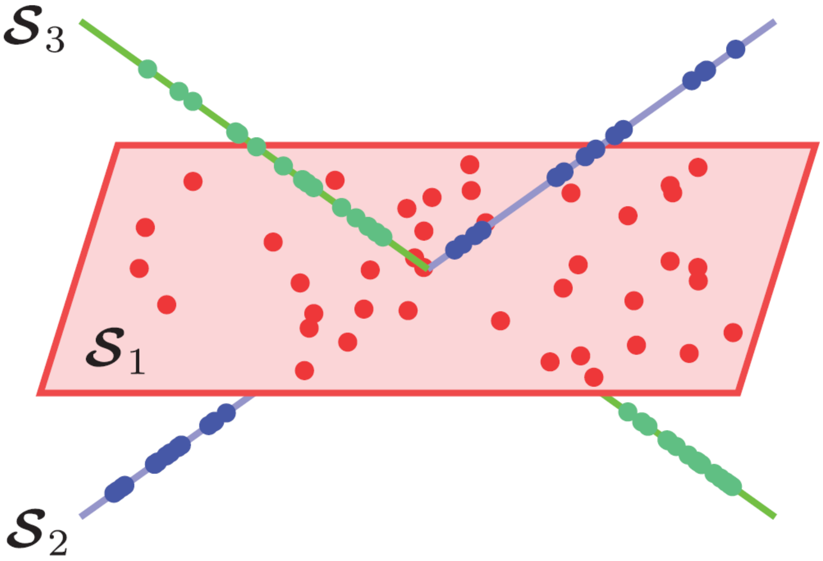


Figure 1. A set of sample points in $\mathbb{R}^3$ from a union of three subspaces: a plane and two lines.

Definition 1 (Subspace clustering, SC [38])**.** *Given a set of samples* $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$*, where d and n are the number of features and samples, respectively. Assume that X are drawn from a union of k subspaces* $\{\mathcal{S}_i\}_{i=1}^k$ *with unknown dimensionality* $\{d_i\}_{i=1}^k$*. Subspace clustering aims to segment the samples according to the underlying subspaces which they are drawn from.*

Due to the advantages of SC, a number of SC methods have been proposed during the past decades. Figure 2 shows the development trend of SC publications and their citations over the past two decades. It can be seen that SC shows a growth trend in general, and the growth trend has dramatically increased since 2013. According to their mechanisms of representing the subspaces, the existing SC methods can be roughly divided into four main categories: algebraic methods, iterative methods, statistical methods, and spectral clustering-based methods. These four categories of SC methods are briefly discussed below (see [39] for details). Algebraic methods include matrix factorization-based methods [40,41] and generalized principal component analysis (GPCA) [42,43], but algebraic methods are usually highly sensitive to noise and outliers, and the complexity of GPCA increases exponentially with the number and dimensions of subspaces. Iterative methods, such as k-subspaces [44–46], alternatively assign the data points to subspaces and fit the subspaces to the corresponding cluster. However, iterative methods usually need to know the number of subspaces and the dimensions of subspaces in advance and are sensitive to the initial point. Statistical methods can be regarded as the probabilistic description based on iterative methods, including probabilistic PCA (PPCA) [47,48], mixture of probabilistic PCA (MPPCA) [47,49], random sample consensus (RANSAC) [50,51], etc. Unfortunately, statistical methods usually require that the dimensions of subspaces are known and equal, and their complexity increases with the number of subspaces. Most recently, spectral clustering-based methods have played a dominant role in SC problems due to their easy implementation and insensitivity to data corruptions and initial points. In the following, we will focus on spectral clustering-based methods.

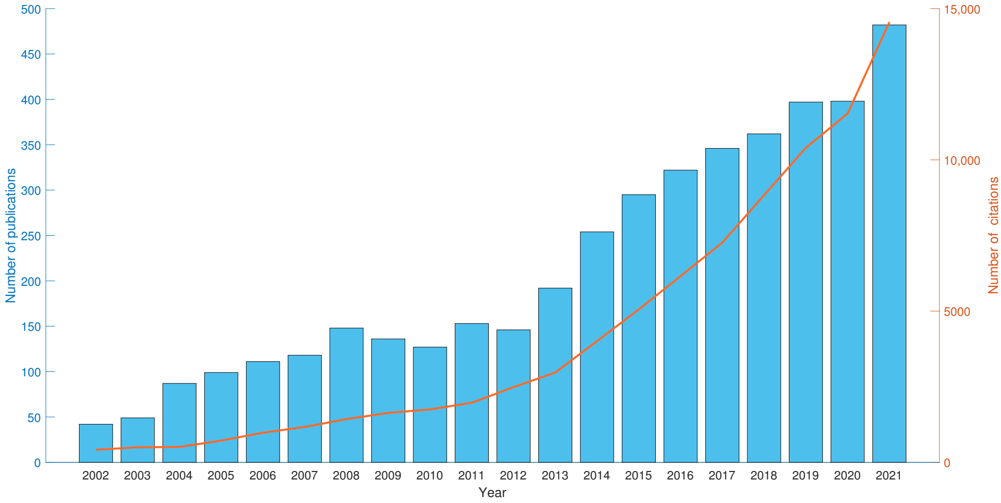


Figure 2. The development trend of SC in the past two decades.

The spectral clustering-based methods usually perform the following two steps. (1) The representation coefficients are learned by a subspace representation model, which is used to construct the affinity matrix. (2) The final clustering result is obtained by performing spectral clustering on the affinity matrix. The basic framework of the spectral clustering-based SC methods is shown in Figure 3. It is worth pointing out that the representation coefficients learned in the first step play a vital role in the clustering effectiveness. However, owing to the complexity and diversity of the inherent unknown structure of the real data, some assumptions must be introduced on the data distributions, such as sparse or low-rank assumptions in the representation learning, thus leading to different ways to construct or learn representation coefficients.

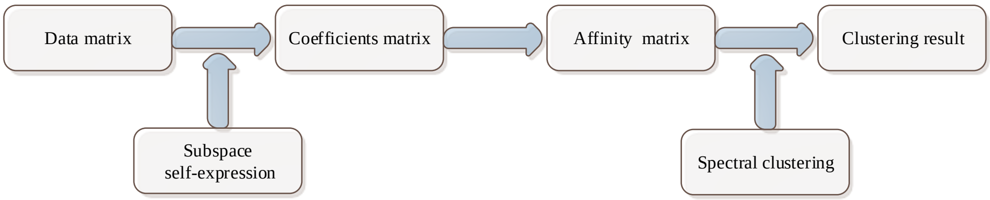


Figure 3. The basic framework of the spectral clustering-based SC methods.

This survey provides a comprehensive review of the recent development of spectral clustering-based SC methods (for convenience, hereinafter referred to as SC). The main contributions of this survey are summarized as follows:

- We systematically review the recent research progress of SC methods and summarize them into three categories and further subcategories according to the strategy of constructing representation coefficient; see Figure 4.
- We introduce the applications of SC methods in face recognition, motion segmentation, handwritten digits recognition, and speech emotion recognition and present several common benchmark datasets.

- We provide some potential research directions which can promote the development of this field.

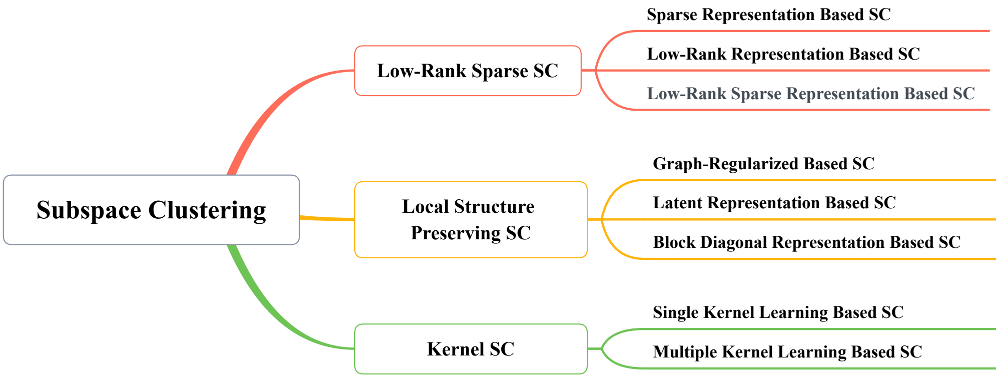

Subspace Clustering
Low-Rank Sparse SC
Sparse Representation Based SC
Low-Rank Representation Based SC
Low-Rank Sparse Representation Based SC
Local Structure Preserving SC
Graph-Regularized Based SC
Latent Representation Based SC
Block Diagonal Representation Based SC
Kernel SC
Single Kernel Learning Based SC
Multiple Kernel Learning Based SC

Figure 4. The categories of SC methods.

The rest of this survey is organized as follows. Sections 2–4 introduce low-rank sparse SC methods, local structure preserving SC methods, and kernel SC methods, respectively; see Table 1. Section 5 introduces the applications of SC methods and presents several common benchmark datasets. Finally, the future research directions and conclusions are presented in Sections 6 and 7, respectively.

Table 1. A brief summary of SC methods.

SC Methods	Subcategories	Brief Description
Low-Rank Sparse SC	SR-Based SC [25,52–61]	Achieve subspace feature selection through sparse representation.
	LRR-Based SC [38,62–72]	Capture the global structure of data through low-rank representation.
	LRSR-Based SC [73–83]	Capture the global structure of data and achieve subspace feature selection by combining low-rank representation with sparse representation.
Local Structure Preserving SC	GR-Based SC [84–91]	Capture the geometric information by combining graph regularization with data self-representation.
	LatR-Based SC [92–98]	Capture latent representation of data by joint low-rank/sparse recovery and salient feature extraction
	BDR-Based SC [99–104]	Pursue the block diagonal structure of coefficient matrix directly by designing special regularization.
Kernel SC	SKL-Based SC [55,105–107]	Be able to process the nonlinear structure of the data.
	MKL-Based SC [108–116]	Be able to process the nonlinear structure of the data and learn a consensus kernel function adaptively.

Remark: SR indicates sparse representation; LRR indicates low-rank representation; LRSR indicates low-rank sparse representation; GR indicates graph-regularized; LatR indicates latent representation; BDR indicates block diagonal representation; SKL indicates single kernel learning; MKL indicates multiple kernel learning.

The symbols used throughout this survey are defined at the end of this section. In this survey, all matrices are denoted by capital letters, such as X, Y; all vectors are represented by boldface letters, such as $\mathbf{x}, \mathbf{y}$; all scalars are represented by lowercase letters, such as x, y. Let $\mathbb{R}^n$ and $\mathbb{R}^{d \times n}$ be the set of n-dimensional vectors and $d \times n$ matrices, respectively. For any matrix $X \in \mathbb{R}^{d\times n}$, let X_{ij} and $\mathbf{x}_j$ denote its (i,j)th element and jth column, respectively. The element-wise l_1 norm is defined as $\|X\|_1 = \sum_{i=1}^{d}\sum_{j=1}^{n}|X_{ij}|$; $\|X\|_F = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{n}|X_{ij}|^2}$ denotes the Frobenius norm; $\|X\|_{2,q} = \left(\sum_{j=1}^{n}\left(\sum_{i=1}^{d}|X_{ij}|^2\right)^{q/2}\right)^{1/q}$ denotes the $l_{2,q}$ norm with $q > 0$; the l_0 pseudo norm is defined as $\|X\|_0 = \#\{X_{ij} \neq 0, \forall i,j\}$, i.e., the total number of non-zero elements in X. Let $r = \mathrm{rank}(X)$ represent the rank of X. Let $\sigma_1(X) \geq \ldots \geq \sigma_r(X) \geq \sigma_{r+1}(X) = \ldots = \sigma_d(X) = 0$ are the singular value of X. $\|X\|_{K_{k,p}} = \left(\sum_{i=1}^{k}\sigma_i(X)^p\right)^{1/p}$ with $p > 0$ and $k \in \{1,2,\ldots,d\}$ denotes the Ky Fan p-k norm. The particular case of the Ky Fan norm with $k = r$ is called the Schatten p norm, i.e., $\|X\|_{S_p} = (\sum_{i=1}^{r}\sigma_i(X)^p)^{1/p}$, when $p = 1$ yields the nuclear norm $\|X\|_*$. In particular, the Schatten 0 pseudo norm is $\|X\|_{S_0} = \#\{\sigma_i(X) \neq 0, \forall i\}$. Let $\mathrm{tr}(X)$ be the trace of matrix X. We denote $\mathrm{diag}(X)$ as a vector whose ith element is the ith diagonal element of matrix X, while $\mathrm{Diag}(\mathbf{x})$ as a matrix whose ith diagonal element is the ith element of vector $\mathbf{x}$. $\odot$ denotes the Hadamard product, i.e., the element-wise product between two matrices. I and $\mathbf{1}$ represent the identity matrix and all one vector, respectively. If the matrix X is positive semi-definite, we denote $X \succeq 0$. If all the elements of X are nonnegative, we denote $X \geq 0$.

2. Low-Rank Sparse SC

In this section, we will review some classical low-rank sparse SC methods. Before starting the main content of this section, we first introduce some basic definitions of SC.

Definition 2 (Independent subspace)**.** *A collection of subspaces $\{\mathcal{S}_i\}_{i=1}^{k}$ is said to be independent, if $dim(\oplus_{i=1}^{k}\mathcal{S}_i) = \sum_{i=1}^{k} dim(\mathcal{S}_i)$, where $dim(\mathcal{S}_i)$ is the dimension of subspace $\mathcal{S}_i$ and $\oplus$ represents the direct sum operation.*

Definition 3 (Disjoint subspace)**.** *A collection of subspaces $\{\mathcal{S}_i\}_{i=1}^{k}$ is said to be disjoint, if $dim(\mathcal{S}_i \oplus \mathcal{S}_j) = dim(\mathcal{S}_i) + dim(\mathcal{S}_j)$ $(\forall i \neq j)$, but $dim(\oplus_{i=1}^{k}\mathcal{S}_i) \neq \sum_{i=1}^{k} dim(\mathcal{S}_i)$.*

It is evident that, when the number of subspaces is more than two, the independent subspace condition is much stronger than the disjoint condition, as the independent subspaces must be disjoint subspaces while the converse is not necessarily true.

Definition 4 (Self-expressiveness property)**.** *Each sample in the union of subspaces can be effectively represented as a linear or affine combination of other samples, i.e., $X = XC$ where $X \in \mathbb{R}^{d\times n}$ is the given data, and $C \in \mathbb{R}^{n\times n}$ is the representation coefficient matrix.*

It is worth pointing out that the representation coefficient matrix C can reflect the similarity between samples to a certain extent which will be used to build the affinity matrix. Therefore, the representation coefficient matrix plays a crucial role in spectral clustering-based SC methods. Ideally, the representation coefficient matrix C should satisfy the following diagonal property.

Definition 5 (Block diagonal property, BDP)**.** *Given a sample matrix X drawn from a union of k disjoint subspaces $\{\mathcal{S}_i\}_{i=1}^{k}$, assume that the samples in the same subspace are aligned together, i.e., $X = [X_1,\ldots,X_K]$ where X_i is a collection of n_i samples drawn from the subspace $\mathcal{S}_i$. The representation coefficient matrix C is said to satisfy the block diagonal property, if*

$$C = \begin{bmatrix} C_1 & 0 & \cdots & 0 \\ 0 & C_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & C_k \end{bmatrix} \quad (1)$$

with $C_i \in \mathbb{R}^{n_i \times n_i}$.

Note that, if the representation coefficient matrix C satisfies the block diagonal property (1), the following conclusions can be drawn: (1) the number of blocks represents the number of subspaces; (2) the size of each block represents the dimension of the corresponding subspace; and (3) the samples of the same block belong to the same subspace.

In practical applications, data often contain noise or outliers, so the data can be expressed as $X = YC + E$, where Y is the clean data or dictionary, generally taken as the observation X itself, and E represents the data corruptions, such as error, noise or outliers. In general, different prior constraints are imposed on the representation coefficient matrix C to obtain an ideal structure. Thus, the self-expressiveness based SC methods could be unified as solving the following optimization problems. Therefore, the self-expressiveness based SC methods can be uniformly described as the following optimization problems:

$$\min_{C,E} \; F(C) + \lambda R(E) \quad \text{s.t.} \;\; X = YC + E, C \in \Omega, \tag{2}$$

where $F(C)$ is a penalty term (or regularization term), which is used to constrain the representation coefficient matrix C to keep an ideal structure, and Ω is the constraint set of C. $\lambda > 0$ is a trade-off parameter. $R(E)$ is an error term describing the difference between the real data and the represented data. Different norms are selected to measure the error according to different noise distributions, which requires some prior knowledge or assumptions. Once the representation coefficient matrix C is determined, the affine matrix A for spectral clustering can be obtained through C. In the existing literature, a commonly used formula is $A = \frac{|C|+|C|^T}{2}$.

2.1. Sparse Representation Based SC

In recent years, sparse representation has become a research hotspot in machine learning, computer vision, applied mathematics, and so on. Sparsity refers to using as few bases as possible to represent data, that is, using as few non-zero representation coefficients as possible. The position of the non-zero representation coefficients indicates that the data are located in the subspace spanned by its corresponding basis. Some classical sparse SC methods are summarized in Table 2.

Table 2. The comparison of classical sparse SC methods.

Methods	**Year**	$\boldsymbol{F(C)}$	$\boldsymbol{R(E)}$	$\boldsymbol{\Omega}$	**Theoretical Results**
SSC [25,52]	2009	$\|C\|_1$	$\|E\|_1$	$\{C \mid \text{diag}(C) = \mathbf{0}\}$	BDP for independent subspaces
SSQP [53]	2011	$\|C^TC\|_1$	$\|E\|_F^2$	$\{C \mid C \geq 0, \text{diag}(C) = \mathbf{0}\}$	BDP for orthogonal subspaces
W-SSC [54]	2012	$\|W \odot C\|_1$	$\|E\|_p$	$\{C \mid \text{diag}(C) = \mathbf{0}\}$	-
GS-graph [55]	2015	$\Gamma_\mu^g(C)$	$\|E\|_1$	-	Grouping effect
l_0-SSC [56]	2016	$\|C\|_0$	$\|E\|_F^2$	$\{C \mid \text{diag}(C) = \mathbf{0}\}$	-
SSCNA [57]	2019	$\Phi_{AL}(C)$	$\|E\|_1$	$\{C \mid C^T\mathbf{1} = \mathbf{1}, \text{diag}(C) = \mathbf{0}\}$	-
SSC-SLp [58]	2019	$\|C\|_p^p$	$\|E\|_F^2$	$\{C \mid C^T\mathbf{1} = \mathbf{1}, \text{diag}(C) = \mathbf{0}\}$	-
SRSSC [59]	2019	$\|W \odot C\|_1 + \mu\|Q \odot C\|_1$	$\|E\|_1$	$\{C \mid C^T\mathbf{1} = \mathbf{1}, \text{diag}(C) = \mathbf{0}\}$	-
SSC+E [60]	2020	$\sum_{i=1}^n \sum_{j=1}^n C_{ij} \ln C_{ij}$	$\|E\|_F^2$	$\{C \mid C^T = C, C \geq 0, \text{diag}(C) = \mathbf{0}\}$	-
AdaptiveSSC [61]	2020	$\|C/W\|_1$	$\|E\|_F^2$	$\{C \mid \text{diag}(C) = \mathbf{0}\}$	-

Remark: - indicates that this item does not exist.

In 2009, Elhamifar and Vidal [25,52] introduced compressed sensing techniques to subspace segmentation and proposed sparse subspace clustering (SSC) to represent each data point as a sparse linear combination of other data points. SSC can be formulated as the following optimization problem:

$$\min_{C,E} \ \|C\|_1 + \lambda \|E\|_1 \quad \text{s.t.} \quad X = XC + E, \mathrm{diag}(C) = \mathbf{0}, \tag{3}$$

where the constraint $\mathrm{diag}(C) = \mathbf{0}$ is used to avoid the trivial solution of representing a data point as a linear combination of itself, i.e., $C = I$. In optimization, l_1-norm is usually regarded as the best convex relaxation of l_0 norm, which can be used to generate sparse solutions due to its nonsmoothness (see Figure 5 for the geometric interpretation). Note that model (3) is a convex optimization, which can be effectively solved by the convex programming algorithms, such as the alternating direction method of multipliers (ADMM) [117] and the semi-smooth Newton augmented Lagrangian method (SSNAL) [118].

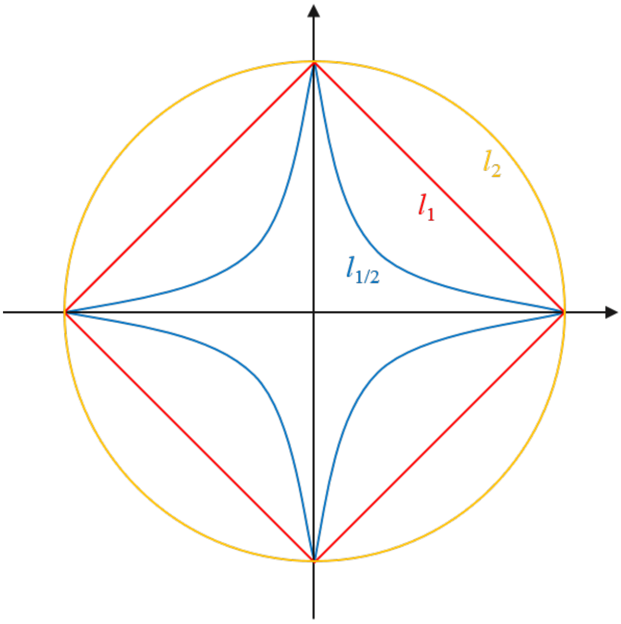


Figure 5. The geometric interpretation of different norms in $\mathbb{R}^2$.

For the SCC method (3) without noise, it has been proved that the representation coefficient matrix obtained by (3) has the block diagonal structure (1) under the assumption of independent subspaces. Elhamifar and Vidal [119] extended the condition of independent subspaces to disjoint subspaces and gave the theoretical boundary for the division of disjoint subspaces. Soltanolkotabi and Candès [120] further extended this condition to the case even though underlying subspaces overlap through geometric analysis. However, these analyses are based on the assumption that the data points are exactly lying in the subspace, which is difficult to satisfy in practice. Therefore, Soltanolkotabi et al. [121] generalized the SCC by using geometric functional analysis and proved that this method can accurately recover the underlying subspaces under the minimum requirements on the direction of the subspace and the number of samples in each subspace. Wang et al. [122] theoretically analyzed the SSC with noise and proved that the modified version of SSC can correctly identify the underlying subspace, even for noisy data. However, the SCC method still has some drawbacks. Since SSC seeks a sparse representation of each data individually, it may not accurately capture the global structure, which may lead to over-segmentation [123]. On the other hand, if the correlation between a group of data points is very high, SSC tends to select only one randomly, which will ignore the correlation structure of data from the same subspace [124].

Since SCC essentially solves a constrained lasso regression problem which is a nonsmooth problem, the computational cost of solving the problem is higher than general smooth problems. Consequently, Wang et al. [53] proposed the following subspace segmentation method via quadratic programming (SSQP)

$$\min_{C,E} \quad \|C^T C\|_1 + \lambda \|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, C \geq 0, \text{diag}(C) = \mathbf{0}, \tag{4}$$

where $\|C^T C\|_1$ is equivalent to $\mathbf{1}^T C^T C \mathbf{1}$, which can be used to enforce the block diagonal structure of C. When the subspaces are orthogonal, it has been proved that the representation coefficient matrix obtained by SSQP (4) without noise has a block diagonal structure.

Pham et al. [54] present a weighted sparse subspace clustering (W-SSC) method by embedding weights into the sparse formulation, which can be formulated as

$$\min_{C,E} \quad \|W \odot C\|_1 + \lambda \|E\|_p \quad \text{s.t.} \quad X = XC + E, \text{diag}(C) = \mathbf{0}, \tag{5}$$

where W is a non-negative weighting matrix, which can take the inverse of a similarity measure between points, such as $W_{ij} = \exp\{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/\sigma\}^{-1}$. $\|E\|_p$ is the penalty of noise E, which can be taken as l_1 norm, $l_{2,1}$ norm or the squared Frobenius norm, respectively, for the random corruptions, sample-specific corruptions and Gaussian white noise. It is worth pointing out that the SSC model (3) is a special case of (5) with $W = I$ and $\|E\|_p = \|E\|_1$. W-SSC uses the spatial relationship among the data to weight the representation coefficient matrix, which improves the performance of SSC significantly. However, W-SSC may fall into a local minimization easily. For this reason, Wang et al. [59] proposed a structural reweight sparse subspace clustering method (SRSSC) by introducing the structural information into W-SSC. If the clustering labels of data $\mathbf{x}_i$ and $\mathbf{x}_j$ are the same, structural matrix $Q_{ij} = 1$; otherwise, $Q_{ij} = 0$. Furthermore, it has been proved that the global optimal solution can be found to be better by considering the structure information. However, there are three parameters needed to be adjusted in SRSSC, which is troublesome and time-consuming. In addition, both of the reweight SSC methods are sensitive to data noise.

Inspired by the traditional graph construction methods, Fang et al. [55] proposed a group sparse graph (GS-graph) subspace clustering method by constructing an informative graph using auto-grouped l_1 regularization, which can be written as

$$\min_{C,E} \quad \Gamma_\mu^g(C) + \lambda \|E\|_1 \quad \text{s.t.} \quad X = XC + E, \tag{6}$$

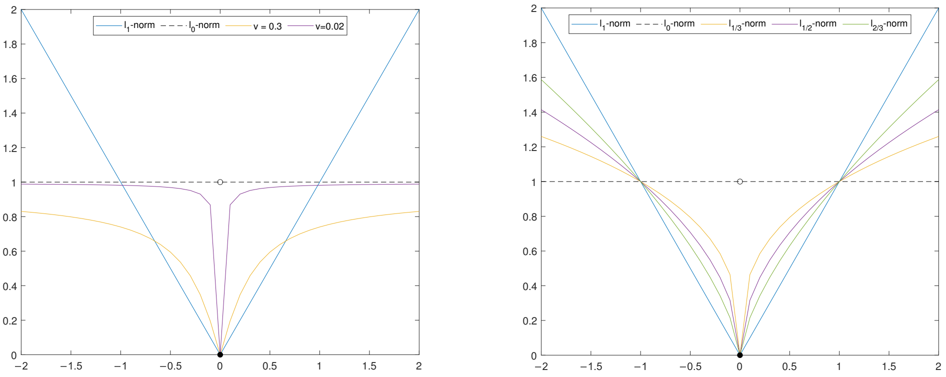
where $\Gamma_\mu^g(C) = \|C\|_1 + \sum_i \sum_{j<k} \max\{|C_{ij}|, |C_{ik}|\}$ is the auto-grouped octagonal shrinkage and clustering algorithm for regression (OSCAR) which can be regarded as a weighted combination of the l_1-norm and a pairwise l_∞-norm [125]. Based on this design, GS-graph can maintain both the sparsity and locality of data simultaneously and is insensitive to noise. Moreover, it has proved that GS-graph (6) has the grouping effect.

Most of the existing sparsity-based SC methods are based on the l_1 norm of the representation coefficient. Although l_1 norm is regarded as the best convex relaxation of l_0 pseudo norm and has achieved many achievements, it does not fully reflect the properties of l_0 norm. Therefore, Yang et al. [56] directly studied the following l_0-induced sparse subspace clustering (l_0-SSC)

$$\min_{C,E} \quad \|C\|_0 + \lambda \|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, \text{diag}(C) = \mathbf{0}. \tag{7}$$

It has proved that the l_0-SSC can give subspace-sparse representation almost surely for arbitrary distinct underlying subspaces. Although the l_0-SSC model (7) is a nonconvex and nonsmooth optimization problem, it can be effectively solved by the proximal gradient descent (PGD) algorithm [56,126] and ADMM [127].

Although the l_0 pseudo norm regularization has been widely regarded as the origin model of sparsity and achieved superior clustering performance, its computational cost is expensive, and the convergence of algorithm is hardly guaranteed. A better strategy is to use some nonconvex functions to approximate l_0 pseudo norm, such as arctangent–logarithmic function [57] and l_p-norm [58]. Dong et al. [57] proposed a sparse subspace clustering via the nonconvex approximation (SSCNA). Specifically, SSCNA used the arctangent–logarithmic function $\Phi_{AL}(\mathbf{x}) = \frac{2}{\pi}\sum_{i=1}^{n} \arctan \frac{ln(|x_i|+1)}{v}$ with $0 < v < 1$ to approximate the l_0-norm. It has proved that the proposed nonconvex approximation is closer to l_0-norm than the l_1-norm and is bounded by l_0 pseudo norm (see Figure 6). Dong et al. [58] formulated the sparse subspace clustering as a smoothed $l_p(0 < p < 1)$ minimization problem (SSC-SLp). It is worth pointing out that the l_p-norm can better approximate the l_0 pseudo norm than the l_1 norm (see Figure 6). In addition, an effective algorithm with convergence was established based on ADMM.



(**a**) Arctangent–logarithmic function

(**b**) l_p-norm

Figure 6. The explanation of nonconvex approximations of l_0 pseudo norm.

To study the theoretical connection between spectral clustering and SSC, Bai and Liang [60] presented a sparse subspace clustering with entropy-norm (SSC+E), which can be written as the following optimization problem:

$$\min_{C,E} \sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij} \ln C_{ij} + \lambda \|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, C^T = C, C \geq 0, \operatorname{diag}(C) = \mathbf{0}, \tag{8}$$

where $\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij} \ln C_{ij}$ is the information entropy of C which can describe the stability of variable predicted and degree of information involvement. Based on the above settings, SSC+E can obtain the closed-form solution of the coefficient matrix. It has proved that the optimal solution of (8) is equivalent to a Gaussian kernel as the sparse representation, that is, spectral clustering with Gaussian kernel can be viewed as SSC+E (8). For this reason, SSC+E obtains a sparse similarity matrix by using the Gaussian kernel, which can avoid the complex computation caused by SSC.

To cope with the challenges of data noise and high-dimensional expression profiles in identify cell types, Zheng et al. [61] proposed an adaptive sparse subspace clustering method (AdaptiveSSC), which can be formulated as the following optimization problem:

$$\min_{C,E} \left\| \frac{C}{W} \right\|_1 + \lambda \|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, \operatorname{diag}(C) = \mathbf{0}, \tag{9}$$

where W is the sample-related weight matrix based on Pearson correlation. If Pearson correlation coefficient between $\mathbf{x}_i$ and $\mathbf{x}_j$ is greater than 0, $W_{ij} = pearson(\mathbf{x}_i, \mathbf{x}_j)$; otherwise, $W_{ij} = 0$. Since AdaptiveSSC uses a data-driven adaptive sparse strategy to maintain the local structure of the data, it is more robust to data noise. However, due to the existence of the l_1-norm, the computing efficiency of AdaptiveSSC is still low, especially for high-dimensional datasets.

2.2. Low-Rank Representation Based SC

As mentioned earlier, SSC can accurately restore the underlying subspace under certain conditions. However, sparse representation-based methods aim to find a sparse representation of each data individually, which may not accurately capture the global structure. Moreover, sparse representation-based approaches may not be robust to noise and outliers. In this situation, low-rank representation came into being and was successfully applied to SC problems. Table 3 summarizes the classical low-rank SC methods.

Table 3. The comparison of classical low-rank SC methods.

Methods	Year	$F(C)$	$R(E)$	Ω	Theoretical Results
LRR [38,62]	2010	$\Vert C\Vert_*$	$\Vert E\Vert_{2,1}$	-	BDP
LRR-PSD [63]	2010	$\Vert C\Vert_*$	$\Vert E\Vert_{2,1}$	$\{C \vert C \succeq 0\}$	-
LSR [64]	2012	$\Vert C\Vert_F^2$	$\Vert E\Vert_F^2$	$\{C \vert \mathrm{diag}(C) = \mathbf{0}\}$	BDP, Grouping effect
CASS [65]	2013	$\sum_j \Vert X\mathrm{Diag}(C_{:,j})\Vert_*$	$\Vert E\Vert_F^2$	-	BDP, Grouping effect
NLRR [66]	2016	$\Vert C\Vert_{K_{k,p}}^p \ (p > 0)$	$\Vert E\Vert_{2,q}^q \ (0 < q < 1)$	-	-
S_q-LRR [67]	2016	$\Vert C\Vert_{S_q}^q \ (0 < q < 1)$	$\Vert E\Vert_1$, $\Vert E\Vert_F^2$ or $\Vert E\Vert_{2,1}$	-	-
LRRSC [68]	2017	$\Vert C\Vert_*$	$\Vert E\Vert_{2,1}$	$\{C \vert C = C^T\}$	-
CAR [69]	2018	$\sum_i \sum_j \vert C_{ij}\vert^{p_{ij}}$	$\Vert E\Vert_F^2$	-	Grouping effect
S_qNM-LRR [70]	2019	$\Vert C\Vert_{S_q}^q \ (q = 1, 2/3 \text{ or } 1/2)$	$\Vert E\Vert_1$, $\Vert E\Vert_F^2$ or $\Vert E\Vert_{2,1}$	-	-
SSRSC [71]	2021	$\Vert C\Vert_F^2$	$\Vert E\Vert_F^2$	$\{C \vert C \geq 0, \mathbf{1}^T C = s\mathbf{1}^T\}$	-
WSLog [72]	2022	$\sum_{i=1}^r \ln(w\sigma_i^p + 1)$	$\Vert E\Vert_{2,1}$	-	-

Remark: - indicates that this item does not exist.

In 2010, Liu et al. [38,62] presented a subspace clustering method based on low-rank representation (LRR) to consider the correlation structure of data. Instead of pursuing a sparse representation in SSC, LRR pursues the lowest-rank representation among data which can better capture the global structure of the data. LRR can be formulated as the following optimization problem:

$$\min_{C,E} \ \|C\|_* + \lambda\|E\|_{2,1} \quad \text{s.t.} \quad X = XC + E, \tag{10}$$

where the nuclear norm is used to approximate the rank of C, and the $l_{2,1}$ norm encourages the columns of E to be zero which can better fit the data corruptions. Liu et al. [38] have proved that, if the subspaces are independent, the representation coefficient matrix obtained from the LRR model (10) without noise has the block diagonal structure. Note that the problem (10) is a convex optimization, which can be effectively solved by the convex programming algorithms, such as the augmented Lagrange multiplier (ALM) method [128], accelerated proximal gradient (APG) [129], ADMM [117,130], and SSNAL [118]. It is worth pointing out Favaro et al. [131] study given the closed form of LRR via the singular value decomposition (SVD) of the data matrix.

As discussed earlier, after obtaining the optimal solution (C^*, E^*), C^* is used to construct the affinity matrix for spectral clustering. Since the affinity matrix is symmetric, it is usually constructed by $A = \frac{|C|+|C|^T}{2}$, which may result in a poor clustering result. To avoid symmetrization post-processing, Ni et al. [63] introduced a symmetric positive semidefinite constraint into LRR (LRR-PSD), which can be formulated as

$$\min_{C,E} \ \|C\|_* + \lambda\|E\|_{2,1} \quad \text{s.t.} \quad X = XC + E, C \succeq 0. \tag{11}$$

To further reduce computational cost in the LRR scheme, Chen et al. [68] proposed a low-rank representation with symmetric constraint (LRRSC) to avoid iterative SVD operation which can significantly decrease the computational cost for the subspace clustering. LRRSC solves the following optimization problem:

$$\min_{C,E} \|C\|_* + \lambda\|E\|_{2,1} \quad \text{s.t.} \quad X = XC + E, C = C^T. \tag{12}$$

It is worth pointing out that the affinity matrix, constructed by making full use of the intrinsically geometrical structure of the data points in the symmetric low-rank matrix, can significantly improve the performance of SC.

For LRR, each iteration must perform the singular value decomposition (SVD) due to the nuclear norm, which leads to high computational complexity and dramatically limits its ability to handle high-dimensional data. To handle this problem, Lu et al. [64] proposed a subspace clustering method via least squares regression (LSR), which was presented by exploiting the correlation of data, which solves the following optimization problem:

$$\min_{C,E} \|C\|_F^2 + \lambda\|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, \text{diag}(C) = \mathbf{0}. \tag{13}$$

It has proved that the optimal solution of LSR satisfies the block diagonal property (1) without noise and with subspace independence. Furthermore, LSR encourages a grouping effect that tends to shrink coefficients of correlated data and group them together. LSR has also been proven to be robust to bounded disturbance. In particular, Zhang et al. [132] proved theoretically that the Frobenius norm can be used as a surrogate of the rank function, and related results can also be found in [133,134]. It is easy to verify that LSR (13) is a strongly convex optimization problem and thus has a closed-form solution. Therefore, compared with the nuclear norm, using the Frobenius norm as a surrogate of the rank function can significantly reduce the computational complexity.

Although LSR has the grouping effect, it cannot produce a sparse solution, which makes the optimal solution obtained by LSR lack interpretability. Lu et al. [65] proposed a correlation adaptive subspace segmentation (CASS) method by using trace Lasso, which can be formulated as

$$\min_{C,E} \sum_j \|X\text{Diag}(C_{:,j})\|_* + \lambda\|E\|_F^2 \quad \text{s.t.} \quad X = XC + E. \tag{14}$$

CASS is a data correlation-dependent method which simultaneously performs automatic data selection and groups correlated data together. Consequently, CASS can be considered as the adaptive balance between SSC and LSR. Lu et al. generalized the EBD conditions and proved that the solution obtained by (14) has the block diagonal structure (1) if the underlying subspaces are independent. Furthermore, CASS has been proven to have the grouping effect, i.e., the coefficients of a group of correlated data are approximately equal.

Wang et al. [69] presented a correlation adaptive regression (CAR) subspace clustering method, which can be formulated as the following optimization problem:

$$\min_{C,E} \sum_i \sum_j |C_{ij}|^{p_{ij}} + \lambda\|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, \tag{15}$$

where $p_{ij} = 1 + r_{i,j}$, with $r_{i,j} = |\mathbf{x}_i^T \mathbf{x}_j|$ being the correlation between data points $\mathbf{x}_i$ and $\mathbf{x}_j$. CAR can be seen as a trade-off between l_1-norm and Frobenius-norm, that is, l_1-norm penalty is imposed on the low correlated data while the Frobenius-norm penalty is imposed on the highly correlated data. Therefore, the CAR model (15) can ensure that the affinity matrix satisfies the inter-subspace sparsity and intracluster uniformity.

To further reveal the inherent correlations among data points, Xu et al. [71] proposed a scaled simplex representation for subspace clustering (SSRSC), which aims to solve the following optimization problem:

$$\min_{C} \|C\|_F^2 + \lambda\|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, C \geq 0, \mathbf{1}^T C = s\mathbf{1}^T, \tag{16}$$

where $0 < s < 1$ is a scalar. Here, the constraint condition $C \geq 0$ is conducive to encouraging data points from the same subspace to represent each other while suppressing the data from different subspaces to represent each other. Thus, the obtained coefficient matrix C is distinctive. The constraint $\mathbf{1}^T C = s\mathbf{1}^T$ can limit the sum of each coefficient vector C_i to be s, which makes the representation more discriminative since C should be non-negative. Note that the SSRSC model (16) is a convex optimization problem with two equality constraints and one inequality constraint. Thus, it can be solved by ADMM [117], SSNAL [118], and so forth.

Although many achievements have been made on the convex nuclear norm, which is a surrogate of the rank function, the nuclear norm minimization may over-penalize large singular values, resulting in a bias. Hence, a solution obtained by the nuclear norm may be suboptimal since it is not a perfect approximation of the matrix rank. In order to achieve a better approximation for the rank function, many nonconvex surrogate functions of rank have been introduced into the SC problems. Jiang et al. [66] proposed a robust subspace segmentation via nonconvex LRR (NLRR) by replacing the nuclear norm with the Ky Fan p-k norm ($0 < p, k \in \{1, \ldots, d\}$) and penalizing the noise via the $l_{2,q}$ norm ($0 < q < 1$) in LRR (10). A proximal iteratively reweighted optimization algorithm (PIRA) is designed to solve the NLRR. Zhang [67] introduced the nonconvex Schatten-q ($0 < q < 1$) regularization to approximate the rank function for the SC problem, called S_q-LRR, which can be solved by linearized alternating direction method with the adaptive penalty (LADMAP) [130] or the generalized iterated shrinkage (GMST) algorithm [135]. Zhang [70] proposed a Schatten-q norm minimization-based LRR (S_pNM-LRR) with $q = 1, 2/3$ and $1/2$, which can be solved by ADMM [136]. Most recently, Shen et al. [72] presented a non-convex low-rank approximation subspace clustering method based on a weighted Schatten-p norm jointed logarithmic constraint (WSLog). By using the logarithmic function to further tighten Schatten-p norm, WSLog can better recover the low rank of data. In addition, WSLog is more robust to noise due to the data-related weight matrix.

2.3. Low-Rank Sparse Representation Based SC

As analyzed above, LRR can capture global information and successfully recover the corrupted data drawn from independent subspaces. However, LRR often fails in the case of disjoint subspaces or overlapping subspaces. SSC is superior to LRR in this respect. However, it is weaker than LRR in capturing the global structure of subspaces from the corrupted data. Therefore, the natural idea is to combine the low-rank representation with sparse representation. Table 4 summarizes the classical low-rank sparse SC methods.

In 2011, Luo et al. [73] proposed a multiple subspace representation (MSR) by combining the sparsity and low-rank of the representation coefficient matrix, which can be formulated as the following optimization problem:

$$\min_{C,E} \ \|C\|_* + \mu\|C\|_1 + \lambda\|E\|_{2,1} \quad \text{s.t.} \quad X = XC + E, \mathbf{1}^T C = \mathbf{1}^T, \tag{17}$$

where $\mathbf{1}^T C = \mathbf{1}^T$ is used to pursue the linear/affine representation of subspaces. Note that the SSC model (3) and LRR model (10) can be regarded as the special case of (5) with $\mu = \infty$ and $\mu = 0$, respectively, and the $l_{2,1}$ norm is more robust against data corruption than the usual Frobenius norm. It has proved that the representation coefficient matrix obtained from the MSR model (17) without noise has the block diagonal structure (1) when the underlying affine subspaces are independent.

A common problem with existing methods is that negative coefficients may be generated, which may cause data to "cancel each other" through complex addition and subtraction [137]. This may result in the potential structure between data not being able to be accurately captured. To deal with this issue, Zhuang et al. [74] proposed a non-negative low-rank and sparse (NNLRS) model by adding a non-negative constraint to the MSR model (17). NNLRS solves the following optimization problem:

$$\min_{C,E} \ \|C\|_* + \mu\|C\|_1 + \lambda\|E\|_{2,1} \quad \text{s.t.} \quad X = XC + E, C \geq 0, \tag{18}$$

where $C \geq 0$ is used to enforce the representation to be non-negative so that the obtained coefficients can be directly used as the graph weights. The non-negativity could potentially enhance the discriminability of the coefficients such that the data points are more likely reconstructed by the data points from the same subspace.

Wang et al. [75] proposed a low-rank sparse subspace clustering (LRSSC) by directly combining SSC (3) and LRR (10). LRSSC solves the following optimization problem:

$$\min_{C,E} \ \|C\|_* + \mu\|C\|_1 + \lambda\|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, \mathrm{diag}(C) = \mathbf{0}. \tag{19}$$

It has proved that the representation coefficient matrices obtained from the LRSSC (19) and the LRSSC with random data have the block diagonal structure (1) when the underlying affine subspaces are independent.

Table 4. The comparison of classical low-rank sparse SC methods.

Methods	Year	$F(C)$	$R(E)$	Ω	Theoretical Results
MSR [73]	2011	$\Vert C\Vert_* + \mu\Vert C\Vert_1$	$\Vert E\Vert_{2,1}$	$\{C \mid \mathbf{1}^T C = \mathbf{1}^T\}$	BDP
NNLRS [74]	2012	$\Vert C\Vert_* + \mu\Vert C\Vert_1$	$\Vert E\Vert_{2,1}$	$\{C \mid C \geq 0\}$	-
LRSSC [75]	2013	$\Vert C\Vert_* + \mu\Vert C\Vert_1$	$\Vert E\Vert_F^2$	$\{C \mid \mathrm{diag}(C) = \mathbf{0}\}$	BDP
LRRLC [76]	2013	$\Vert C\Vert_* + \mu\sum_{i,j} \Vert\mathbf{x}_i - \mathbf{x}_j\Vert_2^2 \lvert C_{ij}\rvert$	$\Vert E\Vert_{2,1}$	-	-
LRSR [77]	2016	$\Vert Y\Vert_* + \Vert C\Vert_* + \mu\Vert J\Vert_1$	$\Vert E\Vert_{2,1}$	$\{C \mid C = J - \mathrm{Diag}(\mathrm{diag}(J))\}$	-
EnSC [78]	2016	$\mu\Vert C\Vert_1 + \frac{1-\mu}{2}\Vert C\Vert_F^2$	$\Vert E\Vert_F^2$	$X = XC + E, \mathrm{diag}(C) = \mathbf{0}$	BDP
CAWR [79]	2017	$\Vert(\mathbf{1}\mathbf{1}^T - U) \odot C\Vert_1 + \frac{\mu}{2}\Vert\sqrt{U} \odot C\Vert_F^2$	$\Vert E\Vert_F^2$	$X = XC + E$	BDP, Grouping effect
GMC-LRSSC [80]	2020	$\psi_B(\sigma(C)) + \mu\psi_B(C)$	$\Vert E\Vert_F^2$	$\{C \mid \mathrm{diag}(C) = \mathbf{0}\}$	-
S_0/l_0-LRSSC [80]	2020	$\Vert C\Vert_{S_0} + \mu\Vert C\Vert_0$	$\Vert E\Vert_F^2$	$\{C \mid \mathrm{diag}(C) = \mathbf{0}\}$	-
SCAT [81]	2020	$\sum_{i,j} \Vert\mathbf{x}_i - \mathbf{x}_j\Vert_2^2 C_{ij} + \frac{\mu}{2}\Vert C\Vert_F^2$	$\Psi(E)$	$\{C \mid C \geq 0, \mathrm{diag}(C) = \mathbf{0}, C^T\mathbf{1} = \mathbf{1}\}$	-
NL-SSLR [82]	2021	$\Vert C\Vert_* + \mu\Vert W \odot C\Vert_1 + \frac{\alpha}{2}\Vert C - \bar{C}_{NL}\Vert_F^2$	$\Vert E\Vert_F^2$	-	-
KSLSR [83]	2022	$\Vert V \odot W\Vert_F^2$	$\Vert E\Vert_F^2$	$\{V \mid \mathrm{diag}(V) = \mathbf{0}, V^T\mathbf{1} \leq k\mathbf{1}\}$	Grouping effect

Remark: - indicates that this item does not exist.

Zheng et al. [76] present a low-rank representation method with local constraint (LRRLC) by combining LRR with local information of data, which can be formulated as

$$\min_{C,E} \ \|C\|_* + \mu\sum_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 |C_{ij}| + \lambda\|E\|_{2,1} \quad \text{s.t.} \quad X = XC + E, \tag{20}$$

where $\sum_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 |C_{ij}|$ is used to measure the data correlation of the representation coefficient C, which changes with the distance between samples. Note that LRRLC is able to capture both the global structure by the nuclear norm and the local structure by the locally constrained regularization term simultaneously.

Wang et al. [77] proposed a low-rank subspace sparse representation (LRSR) method, which can be formulated as

$$\min_{Y,C,J,E} \ \|Y\|_* + \|C\|_* + \mu\|J\|_1 + \lambda\|E\|_{2,1} \quad \text{s.t.} \quad X = YC + E, C = J - \mathrm{Diag}(\mathrm{diag}(J)). \tag{21}$$

Unlike SSC-based and LRR-based models, LRSR (21) aims to recover both the clean dictionary Y and the representation coefficient C simultaneously. It has proved that LRSR can not only recover the low-rank subspaces, but also obtain a relatively sparse segmentation for the disjoint subspaces or even overlapping subspaces.

As mentioned earlier, the l_1-norm regularization can be guaranteed to give a subspace-preserving affinity (i.e., there are no connections between points from different subspaces) under broad conditions. However, the clusters may not be connected (since SSC tends to select only one randomly for a group of highly correlated data points), while Frobenius norm and nuclear norm regularization often improve connectivity, but give a subspace-preserving affinity only for independent subspaces. You et al. [78] proposed a subspace clustering via elastic net regularizer (EnSC), which can be formulated as

$$\min_{C,E} \ \mu\|C\|_1 + \frac{1-\mu}{2}\|C\|_F^2 + \lambda\|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, \text{diag}(C) = \mathbf{0}, \tag{22}$$

where $\mu \in [0,1]$. Note that EnSC (22) can be regarded as a combination of SSC and LSR, and will reduce to each of them when $\mu = 1$ and $\mu = 0$, respectively. It has drawn the theoretical conditions for the subspace preserving property free of noise and with independent subspaces, that is, if the parameter is in a certain range, the optimal solution has a block diagonal structure. It is easy to verify that EnSC (22) is a convex optimization problem that can be solved by the active-set method [78], ADMM [117], SSNAL [118], and so on.

For the image segmentation problem, Wang and Wu [79] proposed a correlation adaptive weighted regression (CAWR) subspace clustering method via combining the correlation weighted l_1-norm and l_2-norm

$$\min_{C,E} \ \|(\mathbf{1}\mathbf{1}^T - U) \odot C\|_1 + \frac{\mu}{2}\|\sqrt{U} \odot C\|_F^2 + \lambda\|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, \tag{23}$$

where $U = X^T X$ is the correlation matrix of the data. It can be seen that CAWR introduces the correlation adaptive weight to each coefficient in the l_1-norm and l_2-norm. It can be seen that CAWR introduces the correlation adaptive weight for both l_1 and l_2 regularization, which can make better use of data information. Many existing models can be regarded as special cases of CAWR, such as SSC [119], LLR [38], and EnSC [78]. It has proved that CAWR has the ability of subspace selection (by l_1 regularization) for uncorrelated data and the grouping effect (by l_2 regularization) for highly correlated data. In addition, the optimal solution obtained by CAWR is proved to have the block diagonal structure under noiseless data and independent subspace. However, CAWR lacks robustness to data corruption [91].

Instead of using the convex approximations of rank and l_0 pseudo norm, Brbić and Kopriva [80] introduced two S_0/l_0 pseudo norm-based nonconvex regularizations for LRSSC. The first method is based on the multivariate generalization of the minimax-concave (GMC) penalty function, which can be formulated as

$$\min_{C,E} \ \psi_B(\sigma(C)) + \mu\psi_B(C) + \lambda\|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, \text{diag}(C) = \mathbf{0}. \tag{24}$$

where $B \in \mathbb{R}^{n\times n}$ is a given matrix, and the GMC penalty is defined by $\psi_B(\mathbf{x}) = \|\mathbf{x}\|_1 - S_B(\mathbf{x})$ with $S_B(\mathbf{x}) = \inf_{\mathbf{v}\in\mathbb{R}^n}\left\{\|\mathbf{v}\|_1 + \frac{1}{2}\|B(\mathbf{x}-\mathbf{v})\|_2^2\right\}$. The GMC-LRSSC can maintain the convexity of low-rank and sparsity-constrained subproblems and achieve better approximation for rank and sparsity than nuclear and l_1-norms. To further approximate the low-rank and sparsity, the Schatten-0 and l_0 pseudo norm regularizations are introduced to LRSSC (S_0/l_0-LRSSC), which is

$$\min_{C,E} \ \|C\|_{S_0} + \mu\|C\|_0 + \lambda\|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, \text{diag}(C) = \mathbf{0}. \tag{25}$$

It is worth pointing out that simultaneous rank and sparsity regularization can be handled by the proximal average method [138,139], which approximates the proximal map of the joint solution by averaging the solutions obtained separately from low-rank and sparsity subproblems. Although GMC-LRSSC (24) and S_0/l_0-LRSSC (25) are nonconvex optimization problems, they can be solved by an improved ADMM [140]. Moreover, the convergence of the two algorithms was established by Brbić and Kopriva [80].

To further reflect the similarities between data points, Zhong and Pun [81] proposed a subspace clustering method with an adaptive transformation matrix (SCAT)

$$\begin{aligned} &\min_{C,E} \sum_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 C_{ij} + \frac{\mu}{2}\|C\|_F^2 + \lambda\Psi(E) \\ &\text{s.t. } X = XC + E, C \geq 0, \operatorname{diag}(C) = \mathbf{0}, C^T\mathbf{1} = \mathbf{1}, \end{aligned} \tag{26}$$

where $\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 C_{ij}$ is used to capture the local structure of data, and the affine constraint $C^T\mathbf{1} = \mathbf{1}$ is used to deal with affine subspaces. $\Psi(E)$ is a function of E which depends on the distribution of data errors. For instance $\|E\|_F^2$ is used for Gaussian residuals while $\|E\|_1$ for Laplacian. Note that SCAT can be regarded as imposing data similarity information into LSR [64]. Hence, SCAT can simultaneously capture the global (due to the LSR) and the local structure (due to the similarity learning) of data.

Most recently, Zhai et al. [82] proposed a novel scalable nonlocal means regularized sketched reweighted sparse and low-rank (NL-SSLR) SC method for the large hyperspectral images (HSIs) clustering. NL-SSLR aims to explore both local and global structural information simultaneously, which solves the following optimization problem:

$$\min_{C,E} \|C\|_* + \mu\|W \odot C\|_1 + \frac{\alpha}{2}\|C - \bar{C}_{NL}\|_F^2 + \lambda\|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, \tag{27}$$

where $\bar{C}_{NL}$ is a nonlocal coefficient matrix obtained by a nonlocal means filter, and W is a weight matrix defined as $W_{ij} = \frac{a}{C_{ij}+b}$ with two small constants a and b. Note that W obtained in this way can better promote sparsity by punishing larger elements, and the two constants a and b are used to avoid overweight, which can be taken as 0.001 in practice. It should be pointed out that the nuclear norm is used to explore the global structure of data, while the reweighted l_1-norm regularization and the nonlocal means filter are utilized to explore the spatial correlation information of data fully.

Since LSR [64] uses Frobenius norm regularization, it cannot obtain sparse solution, which may make it sensitive to data corruptions and data dimensions. To handle this issue, Yang et al. [83] proposed the following k-sparse least squares regression (KSLSR) subspace clustering method via 0-1 integer programming

$$\min_{V,W,E} \|V \odot W\|_F^2 + \lambda\|E\|_F^2 \quad \text{s.t.} \quad X = X(V \odot W) + E, \operatorname{diag}(V) = \mathbf{0}, V^T\mathbf{1} \leq k\mathbf{1}, \tag{28}$$

where $W \in \mathbb{R}^{n\times n}$ is the weight value. $V \in \{0,1\}^{n\times n}$ is the sparse selection factor, where $V_{ij} = 1$ represents that the data $\mathbf{x}_j$ are selected for the linear representation of the data $\mathbf{x}_i$, and $V_{ij} = 0$ indicates that the data $\mathbf{x}_j$ are not selected. $\mathbf{v}_i^T\mathbf{1} \leq k$ indicates that the data $\mathbf{x}_i$ can be linearly represented by at most k data points which can be regarded as a transformation of l_0 pseudo norm. Once V and w are obtained, the coefficient matrix C can be obtained immediately by $C = V \odot W$. It has proved that the KSLSR model (28) has the k-sparsity and grouping effect, thus KSLSR is robust to data corruptions and dimension of data. However, due to the 0-1 constraint on V, KSLSR is an NP-hard combinatorial optimization problem, which is usually difficult to solve directly [141]. To effectively solve KSLSR, problem (28) was transformed into a continuous optimization problem by equivalently replacing the 0-1 constraint on V with a l_p-box intersection [142], i.e., the intersection of a shifted l_p-sphere and a solid cube. Figure 7 shows the geometric illustration of the l_p-box. Then, the KSLSR with continuous constraints can be solved by ADMM.

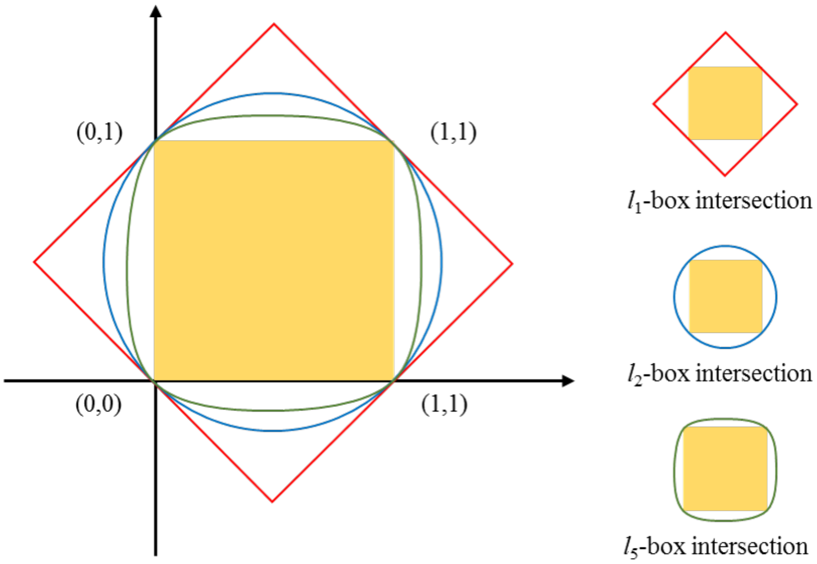


Figure 7. The geometric illustration of the l_p-box in $\mathbb{R}^2$.

3. Local Structure Preserving SC

3.1. Graph-Regularized Based SC

In the previous section, we review some classical low-rank sparse SC methods, such as SSC, LRR, and MSR. Although these methods have reached state-of-the-art performance, they may lose the local structure of data. In other words, the representation coefficients for locally similar features or samples may differ greatly. This may damage the connection of the similarity graph, and ultimately affect the clustering performance. To deal with this issue, a natural idea is to introduce the graph information of the incidence matrix into the subspace representation. Table 5 summarizes the classical graph-regularized SC methods.

Table 5. The comparison of classical graph-regularized SC methods.

Methods	Year	$F(C)$	$R(E)$	Constraint
GLRR [84]	2013	$\|C\|_* + \mu \mathrm{tr}(C^T LC)$	$\|E\|_{2,1}$	$X = XC + E$
LapLRR [85]	2014	$\|C\|_* + \mu \mathrm{tr}(C^T LC)$	$\|E\|_F^2$	$X = XC + E, C \geq 0$
E-SSC [86]	2016	$\|C\|_1 + \frac{\mu}{2} \mathrm{tr}(C^T LC)$	$\|E\|_F^2$	$X = XC + E, \mathrm{diag}(C) = \mathbf{0}, C^T\mathbf{1} = \mathbf{1}$
NSHLRR [87]	2016	$\|C\|_* + \mu_1 \|C\|_1 + \mu_2 \mathrm{tr}(CL_h C^T)$	$\|E\|_1$	$X = XC + E, C \geq 0$
GCLRR [88]	2017	$\|C\|_* + \mu \mathrm{tr}(CLC^T)$	$\|E\|_{2,1}$	$X = XVC + E, V^T V = I$
MLLRR [89]	2019	$\|C\|_* + \mu_1 \|C\|_1 + \mu_2 \mathrm{tr}(CLC^T)$	$\|E\|_{2,1}$	$X = XC + E, C \geq 0$
SGL [90]	2022	$\|C\|_F^2 + \mu \mathrm{tr}(F^T LF)$	$\|E\|_F^2$	$X = XC + E, C \geq 0, C\mathbf{1} = \mathbf{1}, F^T F = I$
GL$l_{1/2}$RSC [91]	2022	$\|C\|_{1/2}^{1/2} + \mu_1 \|L \odot C\|_{1/2}^{1/2} + \mu_2 \|P \odot C\|_{1/2}^{1/2} + \mu_3 \|R \odot C\|_{1/2}^{1/2}$	$\|E\|_F^2$	$X = XC + E, C \geq 0$

To make full use of the correlation of different hyperspectral images and consider their local manifold structure, Lu et al. [84] proposed a graph-regularized low-rank representation (GLRR) subspace clustering method, which can be written as

$$\min_{C,E} \ \|C\|_* + \mu \mathrm{tr}(CLC^T) + \lambda \|E\|_{2,1} \quad \text{s.t.} \quad X = XC + E, \tag{29}$$

where $L = D - W$ is the Laplace matrix of X, and W and D are the incidence matrix and symmetric degree matrix of X, respectively; $\mathrm{tr}(CLC^T)$ is called graph regularization, which can be used to maintain the similar local structures of data [143]. GLRR (29) can be regarded as a generalization of the LRR [38] by considering the local geometrical structure of the data. Hence, GLRR can capture both the global structure (due to the LRR framework) and local structure (due to the graph regularization) of the data. It is easy to verify that the optimization problem (29) is convex and thus can be solved by ALM [128], ADMM [117,130], and other convex optimization algorithms. To further improve the GLRR, Liu et al. [85] proposed a Laplacian regularized LRR (LapLRR) method, which can be regarded as a generalization of GLRR (29) by adding a nonnegative constraint. Chen et al. [86] also introduced the Laplace regularization into SSC [52] and LRR [38], and proposed the enhanced sparse subspace clustering method (E-SSC) and enhanced LRR method (E-LRR).

Yin et al. [87] proposed a non-negative sparse hyper-Laplacian regularized low-rank representation method (NSHLRR), which can be formulated as

$$\min_{C,E} \quad \|C\|_* + \mu_1\|C\|_1 + \mu_2 \mathrm{tr}(CL_hC^T) + \lambda\|E\|_1 \quad \text{s.t.} \quad X = XC + E, C \geq 0, \tag{30}$$

where L_h is the hyper-Laplacian matrix defined in [144]. The Laplacian and the sparsity constraints encourage choosing nearby samples (which most likely belong to the same cluster), rather than the faraway samples (which may belong to other clusters), to represent the sample at the center of the locally linear manifold. Hence, the optimal solution obtained by NSHLRR (30) approximately satisfies the block diagonal structure. By combining graph hyper-Laplacian regularizer with LRR, the NSHLRR can simultaneously capture the global structure and inherent nonlinear geometric information of data. Note that NLSRR can be regarded as an extension of LRRGL [145] by replacing the traditional Laplacian matrix with a hypergraph Laplacian matrix. To improve the robustness of data corruptions, Wang et al. [89] imposed the $l_{1,2}$ regularization to the data corruptions instead of l_1 regularization and applied it to tumor clustering.

Du et al. [88] proposed a graph regularized Compact LRR (GCLRR) method by combining dictionary learning with low-rank representation and introducing graph regularization. The GCLRR can be formulated as

$$\min_{C,V,E} \quad \|C\|_* + \mu \mathrm{tr}(CLC^T) + \lambda\|E\|_{2,1} \quad \text{s.t.} \quad X = XVC + E, V^TV = I, \tag{31}$$

where $V^TV = I$ is used to eliminate the uncertainty caused by the diagonal structure. Instead of using the original data X as a dictionary, the GCLRR used the linear combination of X as the dictionary, i.e., $A = XV$. Thus, the proposed method can realize dictionary learning and low-rank representation simultaneously. In addition, unlike previous SC methods which treat the affinity matrix construction and clustering algorithm separately, GCLRR integrates these two tasks into a single optimization framework to ensure overall optimization. It is worth pointing out that the GCLRR model (31) can capture both the global subspace structure (by the low-rank representation) and the local geometrical structure (by graph regularization).

Although the above graph-regularized based SC methods have shown good performance, these methods are often inefficient and limited in exploring the potential information of data. To deal with these problems, Kang et al. [90] proposed a scalable graph learning (SGL) subspace clustering method which can be written as

$$\min_{C,F,E} \quad \|C\|_F^2 + \mu \mathrm{tr}(F^TLF) + \lambda\|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, C \geq 0, C\mathbf{1} = \mathbf{1}, F^TF = I. \tag{32}$$

Since SGL used the Frobenius norm regularization, it avoided the singular value decomposition, which can increase the computational efficiency. With the help of the graph Laplace information, SGL can capture the local information of the data and handle new data. In addition, the connection between SGL and the k-means clustering was established.

To deal with the challenges brought by diverse visual patterns, noise, and complex background in image processing, Francis et al. [91] proposed the following subspace clustering method by integrating the $l_{1/2}$-norm, l_2-norm and graph Laplacian regularization (GL$l_{1/2}$RSC). The proposed GL$l_{1/2}$RSC can be formulated as

$$\begin{aligned} \min_{C,E} \ & \|C\|_{1/2}^{1/2} + \mu_1 \|L \odot C\|_{1/2}^{1/2} + \mu_2 \|P \odot C\|_{1/2}^{1/2} + \mu_3 \|R \odot C\|_{1/2}^{1/2} + \lambda \|E\|_F^2 \\ \text{s.t.} \ & X = XC + E, C \geq 0, \end{aligned} \tag{33}$$

where P and R can describe the correlation of data, which are defined as $P = \mathbf{11}^T - U$ and $R = \sqrt{U}$ with $U = X^T X$. $C \geq 0$ is used to promote better learning of the local geometrical structure and can deliver more physical interpretation to the data points. Based on the above design, GL$l_{1/2}$RSC has the capabilities of $l_{1/2}$-norm, l_2-norm, which can achieve the subspace selection and subspace grouping simultaneously. Due to the existence of $l_{1/2}$-norm, the obtained representation coefficient matrix C by GL$l_{1/2}$RSC (33) is sparse and has an accurate block diagonal structure. Moreover, GL$l_{1/2}$RSC can retain the local structure of data and is relatively insensitive to noise and outliers.

3.2. Latent Representation Based SC

The previously mentioned methods generally use the observed data itself as the dictionary. However, when the observed date is insufficient or corrupted by noise, the performance of these methods may be inferior. To solve this issue, latent representation SC methods were presented by joint low-rank recovery and salient feature extraction. Table 6 summarizes the classical latent representation SC methods.

To cope with the impact of insufficient data samples, Liu and Yan [92] proposed a Latent low-rank representation (LatLRR) on the basis of LRR, which imposed low-rank constraints on the column representation coefficient matrix and row representation coefficient matrix at the same time. LatLRR can be formulated as

$$\min_{C,H,E} \ \|C\|_* + \|H\|_* + \lambda \|E\|_1 \quad \text{s.t.} \quad X = XC + HX + E, \tag{34}$$

where $H \in \mathbb{R}^{d \times d}$ is used to reflect the information of hidden data. Notice that there is no parameter between C and H in (34) as the strengths of these two parts are automatically balanced. LatLRR aims to use row data information to make up for the corruption of column data information. It has proved that LatLRR (34) can not only deal with insufficient data samples but also robustly extract salient features from corrupted data. However, because the nuclear norm is used as the penalty function, the singular value decomposition (SVD) must be performed twice in each iteration, which leads to high computational complexity and greatly limits its ability to handle high-dimensional data. In addition, Zhang et al. [146] proved that the solution of the noiseless LatLRR (34) is not unique and gave a closed-form solution.

Although the LatLRR method [92] can handle the situation of insufficient data, they may be inaccurate in identifying the underlying subspaces due to the noise, so they may fail to preserve the local structure of data. To further capture the local structure of data, Zhang et al. [93] proposed a regularized low-rank representation (rLRR) method via introducing a data-dependent Laplacian regularization into LatLRR, which can be formulated as

$$\begin{aligned} \min_{C,H,E} \ & \|C\|_* + \|H\|_* + \frac{\mu}{2} \left[\operatorname{tr}\left(C^T L C\right) + \operatorname{tr}\left(H X L C X^T H^T\right) \right] + \lambda \|E\|_{2,1} \\ \text{s.t.} \quad & X = XC + HX + E, Z \geq 0, \end{aligned} \tag{35}$$

where $L = D - W$ is the Laplace matrix of X, and W and D are the incidence matrix and symmetric degree matrix of X, respectively. The Laplacian regularization $\operatorname{tr}\left(C^T L C\right)$ and $\operatorname{tr}\left(H X L C X^T H^T\right)$ are used to preserve the locality and similarity of the observed data and salient data, respectively. It should be noted that the idea of Laplacian regularization has been widely used in machine learning (see, e.g., [147–150]). Note that LatLRR [92] can be regarded as a special case of rLRR with $\mu = 0$. By using Laplace regularization to improve

LatLRR, rLRR can better preserve the local data information. Hence, rLRR can produce more discriminative low-rank coefficients and robust representations.

Table 6. The comparison of classical latent representation SC methods.

Methods	Year	$F(C)$	$R(E)$	Constraint
LatLRR [92]	2011	$\|C\|_* + \|H\|_*$	$\|E\|_1$	$X = XC + HX + E$
rLRR [93]	2014	$\|C\|_* + \|H\|_* + \frac{\mu}{2}\left[\mathrm{tr}\left(C^T LC\right) + \mathrm{tr}\left(HXLCX^T H^T\right)\right]$	$\|E\|_{2,1}$	$X = XC + HX + E, C \geq 0$
FLRR [94]	2018	$\|C\|_F^2 + \|H\|_F^2$	$\|E\|_1$	$X = XC + HX + E$
AS-LRC [95]	2019	$\|C\|_* + \|H\|_{2,1} + \mu_1 \|(\mathbf{1}\mathbf{1}^T - R) \odot C\|_1 + \mu_2 \wp(H, R)$	$\|E\|_1$	$X = XC + HX + E$
FLLRSC [96]	2020	$\|C\|_F^2 + \|H\|_F^2$	$\|E\|_{2,1}$	$\widetilde{X} = \widetilde{X}C + H\widetilde{X} + E$
eLatLRR [97]	2021	$\|C\|_* + \|H\|_* + \frac{\mu}{2}\|HX - HXC\|_F^2$	$\|E\|_1$	$X = XC + HX + E$
LLRRWD [98]	2022	$\mu_1\left(\|C\|_F^2 + \|H\|_F^2\right) + \mu_2 \mathrm{tr}\left((P \odot D)^T C\right)$	$\|E\|_1$	$X = XC + HX + E, C \geq 0, C_{i,i} = 0, C\mathbf{1} = \mathbf{1}$

Inspired by LSR [64], Yu and Wu [94] proposed to replace the nuclear norm with the Frobenius norm. The optimization problem of the Frobenius norm based low-rank representation (FLRR) method is as follows:

$$\min_{C,H,E} \ \|C\|_F^2 + \|H\|_F^2 + \lambda\|E\|_1 \quad \text{s.t.} \quad X = XC + HX + E. \tag{36}$$

It is easy to obtain a closed-form solution of FLRR (36) by virtue of the properties of the Frobenius norm. Consequently, compared with LatLRR (34), the computational complexity of FLRR can be greatly reduced. In addition, it is proved theoretically that the Frobenius norm can be used as a convex surrogate for the rank function. Following [64], FLRR (36) also has block diagonal property and grouping effect, that is, FLRR tends to assign approximately equal coefficients to a set of highly correlated data and group them together. Although FLLRR has achieved a better clustering performance, it is sensitive to outliers and noise due to the Frobenius norm.

To recover the underlying subspaces more accurately, Zhang et al. [95] proposed an adaptive structure-constrained low-rank coding (AS-LRC) method by combining latent representation of data with automatic weighting learning

$$\begin{aligned} \min_{C,H,R,E} \ & \|C\|_* + \|H\|_{2,1} + \mu_1\|(\mathbf{1}\mathbf{1}^T - R) \odot C\|_1 + \mu_2\wp(H, R) + \lambda\|E\|_1 \\ \text{s.t.} \ & X = XC + HX + E, \end{aligned} \tag{37}$$

where R is an auto-weighting matrix, $\wp(H, R) = \|LX - LXR\|_F^2 + \|\mathbf{1}^T - \mathbf{1}^T R\|_F^2 + \|R\|_{2,1}$. Based on the above design, the auto-weighting matrix R is group sparsity, and hence it can better preserve the local information of data. $\|(\mathbf{1}\mathbf{1}^T - R) \odot C\|_1$ can be used to encourage a block diagonal structure of the representation coefficient C. In addition, AS-LRC (37) selects the $l_{2,1}$-norm to encourage a group of sparse features rather than learning low-rank features by nuclear-norm regularization, which is more robust to noise and outliers.

To improve the performance of LRR for hyperspectral band selection, Sun et al. [96] presented a fast and latent low-rank subspace clustering (FLLRSC) method. The FLLRSC can be expressed as the following optimization problem:

$$\min_{C,H,E} \ \|C\|_F^2 + \|H\|_F^2 + \lambda\|E\|_{2,1} \quad \text{s.t.} \quad \widetilde{X} = \widetilde{X}C + H\widetilde{X} + E, \tag{38}$$

where $\widetilde{X} = \Phi^T X \in \mathbb{R}^{m \times n}$ is the projected matrix, and $\Phi = \sqrt{m/n}DFP \in \mathbb{R}^{d \times m}$ is the Hadamard random projection matrix. $D \in \mathbb{R}^{d \times d}$ is a diagonal matrix whose diagonal elements are taken from $-1, 1$. $F \in \mathbb{R}^{d \times d}$ is the Hadamard matrix. $P \in \mathbb{R}^{d \times m}$ is a uniform sampling matrix that randomly samples m columns of DF. Thus, FLLRSC first performs Hadamard projection on the original data to reduce the data dimension, and then performs self-representation on the transformed data to obtain the affinity matrix. Based on this design, FLLRSC can greatly reduce computational cost and solve the problem of insufficient sampling. In addition, the coefficient matrix obtained by FLLRSC has the block diagonal structure and is insensitive to noise.

To reduce the effects of noise, Wu et al. [97] extended the LatLRR (eLatLRR) by introducing the correlations between features in the low-rank decomposition process

$$\min_{C,H,E} \|C\|_* + \|H\|_* + \frac{\mu}{2}\|HX - HXC\|_F^2 + \lambda\|E\|_1 \quad \text{s.t.} \quad X = XC + HX + E, \tag{39}$$

where the term $\|HX - HXC\|_F^2$ is used to establish the relationship between C and H, which can further enhance the data representation. According to the fact that HX is cleaner than X, $\|HX - HXC\|_F^2$ can be used to generate a coefficient matrix C that is more distinctive and satisfies the block diagonal structure. In addition, LatLRR [92] can be considered as a special case of eLatLRR (39) with $\mu = 0$. Then, a non-negative constraint and elastic net regularization are integrated into the eLatLRR, which automatically improves the representation ability of the homogeneous samples while suppressing the effect of noise.

To fully consider the local structure of data, Fu et al. [98] proposed a novel latent LRR with weighted distance penalty (LLRRWD), that is,

$$\begin{aligned} &\min_{C,H,E} \mu_1\left(\|C\|_F^2 + \|H\|_F^2\right) + \mu_2 \text{tr}\left((P \odot D)^T C\right) + \lambda\|E\|_1 \\ &\text{s.t. } X = XC + HX + E, C \geq 0, C_{i,i} = 0, C\mathbf{1} = \mathbf{1}, \end{aligned} \tag{40}$$

where D and P are the Euclidean distance matrix and weight matrix among samples, respectively. If the labels of x_i and x_j are the same, $P_{ij} = 1$; otherwise, $P_{ij} = \theta$ with $\theta > 1$. $\text{tr}\left((P \odot D)^T C\right)$ is called the weighted distance penalty. In particular, if $P_{ij} = 1$, the weighted distance will reduce to the traditional Euclidean distance. Different from the Euclidean distance, the weighted distance can not only maintain the local structure, but also enhance the discrimination of affinity matrix. It is worth noting that FLLRR (36) can be regarded as a special case of LLRRWD (40) with $\mu_2 = 0$. Moreover, a weight matrix is imposed on the sparse error norm to reduce the effect of noise and redundancy.

3.3. Block Diagonal Representation Based SC

As mentioned earlier, a good SC model should have the block diagonal property. The existing methods usually pursue the block diagonal structure of the representation coefficient matrix by imposing different structure priors, such as sparsity and low-rank, which is indirect. Under certain subspace assumptions, the optimal solution obtained by these methods may satisfy the block diagonal structure. However, in practical applications, due to data corruption, the block diagonal property cannot necessarily be guaranteed. The natural idea is to directly impose a block diagonal prior to the representation coefficient matrix. Table 7 summarizes the classical block diagonal representation SC.

To directly pursue the block-diagonal structure, Feng et al. [99] proposed block diagonal SSC and LRR through imposing a rank constraint on the graph Laplacian

$$\text{(BD-SSC)} \min_{C,E} \|C\|_1 + \lambda\|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, \text{diag}(C) = \mathbf{0}, \text{rank}(L_A) = n - k, \tag{41}$$

$$\text{(BD-LRR)} \min_{C,E} \|C\|_* + \lambda\|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, \text{rank}(L_A) = n - k, \tag{42}$$

where L_A is the Laplacian matrix, which is defined as: $L_A(i,j) = -A(i,j)$ if i $\neq$ j; $L_A(i,j) = \sum_{j \neq i} A(i,j)$, otherwise. n and k are the number of samples and categories, respectively. It has been proved that the rank of L_A is equivalent to the number of blocks in the affinity matrix A [151]. Hence, the constraint $\text{rank}(L_A) = n - k$ enforces the obtained matrix C to form a block diagonal affinity matrix, even for noise data.

Table 7. The comparison of classical block diagonal representation SC methods.

Methods	Year	$F(C)$	$R(E)$	Constraint
BD-SSC [99]	2014	$\|C\|_1$	$\|E\|_F^2$	$X = XC + E, \text{diag}(C) = \mathbf{0}, \text{rank}(L_A) = n - k$
BD-LRR [99]	2014	$\|C\|_*$	$\|E\|_F^2$	$X = XC + E, \text{rank}(L_A) = n - k$
BDR [100]	2019	$\|C\|_{[k]}$	$\|E\|_F^2$	$X = XC + E, \text{diag}(C) = \mathbf{0}, C \geq 0, C^T = C$
rBDR [101]	2019	$\|C\|_F^2 + \|H\|_F^2 + \|B_+ - B_- V\|_F^2 + \mu\|V\|_{[k]}$	$\|E\|_{2,1}$	$X = XC + HX + E, \text{diag}(V) = \mathbf{0}, V^T = V, V \geq 0$
LBDR [102]	2022	$\frac{1}{2}\|X - X_{U_1,U_2}\|_F^2 + \mu\|C\|_{[k]}$	$\|E\|_F^2$	$X_{U_1} = X_{U_1}C + E, \text{diag}(C) = \mathbf{0}, C^T = C, C \geq 0$
ABDR [103]	2022	$\Theta(C)$	$\|E\|_F^2$	$X = XC + E$
PBDR [104]	2023	$\|C\|_{[k]} + \mu\|W_2 g(W_1\widetilde{X}) - C\|_F^2$	$\|E\|_F^2$	$\widetilde{X} = \widetilde{X}C + E, \text{diag}(C) = \mathbf{0}, C^T = C, C \geq 0$

Lu et al. [100] constructed a block diagonal matrix induced regularizer and applied it to the SC problem. The proposed block diagonal representation (BDR) SC method aims to solve the following optimization problem:

$$\min_{C,E} \ \|C\|_{[k]} + \lambda\|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, \text{diag}(C) = \mathbf{0}, C \geq 0, C^T = C, \tag{43}$$

where $\|C\|_{[k]} = \sum_{i=n-k-1}^{n} \sigma_i(L_C)$ with Laplacian matrix $L_C = \text{Diag}(C\mathbf{1}) - C$ being the lock diagonal regularizer. For the BDR model (43) without noise, it has been proved that the obtained representation coefficient matrix has the block diagonal structure under the assumption of independent subspaces. It is worth pointing out that the BDR model is a nonconvex optimization problem due to the k-block diagonal regularizer, which can be solved by the alternating minimization algorithm. Although BDR can obtain a relatively ideal block diagonal coefficient matrix for noise data, it sacrifices the convexity of the model and is more complex than SSC and LRR. In addition, the number of categories (or subspaces) in (43) must be given in advance, which is impossible in practical problems.

To cope with the situation of insufficient data samples, Zhang et al. [101] proposed a novel robust block-diagonal adaptive locality-constrained latent representation (rBDLR) method. The rBDLR can be formulated as the following optimization problem:

$$\begin{aligned} \min_{C,H,E,V,\theta} \ & \|C\|_F^2 + \|H\|_F^2 + \|B_+ - B_- V\|_F^2 + \mu\|V\|_{[k]} + \lambda\|E\|_{2,1} \\ \text{s.t.} \quad & X = XC + HX + E, \text{diag}(V) = \mathbf{0}, V^T = V, V \geq 0, \end{aligned} \tag{44}$$

where $B_+ = \left(\sqrt{\alpha}(C + \theta\mathbf{1}^T), \sqrt{\beta}HX\right)^T, B_- = \left(\sqrt{\alpha}I, \sqrt{\beta}HX\right)^T, \theta \in \mathbb{R}^{n\times 1}$ denotes the bias. $\|B_+ - B_- V\|_F^2$ means that V is approximate by $C + \theta\mathbf{1}^T$, which can avoid the overfitting. Specifically, FLRR [94] can be regarded as a special case of rBDLR with $\alpha = 0, \beta = 0$ and removing the constraints on V. By combining the latent representation with the locality-based block-diagonal regularizer, rBDLR can extract the adaptive locality-preserving salient features jointly, and the optimal solution obtained by the rBDLR model (44) has the block diagonal structure.

To capture the nonlinear structure of data, Xu et al. [102] proposed a novel latent block-diagonal representation (LBDR) subspace clustering method which integrates an autoencoder into the block-diagonal representation

$$\begin{aligned} \min_{C,U_1,U_2,E} \ & \frac{1}{2}\|X - X_{U_1,U_2}\|_F^2 + \mu\|C\|_{[k]} + \lambda\|E\|_F^2 \\ \text{s.t.} \quad & X_{U_1} = X_{U_1}C + E, \text{diag}(C) = \mathbf{0}, C^T = C, C \geq 0, \end{aligned} \tag{45}$$

where $U_1 \in \mathbb{R}^{l\times d}$ and $U_2 \in \mathbb{R}^{d\times l}$ are filter matrices, and l is the number of hidden units. $X_{U_1,U_2} \in \mathbb{R}^{d\times n}$ represents the autoencoder for data X, which is defined as $X_{U_1,U_2} = U_2\phi(U_1X)$ with an activation function $\phi(\cdot)$, and $X_{U_1} \in \mathbb{R}^{l\times n}$ represents the latent representation matrix, which is defined as $X_{U_1} = \phi(U_1X)$. Different from the traditional self-representation, LBDR performs self-representation to the reconstructed data. Note that the dimension of the reconstructed data are smaller than the original data, which can greatly reduce the calculation load. Moreover, it has proved theoretically that the LBDR model can project the original data in the nonlinear space into a new linear space.

Inspired by convex biclustering [152], Lin et al. [103] proposed an adaptive block diagonal representation (ABDR) subspace clustering method via coercively fusing both columns and rows of the coefficient matrix. The ABDR can be formulated as

$$\min_{C,E} \ \Theta(C) + \lambda \|E\|_F^2 \quad \text{s.t.} \quad X = XC + E, \tag{46}$$

where $\Theta(C) = \sum_{i,j} W_{ij} \|C_{.i} - C_{.j}\|_2 + \sum_{i,j} W_{ij} \|C_{i.} - C_{j.}\|_2$ with $W_{ij} = \iota^k_{(i,j)} \exp(\phi \|X_{.i} - X_{.j}\|_2^2)$. Here, $\iota^k_{(i,j)}$ is an indicator function whose value is 1 if $X_{.j}$ is the k-KNN of $X_{.i}$ or 0 otherwise. It has been proved that the representation coefficient matrix obtained by the noiseless ABDR has the block diagonal structure without any structure prior. Unlike the traditional BDR methods, ABDR (46) is a convex optimization problem and does not need to give the number of subspaces in advance.

Although the BDR methods mentioned above have attracted extensive attention due to the diagonal property, the high computing cost greatly limits their performance, especially for high-dimensional data. To deal with this problem, Xu et al. [104] proposed the following projective block diagonal representation (PBDR) subspace clustering method

$$\begin{aligned} \min_{C,W_1,W_2,E} \quad & \|C\|_{[k]} + \mu \|W_2 g(W_1 \widetilde{X}) - C\|_F^2 + \lambda \|E\|_F^2 \\ \text{s.t.} \quad & \widetilde{X} = \widetilde{X}C + E, \text{diag}(C) = \mathbf{0}, C^T = C, C \geq 0, \end{aligned} \tag{47}$$

where $\widetilde{X} \in \mathbb{R}^{d \times m}$ is a small subset selected from X by uniform random sampling, $W_1 \in \mathbb{R}^{h \times d}$ and $W_2 \in \mathbb{R}^{m \times h}$ are two weight matrices with h being the number of hidden units, and $g(\cdot)$ is a nonlinear activation function (such as Sigmoid, ReLU, or TanH). It is worth noting that PBDR does not directly process the original data but its subset, which can greatly reduce the dimension of data and thus reduce the computational complexity. It has proved that the optimal solution obtained by the PBDR model (47) has the block diagonal structure. In addition, the proposed PBDR model was further extended to the sparse and low-rank cases, thus improving its ability to capture the global or local structure of data. However, for PBDR, how to select the subset of original data is a difficulty.

4. Kernel SC

4.1. Single Kernel Learning Based SC

Although the previous low-rank sparse SC methods and local structure preserving SC methods have achieved some success, they are limited to the Euclidean distance (due to the linear self-expression among samples) and unable to identify the clusters with nonlinear shaped [153]. To capture the nonlinear structure inherent in many real-world datasets, kernel technology is proposed, which maps the original data to the reproducing kernel Hilbert space (RKHS) [154,155]. In recent years, kernel technology has been successfully applied to the SC problem, which aims to construct the affinity matrix by self-representation in the kernel Hilbert space. The single kernel learning (SKL) based SC methods can be formulated as the following unified framework:

$$\min_{C \in \Omega} \ \frac{1}{2} \|\phi(X) - \phi(X)C\| + F(C), \tag{48}$$

where $\phi(\cdot)$ is a feature map which maps the input data in the original feature space into the reproducing kernel Hilbert space, $F(C)$ is a penalty term (or regularization term), and Ω is the constraint set of C. The KSC model (48) can be equivalently written as

$$\min_{C\in\Omega}\ \frac{1}{2}\mathrm{tr}\Big(\phi(X)^T\phi(X)-\phi(X)^T\phi(X)C-C^T\phi(X)^T\phi(X)+C^T\phi(X)^T\phi(X)C\Big)+F(C). \quad (49)$$

Define the kernel function K with $K_{ij}=\big\langle\phi(\mathbf{x}_i), and\phi(\mathbf{x}_j)\big\rangle$, (49) can be transformed to

$$\min_{C\in\Omega}\ \frac{1}{2}\mathrm{tr}\Big(K-2KC+C^TKC\Big)+F(C). \quad (50)$$

The classical SKL-based SC methods are summarized in Table 8.

Table 8. The comparison of classical SKL-based SC methods.

Methods	Year	Brief Description	Advantage	Disadvantage
KSSC [105]	2014	SKL + SSC	It can handle the nonlinear structure of data and obtain a sparse coefficient matrix.	Its performance greatly depends on the kernel function, and it cannot capture the global structure of data.
KGS-graph [55]	2015	SKL + GS-graph	It can handle the nonlinear structure of data and group highly correlated data together.	Its performance greatly depends on the kernel function, and it cannot capture the global structure of data.
KSCBD [106]	2020	SML + BDR	It can process nonlinear data and obtain a representation coefficient matrix with block diagonal structure.	Its performance greatly depends on the kernel function, and it cannot capture the global structure of data.
ALKBDR [107]	2022	SML + BDR + LRR	It can process nonlinear data and has a stronger generalization ability to process complex data.	Its performance greatly depends on the kernel function, and it has four parameters that need to be adjusted.

Patel and Vidal [105] proposed a kernel sparse subspace clustering (KSSC) method by using the kernel trick to extend the SSC, which can be formulated as

$$\min_{C}\ \mathrm{tr}\Big(K-2KC+C^TKC\Big)+\mu\|C\|_1 \quad \text{s.t.} \quad \mathrm{diag}(C)=0, C^T\mathbf{1}=\mathbf{1}, \quad (51)$$

where K is the kernel Gram matrix which can be chosen as the polynomial or Gaussian kernels. Since KSSC can be regarded as combining the kernel technique with SSC, it can process the nonlinear structure of data and obtain a sparse coefficient matrix. Note that KSSC (51) is a convex optimization problem, so it can be solved by ADMM [117], SSNAL [118], and so forth.

Fang et al. [55] also extend the GS-graph (6) to the kernelized version (named KGS-graph). The KGS-graph aims to solve the following optimization problem:

$$\min_{C}\ \mathrm{tr}\Big(K-2KC+C^TKC\Big)+\Gamma_{\mu}^{g}(C), \quad (52)$$

where $\Gamma_{\mu}^{g}(C)=\|C\|_1+\sum_i\sum_{j<k}\max\{|C_{ij}|,|C_{ik}|\}$. As an extension of the GS-graph (6), the KGS-graph can preserve both the sparsity and locality of data simultaneously and is insensitive to noise. Moreover, it has proved that KGS-graph (52) has the grouping effect.

Yang et al. [106] presented a kernel subspace clustering method with block diagonal prior (KSCBD), which can be formulated as

$$\begin{aligned}&\min_{C,B}\ \mathrm{tr}\Big(K-2KC+C^TKC\Big)+\frac{\mu}{2}\|C-B\|_F^2\\ &\text{s.t.}\ \ B\geq 0, \mathrm{diag}(B)=0, B^T=B, \mathrm{rank}(L_B)=n-k,\end{aligned} \quad (53)$$

where K is the kernel Gram matrix, L_B is the Laplacian matrix corresponding to B, and $\text{rank}(L_B) = n - k$ indicates that the affinity matrix is an ideal graph with k clusters. Since KSCBD embeds the block diagonal prior into the kernel Hilbert space, it can not only process nonlinear data, but the resulting representation coefficient also has a block diagonal structure. However, due to the existence of $n - k$ rank constraints, the problem (53) is NP-hard, which is difficult to solve directly. Therefore, it usually uses the Ky Fans theorem to relax $n - k$ rank constraints and solve the relaxed problem.

Liu et al. [107] presented an adaptive low-rank kernel block diagonal representation (ALKBDR) subspace clustering method, which can be written as

$$\begin{aligned} \min_{\phi(X),C,B} \quad & \text{tr}\|\phi(X) - \phi(X)C\| + \frac{\mu_1}{2}\|C - B\|_F^2 + \mu_2\|B\|_{[K]} + \mu_3\|\phi(X)\|_* + \mu_4\|E\|_1 \\ \text{s.t.} \quad & B \geq 0, \text{diag}(B) = 0, B^T = B, K_G = \phi(X)^T\phi(X) + E, \end{aligned} \tag{54}$$

where K_G is a pre-defined kernel matrix. The ALKBDR model (54) can be seen as combining the kernel learning with low-rank and diagonal representation. Therefore, ALKBDR is not only able to process nonlinear data, but also has a stronger generalization ability to process complex data. However, using only a single pre-defined kernel function greatly limits its clustering performance, and there are four parameters that need to be adjusted, which is tricky and time-consuming.

4.2. Multiple Kernel Learning Based SC

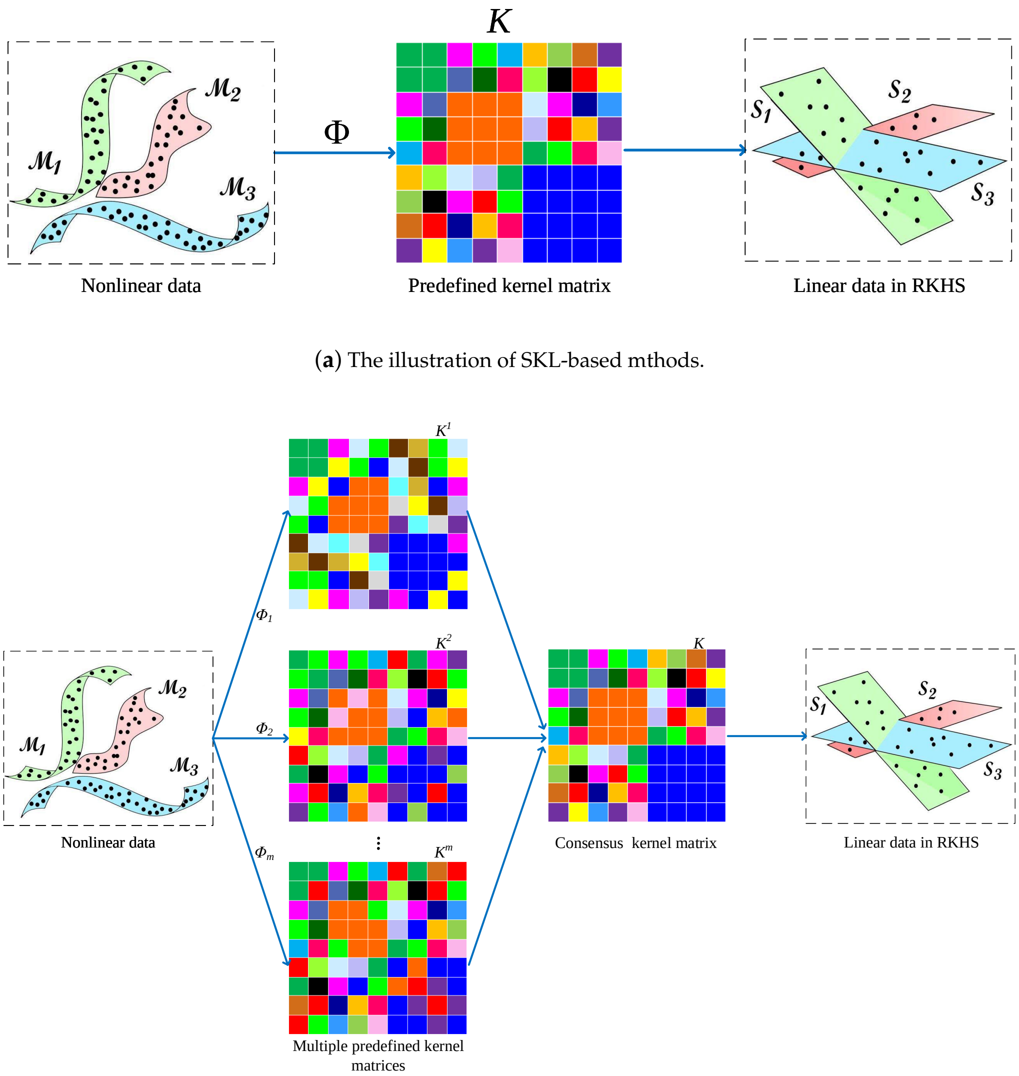
It is worth pointing out that the performance of the kernel clustering will heavily depend on the choice of the kernel function K, which is often affected by prior knowledge about the data [156]. As a result, a good choice of kernel functions is critical for kernel clustering. However, one of the main challenges with kernel methods, in general, is that it is often unclear which kernel is best for a given task. To overcome this challenge, Zhao et al. [157] proposed a multiple kernel learning (MKL) framework by adding an additional procedure to combine a group of pre-specified feature mappings into the kernel clustering. For example, each sample can be mapped to $\phi_{\mathbf{w}}(\mathbf{x}) = [w_1\phi_1(\mathbf{x})^T, w_2\phi_2(\mathbf{x})^T, \cdots w_m\phi_m(\mathbf{x})^T]^T$, where $\{\phi_i(\cdot)\}_{i=1}^m$ is a group of feature mappings and $\mathbf{w} \in \mathbb{R}^m$ specifies the coefficients of each base kernel. Then, the corresponding kernel function can be expressed as

$$\kappa_{\mathbf{w}}(\mathbf{x}_j, \mathbf{x}_k) = \phi_{\mathbf{w}}(\mathbf{x}_j)^T\phi_{\mathbf{w}}(\mathbf{x}_k)^T = \sum_{i=1}^m w_i^2\kappa_i(\mathbf{x}_j, \mathbf{x}_k). \tag{55}$$

In particular, if the pre-specified base kernel functions $\{K^i\}_{i=1}^m$ are known in advance, the kernel matrix $K_{\mathbf{w}}$ can be defined as

$$K_{\mathbf{w}} = \sum_{p=1}^m w_i^2 K^i. \tag{56}$$

Note that, in (56), the l_2-norm constraint is used on the kernel weights to avoid sparse solutions, which may cause redundancy of the pre-specified kernel function. In addition, replacing K_w with a convex combination of pre-specified kernel $\sum_{i=1}^m w_i K^i$ can avoid overfitting and obtain a sparse solution. In order to intuitively compare the SKL-based methods and MKL-based methods, Figure 8 illustrates their implementation steps. Table 9 summarizes some classical MKL-based SC methods.



(**a**) The illustration of SKL-based mthods.

(**b**) The illustration of MKL-based mthods.

Figure 8. The illustration of SKL-based methods and MKL-based methods.

Inspired by multiple kernel clustering [157,158], Kang et al. [108] proposed a joint similarity learning and clustering with multiple kernel (SCMK) method, which can be written as the following optimization problem:

$$\begin{aligned} &\min_{C,P,\mathbf{w}} \ \operatorname{tr}\left(K_{\mathbf{w}} - 2K_{\mathbf{w}}C + C^T K_{\mathbf{w}} C\right) + \mu_1 \|C\|_F^2 + \mu_2 \operatorname{tr}\left(P^T L P\right) \\ &\text{s.t. } \ C^T\mathbf{1} = \mathbf{1}, 0 \le C \le 1, P^T P = I, \sum_{i=1}^{m} \sqrt{w_i} = 1, w_i \ge 0, \end{aligned} \tag{57}$$

where $K_{\mathbf{w}} = \sum_{i=1}^{m} w_i K^i$ with a set of predefined kernels $\{K^i\}_{i=1}^{m}$, L is the Laplace matrix, and P is an indicator matrix. SCMK can be regarded as integrating the MKL, self-representation in the kernel Hilbert space, and local similarity learning in the original space into a unified framework. Therefore, it can capture both the global structure (due to the Frobenius norm regularization) and local structure (due to the graph regularization) of data and has the grouping effect. In addition, it has proved theoretical relationships to kernel k-means, k-means, and spectral clustering methods. However, SCMK (57) is a variable indivisible optimization problem with an orthogonal constraint, which is easy to fall into local solutions.

Table 9. The comparison of classical MKL-based SC methods.

Methods	Year	Brief Description	Advantage	Disadvantage
SCMK [108]	2017	MKL + F-norm penalty + graph penalty	It can capture the global structure of original data and the local structure of reconstructed data and has the grouping effect.	It is a variable indivisible optimization problem with orthogonal constraint, which is easy to fall into local solutions.
LKG [109]	2019	MKL + LRR	It can capture the global structure of the data and find a consensus kernel from the neighborhood of the candidate kernel.	It fails to fully use the structural information of data, and is sensitive to data corruption.
LLMKL [110]	2019	MKL + LRR + graph penalty	It can preserve both global structure and local structure of data, and learn a consensus kernel from the neighborhood of the predefined kernels.	The decomposition form of the kernel function may not be applicable, and it does not fully use the structure information of data.
LAKRSC [111]	2020	MKL + SR	It can learn a consensus kernel by data decomposition and is robust to noise.	The decomposition form of the kernel function may not be applicable and it does not fully use the structure information of data.
SPMKC [112]	2021	MKL + LRR + graph penalty	It can preserve both and local structure of data, and learn a consensus kernel from the neighborhood of the predefined kernels.	It is an optimization problem with orthogonal constraint which is easy to fall into local solutions, and sensitive to data corruptions.
AWLKSC [113]	2021	MKL + LRR + BDR	It can obtain a coefficient matrix with block diagonal structure, and learn a consensus kernel from the neighborhood of the predefined kernels.	The number of categories (or subspaces) must be given in advance.
SLMKC [114]	2021	MKL + AL + F-norm penalty + graph penalty	It can avoid the extra step of generating affinity matrix by representation coefficient, and preserve the local structure of data.	It is an optimization problem with orthogonal constraint which is easy to fall into local solutions, and sensitive to data corruptions.
SPAKS [115]	2022	MKL + F-norm penalty + graph penalty	It can maintain the local structure of data and divide the highly relevant data into the same groups.	It is limited in capturing the global structure of data, and the feature maps or distance matrices in RKHS need to be given in advance.
PMKSC [116]	2022	MKL + multi-view clustering + AL + SEL	It can preserve more discriminative structural information in the kernel space, and learn a consensus kernel from the neighborhood of the predefined kernels.	It does not take full the similarity among multiview data and is an optimization problem with orthogonal constraint which is easy to fall into local solutions.

Remark: AL indicates affinity matrix learning; SEL indicates Self-expressiveness learning.

The existing MKL-based SC methods usually take the linear combination of the predefined kernels as the used kernel matrix, which may be sensitive to noise and limit the representation capability. To solve this problem, a natural idea is to learn a consensus kernel function and data self-representation at the same time. Consequently, Kang et al. [109] proposed a low-rank kernel learning based graph subspace clustering method (LKG), which can be written as

$$\begin{aligned} \min_{C,K,\mathbf{w}} \ & \frac{1}{2}\mathrm{tr}\Big(K - 2KC + C^T KC\Big) + \mu_1\|C\|_* + \mu_2\|K\|_* + \mu_3\|K - \sum_{i=1}^{m} w_i K^i\|_F^2 \\ \text{s.t.} \ & C \geq 0, K \geq 0, w_i \geq 0, \sum_{i=1}^{r} w_i = 0, \end{aligned} \tag{58}$$

where $\{K^i\}_{i=1}^m$ are a set of predefined kernels, and w_i is the weight for the predefined kernel K^i. By imposing the nuclear norm regularization, the obtained kernel matrix can capture the global structure of the data. In contrast to existing methods, LKG aims to learn the low-rank kernel matrix through the similarity of the kernel matrix and find a consensus kernel from the neighborhood of the candidate kernel.

To further consider the correlation between samples, Ren et al. [110] presented a local structural graph and low-rank consensus multiple kernel learning (LLMKL) method for subspace clustering, which can be formulated as

$$\begin{aligned}\min_{C,K,B,E,\mathbf{w}} \; & \frac{1}{2}\mathrm{tr}\left[\left(I-2C+CC^T\right)B^TB\right] \\ & \qquad + \mu_1\|B\|_* + \frac{1}{2}\|K-\sum_{i=1}^{m} w_iK^i\|_F^2 + \mu_3\mathrm{tr}(D^TC) + \lambda\|E\|_1 \\ \text{s.t.} \quad & K = B^TB+E, \mathrm{diag}(C)=0, C^T\mathbf{1}=\mathbf{1}, C\geq 0, w_i \geq 0, \sum_{i=1}^{r} w_i = 1,\end{aligned} \tag{59}$$

where K is the kernel matrix that can be decomposed into B^TB and a sparse noise component E, $\{K^i\}_{i=1}^m$ are a set of predefined kernels, w_i is the weight for the predefined kernel K^i, and D is the similarity matrix of data with $D_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$. LLMKL jointly integrates the MKL technology, the global structure in the kernel space, the local structure in the original space, and the self-representation in the Hilbert space into a unified optimization model. For this reason, LLMKL can preserve both the global structure and local structure of data, and learn a consensus kernel from the neighborhood of the predefined kernels rather than using the linear combination of the predefined kernels directly, which makes it more robust to data noise.

Xue et al. [111] presented a robust subspace clustering method (LAKRSC) based on nonconvex low-rank approximation and adaptive kernel, which can be written as

$$\begin{aligned}\min_{C,B,Z} \; & \frac{1}{2}\mathrm{tr}\left[\left(I-2Z+ZZ^T\right)B^TB\right] + \mu_1\|B\|_{\mathbf{w},S_p}^p + \mu_2\|C\|_1 + \mu_3\sum_{i,j}\psi(E_{ij}) \\ \text{s.t.} \quad & K = B^TB+E, \mathbf{1}^TC=\mathbf{1}, Z = C-\mathrm{diag}(C),\end{aligned} \tag{60}$$

where $\|B\|_{\mathbf{w},S_p} = (\sum_{i=1}^n w_i\sigma_i^p)^{1/p}$ is the weighted Schatten p-norm with $w_i = \alpha/\sigma_i(B)$ for a constant α, and K is the kernel matrix which can be decomposed into B^TB and a data corruption E. $\psi(E_{ij}) = 1 - \exp\left(-E_{ij}^2/(2\varrho^2)\right)$, where ϱ is the size of the Gaussian kernel. It is worth pointing out that LAKRSC uses weighted Schatten p-norm to approximate the rank function, which can better capture the global structure of data. In addition, the kernel function is obtained by data decomposition, which can obtain a better kernel function and better handle data corruption.

Ren and Sun [112] proposed a structure-preserving multiple kernel subspace clustering method (SPMKC), which can be formulated as the following nonconvex optimization problem with an orthogonal constraint, which can be written as

$$\begin{aligned}\min_{C,K,P} \; & \frac{1}{2}\mathrm{tr}\left(K+C^TKC\right) - \mu_1\mathrm{tr}(KC) + \mu_2\mathrm{tr}(P^TLP) + \mu_3\sum_{i=1}^{m} w_i\|K-K^i\|_F^2 + \mu_4\|C\|_F^2 \\ \text{s.t.} \; & C\geq 0, C\mathbf{1}=\mathbf{1}, \mathrm{diag}(C)=0, P^TP=I,\end{aligned} \tag{61}$$

where $\{K^i\}_{i=1}^m$ are a set of predefined kernels, and the weight value w_i is used to control the contribution of ith predefined kernel K^i to K. Note that SPMKC can be regarded as integrating MKL, low-rank representation, and graph regularization into a unified framework. As a result, SPMKC can preserve both the global structure of the input data in the kernel space and the local structure of the original data, and the obtained representation coefficient has the block diagonal structure.

To construct an ideal k-block diagonal affinity matrix, Guo et al. [113] proposed an automatic weighted multiple kernel learning-based (AWLKSC) robust subspace clustering method, which can be formulated as

$$\begin{aligned} \min_{C,K,w} \ & \frac{1}{2}\mathrm{tr}\Big(K - 2KC + C^T KC\Big) + \mu_1 \|C\|_{[k]} + \mu_2 \sum_{i=1}^{m} w_i \|K - K^i\|_F^2 + \mu_3 \|K\|_{w,S_p}^p \\ \text{s.t.} \ & C \geq 0, C = C^T, \mathrm{diag}(C) = 0, \end{aligned} \tag{62}$$

where $\|C\|_{[k]} = \sum_{i=n-k-1}^{n} \sigma_i(L_C)$ with Laplacian matrix $L_C = \mathrm{Diag}(C\mathbf{1}) - C$ is the lock diagonal regularizer, and $\|\cdot\|_{w,S_p}^p$ is the weighted Schatten-p norm. The AWLKSC can be considered as integrating block diagonal constraints, MKL, and low-rank approximation. Therefore, the coefficient matrix obtained by AWLKSC (62) has the block diagonal structure and is robust to noise.

The existing SC methods first learn a coefficient matrix by data self-expressiveness, and then build the affinity graph based on the coefficient matrix. This makes the quality of the affinity matrix largely depends on the coefficient matrix, and thus the obtained affinity graph may perform poorly. To tackle this issue, Ren et al. [114] presented a simultaneous learning self-expressiveness coefficients and affinity matrix multiple kernel clustering (SLMKC) method, which is formulated as

$$\begin{aligned} \min_{C,K,A,P} \ & \mathrm{tr}\Big(K - 2KC + C^T KC\Big) + \mu_1 \Big(\|C\|_F^2 + \|A\|_F^2\Big) \\ & + \mu_2 \mathrm{tr}(CL_A C^T) + \mu_3 \mathrm{tr}(P^T L_C P) + \mu_4 \sum_{i=1}^{m} w_i \|K - K^i\|_F^2 \\ \text{s.t.} \ & A^T \mathbf{1} = \mathbf{1}, A \succeq 0, P^T P = I, \end{aligned} \tag{63}$$

where $\{K^i\}_{i=1}^m$ are a set of predefined kernels, A is the affinity matrix, L_A and L_C are the Laplace matrices of A and C, respectively, and w_i is the weight of the ith candidate kernel K^i which can take as $w_i = 1/(2\sqrt{\|K - k^i\|_F^2 + \xi})$ with ξ being infinitely close to zero. Unlike the existing MKC methods, SLMKC proposes a unified optimization model to simultaneously learn the consensus kernel, the self-expressiveness coefficient matrix, and the affinity matrix. Under this design, SLMKC can avoid the extra step of generating an affinity matrix by the representation coefficient but still maintain the correlation between the coefficient matrix and affinity matrix. Meanwhile, SLMKC can preserve the local structure of the coefficient matrix and affinity matrix due to the Laplacian regularization.

To strengthen the distinguishability of the affinity matrix, Zhang et al. [115] presented the following self-paced smooth multiple kernel subspace clustering (SPAKS) method by combining the feature smoothing regularizer with multiple kernel learning

$$\begin{aligned} \min_{C,\mathbf{w}} \ & \sum_{i=1}^{m} w_i \left(\frac{1}{2}\mathrm{tr}(K^i - 2K^i C + C^T K^i C) + \frac{\mu_1}{2}\mathrm{tr}(D^i C) + \frac{\mu_2}{2}\|C\|_F^2\right) + \mu_3 \sum_{i=1}^{m} (w_i \ln(w_i) - w_i) \\ \text{s.t.} \ & \mathrm{diag}(C) = 0, C \geq 0, C = C^T, \end{aligned} \tag{64}$$

where $\{K^i\}_{i=1}^m$ are a set of predefined kernels, $\{D^i\}_{i=1}^m$ are the distance matrix on RKHS. $\mathrm{tr}(D^i C)$ is the feature smoothing regularizer which can be regarded as a transformation of the graph Laplace regularity. Base on this design, SPAKS can enable the self-representation coefficients of data points draw from the same subspace to be closer, thus maintaining the local structure of data. In addition, due to the existence of the Frobenius norm, SPAKS can divide the highly relevant data into the same groups. However, SPAKS is limited in capturing the global structure of data, and the feature maps or distance matrices in RKHS need to be given in advance.

Inspired by projective clustering, Sun et al. [116] proposed a projective multiple kernel subspace clustering method (PMKSC), which can be written as

$$\begin{aligned} &\min_{H^i,C^i,C} \sum_{i=1}^{m}\Big(\mathrm{tr}(K^i(I - H^i {H^i}^T)) + \mu_1\|H^i - C^iH^i\|_F^2 + \mu_2\|C^i - C\|_F^2\Big) + \mu_3\|C|_F^2 \\ &\text{s.t. } {H^i}^T H^i = I, S^i \geq 0, C^i\mathbf{1} = \mathbf{1}, \mathrm{diag}(C^i) = 0, \forall\, i, \end{aligned} \tag{65}$$

where $\{K^i\}_{i=1}^m$ are a set of predefined kernels, $\{H^i\}_{i=1}^m$ are the projected kernel partition matrix, $\{C^i\}_{i=1}^m$ are the projective sub-graph corresponding to each basic kernel partition, which is used to preserve the complex relationship between each view representation, and C is the fusion graph which can be directly used as the affinity matrix for spectral clustering. Figure 9 shows the flowchart of PMKSC. By combining the kernel self-expressiveness with multiview data, PMKSC can preserve more discriminative structural information in RKHS, and be robust to data noise and redundancy. In addition, PMKSC incorporated affinity matrix learning into the unified model, avoiding the extra step of generating an affinity matrix by a representation coefficient, which helps improve clustering performance.

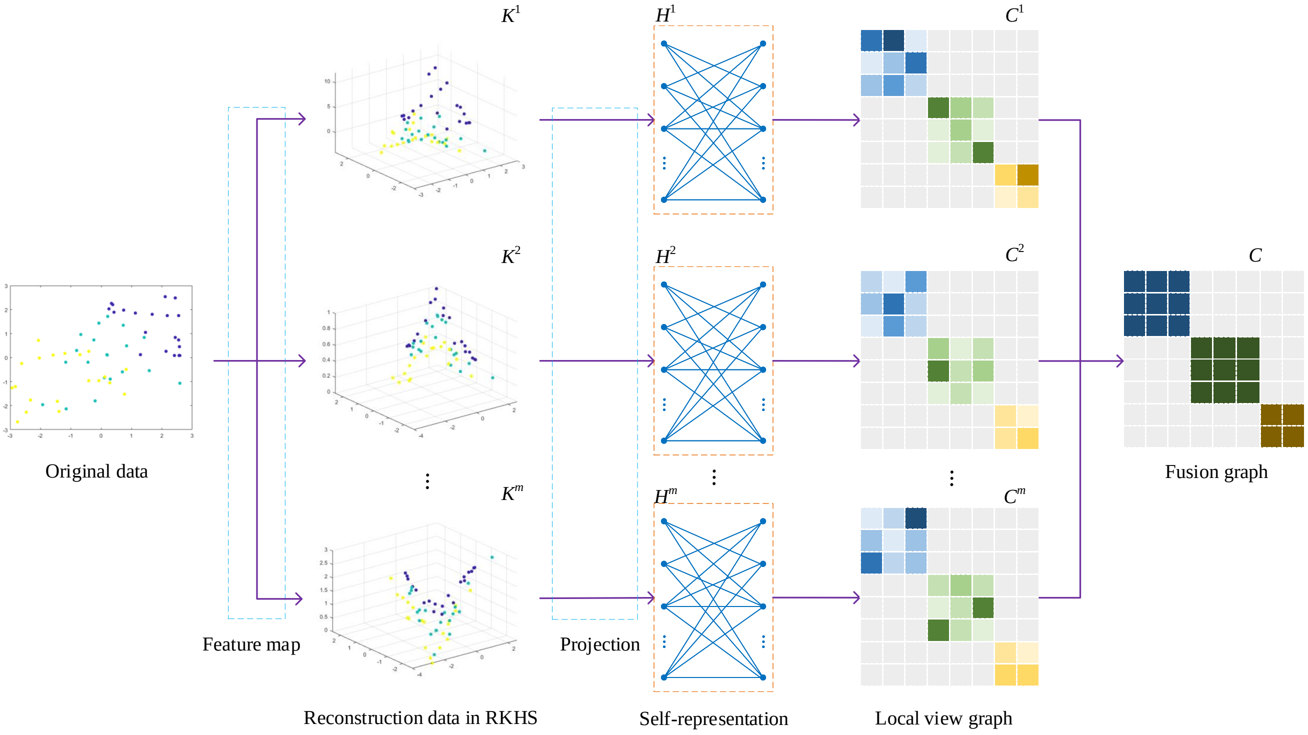


Figure 9. The flowchart of PMKSC.

5. Application

SC is an effective unsupervised learning problem for high-dimensional data mining, which has been successfully applied to computer vision, pattern recognition, and other fields, such as face recognition [25–27], motion segmentation [28–30], image processing [31–33], and speech emotion recognition [34,35]. In this section, we aim to introduce the application fields of SC in detail and introduce some benchmark datasets.

5.1. Face Recognition

In recent years, face recognition is a biological recognition technology based on human facial feature information, which has been a research hotspot in the field of computer vision and pattern recognition. It has been proved that face images under different lighting or expression transformation can be approximated by a low dimensional subspace. Specifically, a group of face images taken from multiple people can be regarded as a combination of nine-dimensional linear subspaces [22,23]. Therefore, SC has been successfully applied to face recognition. Next, we introduce two benchmark datasets for face recognition, i.e., ORL face dataset and Extended Yale B face dataset. Figures 10 and 11 show some typical images of these two face datasets, respectively.

- ORL face dataset (https://cam-orl.co.uk/facedatabase.html, accessed on 5 October 2022) consists of frontal face images collected by 40 individuals (10 samples per person) under different facial expressions (smiling or not smiling and open or close eyes) and facial details (glasses or no glasses) with 112 × 92 pixels for each image [159]. To reduce the computational cost, these images have been downsampled into 32 × 32 pixels [160], so each image can be regarded as a column vector of 1024 dimensions.
- Extended Yale B face dataset (https://paperswithcode.com/dataset/extended-yale-b-1, accessed on 5 October 2022) consists of frontal facial images collected by 38 individuals under 64 different lighting conditions with 192 × 168 pixels for each image [161]. Following [160], the original images can be downsampled into 32 × 32 pixels, so each image can be formed as a 1024-dimensional vector.

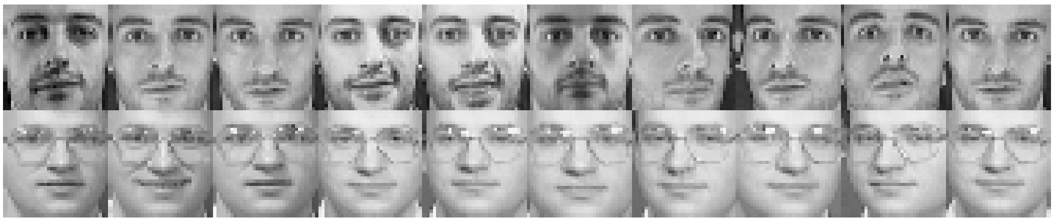


Figure 10. The typical images of the ORL face dataset.

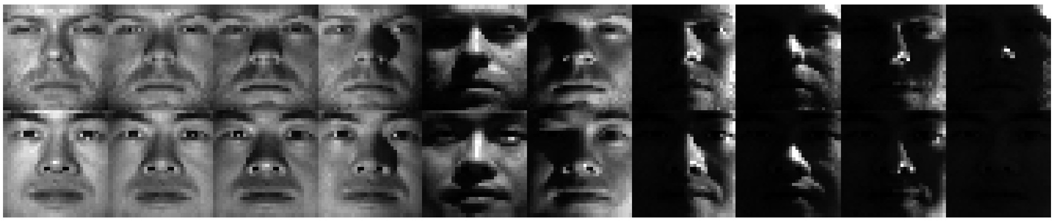


Figure 11. The typical images of the Extended YaleB face dataset.

5.2. Motion Segmentation

Motion segmentation refers to dividing the video sequence into multiple space-time regions according to the motion tracks of different rigid objects. In other words, clustering the feature points on the objects with rigid motion in the video, so that each category corresponds to an independently moving rigid object, and then the motion trajectory is obtained. In particular, the coordinates of points in the trajectory of a moving object can form a low-dimensional subspace, whose dimension is at most 4 [25].

Hopkins155 motion segmentation database (http://www.vision.jhu.edu/data/hopkins155/, accessed on 5 October 2022) is a common benchmark dataset for motion segmentation. This dataset contains 155 video sequences where 120 video sequences contain two motions, and 35 video sequences have three motions [29]. For each sequence, a tracker is used to extract the point trajectories, and the outliers are extracted manually. On average, each sequence with 2 motions has 266 feature trajectories and 30 frames, and each sequence with 3 motions has 398 feature trajectories and 29 frames. It is worth pointing out that each sequence is a sole dataset (i.e., data matrix X), so there are in total 155 SC tasks. Figure 12 show some typical images of the Hopkins155 motion segmentation database.

Figure 12. The typical images of the Hopkins155 motion segmentation database.

5.3. Handwritten Digits Clustering

The images of handwritten digits reside in the subspaces of dimension 12 [162], so the problem of clustering images of handwritten digits can be solved by the SC methods. Next, we introduce two benchmark datasets for handwritten digits clustering, i.e., the USPS dataset and MNIST dataset. Figure 13 shows some images of these two datasets.

- USPS database (http://gaussianprocess.org/gpml/data/, accessed on 5 October 2022) contains 9298 grey scale images of handwritten digits 0–9 [163]. Each digit image of USPS database has 16 × 16 pixels, so that each image can be represented by a 256-dimensional vector.
- MNIST database (http://yann.lecun.com/exdb/mnist/, accessed on 5 October 2022) contains 70,000 grey scale images of handwritten digits 0–9 [164]. Each gray image of MNIST database has 28 × 28 pixels, so each image can be vectorized into a column vector of 784 dimensions.

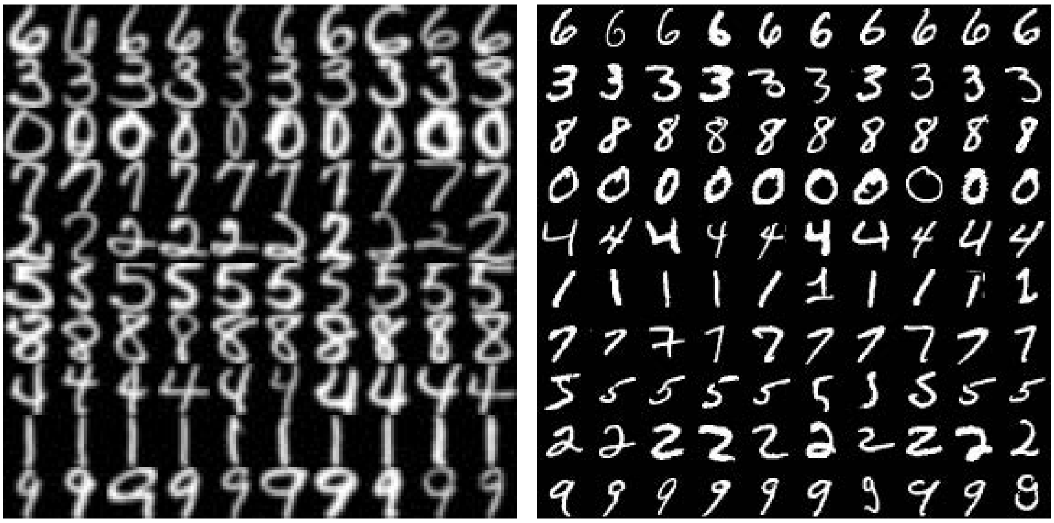


Figure 13. Handwritten digit images from the USPS database (**left**) and MNIST database (**right**).

5.4. Speech Emotion Recognition

Speech emotion recognition aims to analyze speech data and speculate the possible emotions, such as anger, disgust, fear, happiness, neutrality, sadness, and surprise. It can be implemented in the following two steps: (1) Extracting features that can effectively express the emotional content of speech; (2) Clustering the extracted features. Here, we introduce three commonly used benchmark datasets. For more information, see [165].

- The GEneva multimodal emotion portrayals (GEMEP) [166] is a French content corpus containing 1260 emotional speeches delivered by 10 professional actors (5 female) under 18 speech emotion categories. These emotional categories cover the well-known “six major” emotions, as well as the nuances of these emotions (e.g., panic, fear).
- The airplane behavior corpus (ABC) [167] is a German content corpus crafted for the special target application of public transport surveillance. ABC contains 430 corpora delivered by 8 German speakers (4 female) under 6 speech emotion categories (aggressive, cheerful, intoxicated, nervous, neutral, and tired). The numbers of samples of six emotions are 95, 105, 33, 93, 79, and 25, respectively.

- The eNTERFACE corpus [166] is an open English audio-visual emotion database. It consists of 1277 corpora made by 42 English speakers (8 female) from 14 countries under six basic emotions (anger, disgust, fear, happiness, sadness, and surprise). The numbers of samples of six emotions are 215, 215, 215, 207, 210, and 215, respectively.

6. Research Prospect for Subspace Clustering

The emergence of high-dimensional data has promoted the development of traditional machine learning. Although subspace clustering has been developed to varying degrees in many practical applications, there are still several directions to be solved in the future.

6.1. Deep Subspace Clustering

With the development of deep learning technology in recent years, it has been widely used in various fields due to its ability to effectively explore the deep features of data. Deep subspace clustering (DSC) has emerged, which performs subspace clustering based on the low-dimensional features of the original data learned by deep learning techniques.

At present, convolutional neural network [168–171], generative adversarial network [172–175], deep auto-encoder [176–179], and other mainstream deep learning methods have been successfully applied to subspace clustering, and achieved good clustering performance in many fields. Although the DSC methods have improved the clustering accuracy to a certain extent, how to effectively mine the subspace structure inside the data and obtain a more robust data representation still needs further research. In addition, the DSC methods are limited by computationally time-consuming and large memory consumption, so how to quickly solve the DSC needs to be urgently settled.

6.2. Data Outliers Detection

As Hampel et al. [180] pointed out, a routine dataset may contain about 1–10% (or more) of outliers, while for high-dimensional data, this proportion may be larger. In the process of data processing, data outliers have a high impact on the performance of models, and affect the generalization ability of models, even leading to the model being unable to extract effective features. Therefore, it is essential to identify and separate data outliers.

At present, statistical researchers often use robust regression (such as least trimmed squares regression [181], Huber regression [182], and least absolute deviation [183]) or mean-shift model [184] to solve this problem. Up to now, a subspace clustering method with outlier detection has not been found. Consequently, how to combine the subspace clustering with data outliers detection is meaningful research.

6.3. Tuning Parameters Selection

In the subspace clustering methods, there exists exists some tuning parameters, especially in the kernel subspace clustering methods. A good tuning parameter can often determine the clustering performance of the these methods. At present, grid search (such as cross validation), the Bayesian method, or the empirical method are usually used to adjust the tuning parameters; however, this is a time-consuming process. Hence, it is interesting to study how to choose the regulating parameters.

6.4. Kernel Functions Selection

As mentioned earlier, the performance of kernel subspace clustering methods depends heavily on the choice of kernel functions. Therefore, a good choice of kernel functions is crucial for kernel subspace clustering. However, a major challenge of kernel methods is that it is usually unclear which kernel is optimal. Although many kernel function construction methods or multiple kernel learning methods have been proposed, this problem has still not been well resolved. As a consequence, studying the selection of kernel functions is also a promising direction.

6.5. Theory Analysis

According to the previous description, it can be found that most of the subspace clustering methods lack theoretical support, and only a small part of them have investigated the block diagonal property of the self-representation coefficient matrix and grouping effect. In addition, no articles studying the statistical properties of the model (for example, the error bound theory and consistency) and the clustering recovery theory have been found. Studying the theoretical properties of the models can help us understand and analyze them, and thus better apply them to practical problems. Therefore, it is very meaningful to further study the theoretical properties of subspace clustering methods.

7. Conclusions

In this survey, we have reviewed the development of SC methods in the past two decades. According to the strategy of the constructing representation coefficient, we divide the classical SC methods into three categories, i.e., low-rank sparse SC methods, local structure preserving SC methods, and kernel SC methods. Among them, the low-rank sparse SC methods can capture the global structure of data and achieve subspace feature selection by low-rank representation and sparse representation. Local structure preserving SC methods can better capture the geometric information of data. Moreover, kernel SC methods can not only capture the geometric information of data, but also cope with the challenges brought by the nonlinearity of data. Then, the application fields of SC and the commonly used benchmark datasets are introduced. Finally, we have discussed several interesting and meaningful future research directions.

Author Contributions: Conceptualization, W.Q., X.X., H.C. and L.K.; methodology, W.Q., X.X., H.C. and L.K.; validation, X.X. and L.K.; formal analysis, W.Q., H.C. and L.K.; investigation, W.Q., H.C.; writing—original draft preparation, W.Q., X.X., H.C. and L.K.; writing—review and editing, W.Q., X.X., H.C. and L.K.; visualization, W.Q.; supervision, X.X. and H.C.; project administration, X.X. and L.K.; funding acquisition, L.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 12071022 and Grant 12001019; and in part by the Project Funded by China Postdoctoral Science Foundation under Grant 2022M723327.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Driver, H.E.; Kroeber, A.L. *Quantitative Expression of Cultural Relationships*; University of California Press: Berkeley, CA, USA, 1932; Volume 31.
2. Zubin, J. A technique for measuring like-mindedness. *J. Abnorm. Soc. Psychol.* **1938**, *33*, 508. [CrossRef]
3. Cattell, R.B. The description of personality: Basic traits resolved into clusters. *J. Abnorm. Soc. Psychol.* **1943**, *38*, 476. [CrossRef]
4. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 27 December 1965–7 January 1966; University of California Press: Berkeley, CA, USA , 1967; pp. 281–297.
5. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [CrossRef]
6. Kodinariya, T.M.; Makwana, P.R. Review on determining number of cluster in *k*-means clustering. *Int. J. Adv. Res. Comput. Sci. Manag. Stud.* **2013**, *1*, 90–95.
7. Bai, L.; Liang, J.; Cao, F. A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters. *Inf. Fusion* **2020**, *61*, 36–47. [CrossRef]
8. Donath, W.E.; Hoffman, A.J. Lower bounds for the partitioning of graphs. *IBM J. Res. Dev.* **1973**, *17*, 420–425. [CrossRef]
9. Fiedler, M. Algebraic connectivity of graphs. *Czechoslov. Math. J.* **1973**, *23*, 298–305. [CrossRef]
10. Ng, A.; Jordan, M.; Weiss, Y. On spectral clustering: Analysis and an algorithm. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 3–8 December 2001; pp. 849–856.

11. Janani, R.; Vijayarani, S. Text document clustering using spectral clustering algorithm with particle swarm optimization. *Expert Syst. Appl.* **2019**, *134*, 192–200. [CrossRef]
12. Fraley, C.; Raftery, A.E. Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.* **2002**, *97*, 611–631. [CrossRef]
13. Handcock, M.S.; Raftery, A.E.; Tantrum, J.M. Model-based clustering for social networks. *J. R. Stat. Soc. Ser. A Stat. Soc.* **2007**, *170*, 301–354. [CrossRef]
14. Bouveyron, C.; Celeux, G.; Murphy, T.B.; Raftery, A.E. *Model-Based Clustering and Classification for Data Science: With Applications in R*; Cambridge University Press: Cambridge, UK, 2019.
15. Johnson, S.C. Hierarchical clustering schemes. *Psychometrika* **1967**, *32*, 241–254. [CrossRef]
16. Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: An overview, II. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2017**, *7*, e1219. [CrossRef]
17. Jafarzadegan, M.; Safi-Esfahani, F.; Beheshti, Z. Combining hierarchical clustering approaches using the PCA method. *Expert Syst. Appl.* **2019**, *137*, 1–10. [CrossRef]
18. Comito, C.; Pizzuti, C.; Procopio, N. Online clustering for topic detection in social data streams. In Proceedings of the IEEE 28th International Conference on Tools with Artificial Intelligence, San Jose, CA, USA, 6–8 November 2016; pp. 362–369.
19. Comito, C. How covid-19 information spread in us the role of twitter as early indicator of epidemics. *IEEE Trans. Serv. Comput.* **2022**, *15*, 1193–1205. [CrossRef]
20. Bellman, R. Dynamic programming and Lagrange multipliers. *Nat. Acad. Sci.* **1956**, *42*, 767–769. [CrossRef]
21. Muja, M.; Lowe, D.G. Scalable nearest neighbor algorithms for high dimensional data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 2227–2240. [CrossRef] [PubMed]
22. Basri, R.; Jacobs, D.W. Lambertian reflectance and linear subspaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 218–233. [CrossRef]
23. Lee, K.C.; Ho, J.; Kriegman, D.J. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 684–698.
24. Agrawal, R.; Gehrke, J.; Gunopulos, D.; Raghavan, P. Automatic subspace clustering of high dimensional data for data mining applications. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Seattle, WA, USA, 1–4 June 1998; pp. 94–105.
25. Elhamifar, E.; Vidal, R. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2765–2781. [CrossRef]
26. Lai, Z.; Mo, D.; Wen, J.; Shen, L.; Wong, W.K. Generalized robust regression for jointly sparse subspace learning. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 756–772. [CrossRef]
27. Liao, M.; Gu, X. Face recognition approach by subspace extended sparse representation and discriminative feature learning. *Neurocomputing* **2020**, *373*, 35–49. [CrossRef]
28. Tomasi, C.; Kanade, T. Shape and motion from image streams under orthography: A factorization method. *Int. J. Comput. Vis.* **1992**, *9*, 137–154. [CrossRef]
29. Tron, R.; Vidal, R. A benchmark for the comparison of 3-d motion segmentation algorithms. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
30. Xiao, T.; Liu, P.; Zhao, W.; Liu, H.; Tang, X. Structure preservation and distribution alignment in discriminative transfer subspace learning. *Neurocomputing* **2019**, *337*, 218–234. [CrossRef]
31. Hong, W.; Wright, J.; Huang, K.; Ma, Y. Multiscale hybrid linear models for lossy image representation. *IEEE Trans. Image Process.* **2006**, *15*, 3655–3671. [CrossRef] [PubMed]
32. Yang, A.Y.; Wright, J.; Ma, Y.; Sastry, S.S. Unsupervised segmentation of natural images via lossy data compression. *Comput. Vis. Image. Underst.* **2008**, *110*, 212–225. [CrossRef]
33. Cai, Y.; Zhang, Z.; Cai, Z.; Liu, X.; Jiang, X.; Yan, Q. Graph convolutional subspace clustering: A robust subspace clustering framework for hyperspectral image. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4191–4202. [CrossRef]
34. Song, P.; Zheng, W. Feature selection based transfer subspace learning for speech emotion recognition. *IEEE Trans. Affect. Comput.* **2018**, *11*, 373–382. [CrossRef]
35. Zhang, W.; Song, P. Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 307–318. [CrossRef]
36. Günnemann, S.; Boden, B.; Seidl, T. Finding density-based subspace clusters in graphs with feature vectors. *Data Min. Knowl. Discov.* **2012**, *25*, 243–269. [CrossRef]
37. Chen, Y.; Jalali, A.; Sanghavi, S.; Xu, H. Clustering partially observed graphs via convex optimization. *J. Mach. Learn. Res.* **2014**, *15*, 2213–2238.
38. Liu, G.; Lin, Z.; Yu, Y. Robust subspace segmentation by low-rank representation. *IEEE Trans. Cybern.* **2010**, *44*, 663–670.
39. Vidal, R. Subspace clustering. *IEEE Signal Process. Mag.* **2011**, *28*, 52–68. [CrossRef]
40. Boult, T.E.; Brown, L.G. Factorization-based segmentation of motions. In Proceedings of the IEEE Workshop on Visual Motion, Princeton, NJ, USA, 7–9 October 1991; pp. 179–180.
41. Costeira, J.P.; Kanade, T. A multibody factorization method for independently moving objects. *Int. J. Comput. Vis.* **1998**, *29*, 159–179. [CrossRef]

42. Vidal, R.; Ma, Y.; Sastry, S. Generalized principal component analysis (GPCA). *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1945–1959. [CrossRef] [PubMed]
43. Tsakiris, M.C.; Vidal, R. Algebraic clustering of affine subspaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 482–489. [CrossRef]
44. Tseng, P. Nearest q-flat to m points. *J. Optim. Theory Appl.* **2000**, *105*, 249–252. [CrossRef]
45. Ho, J.; Yang, M.H.; Lim, J.; Lee, K.C.; Kriegman, D. Clustering appearances of objects under varying illumination conditions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Madison, WI, USA, 18–20 June 2003; pp. 1–8.
46. Rodrigues, É.O.; Torok, L.; Liatsis, P.; Viterbo, J.; Conci, A. k-MS: A novel clustering algorithm based on morphological reconstruction. *Pattern Recognit.* **2017**, *66*, 392–403. [CrossRef]
47. Tipping, M.E.; Bishop, C.M. Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1999**, *61*, 611–622. [CrossRef]
48. Adler, A.; Elad, M.; Hel-Or, Y. Probabilistic subspace clustering via sparse representations. *IEEE Signal Process. Lett.* **2012**, *20*, 63–66. [CrossRef]
49. Gruber, A.; Weiss, Y. Multibody factorization with uncertainty and missing data using the EM algorithm. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 27 June–2 July 2004; pp. 707–714.
50. Fischler, M.A.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]
51. Eldar, Y.C.; Mishali, M. Robust recovery of signals from a structured union of subspaces. *IEEE Trans. Inf. Theory* **2009**, *55*, 5302–5316. [CrossRef]
52. Elhamifar, E.; Vidal, R. Sparse subspace clustering. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 2790–2797.
53. Wang, S.; Yuan, X.; Yao, T.; Yan, S.; Shen, J. Efficient subspace segmentation via quadratic programming. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 7–11 August 2011; pp. 519–524.
54. Pham, D.S.; Budhaditya, S.; Phung, D.; Venkatesh, S. Improved subspace clustering via exploitation of spatial constraints. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 550–557.
55. Fang, Y.; Wang, R.; Dai, B.; Wu, X. Graph-based learning via auto-grouped sparse regularization and kernelized extension. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 142–154. [CrossRef]
56. Yang, Y.; Feng, J.; Jojic, N.; Yang, J.; Huang, T.S. ℓ_0-sparse subspace clustering. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 731–747.
57. Dong, W.; Wu, X.J.; Kittler, J.; Yin, H.F. Sparse subspace clustering via nonconvex approximation. *Pattern Anal. Appl.* **2019**, *22*, 165–176. [CrossRef]
58. Dong, W.; Wu, X.j.; Kittler, J. Sparse subspace clustering via smoothed l_p minimization. *Pattern Recognit. Lett.* **2019**, *125*, 206–211. [CrossRef]
59. Wang, P.; Han, B.; Li, J.; Gao, X. Structural reweight sparse subspace clustering. *Neural Process. Lett.* **2019**, *49*, 965–977. [CrossRef]
60. Bai, L.; Liang, J. Sparse subspace clustering with entropy-norm. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 13–18 July 2020; pp. 561–568.
61. Zheng, R.; Liang, Z.; Chen, X.; Tian, Y.; Cao, C.; Li, M. An adaptive sparse subspace clustering for cell type identification. *Front. Genet.* **2020**, *11*, 407. [CrossRef]
62. Liu, G.; Lin, Z.; Yan, S.; Sun, J.; Yu, Y.; Ma, Y. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 171–184. [CrossRef]
63. Ni, Y.; Sun, J.; Yuan, X.; Yan, S.; Cheong, L.F. Robust low-rank subspace segmentation with semidefinite guarantees. In Proceedings of the IEEE International Conference on Data Mining Workshops, Sydney, Australia, 13–17 December 2010; pp. 1179–1188.
64. 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 Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 347–360.
65. Lu, C.; Feng, J.; Lin, Z.; Yan, S. Correlation adaptive subspace segmentation by trace lasso. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1345–1352.
66. Jiang, W.; Liu, J.; Qi, H.; Dai, Q. Robust subspace segmentation via nonconvex low rank representation. *Inform. Sci.* **2016**, *340*, 144–158. [CrossRef]
67. Zhang, X.; Xu, C.; Sun, X.; Baciu, G. Schatten-q regularizer constrained low rank subspace clustering model. *Neurocomputing* **2016**, *182*, 36–47. [CrossRef]
68. Chen, J.; Mao, H.; Sang, Y.; Yi, Z. Subspace clustering using a symmetric low-rank representation. *Knowl. Based Syst.* **2017**, *127*, 46–57. [CrossRef]
69. Wang, W.; Zhang, B.; Feng, X. Subspace segmentation by correlation adaptive regression. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 2612–2621.
70. Zhang, H.; Yang, J.; Shang, F.; Gong, C.; Zhang, Z. LRR for subspace segmentation via tractable schatten-p norm minimization and factorization. *IEEE Trans. Cybern.* **2019**, *49*, 1722–1734. [CrossRef] [PubMed]

71. Xu, J.; Yu, M.; Shao, L.; Zuo, W.; Meng, D.; Zhang, L.; Zhang, D. Scaled simplex representation for subspace clustering. *IEEE Trans. Cybern.* **2021**, *51*, 1493–1505. [CrossRef] [PubMed]
72. Shen, Q.; Chen, Y.; Liang, Y.; Yi, S.; Liu, W. Weighted Schatten p-norm minimization with logarithmic constraint for subspace clustering. *Signal Process.* **2022**, *198*, 108568. [CrossRef]
73. Luo, D.; Nie, F.; Ding, C.; Huang, H. Multi-subspace representation and discovery. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Athens, Greece, 5–9 September 2011; pp. 405–420.
74. Zhuang, L.; Gao, H.; Lin, Z.; Ma, Y.; Zhang, X.; Yu, N. Non-negative low rank and sparse graph for semi-supervised learning. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2328–2335.
75. Wang, Y.X.; Xu, H.; Leng, C. Provable subspace clustering: When LRR meets SSC. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 64–72.
76. Zheng, Y.; Zhang, X.; Yang, S.; Jiao, L. Low-rank representation with local constraint for graph construction. *Neurocomputing* **2013**, *122*, 398–405. [CrossRef]
77. Wang, J.; Shi, D.; Cheng, D.; Zhang, Y.; Gao, J. LRSR: Low-rank-sparse representation for subspace clustering. *Neurocomputing* **2016**, *214*, 1026–1037. [CrossRef]
78. You, C.; Li, C.G.; Robinson, D.P.; Vidal, R. Oracle based active set algorithm for scalable elastic net subspace clustering. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3928–3937.
79. Wang, W.; Wu, C. Image segmentation by correlation adaptive weighted regression. *Neurocomputing* **2017**, *267*, 426–435. [CrossRef]
80. Brbić, M.; Kopriva, I. ℓ_0-motivated low-rank sparse subspace clustering. *IEEE Trans. Cybern.* **2020**, *50*, 1711–1725. [CrossRef]
81. Zhong, G.; Pun, C.M. Subspace clustering by simultaneously feature selection and similarity learning. *Knowl. Based Syst.* **2020**, *193*, 105512. [CrossRef]
82. Zhai, H.; Zhang, H.; Zhang, L.; Li, P. Nonlocal means regularized sketched reweighted sparse and low-rank subspace clustering for large hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4164–4178. [CrossRef]
83. Yang, T.; Zhou, S.; Zhang, Z. The k-sparse LSR for subspace clustering via 0-1 integer programming. *Signal Process.* **2022**, *199*, 108622. [CrossRef]
84. Lu, X.; Wang, Y.; Yuan, Y. Graph-regularized low-rank representation for destriping of hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 4009–4018. [CrossRef]
85. Liu, J.; Chen, Y.; Zhang, J.; Xu, Z. Enhancing low-rank subspace clustering by manifold regularization. *IEEE Trans. Image Process.* **2014**, *23*, 4022–4030. [CrossRef]
86. Chen, W.; Zhang, E.; Zhang, Z. A Laplacian structured representation model in subspace clustering for enhanced motion segmentation. *Neurocomputing* **2016**, *208*, 174–182. [CrossRef]
87. Yin, M.; Gao, J.; Lin, Z. Laplacian regularized low-rank representation and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 504–517. [CrossRef]
88. Du, S.; Ma, Y.; Ma, Y. Graph regularized compact low rank representation for subspace clustering. *Knowl. Based Syst.* **2017**, *118*, 56–69. [CrossRef]
89. Wang, J.; Liu, J.X.; Zheng, C.H.; Wang, Y.X.; Kong, X.Z.; Wen, C.G. A mixed-norm Laplacian regularized low-rank representation method for tumor samples clustering. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2019**, *16*, 172–182. [CrossRef]
90. Kang, Z.; Lin, Z.; Zhu, X.; Xu, W. Structured graph learning for scalable subspace clustering: From single view to multiview. *IEEE Trans. Cybern.* **2022**, *52*, 8976–8986. [CrossRef]
91. Francis, J.; Baburaj, M.; George, S.N. An $l_{1/2}$ and graph regularized subspace clustering method for robust image segmentation. *ACM Trans. Multimedia Comput. Commun. Appl.* **2022**, *18*, 1–24. [CrossRef]
92. Liu, G.; Yan, S. Latent low-rank representation for subspace segmentation and feature extraction. In Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1615–1622.
93. Zhang, Z.; Yan, S.; Zhao, M. Similarity preserving low-rank representation for enhanced data representation and effective subspace learning. *Neural Netw.* **2014**, *53*, 81–94. [CrossRef]
94. Yu, S.; Yiquan, W. Subspace clustering based on latent low rank representation with Frobenius norm minimization. *Neurocomputing* **2018**, *275*, 2479–2489. [CrossRef]
95. Zhang, Z.; Wang, L.; Li, S.; Wang, Y.; Zhang, Z.; Zha, Z.; Wang, M. Adaptive structure-constrained robust latent low-rank coding for image recovery. In Proceedings of the IEEE International Conference on Data Mining, Beijing, China, 8–11 November 2019; pp. 846–855.
96. Sun, W.; Peng, J.; Yang, G.; Du, Q. Fast and latent low-rank subspace clustering for hyperspectral band selection. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 3906–3915. [CrossRef]
97. Wu, M.; Wang, S.; Li, Z.; Zhang, L.; Wang, L.; Ren, Z. Joint latent low-rank and non-negative induced sparse representation for face recognition. *Appl. Intell.* **2021**, *51*, 8349–8364. [CrossRef]
98. Fu, Z.; Zhao, Y.; Chang, D.; Wang, Y.; Wen, J. Latent low-rank representation with weighted distance penalty for clustering. *IEEE Trans. Cybern.* **2022**. [CrossRef]

99. Feng, J.; Lin, Z.; Xu, H.; Yan, S. Robust subspace segmentation with block-diagonal prior. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3818–3825.
100. Lu, C.; Feng, J.; Lin, Z.; Mei, T.; Yan, S. Subspace clustering by block diagonal representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 487–501. [CrossRef]
101. Zhang, Z.; Ren, J.; Li, S.; Hong, R.; Zha, Z.; Wang, M. Robust subspace discovery by block-diagonal adaptive locality-constrained representation. In Proceedings of the ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1569–1577.
102. Xu, Y.; Chen, S.; Li, J.; Han, Z.; Yang, J. Autoencoder-based latent block-diagonal representation for subspace clustering. *IEEE Trans. Cybern.* **2022**, *52*, 5408–5418. [CrossRef]
103. Lin, Y.; Chen, S. Convex subspace clustering by adaptive block diagonal representation. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**. [CrossRef]
104. Xu, Y.; Chen, S.; Li, J.; Xu, C.; Yang, J. Fast subspace clustering by learning projective block diagonal representation. *Pattern Recognit.* **2023**, *135*, 109152. [CrossRef]
105. Patel, V.M.; Vidal, R. Kernel sparse subspace clustering. In Proceedings of the IEEE International Conference on Image Processing, Paris, France, 27–30 October 2014; pp. 2849–2853.
106. Yang, Y.; Wang, T. Kernel subspace clustering with block diagonal prior. In Proceedings of the International Conference on Machine Learning, Big Data and Business Intelligence, Chengdu, China, 23–25 October 2020; pp. 367–370.
107. Liu, M.; Wang, Y.; Sun, J.; Ji, Z. Adaptive low-rank kernel block diagonal representation subspace clustering. *Appl. Intell.* **2022**, *52*, 2301–2316. [CrossRef]
108. Kang, Z.; Peng, C.; Cheng, Q. Twin learning for similarity and clustering: A unified kernel approach. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 2080–2086.
109. Kang, Z.; Wen, L.; Chen, W.; Xu, Z. Low-rank kernel learning for graph-based clustering. *Knowl. Based Syst.* **2019**, *163*, 510–517. [CrossRef]
110. Ren, Z.; Li, H.; Yang, C.; Sun, Q. Multiple kernel subspace clustering with local structural graph and low-rank consensus kernel learning. *Knowl. Based Syst.* **2020**, *188*, 105040.
111. Xue, X.; Zhang, X.; Feng, X.; Sun, H.; Chen, W.; Liu, Z. Robust subspace clustering based on non-convex low-rank approximation and adaptive kernel. *Inf. Sci.* **2020**, *513*, 190–205. [CrossRef]
112. Ren, Z.; Sun, Q. Simultaneous global and local graph structure preserving for multiple kernel clustering. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 1839–1851. [CrossRef]
113. Guo, L.; Zhang, X.; Liu, Z.; Xue, X.; Wang, Q.; Zheng, S. Robust subspace clustering based on automatic weighted multiple kernel learning. *Inf. Sci.* **2021**, *573*, 453–474. [CrossRef]
114. Ren, Z.; Lei, H.; Sun, Q.; Yang, C. Simultaneous learning coefficient matrix and affinity graph for multiple kernel clustering. *Inf. Sci.* **2021**, *547*, 289–306.
115. Zhang, Q.; Kang, Z.; Xu, Z.; Huang, S.; Fu, H. Spaks: Self-paced multiple kernel subspace clustering with feature smoothing regularization. *Knowl. Based Syst.* **2022**, *253*, 109500. [CrossRef]
116. Sun, M.; Wang, S.; Zhang, P.; Liu, X.; Guo, X.; Zhou, S.; Zhu, E. Projective multiple kernel subspace clustering. *IEEE Trans. Multimedia* **2022**, *24*, 2567–2579. [CrossRef]
117. Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **2011**, *3*, 1–122. [CrossRef]
118. Li, X.; Sun, D.; Toh, K.C. A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems. *SIAM J. Optim.* **2018**, *28*, 433–458. [CrossRef]
119. Elhamifar, E.; Vidal, R. Clustering disjoint subspaces via sparse representation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 1926–1929.
120. Soltanolkotabi, M.; Candès, E.J. A geometric analysis of subspace clustering with outliers. *Ann. Stat.* **2012**, *40*, 2195–2238. [CrossRef]
121. Soltanolkotabi, M.; Elhamifar, E.; Candès, E.J. Robust subspace clustering. *Ann. Stat.* **2014**, *42*, 669–699. [CrossRef]
122. Wang, Y.X.; Xu, H. Noisy sparse subspace clustering. *J. Mach. Learn. Res.* **2016**, *17*, 320–360.
123. Nasihatkon, B.; Hartley, R. Graph connectivity in sparse subspace clustering. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 2137–2144.
124. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 301–320. [CrossRef]
125. Bondell, H.D.; Reich, B.J. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **2008**, *64*, 115–123. [CrossRef]
126. Pourkamali-Anaraki, F.; Folberth, J.; Becker, S. Efficient solvers for sparse subspace clustering. *Signal Process.* **2020**, *172*, 107548. [CrossRef]
127. Chen, H.; Kong, L.; Li, Y. Nonconvex clustering via ℓ_0 fusion penalized regression. *Pattern Recognit.* **2022**, *128*, 108689. [CrossRef]
128. Lin, Z.; Chen, M.; Ma, Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv* **2010**, arXiv:1009.5055.
129. Toh, K.C.; Yun, S. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pac. J. Optim.* **2010**, *6*, 615–640.

130. Lin, Z.; Liu, R.; Su, Z. Linearized alternating direction method with adaptive penalty for low-rank representation. In Proceedings of the Advances in Neural Information Processing Systems, Granada, Spain, 12–14 December 2011; pp. 612–620.
131. Favaro, P.; Vidal, R.; Ravichandran, A. A closed form solution to robust subspace estimation and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 1801–1807.
132. Zhang, H.; Yi, Z.; Peng, X. fLRR: Fast low-rank representation using Frobenius-norm. *Electron. Lett.* **2014**, *50*, 936–938. [CrossRef]
133. Peng, X.; Yu, Z.; Yi, Z.; Tang, H. Constructing the L2-graph for robust subspace learning and subspace clustering. *IEEE Trans. Cybern.* **2017**, *47*, 1053–1066. [CrossRef]
134. Peng, X.; Lu, C.; Yi, Z.; Tang, H. Connections between nuclear-norm and frobenius-norm-based representations. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 218–224. [CrossRef] [PubMed]
135. Zuo, W.; Meng, D.; Zhang, L.; Feng, X.; Zhang, D. A generalized iterated shrinkage algorithm for non-convex sparse coding. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 217–224.
136. Magnússon, S.; Weeraddana, P.C.; Rabbat, M.G.; Fischione, C. On the convergence of alternating direction lagrangian methods for nonconvex structured optimization problems. *IEEE Trans. Control Netw. Syst.* **2016**, *3*, 296–309. [CrossRef]
137. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791. [CrossRef] [PubMed]
138. Yu, Y.L. Better approximation and faster algorithm using the proximal average. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 458–466.
139. Yu, Y.; Zheng, X.; Marchetti-Bowick, M.; Xing, E. Minimizing nonconvex non-separable functions. In Proceedings of the Artificial Intelligence and Statistics, San Diego, CA, USA, 9–12 May 2015; pp. 1107–1115.
140. Wang, Y.; Yin, W.; Zeng, J. Global convergence of ADMM in nonconvex nonsmooth optimization. *J. Sci. Comput.* **2019**, *78*, 29–63. [CrossRef]
141. Garey, M.R.; Johnson, D.S. *Computers and Intractability: A Guide to the Theory of NP-Completeness*; W. H. Freeman: New York, NY, USA, 1979.
142. Wu, B.; Ghanem, B. ℓ_p-box ADMM: A versatile framework for integer programming. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1695–1708. [CrossRef] [PubMed]
143. Belkin, M.; Niyogi, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 3–8 December 2001; pp. 585–591.
144. Zhou, D.; Huang, J.; Schölkopf, B. Learning with hypergraphs: Clustering, classification, and embedding. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 4–7 December 2006; pp. 1601–1608.
145. Wang, Y.; Zhang, W.; Wu, L.; Lin, X.; Fang, M.; Pan, S. Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering. In Proceedings of the International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016; pp. 2153–2159.
146. Zhang, H.; Lin, Z.; Zhang, C. A counterexample for the validity of using nuclear norm as a convex surrogate of rank. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Prague, Czech Republic, 23–27 September 2013; pp. 226–241.
147. Cai, D.; He, X.; Han, J.; Huang, T.S. Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1548–1560.
148. Zheng, M.; Bu, J.; Chen, C.; Wang, C.; Zhang, L.; Qiu, G.; Cai, D. Graph regularized sparse coding for image representation. *IEEE Trans. Image Process.* **2011**, *20*, 1327–1336. [CrossRef]
149. Slepcev, D.; Thorpe, M. Analysis of p-Laplacian regularization in semisupervised learning. *SIAM J. Math. Anal.* **2019**, *51*, 2085–2120. [CrossRef]
150. Tang, K.; Xu, K.; Jiang, W.; Su, Z.; Sun, X.; Luo, X. Selecting the best part from multiple Laplacian autoencoders for multi-view subspace clustering. *IEEE Trans. Knowl. Data. Eng.* **2022**. [CrossRef]
151. Von Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **2007**, *17*, 395–416. [CrossRef]
152. Chi, E.C.; Allen, G.I.; Baraniuk, R.G. Convex biclustering. *Biometrics* **2017**, *73*, 10–19. [CrossRef] [PubMed]
153. Schölkopf, B.; Smola, A.; Müller, K.R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* **1998**, *10*, 1299–1319. [CrossRef]
154. Girolami, M. Mercer kernel-based clustering in feature space. *IEEE Trans. Neural Netw.* **2002**, *13*, 780–784. [CrossRef] [PubMed]
155. Zhang, L.; Zhou, W.; Jiao, L. Kernel clustering algorithm. *Chin. J. Comput.* **2002**, *25*, 587–590.
156. Shawe-Taylor, J.; Cristianini, N. *Kernel Methods for Pattern Analysis*; Cambridge University Press: Cambridge, UK, 2004.
157. Zhao, B.; Kwok, J.T.; Zhang, C. Multiple kernel clustering. In Proceedings of the SIAM International Conference on Data Mining, Sparks, NV, USA, 30 April–2 May 2009; pp. 638–649.
158. Huang, H.C.; Chuang, Y.Y.; Chen, C.S. Multiple kernel fuzzy clustering. *IEEE Trans. Fuzzy Syst.* **2011**, *20*, 120–134. [CrossRef]
159. Samaria, F.S.; Harter, A.C. Parameterisation of a stochastic model for human face identification. In Proceedings of the IEEE Workshop on Applications of Computer Vision, Sarasota, FL, USA, 5–7 December 1994; pp. 138–142.
160. Yin, M.; Xie, S.; Wu, Z.; Zhang, Y.; Gao, J. Subspace clustering via learning an adaptive low-rank graph. *IEEE Trans. Image Process.* **2018**, *27*, 3716–3728. [CrossRef]
161. Georghiades, A.S.; Belhumeur, P.N.; Kriegman, D.J. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 643–660. [CrossRef]

162. Hastie, T.; Simard, P.Y. Metrics and models for handwritten character recognition. *Stat. Sci.* **1998**, *13*, 54–65. [CrossRef]
163. Hull, J.J. A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell.* **1994**, *16*, 550–554. [CrossRef]
164. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
165. Schuller, B.; Vlasenko, B.; Eyben, F.; Rigoll, G.; Wendemuth, A. Acoustic emotion recognition: A benchmark comparison of performances. In Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding, Merano, Italy, 13–17 December 2009; pp. 552–557.
166. Martin, O.; Kotsia, I.; Macq, B.; Pitas, I. The eNTERFACE'05 audio-visual emotion database. In Proceedings of the International Conference on Data Engineering Workshops, Atlanta, GA, USA, 3–7 April 2006; pp. 1–8.
167. Schuller, B.; Arsic, D.; Rigoll, G.; Wimmer, M.; Radig, B. Audiovisual behavior modeling by combined feature spaces. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, HI, USA, 16–20 April 2007; pp. 733–736.
168. Kheirandishfard, M.; Zohrizadeh, F.; Kamangar, F. Multi-level representation learning for deep subspace clustering. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2020; pp. 2039–2048.
169. Hu, X.; Li, T.; Zhou, T.; Peng, Y. Deep spatial-spectral subspace clustering for hyperspectral images based on contrastive learning. *Remote Sens.* **2021**, *13*, 4418. [CrossRef]
170. Liu, M.; Wang, Y.; Ji, Z. Self-supervised convolutional subspace clustering network with the block diagonal regularizer. *Neural Process. Lett.* **2021**, *53*, 3849–3875. [CrossRef]
171. Han, T.; Niu, S.; Gao, X.; Yu, W.; Cui, N.; Dong, J. Deep low-rank graph convolutional subspace clustering for hyperspectral image. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
172. Zhou, P.; Hou, Y.; Feng, J. Deep adversarial subspace clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1596–1604.
173. Yu, Z.; Zhang, Z.; Cao, W.; Liu, C.; Chen, C.P.; Wong, H.S. Gan-based enhanced deep subspace clustering networks. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 3267–3281. [CrossRef]
174. Li, S.; Liu, L.; Liu, J.; Song, W.; Hao, A.; Qin, H. SC-GAN: Subspace clustering based GAN for automatic expression manipulation. *Pattern Recognit.* **2023**, *134*, 109072. [CrossRef]
175. Yang, X.; Yan, J.; Cheng, Y.; Zhang, Y. Learning Deep Generative Clustering via Mutual Information Maximization. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**. [CrossRef]
176. Ji, P.; Zhang, T.; Li, H.; Salzmann, M.; Reid, I. Deep subspace clustering networks. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 24–33.
177. Guo, X.; Liu, X.; Zhu, E.; Zhu, X.; Li, M.; Xu, X.; Yin, J. Adaptive self-paced deep clustering with data augmentation. *IEEE Trans. Knowl. Data. Eng.* **2020**, *32*, 1680–1693. [CrossRef]
178. Lv, J.; Kang, Z.; Lu, X.; Xu, Z. Pseudo-supervised deep subspace clustering. *IEEE Trans. Image Process.* **2021**, *30*, 5252–5263. [CrossRef]
179. Wang, J.; Jiang, J. Unsupervised deep clustering via adaptive GMM modeling and optimization. *Neurocomputing* **2021**, *433*, 199–211. [CrossRef]
180. Hampel, F.R.; Ronchetti, E.M.; Rousseeuw, P.; Stahel, W.A. *Robust Statistics: The Approach Based on Influence Functions*; Wiley-Interscience: New York, NY, USA, 1986.
181. Rousseeuw, P.J.; Leroy, A.M. *Robust Regression and Outlier Detection*; John Wiley & Sons: Hoboken, NJ, USA, 2005.
182. Chen, B.; Zhai, W.; Kong, L. Variable selection and collinearity processing for multivariate data via row-elastic-net regularization. *Adv. Stat. Anal.* **2022**, *106*, 79–96. [CrossRef]
183. Wang, L. The L1 penalized LAD estimator for high dimensional linear regression. *J. Multivar. Anal.* **2013**, *120*, 135–151. [CrossRef]
184. She, Y.; Owen, A.B. Outlier detection using nonconvex penalized regression. *J. Am. Stat. Assoc.* **2011**, *106*, 626–639. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.