Wang et al. BMC Bioinformatics 2019, 20(Suppl 22):718
https://doi.org/10.1186/s12859-019-3231-5

BMC Bioinformatics

RESEARCH Open Access

Multi-cancer samples clustering via graph regularized low-rank representation method under sparse and symmetric constraints

Juan Wang, Cong-Hai Lu, Jin-Xing Liu*, Ling-Yun Dai and Xiang-Zhen Kong

From International Conference on Data Science, Medicine and Bioinformatics
Nanning, China. 22-24 June 2019

Abstract

Background: Identifying different types of cancer based on gene expression data has become hotspot in bioinformatics research. Clustering cancer gene expression data from multiple cancers to their own class is a significance solution. However, the characteristics of high-dimensional and small samples of gene expression data and the noise of the data make data mining and research difficult. Although there are many effective and feasible methods to deal with this problem, the possibility remains that these methods are flawed.

Results: In this paper, we propose the graph regularized low-rank representation under symmetric and sparse constraints (sgLRR) method in which we introduce graph regularization based on manifold learning and symmetric sparse constraints into the traditional low-rank representation (LRR). For the sgLRR method, by means of symmetric constraint and sparse constraint, the effect of raw data noise on low-rank representation is alleviated. Further, sgLRR method preserves the important intrinsic local geometrical structures of the raw data by introducing graph regularization. We apply this method to cluster multi-cancer samples based on gene expression data, which improves the clustering quality. First, the gene expression data are decomposed by sgLRR method. And, a lowest rank representation matrix is obtained, which is symmetric and sparse. Then, an affinity matrix is constructed to perform the multi-cancer sample clustering by using a spectral clustering algorithm, i.e., normalized cuts (Ncuts). Finally, the multi-cancer samples clustering is completed.

Conclusions: A series of comparative experiments demonstrate that the sgLRR method based on low rank representation has a great advantage and remarkable performance in the clustering of multi-cancer samples.

Keywords: Affinity matrix, Gene expression data, Graph regularization, Symmetric constraint, Low-rank representation, Spectral clustering

Background

Currently, cancer is one of the most prevalent human diseases, and cancer seriously threatens the quality of human life [1]. The number of variety cancers is increasing, which makes it difficult to effect a radical cure of cancer. A good performing cancer diagnosis method can help doctors to formulate treatment strategies for patients effectively and in a timely manner. In addition, cancer clustering based on gene expression data has become one of the frontiers of bioinformatics research. In the field, it provides an effective way to further explore gene expression data. For example, it can be used to classify cancer [2], select genes [3] and discover cancer linked biomarker genes [4]. In this paper, we propose a methodology to processing gene expression data for identifying different types of cancer.

Since the start of the twenty-first century, the volume of high dimensional and complex gene expression data has exploded with the advent and development of gene

* Correspondence: sdcavell@126.com
School of Information Science and Engineering, Qufu Normal University, Rizhao, China

detection technology such as DNA microarray technology [5]. So far, the researchers have proposed many well-performing methods and used them for gene expression data mining, such as K-means clustering [6], nonnegative matrix factorization (NMF) [7] and principal component analysis (PCA) [8]. More recently, because of the high dimensional nature of gene expression data, the low-rank representation (LRR) method has become a popular and promising method since its prototype was proposed by Liu et al. [9]. The LRR method can preserve the subspace structure of the raw dataset in a lowest rank representation matrix. Theoretically, the lowest rank representation matrix is a block-diagonal matrix with a well grouped effect, and this matrix can well capture the global structural information of high-dimensional dataset [10]. And then, the clustering method, such as spectral clustering method, is used to cluster the lowest rank representation matrix to realize the subspace segmentation. The LRR clustering method has been adopted widely in many fields due to the advantages of the lowest rank representation matrix, such as facial recognition [11], genetic microarray data clustering [12], image clustering [13] and subspace segmentation [14]. And, LRR method achieves good results in processing high-dimensional datasets.

In general, high-dimensional data always have noisy and outliers because of the complexity of the collection process. And, the noisy and outliers inevitably impairs the intrinsic structure of the data space. Therefore, the outliers and noise cause difficulties during processing the raw data. Especially in the LRR method, the high-dimensional data are usually used in the form of a dictionary matrix, which inevitably adversely affects grouped effect of lowest rank representation matrix. As described in [9], the LRR method may fail to obtain a block-diagonal lowest rank representation matrix in complex applications, which makes it difficult to integrate the lowest rank representation matrix with other information. To alleviate this problem, the commonly used solution is to combine the LRR method with the spectral clustering method (the Ncuts clustering method is often adopted) to get the final clustering result. The LRR method and the spectral clustering method are linked by an affinity matrix which is constructed based on the lowest rank representation matrix. And, the affinity matrix has better grouping effects. In general, in order to construct the affinity matrix, a symmetric operation step is usually performed to establish a similarity-based undirected graph. However, this simple symmetric operation inevitably leads to the loss of important structural information of the raw dataset. To tackle this disadvantage, Ni et al. proposed an approach named the low-rank representation with positive semi-definite (LRR-PSD) to obtain a symmetric positive semi-definite (PSD) matrix [15]. In the LRR-PSD method, an affinity matrix is constructed based on the PSD matrix. This method inspired Chen et al. to propose a low-rank representation with symmetric constraint (LRRSC) method for learning a symmetric lowest rank representation matrix [16]. In this method, the affinity matrix is constructed according to the angular information of the principal directions of the symmetric lowest rank representation matrix. The obtained affinity matrix is better than the matrix which is obtained by simple symmetric operation.

However, compared with the sparse representation method, which considers the sparsest representation of each data point or data vector individually, the LRR method mainly focuses on the global structural information of data [9]. That leads to the LRR method ignoring the local geometrical structural information of data. Because it is shown that the intrinsic local geometrical structures within the high-dimensional data are important for the subspace clustering model [17], some researchers introduce nonlinear dimensionality reduction methods into the LRR, such as the manifold learning theory.

At present, many well-established nonlinear dimensionality reduction methods have been proposed since Tenenbaum et al. and Roweis et al. proposed isometric mapping (ISOMAP) [18] and locally linear embedding (LLE) [19], respectively. The typical methods include the locality preserving projection (LPP) [20], local tangent space alignment (LTSA) [21], Laplacian eigenmap (LE) [22] and Riemannian normal coordinates (RNC) [23]. They can generate a low-dimensional subspace according to the submanifold of the observation datasets. Furthermore, the manifold learning method treats the local geometrical structures of data points as submanifolds. Inspired by the local invariance [20], the manifold learning method estimates the geometrical structures of the submanifold using random data points [24]. The method can map the submanifolds from the high-dimensional space to the low-dimensional space. Therefore, the local geometrical structural information of the raw high-dimensional dataset can be preserved in the low-dimensional space [25].

In order to improve the original LRR method, some researchers combine manifold learning theory with the LRR. For instance, Yin et al. proposed a novel model called the nonnegative sparse hyper-Laplacian regularized LRR (NSHLRR) that can acquire the inherent information within dataset [24]. Motivated by the NSHLRR, Wang et al. proposed the Laplacian regularized low-rank representation (LLRR) method to identify differently expressed genes [26]. More recently, Wang et al.

presented a tumor sample clustering method named the Mixed-norm Laplacian Regularized low-rank representation (MLLRR) [27].

Motivated by the above methods, in order to obtain a better lowest rank matrix that can avoid the simple symmetric operation and preserve the intrinsic local geometrical structures within the raw high-dimensional dataset, we introduce symmetric sparse constraints and graph regularization based on manifold learning into the LRR method, and propose the graph regularized low-rank representation method under combined the sparse and symmetric constraints, or short sgLRR method. The sgLRR method can obtain a lowest rank representation matrix that can well capture the global structure information of the high-dimensional raw dataset and meanwhile preserve the intrinsic local geometrical structures within the dataset. Furthermore, the sgLRR method weakens the adverse effect of noise in the raw dataset by strengthening the symmetric constraint to the lowest rank representation matrix. The obtained lowest rank representation matrix is an excellent basis for constructing the affinity matrix. To take full advantage of the lowest rank representation matrix, we consider the angular information of the principal directions of the lowest rank representation matrix. Therefore, in contrast to the traditional approach, we perform skinny singular value decomposition (SVD) operations on the lowest rank representation to construct the affinity matrix. Finally, based on the affinity matrix, we adopt a spectral clustering algorithm to obtain the clustering results.

We adopt the sgLRR method for multi-cancer sample clustering based on gene expression data. Our experiment design is carried out in the following three steps: Step One: the sgLRR method is used to process multi-cancer sample gene expression dataset. And, we can obtain a lowest rank representation matrix. Step Two: an affinity matrix is constructed by exploiting the obtained lowest rank representation matrix. Step Three: based on the affinity matrix, we adopt a spectral clustering algorithm, i.e., Ncuts method, to perform the multi-cancer sample clustering. Compared with a lot of related methods, the sgLRR method has better performance in multi-cancer sample clustering.

In summary, the main contributions of our work are as follows:

(1) We introduce the symmetric sparse constraints and graph regularization based on manifold learning into the original LRR method and develop a novel method named the sgLRR. The regularized graph is better for preserving the local geometrical structure of raw high-dimensional data. And, the symmetric constraint weakens the effect of noise in the raw dataset. Therefore, we use sgLRR method to get a better lowest rank presentation matrix that has better grouping effect for the subspace clustering.
(2) Based on the lowest rank presentation matrix, we construct an affinity matrix to further improve its grouping effect. As the link of sgLRR method and Ncuts clustering method, the affinity matrix makes full useful of the angular information of the principal directions of the lowest rank representation matrix.
(3) By combining sgLRR with the Ncuts clustering method, we apply the sgLRR method to multi-cancer sample clustering, and extensive experiments are conducted on gene expression data. Compared with other methods, our methodology has better performance in multi-cancer sample clustering.

The remainder of this paper is outlined as follows: The section 2 summarizes the LRR method, and gives a brief review of some related work. And then, we describe the proposed sgLRR method in detail. In section 3, based on The Cancer Genome Atlas (TCGA) dataset [28], we perform a large number of comparative experiments to demonstrate the sgLRR method with better performance on the multi-cancer sample clustering. And, we discuss and analysis the experimental results from different aspects. In section 4, we describe the corresponding discussion. In the section 5, we summarize our work for the full paper.

Methods

First, we review the related work about the low-rank representation. And then, then we introduced our proposed approach.

Related work

In recent years, the LRR method and its improved algorithms have been widely used in many fields. Furthermore, the group theory based on manifold learning has also captured the attention of the researchers. In subsections ***Low-Rank Representation*** and ***The Symmetric Constraint for the Low-Rank Representation***, we review the LRR method [9] and the symmetric constraint for the low-rank representation [16], respectively. Then, in subsection ***Manifold Learning for Graph Regularization***, we give a detail introduction to graph regularization based on manifold learning [29, 30].

Low-Rank Representation Learning a lowest rank representation matrix of the observation dataset is the aim of LRR method [9]. Given an observation data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ with no error. And, there is an overcomplete dictionary matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_k] \in \mathbb{R}^{m \times k}$ and a union of multiple low-dimensional independent subspaces. It is assumed that the subspaces can be

linearly spanned by the dictionary matrix **A**. Therefore, the given observation data **X** can be represented by these low-dimensional subspaces, and the relationship between data **X** and matrix **A** is **X** = **AZ**. In other words, data **X** is a linear combination of the dictionary matrix **A**. The function of the LRR method is as follows:

$$\min_{\mathbf{Z}} \ rank(\mathbf{Z}) \quad \text{s.t.}\mathbf{X} = \mathbf{AZ}. \tag{1}$$

Here, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ is the observation data matrix. m is the total number of features, and n is the total number of the samples. $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_k]$ is the overcomplete dictionary matrix, and $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n] \in \mathbb{R}^{k \times n}$ is the low rank representation matrix. The element $\mathbf{z}_i$ of matrix **Z** is the mapping relationship from $\{\mathbf{x}_i | \ \mathbf{x}_i \in \mathbf{X}^{m \times n}, 1 \le i \le n\}$ to the dictionary matrix **A**. In general, the matrix **Z** is also called the coefficient matrix, and it is a new expression form of **X** that is based on the dictionary matrix **A**. The purpose of the LRR method is to find the lowest rank representation matrix $\mathbf{Z}^*$.

In practical analysis, the observation data matrix **X** is usually selected as the dictionary matrix **A**, which is a very important aspect of the LRR [9, 15, 26, 27]. According to the matrix multiplication rule, the lowest rank representation matrix $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n] \in \mathbb{R}^{n \times n}$ is a square matrix. Equation (1) can be rewritten as follows:

$$\min_{\mathbf{Z}} \ rank(\mathbf{Z}) \quad \text{s.t.}\mathbf{X} = \mathbf{XZ}. \tag{2}$$

In this case, the matrix $\mathbf{Z}^*$ represents the relation of each sample of **X**. In other words, the element $\mathbf{z}_{ij}^*$ of matrix $\mathbf{Z}^*$ represents the similarity between the samples $\mathbf{x}_i$ and $\mathbf{x}_j$. Therefore, the element $\mathbf{z}_{ij}^*$ should be equal to the element $\mathbf{z}_{ji}^*$. That is, the matrix $\mathbf{Z}^*$ is a symmetric matrix when the observation data matrix **X** is clean.

Because the rank function is nonconvex, no closed expression can be found. Therefore, Eq. (2) is very difficult to solve. The related research has shown that the nuclear-norm of a matrix is a minimal convex envelope of the rank of the matrix [31–33]. Therefore, Eq. (2) is equivalent to the following nuclear-norm convex optimization problem:

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_* \quad \text{s.t.}\mathbf{X} = \mathbf{XZ}. \tag{3}$$

Here, $\|\cdot\|_*$ is the nuclear-norm. It is the sum of all singular values of the matrix **Z**, which is the minimal convex envelope of the rank function [17]. In the actual situation, the observation data are inevitably polluted by noise or outliers under certain special circumstances. Therefore, a certain regularization constraint $\|\cdot\|_l$ is usually added to (3) to balance the interference. The improved formula is as follows:

$$\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \gamma\|\mathbf{E}\|_l \quad \text{s.t.}\mathbf{X} = \mathbf{XZ} + \mathbf{E}. \tag{4}$$

Here, the matrix **E** denotes the noise or outliers. The parameter $\gamma > 0$ is to balance the adaptability of each part in (4), and $\|\cdot\|_l$ is regularization constraint. In general, the appropriate regularization constraint $\|\cdot\|_l$ is selected according to different types of noise and outliers in real environments. For example, $\|\cdot\|_{2,1}$, i.e., the $l_{2,1}$ norm, is used to extract the sample-specific corruptions and small noise or outliers, and $\|\cdot\|_0$, i.e., the l_0 norm, is used to deal with the significant noise or outliers [27]. Solving the l_0 norm is an NP-hard problem. Therefore, it is usually replaced by $\|\cdot\|_1$, i.e., the l_1 norm.

The above is a description of the classic original LRR method. The LRR method deals with the observation data from a holistic perspective. That is, the global structural features of the observation data are represented by the lowest rank representation matrix $\mathbf{Z}^*$. In addition, the LRR method maps the structures of the observation data from high-dimensional spaces to low-dimensional spaces. It reduces the difficulty of processing high-dimensional observation data.

The Symmetric Constraint for the Low-Rank Representation With clean observation data, based on function (3), the most ideal matrix $\mathbf{Z}^*$ is a block diagonal and strictly symmetric matrix, as shown in Fig. 1. However, according to function (4), the lowest rank representation matrix $\mathbf{Z}^*$ is not strictly symmetric when using real data with noise and outliers, as shown in Fig. 2 [34].. In other words, because the element $\mathbf{Z}_{ij}$ is not equal to the element $\mathbf{Z}_{ji}$, the degree of similarity of the i-th sample to the j-th sample is not equal to the degree of similarity of the i-th of sample to the j-th sample. One question worth considering is which of the two elements is more suitable to be used to reflect the similarity between the two samples.

In general, an affinity matrix is usually constructed using symmetric operation, i.e., $(|\mathbf{Z}^*| + |\mathbf{Z}^*|^{\mathrm{T}})/2$, to reflect the similarity of samples. Then, based on the affinity matrix, spectral clustering algorithms generally use the Ncuts clustering method for subspace clustering. To avoid symmetric operations, Chen et al. imposed a symmetric constraint onto the LRR to obtain a symmetric lowest rank representation matrix [16]. The improved method was named the low-rank representation with symmetric constraint (LRRSC) and it can be expressed as follows:

$$\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \gamma\|\mathbf{E}\|_l \quad \text{s.t.}\mathbf{X} = \mathbf{XZ} + \mathbf{E}, \mathbf{Z} = \mathbf{Z}^{\mathrm{T}}. \tag{5}$$

The symmetric lowest rank representation matrix $\mathbf{Z}^*$ can greatly preserve the subspace structures of the observation data. Therefore, the affinity matrix based on

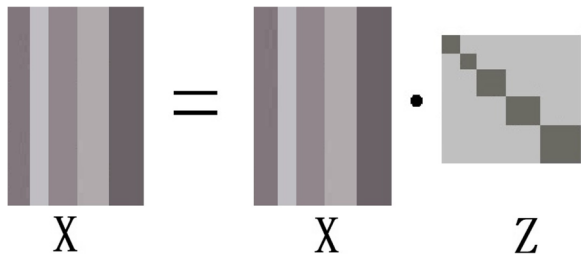

Fig. 1 The LRR method with clean data

the principal direction angular information of the symmetric lowest rank representation matrix $\mathbf{Z}^*$ can effectively reflect the similarity between samples. However, the LRRSC method does not consider the local geometrical structural information. It may lose important information when obtaining the lowest rank representation matrix. In the next section, we use manifold learning with graph regularization to solve this problem.

Manifold Learning for Graph Regularization In actual situations, the given observation data $\mathbf{X} \in \mathbb{R}^{m \times n}$ are usually high-dimensional. Thus, the local geometrical structural information exists at each data point and at its k-nearest-neighboring data points. Capturing the local geometrical structural information is important for the performance of subspace clustering. Fortunately, graph regularization based on manifold learning provides a feasible option to achieve this aim [29]. This approach can preserve the intrinsic local geometrical structures that are embedded in the high-dimensional data space.

In graph theory, the "manifold assumption" is that data points near local geometrical structures should keep their proximity under a new basis [35]. If we map the adjacent data points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the high-dimensional space to the low-dimensional space, their mapping data points $\mathbf{z}_i$ and $\mathbf{z}_j$ should be close in the low-dimensional space. Therefore, the local geometrical structural information of the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ can be represented in the low-dimensional space. In other words, if the characteristics of the data points are similar in the high-dimensional space, their mapping data points can be clustered into the same class in the low-dimensional space.

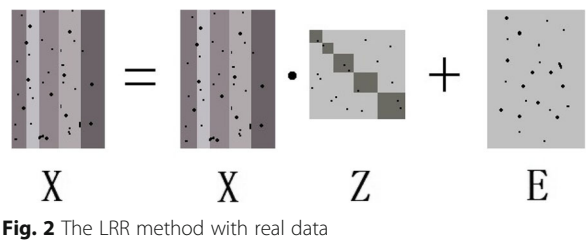

Fig. 2 The LRR method with real data

We take each data point as a vertex. The data points are defined by the column of observation data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$. Therefore, the number of vertices is n. All n vertices form a graph $\mathbf{G}$. The weight of the connected edge of vertices i and j in the graph $\mathbf{G}$ is represented by $\mathbf{w}_{ij}$. The assignment rule of $\mathbf{w}_{ij}$ is as follows:

$$\mathbf{w}_{ij} = \begin{cases} 1 \text{ if } \mathbf{x}_i \in \mathbf{N}_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathbf{N}_k(\mathbf{x}_i) \\ 0 \text{ otherwise} \end{cases}, \tag{6}$$

where $\mathbf{N}_k(\mathbf{x}_i)$ denotes the set of the k-nearest-neighbors of $\mathbf{x}_i$. As suggest in [26, 27], we select the $k = 5$ as the nearest neighbors for the experimental datasets. In addition, all elements $\mathbf{w}_{ij}$ make up a symmetric weight matrix $\mathbf{W}$.

In the low-dimensional space, the new relationship of the data points is as follows:

$$\min_{\mathbf{Z}} \sum\nolimits_{ij} \left\| \mathbf{z}_i - \mathbf{z}_j \right\|^2 \mathbf{w}_{ij}. \tag{7}$$

After a linear algebraic transformation, the above optimization problem (7) can be written as follows:

$$\min_{\mathbf{Z}} tr\left(\mathbf{Z}^{\mathrm{T}} \mathbf{L} \mathbf{Z}\right). \tag{8}$$

Here, $tr(\cdot)$ is the trace of the matrix. $\mathbf{L}$ is called the graph Laplacian matrix. It is defined by $\mathbf{L} = \mathbf{D} - \mathbf{W}$. The matrix $\mathbf{D}$ is a diagonal matrix, and the element $\mathbf{d}_{ii}$ of $\mathbf{D}$ is sum of the i-th row of $\mathbf{W}$, i.e., $\mathbf{d}_{ii} = \sum_j^n \mathbf{w}_{ij}$.

sgLRR methodology

In this section, we introduce our method for multi-cancer sample clustering. First, we obtain the objective function according to the problem's formulation and solve the function using the linearized adaptive direction method with the adaptive penalty (LADMAP) method [36]. Then, we provide the complete algorithm for easier understanding. Second, we combined our method with the Ncuts clustering method by learning an affinity matrix. Finally, the proposed method is used for the subspace segmentation of multi-cancer sample clustering.

Problem formulation and the solution

In this subsection, our goal is to propose a novel LRR model to preserve the intrinsic local geometrical structures of the observation data and simultaneously weaken the effects of noise and outliers in the dictionary matrix. We introduce graph regularization based on manifold learning and the symmetric constraint into the original LRR method. It is as follows:

$$\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \beta tr\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}\right) + \gamma \|\mathbf{E}\| \tag{9}$$

where β and γ are penalty parameters, $\mathbf{L}$ is the Laplacian matrix, and $\|\cdot\|_1$ is the regularization constraint. $\|\mathbf{E}\|_1$ is

the sum of the absolute values of each element in the matrix $\mathbf{E}$.

In addition, according to the sparsity-based clustering method, e.g., sparse coding combined with clustering, the sparsity constraint can be thought of as a strategy for information localization [37]. Thus, the coefficient matrix with the sparsity constraint can improve the performance of subspace clustering. Namely, by combining the low-rank and sparse data, the within-class affinities are dense, and the between-class affinities are zeros. So, we introduce the sparsity constraint into Eq. (9), and the finally objective function of our method is as follows:

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \lambda\|\mathbf{Z}\|_1 + \beta tr\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}\right) + \gamma\|\mathbf{E}\| \\ &\text{s.t.}\mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E}, \mathbf{Z} = \mathbf{Z}^{\mathrm{T}}, \end{aligned} \tag{10}$$

where λ is the penalty parameter, and $\|\mathbf{Z}\|_1$ is the sparsity constraint for the low-rank representation matrix $\mathbf{Z}$.

We call the objective function in (10) the graph regularized low-rank representation under combined the sparse and symmetric constraints (sgLRR) method. To obtain a globally optimal solution, we adopt the LADMAP to solve problem (10).

First, we introduce the auxiliary variable $\mathbf{J}$ to separate variables. It is as follows:

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \lambda\|\mathbf{J}\|_1 + \beta tr\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}\right) + \gamma\|\mathbf{E}\| \\ &\text{s.t.}\mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E}, \mathbf{Z} = \mathbf{Z}^{\mathrm{T}}, \mathbf{Z} = \mathbf{J}. \end{aligned} \tag{11}$$

Second, problem (11) can be converted into an unconstrained optimization problem by using the augmented Lagrange multiplier method (ALM) [38]. The formula is rewritten as follows:

$$\begin{aligned} \ell(\mathbf{Z},\mathbf{E},\mathbf{J},\mathbf{Y}_1,\mathbf{Y}_2) = {} & \|\mathbf{Z}\|_* + \lambda\|\mathbf{J}\|_1 + \beta tr\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}\right) + \gamma\|\mathbf{E}\| \\ & + \langle \mathbf{Y}_1, \mathbf{X}-\mathbf{X}\mathbf{Z}-\mathbf{E}\rangle + \langle \mathbf{Y}_2, \mathbf{Z}-\mathbf{J}\rangle \\ & + \frac{\mu}{2}\left(\|\mathbf{X}-\mathbf{X}\mathbf{Z}-\mathbf{E}\|_F^2 + \|\mathbf{Z}-\mathbf{J}\|_F^2\right), \end{aligned} \tag{12}$$

Here, $\|\cdot\|_F$ is the Frobenius-norm; λ, β, λ and μ are the penalty parameters; $\langle \mathbf{A}, \mathbf{B}\rangle = tr(\mathbf{A}^T\mathbf{B})$ represents the Euclidean inner product between the two matrices, and $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are Lagrangian multipliers.

According to the LADMAP method, problem (12) is divided into three problems. They are as follows:

$$\begin{aligned} \ell_1(\mathbf{Z}) = {} & \|\mathbf{Z}\|_* + \beta tr\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}\right) + \langle \mathbf{Y}_1, \mathbf{X}-\mathbf{X}\mathbf{Z}-\mathbf{E}\rangle \\ & + \langle \mathbf{Y}_2, \mathbf{Z}-\mathbf{J}\rangle + \frac{\mu}{2}\left(\|\mathbf{X}-\mathbf{X}\mathbf{Z}-\mathbf{E}\|_F^2 + \|\mathbf{Z}-\mathbf{J}\|_F^2\right), \end{aligned} \tag{13}$$

$$\begin{aligned} \ell_2(\mathbf{E}) = {} & \gamma\|\mathbf{E}\|_1 + \langle \mathbf{Y}_1, \mathbf{X}-\mathbf{X}\mathbf{Z}-\mathbf{E}\rangle \\ & + \frac{\mu}{2}\|\mathbf{X}-\mathbf{X}\mathbf{Z}-\mathbf{E}\|_F^2, \end{aligned} \tag{14}$$

$$\ell_3(\mathbf{J}) = \lambda\|\mathbf{J}\|_1 + \langle \mathbf{Y}_2, \mathbf{Z}-\mathbf{J}\rangle + \frac{\mu}{2}\|\mathbf{Z}-\mathbf{J}\|_F^2. \tag{15}$$

Problem (13) can be replaced by solving the following problem (16):

$$\min\|\mathbf{Z}\|_* + \langle \nabla_{\mathbf{Z}}\mathbf{q}(\mathbf{Z}_k), \mathbf{Z}-\mathbf{Z}_k\rangle + \frac{\eta_1}{2}\|\mathbf{Z}-\mathbf{Z}_k\|_F^2, \tag{16}$$

where $\nabla_{\mathbf{Z}}\mathbf{q}(\mathbf{Z}_k) = \beta(\mathbf{Z}_k\mathbf{L}^{\mathrm{T}} + \mathbf{Z}_k\mathbf{L}) + \mu_k(\mathbf{Z}_k - \mathbf{J}_k + \mathbf{Y}_2^k/\mu_k) + \mu_k\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{Z}_k - \mathbf{X} + \mathbf{E}_k - \mathbf{Y}_1^k/\mu_k)$, $\eta_1 = 2\beta\|\mathbf{L}\|_2 + \mu_k(1 + \|\mathbf{X}\|_2^2)$.

We use the following *Lemma-1* to solve problem (16). Chen et al. have given the rigorous mathematical derivations and detailed proofs for this theorem [16].

Lemma 1 Given a square matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$, there is a unique closed form solution to solve optimization problem (17).

$$\begin{aligned} \mathbf{P}^* = {} & \arg\min_{\mathbf{P}} \frac{1}{\mu}\|\mathbf{P}\|_* + \frac{1}{2}\|\mathbf{P}-\mathbf{Q}\|_F^2 \quad \text{s.t.}\mathbf{P} \\ & = \mathbf{P}^{\mathrm{T}}. \end{aligned} \tag{17}$$

It is as follows:

$$\mathbf{P}^* = \mathbf{U}_r\left(\mathbf{\Sigma}_r - \frac{1}{\mu}\cdot\mathbf{I}_r\right)\mathbf{V}_r^{\mathrm{T}}. \tag{18}$$

Here, $\mathbf{\Sigma}_r$, $\mathbf{U}_r$ and $\mathbf{V}_r$ are obtained using $\tilde{\mathbf{Q}} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^{\mathrm{T}}$, which is the skinny SVD of the symmetric matrix $\tilde{\mathbf{Q}}$. In addition, $\mathbf{\Sigma}_r = diag(\sigma_1, \sigma_2, ..., \sigma_r)$ with $\{r : \sigma_r > \frac{1}{\mu}\}$ are the positive singular values of matrix $\tilde{\mathbf{Q}}$. $\mathbf{U}_r$ and $\mathbf{V}_r$ are the singular vectors of matrix $\tilde{\mathbf{Q}}$. Matrix $\tilde{\mathbf{Q}}$ is obtained by $\tilde{\mathbf{Q}} = (\mathbf{Q} + \mathbf{Q}^{\mathrm{T}})/2$ and the skinny SVD only keeps the positive singular values of the normal SVD. $\mathbf{I}_r$ is an identity matrix.

According to *Lemma-1*, we set $\mathbf{Q} = \mathbf{Z}_k - \nabla_{\mathbf{Z}}\mathbf{q}(\mathbf{Z}_k)/\eta_1$. Then, we solve problem (16) by using the iterative formula (19):

$$\mathbf{Z}_{k+1}^* = \Theta_{\frac{1}{\eta_1}}\left(\frac{\mathbf{Q} + \mathbf{Q}^{\mathrm{T}}}{2}\right). \tag{19}$$

Here, $\Theta_\varepsilon(\mathbf{A}) = \mathbf{U}_r\mathbf{S}_\varepsilon(\mathbf{\Sigma}_r - \frac{1}{\mu_k}\cdot\mathbf{I}_r)\mathbf{V}_r^{\mathrm{T}}$ and $\mathbf{S}_\varepsilon(x) = \mathrm{sgn}(x)\max(|x| - \varepsilon, 0)$.

We update $\mathbf{E}$ and $\mathbf{J}$ by minimizing $\ell_2(\mathbf{E})$ and $\ell_3(\mathbf{J})$. And, $\mathbf{E}$ and $\mathbf{J}$ are independent of each other in this minimization problem. And then, based on a singular value thresholding algorithm, we obtain the iterative formulas of $\mathbf{E}$ and $\mathbf{J}$. We set $\frac{\partial\ell_2}{\partial\mathbf{E}} = 0$ and $\frac{\partial\ell_3}{\partial\mathbf{J}} = 0$, respectively. Then, we obtain the following equations.

$$\frac{\partial \ell_2}{\partial \mathbf{E}_k} = \mu_k\left(\mathbf{E}_k - \mathbf{X} + \mathbf{X}\mathbf{Z}_{k+1} - \mathbf{Y}_1^k/\mu_k\right) = 0 \Rightarrow \mathbf{E}_k = \mathbf{X} - \mathbf{X}\mathbf{Z}_{k+1} + \mathbf{Y}_1^k/\mu_k, \quad (20)$$

$$\frac{\partial \ell_3}{\partial \mathbf{J}_k} = \mu_k\left[\mathbf{J}_k - \left(\mathbf{Z}_{k+1} + \mathbf{Y}_2^k/\mu_k\right)\right] = 0 \Rightarrow \mathbf{J}_k = \mathbf{Z}_{k+1} + \mathbf{Y}_2^k/\mu_k. \quad (21)$$

According to the NSHLRR method [24] and the singular value thresholding algorithm [39], the iterative formulas of **E** and **J** are as follows:

$$\mathbf{E}_{k+1} = \Psi mfrac\left(\mathbf{X} - \mathbf{X}\mathbf{Z}_{k+1} + \mathbf{Y}_1^k/\mu_k\right)\frac{\gamma}{\mu_k}, \quad (22)$$

$$\mathbf{J}_{k+1} = \max\left\{\Psi_{\frac{\lambda}{\mu_k}}\left(\mathbf{Z}_{k+1} + \frac{1}{\mu_k}\mathbf{Y}_2^k\right), 0\right\}. \quad (23)$$

Based on the above, we discuss the time complexity of sgLRR compared to the original LRR. As described in [36], the complexity of LADMAP method for LRR is $O(rmn)$, where r is the rank of the matrix **Z**, m and n is the size of observation data matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$. For sgLRR method, the construction of the k-nearest neighbor graph is $O(n^2m)$. Therefore, the complexity of sgLRR is $O(rmn + n^2m)$.

Finally, **Algorithm 1** provides the complete sgLRR algorithm.

Algorithm 1: Solving sgLRR method by LADMAP

Input: observation data $\mathbf{X}$, the parameters λ, β, and λ, the number of k-nearest-neighbors.

Output: $\mathbf{Z}^*$

Initialization: $\mathbf{Z}_0 = \mathbf{E}_0 = \mathbf{J}_0 = \mathbf{Y}_1^0 = \mathbf{Y}_2^0 = 0$, $\rho_0 = 2.5$, $\mu_0 = 10^{-6}$, $\mu_{\max} = 10^6$, $\varepsilon_1 = 10^{-6}$, $\varepsilon_2 = 10^{-2}$, $\eta_1 = 1.25 \times \|\mathbf{X}\|^2$, $\mathbf{L}$.

While not convergence do:

Updating $\mathbf{Z}_{k+1}$ as (19);

Updating $\mathbf{E}_{k+1}$ as (22);

Updating $\mathbf{J}_{k+1}$ as (23);

Updating $\mathbf{Y}_1$ and $\mathbf{Y}_2$:

$\mathbf{Y}_1^{k+1} = \mathbf{Y}_1^k + \mu_k(\mathbf{X} - \mathbf{X}\mathbf{Z}_{k+1} - \mathbf{E}_{k+1})$,

$\mathbf{Y}_2^{k+1} = \mathbf{Y}_2^k + \mu_k(\mathbf{Z}_{k+1} - \mathbf{J}_{k+1})$.

Updating μ_{k+1}: $\mu_{k+1} = \min(\mu_{\max}, \rho_k\mu_k)$.

where $\rho_k = \begin{cases} \rho_0, & \text{if } \mu_k \max\{\eta_1\|\mathbf{Z}_k - \mathbf{Z}_{k+1}\|, \|\mathbf{J}_k - \mathbf{J}_{k+1}\|, \|\mathbf{E}_k - \mathbf{E}_{k+1}\|\} < \varepsilon_2 \\ 1, & \text{otherwise} \end{cases}$.

Checking convergence,

If $\|\mathbf{X} - \mathbf{X}\mathbf{Z}_{k+1} - \mathbf{E}_{k+1}\|/\|\mathbf{X}\| < \varepsilon_1$ or $\mu_{k+1} \cdot \max\{\eta_1\|\mathbf{Z}_k - \mathbf{Z}_{k+1}\|, \|\mathbf{J}_k - \mathbf{J}_{k+1}\|, \|\mathbf{E}_k - \mathbf{E}_{k+1}\|\} < \varepsilon_2$

End while.

sgLRR method combined with the Ncuts clustering method

We obtain the lowest rank representation matrix $\mathbf{Z}^*$ by **Algorithm 1**. The obtained matrix $\mathbf{Z}^*$ inherits and improves the grouping effect of the LRR method. The symmetric property of matrix $\mathbf{Z}^*$ strictly reflects the similarity of the data samples, and the data samples that belong to the same group highlight the same subspace of matrix $\mathbf{Z}^*$. However, as mentioned in [9], the complex application may fail in the lowest rank representation matrix $\mathbf{Z}^*$ and in fully using the information within matrix $\mathbf{Z}^*$. Therefore, we combine the sgLRR method with the Ncuts clustering method to guarantee correct segmentation results.

First, we learn an affinity matrix **H** that is the link between the sgLRR method and the Ncuts clustering method. The affinity matrix **H** utilizes the angular similarity information of the principal direction of matrix $\mathbf{Z}^*$, and matrix **H** is a similar undirected graph that further improves the grouping effect. The process below can be defined as learning matrix **H**.

1. The matrix $\mathbf{Z}^*$ is decomposed into $\mathbf{Z}^* = \mathbf{U}^*\mathbf{\Sigma}^*(\mathbf{V}^*)^{\mathrm{T}}$ using skinny SVD.
2. Define the matrix $\mathbf{M} = \mathbf{U}^*(\mathbf{\Sigma}^*)^{1/2}$ or the matrix $\mathbf{N} = (\mathbf{\Sigma}^*)^{1/2}(\mathbf{V}^*)^{\mathrm{T}}$. Because the matrix $\mathbf{Z}^*$ is a symmetrical matrix, both matrix **M** and matrix **N** are equivalent for leaning the affinity matrix **H**.
3. The element of the affinity matrix **H** is calculated using function (24).

$$\mathbf{H}_{ij} = \left(\frac{\mathbf{m}_i^{\mathrm{T}}\mathbf{m}_j}{\|\mathbf{m}_i\|_2\|\mathbf{m}_j\|_2}\right)^2 \quad \text{or} \quad \mathbf{H}_{ij} = \left(\frac{\mathbf{n}_i^{\mathrm{T}}\mathbf{n}_j}{\|\mathbf{n}_i\|_2\|\mathbf{n}_j\|_2}\right)^2, \quad (24)$$

where $\mathbf{m}_i$ is the i-th row of **M**, and $\mathbf{n}_i$ is the i-th row of **N**.

Next, we adopt the Ncuts clustering method to produce the final data sample clustering results. The Ncuts clustering method was proposed by Shi et al. and is closely related to graph theory [40]. This approach can well reflect the degree of similarity within classes and the degree of dissimilarity between classes. This approach has been successfully applied in image segmentation and has numerous successful examples in different fields and datasets, such as gene expression overlapping clustering based on the penalized weighted normalized cut [41].

Finally, we briefly summarize the process of the multi-cancer sample clustering algorithm. It is as follows.

Algorithm 2: Clustering multi-cancer samples based on the sgLRR method

Input: the observation data $\mathbf{X}$, i.e., the gene expression data, the number of types cancers k.

Step:

1) Obtain the lowest rank representation matrix $\mathbf{Z}^*$ by **Algorithm 1.**
2) Learn an affinity matrix $\mathbf{H}$ by the function (24).
3) Use the Ncuts clustering method based on the affinity matrix $\mathbf{H}$ to perform multi-cancer sample clustering.

Output: multi-cancer sample clustering results.

End

Results

Datasets

As the biggest cancer genomic profile database, The Cancer Genome Atlas (TCGA) provides publicly

available datasets with over 30 types of cancers using high-throughput sequencing technology and integrated multidimensional analyses to help improve the diagnosis, prevention, and treatment of cancer [28].

We use five real gene expression datasets that were downloaded from the TCGA to construct the integrated datasets for the experiments. The five original datasets are the cholangiocarcinoma (CHOL) dataset, the head and neck squamous cell carcinoma (HNSC) dataset, the colon adenocarcinoma (COAD) dataset, the esophageal carcinoma (ESCA) dataset and the pancreatic adenocarcinoma (PAAD) dataset. Each dataset consists a different number of cancer samples and normal samples, and each sample contains 20,502 genes. Table 1 lists the distribution number of the samples for each dataset.

As listed in the Table 1, we use all the cancer samples of each dataset to construct six integrated datasets. And, the six integrated datasets are named the CO-CH (COAD-CHOL) dataset, the PA-ES (PAAD-ESCA) dataset, the CH-HN-CO (CHOL-HNSC-COAD) dataset, the ES-CH-HN (ESCA-CHOL-HNSC) dataset, the ES-CO-PA-HN (ESCA-COAD-PAAD-HNSC) dataset and the CO-CH-ES-HN (COAD-CHOL-ESCA-HNSC) dataset, respectively. The characteristics of the datasets are as follows: the CO-CH dataset contain all 298 cancer samples from COAD and CHOL; the PA-ES dataset contain all 359 cancer samples from PAAD and ESCA; the CH-HN-CO dataset contain all 696 cancer samples from CHOL, HNSC and COAD; the ES-CH-HN dataset contain all 617 cancer samples ESCA, CHOL and HNSC; the CO-CH-ES-HN dataset contain all 879 cancer samples from COAD, CHOL, ESCA and HNSC; and, the ES-CO-PA-HN dataset contain all 1019 cancer samples from ESCA, COAD, PAAD and HNSC. The distribution of the six datasets are summarized and listed in Table 2. We conduct experiments on the basis of the six datasets to prove the performance of sgLRR method.

Table 1 The distribution of the samples in the five datasets

Gene Expression Datasets	The Distribution of the Samples in the Datasets		
	Cancer Samples	Normal Samples	Total of Number
COAD	262	19	281
ESCA	183	9	192
CHOL	36	9	45
PAAD	176	4	180
HNSC	398	20	418

Note: The Gene Expression Datasets represent the different cancer sample data: *COAD* colon adenocarcinoma, *ESCA* esophagus cancer, *CHOL* cholangiocarcinoma, *PAAD* pancreatic adenocarcinoma, *HNSC* head and neck squamous cell carcinoma

Measurement metrics for the experiment results

In this article, we use multiple measures to strictly analyze the clustering results. The clustering results are mainly evaluated by the Accuracy (Acc) [42], Matthews Correlation Coefficient (MCC) [43], Rand Index (RI) [44] and Normalized Mutual Information (NMI) [45]. Next, we concisely introduce them.

Accuracy

The Accuracy (Acc) evaluates the clustering results on the global level by calculating the matching degree of experimental result labels and actual labels. The values of the Acc ranges from 0 to 1, and the higher value is, the better the clustering results is. The specific formula is as follows.

$$Acc = \frac{\sum_{i=1}^{n} \delta(p_i, map(q_i))}{n} \times 100\%. \tag{25}$$

Here, $\delta(p_i, map(q_i))$ is defined as follows:

$$\delta(p_i, map(q_i)) = \begin{cases} 1, \text{if } p_i = map(q_i) \\ 0, \text{otherwise} \end{cases}, \tag{26}$$

where n is the number of data samples, p_i is the real label for the i-th sample, and q_i is the experimental result of the i-th sample. $map(q_i)$ is the mapping function that can match the clustering result to the real label using the Kuhn-Munkres approach [46].

Matthews correlation coefficient

The Matthews Correlation Coefficient (MCC) is widely used performance measure in biomedical research to handle imbalanced datasets [43, 47–49]. In general, MCC represents a comprehensive evaluation measure which has a better balance of both aspects of the accuracy and coverage than the individual precision and recall values [49]. The MCC is defined in terms of TP (True Positive), FP (False Positive) and TN (True Negative), FN (False Negative), and its formula is as follows.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \times 100\%, \tag{27}$$

where TP is the number of true positives, where the data points that actually belong to the same cluster are grouped into the same cluster in the experiment results. TN is the number of true negatives, where the data points that actually belong to the same cluster are grouped into the different clusters in the experiment results. FP is the number of false positives, where the data points that actually belong to different clusters are grouped into the same cluster in the experiment results. FN is the number of false negatives, where the data points that actually belong to different clusters are

Table 2 The distribution of the six datasets

Datasets	The number of samples of each type of cancer	Total number of samples	subspace number
CO-CH	262–36	298	2
PA-ES	176–183	359	2
CH-HN-CO	36–398-262	696	3
ES-CH-HN	183–36-398	617	3
CO-CH-ES-HN	262–36–183-398	879	4
ES-CO-PA-HN	183–262–176-398	1019	4

Note: The datasets represent different integrated datasets. The characteristics of each dataset are described in the previous passage

grouped into the different clusters in the experiment results. The Fig. 3 shows *TP*, *TN*, *FP*, *FN* clearly.

The value of MCC is took in the interval $[-1, 1]$, with 1 representing a complete agreement, −1 indicating a complete disagreement, and 0 indicating that the clustered result was uncorrelated with the ground truth [50]. For the multi-class dataset clustering with k classes, the MCC can be calculated by the confusion matrix. And, the confusion matrix is a matrix $\mathbf{C} = (C_{ij})_{k\times k}$ with the size of $k \times k$, where C_{ij} represents the number of samples, which in actually belongs to the class i, are clustered to be in the cluster j. And, the confusion matrix and the item of *TP, FP, TN, FN* for multi-cancer samples clustering are defined as the Fig. 4.

Rand index

The Rand Index (RI) is an objective criterion for the evaluation of clustering methods. From a mathematical standpoint, the RI is related to the Acc, but it is applicable even when class labels are not used [44]. Given the set of n data points $S = \{O_1, O_2, ..., O_n\}$ that are to be clustered, the specific partitions $V = \{v_1, v_2, ..., v_r\}$ and $U = \{u_1, u_2, ..., u_c\}$ are the clustering results of S that are divided into r and c disjointed sets, respectively. If V represents the true results and U represents the experiment results, then RI is defined as follows.

$$RI = \frac{a+d}{a+b+c} \times 100\%, \tag{28}$$

where $\begin{cases} a \text{ is the total of the data pairs that exist in the same cluster for } V \text{ and } U. \\ b \text{ is the total of the data pairs that exist in the different clusters both for } V \text{ and } U. \\ c \text{ is the case with different the } a \text{ and } b. \end{cases}$

The value of the RI ranges from 0 to 1, and the higher value is, the better the clustering results is.

Normalized mutual information

The Normalized Mutual Information (NMI) is commonly used in clustering to measure the similarity of two clustering results [45]. There are the clusters $\Xi = [\xi_i]_k$ obtained by the clustering algorithm and the true inherent classes $\Omega = [\omega_i]_k$. The NMI is defined as follows.

$$NMI(\Xi, \Omega) = 2 \times \frac{I(\Xi; \Omega)}{(H(\Xi) + H(\Omega))} \times 100\%. \tag{29}$$

Here, $I(\Xi; \Omega) = \sum_{\xi_i \in \Xi}^{k} \sum_{\omega_i \in \Omega}^{k} p(\xi, \omega) \log(\frac{p(\xi,\omega)}{p(\xi)p(\omega)})$ is the mutual information, and $H(\Xi) = \sum_{i=1}^{k} p(\xi_i) \times I(\xi_i) = \sum_{i=1}^{k} p(\xi_i) \times \log_2(\frac{1}{p(\xi_i)})$, where $p(\xi_i)$ $(p(\omega_i))$ is the probability of an object being in *cluster* ξ_k (class ω_k), and $p(\xi_k, \omega_j)$ is the joint probability that an object lies in cluster ξ_k and class ω_k. The value of the NMI ranges from 0 to 1, and the higher value is, the better the clustering results is.

Experiment result and discussion

In this subsection, we cluster the multi-cancer samples using the sgLRR method and compare the results with other related methods to analyze the performance. The related methods in the comparative experiments include K-means, T-SNE, LLE, NMF [42], PCA [33], LRR [9], LLRR [26] and MLLRR [27]. And, the experiments are carried out on the six integrated datasets. We obtain the clustering results of the sgLRR method and the comparison methods. For the compared methods: K-means, T-SNE, LLE, NMF, PCA, LRR, LLRR, and MLLRR, they are the traditional existing clustering and dimensional reduction methods. And, we categorize these methods into three classes. The first kind of method is the classic method for clustering, including K-means, NMF and PCA. The T-SNE and LLE belong to the second kind of

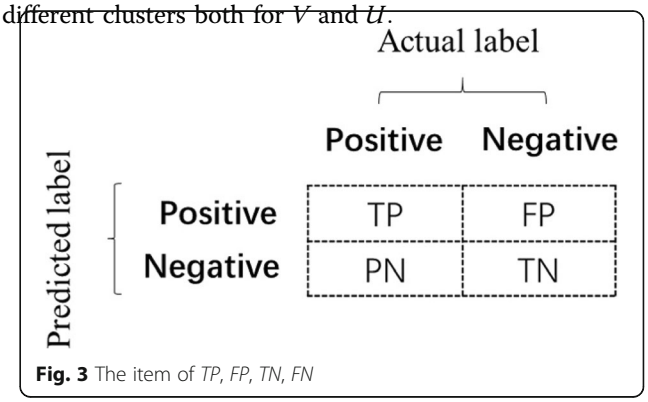

Fig. 3 The item of *TP, FP, TN, FN*

method for dimensional reduction. T-SNE and LLE both are well-established manifold learning methods. Among them, T-SNE is almost the dominant one in bioinformatics, especially for expression data. The third kind of method include LRR, LLRR, MLRR and our proposed sgLRR method. These methods belong to the subspaces clustering methods to dimensional reduction by low rank representation the raw datasets.

In addition, K-means clustering algorithm is usually used to obtain the final clustering result in spectral clustering. In this paper, K-means method uses the K-means + + algorithm for cluster center initialization and squared Euclidean distance by default, and in K-means + + algorithm, the initial cluster center is randomly selected [51]. Therefore, if K-means method is used to repeat the experiment with the same dataset, the results of these experiments will not be identical, and there will be minor differences. This difference will affect our performance evaluation of clustering methods. In our experiments, in order to improve the reasonable of results and reduce the difference, we repeat clustering experiment 50 times. And, the mean of results is taken as the measurements of clustering results.

The experimental results are listed in the Table 3. And, we highlight the best clustering results in bold. Of all the best results, except for the few results, the results obtained by sgLRR method are overwhelmingly superior in the nine experimental methods. In the next, based on the clustering results, we detailed discuss and analyze the advantage of sgLRR method which is different with the above comparison methods. And, the details are as follows.

1. For the most of metrics results in the Table 3, the LRR, LLRR, MLRR and sgLRR methods are better than the first kind method, including K-means, NMF and PCA. Furthermore, the performances of LLRR, MLRR and sgLRR methods improve as the number of cancer types increasing. Notably, the best clustering results are mainly obtained by the sgLRR method. From an overall standpoint, the experimental results demonstrate that the methods for the low-rank representation are better for multiple subspace clustering than the classical clustering method. One mainly reason is that the low-rank representation methods with the characteristics of capturing the subspace structure of datasets. Therefore, the gene expression data structures of each type cancer are stored in their respective low dimensional subspaces, which makes the different types of cancer dataset more separable.
2. In order to demonstrate dimensionality reduction datasets of sgLRR with the better performance on the multi-cancer gene expression datasets, we compare sgLRR method with the second class of method: T-SNE and LLE. At first, we visualize the dimensionality reduction datasets which are obtained by the T-SNE, LLE and sgLRR methods, as shown in Fig. 5. And, the data points are colored according to their actual labels. As shown in the Fig. 5, in the dimensionality reduction data obtained by the T-SNE method, there are several overlaps between clusters of different types of cancer samples such as I-2, I-4, I-5 and I-6. In the dimensionality reduction data obtained by the LLE method, the separability of clusters of different types of cancers is not obvious such as II-3, II-5 and II-6. In the dimensionality reduction data obtained by the sgLRR method, the independence of different cancer sample clusters is obvious such as III-2, III-4 and III-6, and the data subspace has better separability effect than T-SNE and LLE methods.

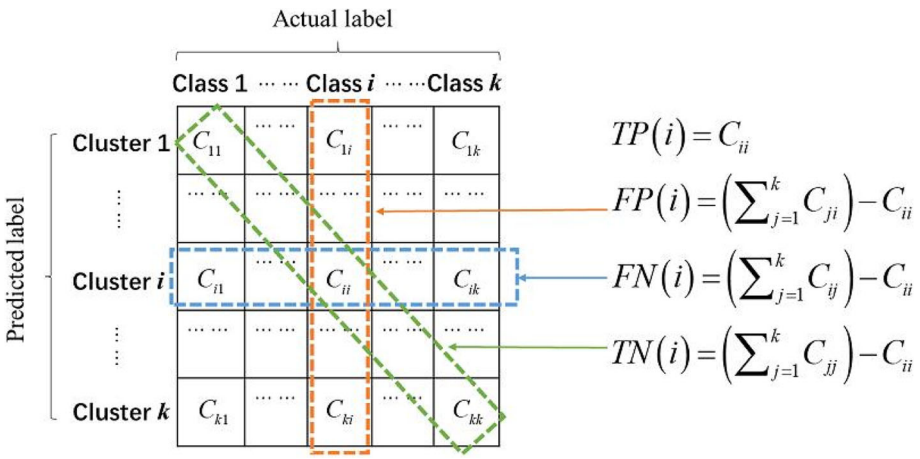

Fig. 4 Confusion matrix and the item of *TP*, *FP*, *TN*, *FN* for multi-cancer dataset clustering

Table 3 The clustering results of all methods on the different integrated datasets

Datasets	Measure	Method								
		K-means	T-SNE	LLE	NMF	PCA	LRR	LLRR	MLLRR	sgLRR
CO-CH	ACC	95.40	89.31	71.14	93.14	93.63	97.99	**99.66**	98.99	98.66
	MCC	80.06	70.21	88.58	75.53	72.99	46.76	91.06	90.78	**95.20**
	RI	91.43	49.85	63.66	87.18	88.07	96.04	**99.32**	98.00	97.34
	NMI	66.46	51.01	75.71	54.12	54.47	41.92	80.51	77.04	**87.51**
PA-ES	ACC	98.25	91.84	77.26	**99.16**	**99.16**	96.38	**99.16**	**99.16**	**99.16**
	MCC	84.73	83.74	62.04	**98.34**	**98.34**	77.98	96.71	96.71	**98.34**
	RI	97.37	84.97	66.46	**98.35**	98.33	93.00	98.34	98.34	98.34
	NMI	81.06	59.49	41.43	**93.86**	**93.86**	65.43	89.39	89.39	**93.86**
CH-HN-CO	ACC	89.22	83.03	80.99	76.99	85.77	95.86	97.99	84.77	**98.28**
	MCC	67.91	65.69	60.77	82.25	69.17	61.23	68.31	80.96	**87.25**
	RI	90.00	87.16	82.85	80.19	87.76	94.70	96.67	84.66	**97.48**
	NMI	73.55	78.43	69.07	76.22	72.59	68.25	73.81	77.83	**80.99**
ES-CH-HN	ACC	85.56	52.35	61.65	84.52	80.03	82.17	93.19	94.32	**96.11**
	MCC	66.26	32.62	42.01	67.44	66.08	43.04	61.18	66.49	**91.32**
	RI	82.67	60.97	63.89	80.25	78.15	72.36	89.07	90.30	**93.10**
	NMI	56.77	30.77	33.26	47.35	72.59	36.92	52.38	57.20	**78.75**
CO-CH-ES-HN	ACC	86.89	60.52	63.04	82.31	81.32	79.24	92.48	87.94	**94.17**
	MCC	70.00	65.30	79.30	51.78	71.30	52.44	78.05	73.03	**91.71**
	RI	89.43	74.59	71.98	85.07	86.57	81.58	91.42	88.95	**93.34**
	NMI	71.04	48.43	52.00	57.67	69.12	54.76	74.42	69.82	**80.27**
ES-CO-PA-HN	ACC	86.89	85.83	67.76	82.31	81.32	79.24	92.48	87.94	**94.17**
	MCC	79.51	78.52	77.40	84.23	**89.82**	59.21	88.38	83.63	85.49
	RI	89.43	91.45	78.06	85.07	86.57	81.58	91.42	88.95	**93.34**
	NMI	76.15	81.42	55.93	72.97	79.53	61.82	**81.24**	76.58	76.23

Note: The best clustering results are highlighted in bold

Therefore, we come to the conclusion that the separability among different types cancers in the dimensionality reduction data of three methods is best by sgLRR method, followed by T-SNE method, and finally by LEE method. Moreover, sgLRR method makes the data points more separable between classes than T-SNE method. That is due to that sgLRR method combines the low-rank representation method with the graph regularization constraint based on manifold learning, which enhances the separable between different types of cancer data in dimensionality reduction datasets.

3. In the third kind of method, what LRR LLRR MLLRR and sgLRR methods have in common is that they represent global structure of the raw dataset by a low-rank matrix with low dimensional subspaces. However, comparing the LRR method with the LLRR, MLLRR and sgLRR methods, the clustering results of most datasets are better than the LRR method. This is because LLRR, MLLRR and sgLRR methods with graph regularization based on manifold learning can capture the inherent geometric structural information of datasets. The results suggest that introducing graph regularization based on manifold learning can improve the clustering performance of the method. Moreover, we can find that the most of metrics of the sgLRR are higher than those of the LLRR, and they are also the best in all comparison methods. This is because the symmetry constraint weakens the effect of noise in the genetic expression data, and it makes the lowest rank representation matrix that is obtained by the sgLRR better for preserving the similarity among the cancer samples than the LLRR.

In addition, the affinity matrix that is constructed based on the lowest rank representation matrix also plays a key role in the clustering. To briefly and clearly explain the contribution of the affinity matrix, we randomly select three typical datasets and give the heat maps based on matrix $\mathbf{Z}^*$ and matrix $\mathbf{H}^*$ for each respective dataset. The heat maps of the three selected datasets (CO-CH, CH-

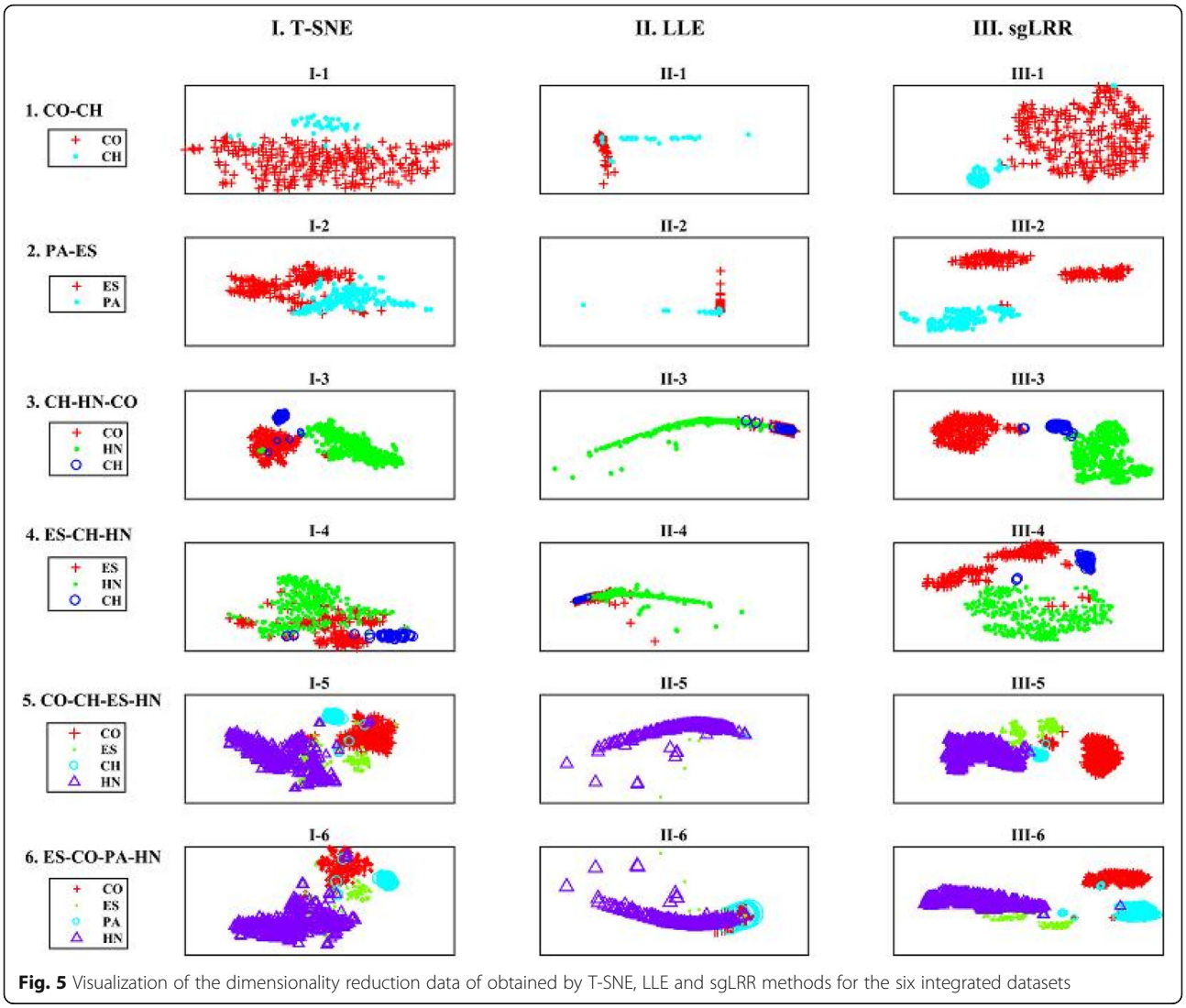

Fig. 5 Visualization of the dimensionality reduction data of obtained by T-SNE, LLE and sgLRR methods for the six integrated datasets

HN-CO and CO-CH-ES-HN) are as shown Fig. 6. In addition, in Fig. 6, the larger that the matrix element is, the brighter the corresponding position on the heat map. As we can see from Fig. 6, it is obvious that the grouping effect of matrix $\mathbf{H}^*$ is better than that of $\mathbf{Z}^*$.

4. In this part, we analyze the advantages of sgLRR method from the relationship between subspace number contained in datasets and method performance. According to Table 2, each of the six integrated datasets contains different types and amounts of cancer. In other word, the different integrated datasets contain different number of subspaces. And, there is reason to believe that the complexity of the internal geometric structures for the dataset has a notable positive correlation with the number of subspaces. Therefore, among the six integrated datasets, the CO-CH-ES-HN and ES-CO-PA-HN are the most complex, followed by the CH-HN-CO and ES-CH-HN that contain three types of cancer, and finally, followed by the integrated datasets CO-CH and PA-ES that contain two types of cancer. Based on the Table 3, we can see that the sgLRR method is better than the other methods as the subspace number increases. Specifically, the metrics of the sgLRR on the ES-CH-HN dataset are 2.92 (ACC), 30.14 (MCC), 4.03 (RI) and 26.37 (NMI) percentage points higher than those of the LLRR. Furthermore, for the CO-CH-ES-HN dataset, the percentages of the sgLRR are 1.69 (ACC), 13.66 (MCC), 1.92 (RI) and 5.85 (NMI) higher than those of the LLRR. This proves that the sgLRR method is more suitable for multi-cancer sample clustering than comparison methods.

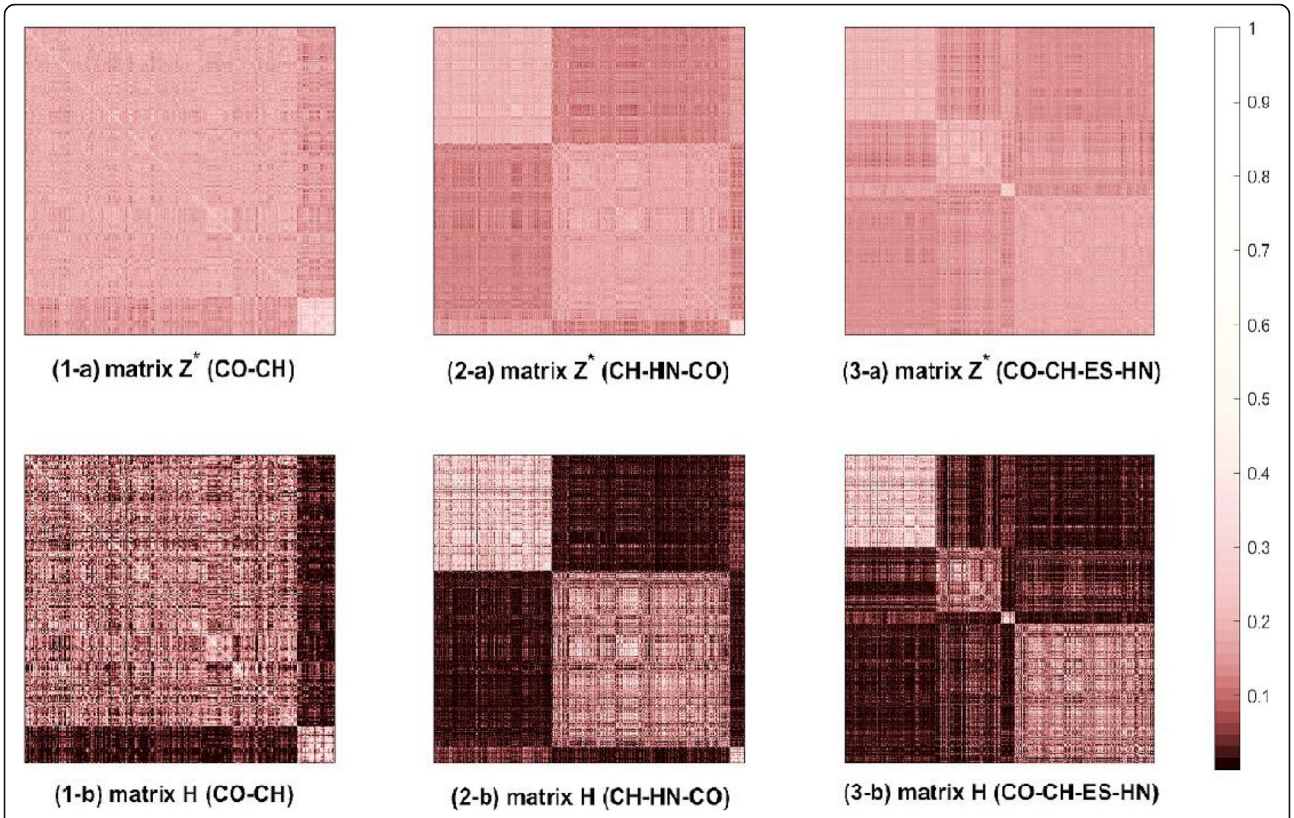

Fig. 6 The heat maps intuitively compare the grouping effects of matrices $\mathbf{Z}^*$ and $\mathbf{H}$. (*1-a*) and (*1-b*) are the heat maps based on the matrices $\mathbf{Z}^*$ and $\mathbf{H}$ for the CO-CH dataset, respectively; (*2-a*) and (*2-b*) are the heat maps based on the matrices $\mathbf{Z}^*$ and $\mathbf{H}$ for the CH-HN-CO dataset, respectively; and (*3-a*) and (*3-b*) are the heat maps based on the matrices $\mathbf{Z}^*$ and $\mathbf{H}$ for the CO-CH-ES-HN dataset, respectively

Through the above analysis, we can conclude that the combination of graph regularization based on manifold learning and the symmetry constraint plays a significant role in the sgLRR and achieves satisfactory results in multi-cancer samples clustering.

Discussion

Based on the comparison and demonstration of the above experimental results, our proposed sgLRR method has advantages over other methods. The sgLRR method based on low rank representation has a great advantage in multi-subspace clustering. By means of symmetric constraint and sparse constraint, the influence of data noise on low-rank representation is alleviated. Meanwhile, the local geometric structure of data is retained through graph regularization constraint based on manifold learning, which improves the clustering effect in subspace clustering. Compared with other methods based on low rank representation, our method takes into account various factors that affect subspace clustering and improves the performance of the method. These advantages have been demonstrated in experiments with gene expression data from multiple cancers.

Conclusions

In this paper, we introduce graph regularization based on manifold learning and symmetric sparse constraints into the original LRR and propose a novel method called the sgLRR. The original LRR method can capture the global geometrical information of the whole observation data. The lowest rank representation matrix $\mathbf{Z}^*$ of the sgLRR method has the properties of the traditional LRR method and can capture the intrinsic local geometric information within data. In addition, the symmetry constraint weakens the effect of noise in the dictionary matrix and makes the lowest rank representation matrix $\mathbf{Z}^*$ strictly and accurately preserve the similarity between samples.

We adopt the sgLRR method for multi-cancer samples clustering based on gene expression dataset. First, we use the sgLRR to obtain the lowest rank representation matrix $\mathbf{Z}^*$. Then, based on the angular similarity information of the lowest rank representation matrix $\mathbf{Z}^*$, we learn an affinity matrix $\mathbf{H}$ by using a unitary matrix that is obtained using skinny SVD. The results prove that the affinity matrix has a better grouping effect. Finally, based on the affinity matrix $\mathbf{H}$, the spectral clustering algorithm (Ncuts) is used to obtain the clustering results.

We compare the experimental results from other methods to the sgLRR, including the K-means, T-SNE, LLE, NMF, PCA, LRR, LLRR and MLLRR methods. The experimental results show that the sgLRR method is a novel efficient method for multi-cancer sample clustering. The sgLRR method performs well on the dataset, which contain multiple subspaces. In future work, we will further study the sgLRR method. For example, the current method can be extended to identify characteristic cancer genes or to analyze cancer pathways.

Abbreviations

ALM: The augmented Lagrange multiplier method;; LADMAP: The linearized adaptive direction method with the adaptive penalty method; LRR: Low-rank representation; Ncuts: Normalized cuts; sgLRR: The graph regularized low-rank representation under symmetric and sparse constraints method; SVD: Singular value decomposition; TCGA: The Cancer Genome Atlas

Acknowledgements

Thanks go to the editor and the anonymous reviewers for their comments and suggestions.

About this supplement

This article has been published as part of *BMC Bioinformatics Volume 20 Supplement 22, 2019: Decipher computational analytics in digital health and precision medicine*. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-22.

Authors' contributions

CHL and JW jointly conceptualized the algorithm, designed the sgLRR method and carried out the experiments. JXL and LYD gave statistical and computational advice for the project and analysed the results of the methods. XZK participated in designing evaluation criteria, and drafted the manuscript and polished the English expression. All authors read and approved the final manuscript.

Funding

Publication costs are funded by the NSFC under grant Nos. 61572284, 61702299, 61972226 and 61902215. The funder played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The datasets that support the findings of this study can be found in the [The Cancer Genome Atlas (TCGA)] https://cancergenome.nih.gov/.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Published: 30 December 2019

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.
2. Feng C, Xu Y, Liu J, Gao Y, Zheng C. Supervised discriminative sparse PCA for com-characteristic gene selection and tumor classification on multiview biological data. IEEE Trans Neural Netw Learn Syst. 2019;30(10):2926–37.
3. Liu J-X, Feng C-M, Kong X-Z, Xu Y. Dual graph-Laplacian PCA: a closed-form solution for bi-clustering to find "checkerboard" structures on gene expression data. IEEE Access. 2019, 7:151329–38.
4. Sadhu A, Bhattacharyya B. Discovery of cancer linked biomarker genes through common subcluster mining. In: 2016 international conference on bioinformatics and systems biology (BSB): Mar, Allahabad, India 2016. p. 1–5.
5. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. The sequence of the human genome. Science (New York, NY). 2001;**291**(5507):1304–51.
6. Mavroeidis D, Marchiori E. Feature selection for k-means clustering stability: theoretical analysis and an algorithm. Data Min Knowl Disc. 2014;28(4):918–60.
7. Zheng CH, Ng TY, Zhang L, Shiu CK, Wang HQ. Tumor classification based on non-negative matrix factorization using gene expression data. IEEE Trans NanoBiosci. 2011;10(2):86–93.
8. Pooladi M, Tavirani MR, Hashemi M, HesamiTackallou S, Abad SKR, Moradi A, Zali AR, Mousavi M, Dalvand LF, Rakhshan A, et al. Cluster and principal component analysis of human glioblastoma multiforme (GBM) tumor proteome. Iran J Cancer Prevent. 2014;7(2):87–95.
9. Liu GC, Lin ZC, Yu Y. Robust subspace segmentation by low-rank representation. In: Proceedings of the 27th international conference on machine learning (ICML-10): 2010. 2010.
10. Lu C, Feng J, Lin Z, Mei T, Yan S. Subspace clustering by block diagonal representation. IEEE Trans Pattern Anal Mach Intell. 2018:1–1.
11. Chen CF, Wei CP, Wang YF. Low-rank matrix recovery with structural incoherence for robust face recognition. In: 2012 IEEE conference on computer vision and pattern recognition: Jun. 2012. p. 2618–25.
12. Cui Y, Zheng CH, Yang J. Identifying subspace gene clusters from microarray data using low-rank representation. PLoS One. 2013;8(3):e59377.
13. Zhang ZY, Zhao KK. Low-rank matrix approximation with manifold regularization. IEEE Trans Pattern Anal Mach Intell. 2013;35(7):1717–29.
14. Liu GC, Yan SC. Latent low-rank representation for subspace segmentation and feature extraction. In: 2011 international conference on computer vision: Nov.; Barcelona, Spain 2011. p. 1615–1622.
15. Ni YZ, Sun J, Yuan XT, Yan SC, Cheong LF. Robust low-rank subspace segmentation with semidefinite guarantees. In: Proceedings of the 2010 IEEE international conference on data mining workshops (ICDMW '10): Dec.; Sydney, NSW, Australia, IEEE Computer Society 2010. p. 1179–1188.
16. Chen J, Mao H, Sang Y, Yi Z. Subspace clustering using a symmetric low-rank representation. Knowl-Based Syst. 2017;127:46–57.
17. Yin M, Gao JB, Lin ZC, Shi QF, Guo Y. Dual graph regularized latent low-rank representation for subspace clustering. IEEE Trans Image Process. 2015; 24(12):4918–33.
18. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science (New York, NY). 2000;290(5500): 2319–23.
19. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science (New York, NY). 2000;290(5500):2323–6.
20. He X. Locality preserving projections. Chicago: University of Chicago; 2005.
21. Zhang Z, Zha H. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM J Sci Comput. 2006;26(1):313–38.
22. Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Proceedings of the 14th international conference on neural information processing systems: natural and synthetic: Dec.; Vancouver, British Columbia, Canada. MIT Press 2001. p. 585–591.
23. Lin T, Zha H, Lee SU. Riemannian manifold learning for nonlinear dimensionality reduction. In: Leonardis A, Bischof H, Pinz A, editors. Computer vision – ECCV 2006. Berlin/Heidelberg: Springer; 2006. p. 44–55.
24. Yin M, Gao J, Lin Z. Laplacian regularized low-rank representation and its applications. IEEE Trans Pattern Anal Mach Intell. 2016;38(3):504–17.
25. He XF, Cai D, Shao YL, Bao HJ, Han JW. Laplacian regularized Gaussian mixture model for data clustering. IEEE Trans Knowl Data Eng. 2011;23(9): 1406–18.
26. Wang YX, Liu JX, Gao YL, Zheng CH, Shang JL. Differentially expressed genes selection via Laplacian regularized low-rank representation method. Comput Biol Chem. 2016;65:185–92.
27. Wang J, Liu JX, Zheng CH, Wang YX, Kong XZ, Weng CG. A mixed-norm Laplacian regularized low-rank representation method for tumor samples clustering. IEEE/ACM Trans Comput Biol Bioinform. 2018;7:1–1.
28. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Poznan, Poland). 2015;19(1A):A68–77.
29. Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. 2003;15(6):1373–96.

30. Sun SL, Hussain Z, Shawe-Taylor J. Manifold-preserving graph reduction for sparse semi-supervised learning. Neurocomputing. 2014;124:13–21.
31. Candès EJ, Recht B. Exact matrix completion via convex optimization. Found Comput Math. 2009;9:717–72.
32. Keshavan RH, Montanari A, Oh S. Matrix completion from noisy entries. In: Proceedings of the 22nd international conference on neural information processing systems. 2009. p. 952–60.
33. Candès EJ, Li X, Ma Y, Wright J. Robust principal component analysis? ACM. 2011;58(3):1–37.
34. Liu GC, Lin ZC, Yan SC, Sun J, Yu Y, Ma Y. Robust recovery of subspace structures by low-rank representation. IEEE Trans Pattern Anal Mach Intell. 2013;35(1):171–84.
35. Dai LY, Feng CM, Liu JX, Zheng CH, Yu J, Hou MX. Robust nonnegative matrix factorization via joint graph Laplacian and discriminative information for identifying differentially expressed genes. Complexity. 2017;2017:11.
36. Lin Z, Liu R, Su Z. Linearized alternating direction method with adaptive penalty for low-rank representation. Adv Neural Inf Proces Syst. 2011:612–20.
37. Oktara Y, Turkan M. A review of sparsity-based clustering methods. Signal Process. 2018;148:20–30.
38. Lin Z, Chen M, Ma Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. Eprint Arxiv. 2010;
39. Cai J-F, Candès EJ, Shen Z. A singular value thresholding algorithm for matrix completion. SIAM J Optim. 2008;20(4):1956–82.
40. Shi JB, Malik J. Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell. 2000;22(8):888–905.
41. Teran Hidalgo SJ, Zhu T, Wu M, Ma S. Overlapping clustering of gene expression data using penalized weighted normalized cut. Genet Epidemiol. 2018;42(8):796–811.
42. Zheng CH, Huang DS, Zhang L, Kong XZ. Tumor clustering using nonnegative matrix factorization with gene selection. IEEE Trans Inf Technol Biomed. 2009;13(4):599–607.
43. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. 1975;405(2):442–51.
44. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66(336):846–50.
45. Manning C, Raghavan P, Schütze H. Introduction to information retrieval, vol. 1. Cambridge: Cambridge University Press; 2008.
46. Zhu H, Zhou MC, Alkins R. Group role assignment via a Kuhn-Munkres algorithm-based solution. IEEE Trans Syst Man Cybernet Part A Syst Hum. 2012;42(3):739–50.
47. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS One. 2017;12(6): e0177678.
48. Liu G, Mercer TR, Shearwood A-MJ, Siira SJ, Hibbs ME, Mattick JS, Rackham O, Filipovska A. Mapping of mitochondrial RNA-protein interactions by digital RNase footprinting. Cell Rep. 2013;5(3):839–48.
49. Gu Q, Zhu L, Cai Z. Evaluation measures of the classification performance of imbalanced data sets. Commun Comput Inform Sci. 2009;51:461–71.
50. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics (Oxford, England). 2000;16(5):412–24.
51. Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms; New Orleans, Louisiana. 1283494: Society for Industrial and Applied Mathematics 2007. p. 1027–35.

Publisher's Note