Wang *et al. BMC Bioinformatics* (2021) 22:334
https://doi.org/10.1186/s12859-021-04220-6

BMC Bioinformatics

RESEARCH Open Access

Multi-view manifold regularized compact low-rank representation for cancer samples clustering on multi-omics data

Juan Wang[1*], Cong-Hai Lu[1], Xiang-Zhen Kong[1*], Ling-Yun Dai[1], Shasha Yuan[1] and Xiaofeng Zhang[2]

From The International Conference on Data Science, Analytics, and Engineering (IDSAE) 2020/2021
Virtual. 24-25 January 2021

*Correspondence:
wangjuansdu@163.com;
kongxzhen@163.com
[1] School of Computer Science, Qufu Normal University, Rizhao 276826, China
Full list of author information is available at the end of the article

Abstract

Background: The identification of cancer types is of great significance for early diagnosis and clinical treatment of cancer. Clustering cancer samples is an important means to identify cancer types, which has been paid much attention in the field of bioinformatics. The purpose of cancer clustering is to find expression patterns of different cancer types, so that the samples with similar expression patterns can be gathered into the same type. In order to improve the accuracy and reliability of cancer clustering, many clustering methods begin to focus on the integration analysis of cancer multi-omics data. Obviously, the methods based on multi-omics data have more advantages than those using single omics data. However, the high heterogeneity and noise of cancer multi-omics data pose a great challenge to the multi-omics analysis method.

Results: In this study, in order to extract more complementary information from cancer multi-omics data for cancer clustering, we propose a low-rank subspace clustering method called multi-view manifold regularized compact low-rank representation (MmCLRR). In MmCLRR, each omics data are regarded as a view, and it learns a consistent subspace representation by imposing a consistence constraint on the low-rank affinity matrix of each view to balance the agreement between different views. Moreover, the manifold regularization and concept factorization are introduced into our method. Relying on the concept factorization, the dictionary can be updated in the learning, which greatly improves the subspace learning ability of low-rank representation. We adopt linearized alternating direction method with adaptive penalty to solve the optimization problem of MmCLRR method.

Conclusions: Finally, we apply MmCLRR into the clustering of cancer samples based on multi-omics data, and the clustering results show that our method outperforms the existing multi-view methods.

Keywords: Low-rank subspace clustering, Concept factorization, Manifold regularization, Cancer multi-omics Data

Background

Cancer is a series of complex diseases with high heterogeneity. Nowadays, cancer has gradually become one of the most common and fatal diseases worldwide. Medical studies show that gene variation and mutation are the main factors leading to the formation and development of cancer diseases [1, 2]. Moreover, the abnormality and mutation mechanism of gene will lead to the pathological difference of cancer, thus forming different tumor types. As diagnosis of cancer is very important for the determination of cancer therapeutic schedule or regime, the identification of cancer types has attracted much attention in cancer research [3].

Sequencing technology has opened the omics era of life science and is leading and changing the development of the whole field of cancer research [2, 4]. With the development and popularization of sequencing technology, genomics has made great progress. The generation of massive cancer genomics data provides an effective avenue to investigate the pathogenesis of cancer at the genomic scale. As the most commonly used database for analyzing cancer sequencing data, The Cancer Genome Atlas (TCGA) can provide a variety of cancer genomics data, such as gene expression data, DNA methylation data, copy number variation data, gene regulation data and some clinical medical data [5]. These cross omic measurements provide valuable opportunities for systematic and in-depth study of cancer. In the past decade, TCGA data sets have been widely used in the study of individual cancer type and pan-cancer [6, 7]. And these studies based on TCGA data have contributed to the accumulation and discovery of cancer knowledge.

In the field of bioinformatics, machine learning algorithms play an important role in disease diagnosis, pathogenic factors discovery and treatment outcome prediction, etc. [8, 9]. As an exploratory algorithm in machine learning, clustering algorithm is often used to identify cancer types. In caner classification, the purpose of clustering algorithm is to find sample groups with similar expression patterns by analyzing omics data, so as to classify cancer patients or sample tissues. So far, many classical methods have been proposed for the detection of cancer categories. Gao et al. proposed sparse non-negative matrix factorization to identify cancer class based on gene expression profile [10]. In [11], Ye et al. applied independent component analysis (ICA) into tumor clustering. In [12], the penalized matrix decomposition method was proposed to cluster tumor according to meta samples based on gene expression data. In [13], Nguyen et al. used partial least squares for classification of multiple types of cancer. As in references [10–13], most studies use gene expression data to classify cancer types. With the deepening of cancer research, methylation profile is found to be different among tumor types and can be used as a powerful tool for sample identification [14, 15]. In addition, studies shown that copy number abnormality, as an important gene mutation, can lead to the abnormal growth of tissue cells and play an important role on genetic diversity and evolution [16, 17]. Therefore, these data can also be used as feature sources for cancer type recognition. For example, Polovinkin et al. used DNA methylation data to study the oncological diseases diagnosis, and achieved high accuracy in the classification of different types of cancer patients [18]. Virmani distinguished different subtypes of lung cancer based on DNA methylation markers [19].

All of the above studies indicate that a variety of mutation mechanisms contribute to the occurrence and development of cancer [20]. In order to investigate cancer type identification more accurately, it is necessary to analyze the cancer multi-omics data comprehensively. However, the heterogeneity, high noise, high feature dimensionality and small sample volume, and the differences in measurement and data types of different omics data bring a great challenge to the integrated analysis of multi-omics data [21]. To this end, a variety of integration and analysis algorithms have been proposed. These approaches are mainly divided into two categories. One is network-based methods. For example, Ma et al. presented Affinity Network Fusion (ANF) method to cluster patient using gene expression, miRNA expression and DNA methylation data [22]. Wang et al. developed Similarity Network Fusion (SNF) model to integrate microRNA expression, DNA methylation and mRNA expression data for cancer subtypes identifying [23]. The other is based on matrix decomposition methods. For example, Strazar et al. came up with an Integrative Orthogonality-regularized Nonnegative Matrix Factorization (iONMF) to deal with important information from multiple data sources [24]. Liu et al. presented Block-Constraint Robust Principal Component Analysis (BCRPCA) model to integrate and analysis TCGA data [25].

Recently, the low-rank representation method, namely LRR, was proposed to solve the problem of subspace clustering [26]. In LRR, the rank of representation matrix is considered as equivalent to the dimension of the low-dimensional subspace. LRR strengthens the correlation of representation vectors by enforcing low-rank constraint on the representation matrix. Benefiting from its pleasing efficacy in the acquisition of global structure of high-dimensional data, LRR is considered as a vigorous method and has received a great deal of attention. As a result, many improved methods based on LRR are developed, such as Latent Low-Rank Representation (LatLRR) [27], Structure-Constrained LRR (SC-LRR) [28], Non-negative Spare Hyper-Laplacian regularized LRR (NSHLRR) [29], graph regularized LRR under sparse and symmetric constraints (sgLRR) [30], and Laplacian regularized LRR (LLRR) [31]. However, these methods are only suitable to process single type data. When processing multi-view feature data, these methods may ignore the complementary information between views, thus reducing the learning performance of the algorithm. In order to deal with multi-view data, Brbić et al. developed Multi-view Low-Rank Sparse Subspace Clustering (MLRSSC) [32]. In MLRSSC model, a consistent low-rank affinity matrix is constructed from multi-view data to jointly learn subspace representation. The experimental results on both simulated and real datasets show that MLRSSC method has excellent clustering performance. In [32], it is shown that the MLRSSC framework is suitable for multimodal data, which is crucial to the analysis of heterogeneous multi-omics data. However, MLRSSC method does not consider the influence of local structure on manifold structure learning. Moreover, like most of the existing LRR based methods, it directly uses the observation data as the dictionary matrix to describe the subspaces of data. Since omics data of cancer are usually high-dimension and small sample, using observation data as spatial mapping benchmark will lead to insufficient expression of low-dimensional subspace, thus degrading the learning ability of LRR algorithm on data subspaces.

In light to the shortcomings described above, we present Multi-view Manifold Regularized Compact Low-Rank Representation method, which is called MmCLRR for short. Unlike most LRR based approaches, in MmCLRR, the concept factorization [33] idea is introduced to model dictionary matrix. Specifically, we consider the dictionary as a set of concepts, and each concept corresponds to a low-dimensional subspace, that is, the cluster center. According to concept factorization, the dictionary is modeled as a linear combination of original data. The dictionary matrix constructed by concept can enhance the description of the low-dimensional mapping space and help to obtain the structure of subspace accurately. Besides, the manifold regularization is also imposed on the low-rank affinity matrix to defend the local geometrical structure of each view. Similar to MLRSSC, the ultimate goal of MmCLRR is to achieve the consistent low-rank coefficient matrix from multi-view data. In MmCLRR, we jointly obtain the low-rank representation of multi-view by balancing the consistency of different views. At the same time, the balanced constraint on low-rank representation can avoid the noise propagation in the mapping process.

The key contributions of this study are summarized as below.

1. A multi-view based clustering analysis method named MmCLRR is proposed. Against specified fixed dictionary matrix used in most LRR methods, in MmCLRR, we adopt concept factorization to model the dictionary matrix. Concept factorization makes the dictionary update continuously during optimization, which enhances the completeness of dictionary and breaks through the bottleneck of using fixed dictionary matrix to describe subspace in LRR. In addition, we apply manifold regularization to further preserve the local topology of the data in the projecting. Benefiting by concept factorization and manifold regularization, the proposed method can capture the inherent subspace structure located in each view, and identify the latent subspace hidden in multi-view.
2. We apply MmCLRR to model cancer multi-omics data, and further propose a new cancer clustering framework based on multi omics data. This will make the clustering study of cancer get rid of the limitation of single omics data, and greatly promote the development of multi-omics data in cancer clustering research.
3. The clustering framework of MmCLRR is used to study cancer clustering, and many experiments of cancer samples clustering based on multi-omics data are provided. The experimental results indicate that it is feasible to cluster cancer using multi-omics data. These results also demonstrate the effectiveness of MmCLRR in cancer clustering.

The remainder of this article is schemed as follows. In Sect. 2 a brief overview of the related work including LRR, manifold regularization as well as concept factorization is given. In Sect. 3, the developed MmCLRR method and its model on cancer multi-omics data are elaborated. The experiment results and the performance analysis based on MmCLRR and several comparison methods are demonstrated in Sect. 4. The conclusion of this work is given in Sect. 5.

Methods

LRR and MLRSSC

LRR is an important method of subspace clustering firstly developed by Liu et al. [34]. The main idea of LRR is to regard high-dimensional data as a mapping from low-dimensional space. For specific high-dimensional data, the corresponding low-dimensional space is usually a combination of several independent subspaces. In other words, high-dimensional data can be regarded as the mapping combination of these low-dimensional subspaces. The tenet of LRR is to seek the subspace structure contained in high-dimensional observed data by calculating the mapping coefficient. Because the dimension of the low-dimensional subspace is far lower than that of the original observation data, the mapping coefficient of the high-dimensional data is low rank. Therefore, LRR aims to obtain the lowest rank coefficient matrix by optimizing the rank minimization problem. For observation data **X**, the object of LRR is defined as follows.

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

Here, **A** is the projection basis from high-dimensional space to low-dimensional space, often known as dictionary. The high-dimensional observation data can be formed by a linear combination of **A**, and the coefficients of linear combination constitutes matrix **Z**. So **Z** is called coefficient matrix, also named as low-rank representation matrix or low-rank affinity matrix. Supposing $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n]$, where n is the number of data points, then the column vector $\mathbf{z}_j$ is also thought as the mapping representation of the original data points j in each low-dimensional subspace. Therefore, matrix **Z** contains abundant subspace structure information for subspace segmentation.

In practice, the original high-dimensional data are directly regarded as **A**. And the nuclear norm is used as the surrogate of rank function to obtain the convex optimization of problem (1). The deformation of the optimal problem of LRR is as follows.

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_*, \ s.t. \ \ \mathbf{X} = \mathbf{XZ}. \tag{2}$$

Here, $\|\cdot\|_*$ is the nuclear norm and $\|\mathbf{Z}\|_* = \sum_i \sigma_i$, where σ_i is the singular value of **Z**. At this point, the elements of **Z** can be regarded as the similar expression between the original data points in the mapping space. In subspace segmentation, the data points with high similarity expression are approximately from the same subspace, so these data points are clustered into the same class.

Generally, the observations from the real world are noisy. In order to reduce the influence of noise on subspace learning, an error item is usually added to the object of LRR. For random noise, we often employ l_1-norm to characterize the error term. To this end, the optimization problem (2) can be transformed as:

$$\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \alpha\|\mathbf{E}\|_1, \ s.t. \ \ \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \tag{3}$$

where **E** indicates the error, $\|\cdot\|_1$ denotes l_1-norm which is a regularization strategy to make a matrix sparse and the l_1-norm of matrix **E** is defined as $\|\mathbf{E}\|_1 = \sum_i \sum_j |e_{ij}|$, α is a super parameter to balance the noise. After LRR decomposing, the minimizer $\mathbf{E}^*$ and $\mathbf{Z}^*$ can be acquired. Among them, $\mathbf{E}^*$ can be used for noise removal [35, 36] or feature selection, $\mathbf{Z}^*$ can be used for subspace clustering [37] or classification [38–40], and $\mathbf{XZ}^*$ can be used for the low-rank recovery of original data [41].

MLRSSC is a multi-view clustering framework. It jointly learns a subspace representation by constructing a consistent similarity matrix shared by multi-view data. Given a dataset $\mathbf{X} = \left\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(m_v)}\right\}$ containing m_v views, $\mathbf{X}^{(v)} \in R^{M^{(v)} \times N}$ corresponds to view v. Here, N denotes the number of samples, and all views are from the same sample group. $M^{(v)}$ denotes the feature number of view v, and each view has its own features. In MLRSSC, for the purpose of learning a joint representation matrix, the regularization item is introduced to ensure the agreement between affinity matrices of pairwise views. At the same time, MLRSSC encourages the sparsity of low-rank representation. The objective function of MLRSSC is as follows.

$$\begin{aligned} &\min \sum_{v=1}^{m_v} \left(\beta_1 \left\|\mathbf{C}^{(v)}\right\|_* + \beta_2 \left\|\mathbf{C}^{(v)}\right\|_1\right) + \sum_{1 \le v, w \le m_v, v \ne w} \lambda^{(v)} \left\|\mathbf{C}^{(v)} - \mathbf{C}^{(w)}\right\|_F^2 \\ &\quad s.t.\ \ \mathbf{X}^{(v)} = \mathbf{X}^{(v)}\mathbf{C}^{(v)}, \operatorname{diag}(\mathbf{C}^{(v)}) = 0, v = 1, \ldots, m_v. \end{aligned} \tag{4}$$

Here, $\mathbf{C}^{(v)}$ is the low-rank representation corresponding to view v. β_1, β_2 and $\lambda^{(v)}$ are parameters to balance low rank, sparse constraints and the consistency across views, respectively.

Manifold regularization

Usually, the naturally generated data are approximately regarded as to be located in a certain manifold. Many studies have shown that the manifold structure of data is very important to the low-dimensional space learning or low-dimensional representation [42, 43]. However, these data are usually from high-dimensional space and have insufficient sample size, which makes it very difficult to obtain the global structure of the data manifold accurately. In manifold theory, each small enough part of a manifold is considered to come from Euclidean space and the manifold can be regarded as the adhesion of these small parts. So, researchers focus on preserving the local structure information of manifold to learn the topological properties from scattered data. In practice, the nearest neighbor graph based on data points is used to model the local geometry of the data manifold [44]. Given $\mathbf{X} = [\mathrm{x}_1, \mathbf{x}_2, \ \ldots \ , \mathrm{x}_n]$ from an underlying submanifold of high-dimensional space, n is the number of data points, we can construct a nearest neighbor graph $\boldsymbol{G}$ with n nodes. In $\boldsymbol{G}$, each node corresponds to a sample point, and the sample points are connected by edges. More specifically, we first determine the k-nearest neighbors of each data point by calculating the Euclidean distance between the data points, and then assign the weights of the connecting edges between the data points. There are three main ways to assign the weights of edges. For more details, please refer [45]. In this paper, we use Gaussian Kernel to calculate the weights. For the edge connecting data points x_i and x_j, the according weight is set as

$$H_{ij} = \begin{cases} \mathrm{e}^{-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2t^2}} & \text{if } \mathbf{x}_j \in N_k(\mathbf{x}_i) \text{ or } \mathbf{x}_i \in N_k\left(\mathbf{x}_j\right) \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

Here, k is the number of nearest neighbors. $N_k\left(\mathbf{x}_j\right)$ denotes the set of k nearest neighbors based on x_j. For high-dimensional data $\mathbf{X}$, all the weights of the edges between data points form a symmetric weight matrix, which is denoted as $\mathbf{H}$. Because $\mathbf{H}$ contains the local structure information of the submanifold in which the observed data are located, based on **H**, every data point of the observation data can be represented as a linear combination of its nearest neighbors.

According to the basic assumption of manifold theory, namely, if two data points in the data manifold are close to each other, their mappings of the two data points in a new coordinates are still close [46], we can minimize the objective as shown in formula (6) to preserve the inherent local structure of high-dimensional data.

$$\begin{aligned}&\sum_{i,j}\left\|\mathbf{z}_i-\mathbf{z}_j\right\|^2 H_{ij}\\&=\sum_i \mathbf{z}_i^T\mathbf{z}_i D_{ii}-\sum_{i,j}\mathbf{z}_i^T\mathbf{z}_j H_{ij}\\&=tr\left(\mathbf{Z}(\mathbf{D}-\mathbf{H})\mathbf{Z}^T\right)\\&=tr\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^T\right).\end{aligned} \tag{6}$$

Here, $\mathbf{z}_i$ is the mapping expression of data point x_i. The matrix **D** is diagonal, and its diagonal element is defined as $D_{ii}=\sum_j H_{ji}$. $\mathbf{L}=\mathbf{D}-\mathbf{H}$ is named as graph Laplacian matrix [47]. $tr(\cdot)$ denotes the trace function. The manifold regularization is widely used to enhance various algorithms [48–50].

Concept factorization

The basic idea of concept factorization is that each prominent concept in the observation data set can be represented by associating data points with similar concepts [33]. Namely, each concept can be represented by the linear combination of the whole data points. The vectors generated by this linear combination characterizes the key concepts shared by relevant data points. Given data set $\mathbf{X}=[x_1,\mathbf{x}_2,\ \ldots\ ,x_n]$, x_i denotes data point i, then the concept $\mathbf{R}_c$ can be represented mathematically as follows.

$$R_c=\sum_{i=1}^{n} w_{ic}\mathbf{x}_i. \tag{7}$$

Here, w_{ic} is an association coefficient, showing the degree of association of x_i with concept $\mathbf{R}_c$.

On the other hand, the data point in the observation data can also be approximated by linear union of these concepts, in mathematics, which can be expressed in the following formula.

$$\mathbf{x}_i=\sum_c m_{ic}\mathbf{R}_c, \tag{8}$$

where m_{ic} is overlap coefficient that indicates how well x_i overlaps the concept $\mathbf{R}_c$. We denote the association coefficient matrix composed of coefficient w_{ic} as **W**, and the overlap coefficient matrix formed by m_{ic} as **M**. In mathematics, the idea of concept factorization can be formulated as follows.

$$\mathbf{X}\approx\mathbf{X}\mathbf{W}\mathbf{M}^T. \tag{9}$$

In Eq. (9), $\mathbf{XW}$ can be seen as center of concept, and $\mathbf{M}$ can be regarded as the projection of original data point on concept center. After concept factorization, we can find the prominent concepts in a given dataset and cluster membership for each data point. Due to the excellent performance of concept factorization in concept discovery, it has been widely concerned and applied into clustering research [51, 52].

The proposed MmCLRR method

In this part, the proposed Multi-view Manifold Regularized Compact Low-Rank Representation (MmCLRR) method and its solution are elaborated. And then the model of MmCLRR based on cancer multi-omics data is given.

Problem formulation and the solution

Most LRR-based methods select observed data as dictionary to learn the low-rank representation of high-dimensional data. The noise contained in the data and the insufficient sample size will lead to the incompleteness of the dictionary, which will directly affect the mapping expression of the original data in the low-dimensional space. To this end, we introduce concept decomposition into MLRSSC method to reconstruct dictionary matrix using the linear combination of original sample points. Meanwhile, in view of the advantages of manifold regularization in exploring the local structure of manifold, we further introduce manifold regularization into our method. In MmCLRR, we combine the sparse LRR model with the data dictionary modeling and manifold regularization constraints to obtain the subspace structure information comprehensively. Given a dataset with m_v views $\mathbf{X} = \left\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(m_v)}\right\}$, where $\mathbf{X}^{(v)}$ represents the v-th feature view, the MmCLRR method can be formulated as

$$\begin{aligned} &\min \sum_{v=1}^{m_v}\left[\left\|\mathbf{Z}^{(v)}\right\|_* + \gamma_1\left\|\mathbf{Z}^{(v)}\right\|_1 + \gamma_2\left\|\mathbf{E}^{(v)}\right\|_{2,1} + \gamma_3 tr\left(\mathbf{Z}^{(v)}\mathbf{L}^{(v)}\mathbf{Z}^{(v)T}\right)\right] \\ &\quad + \sum_{1\le v,w\le m_v, v\ne w} \gamma^{(v)}\left\|\mathbf{Z}^{(v)} - \mathbf{Z}^{(w)}\right\|_F^2 s.t.\ \ \mathbf{X}^{(v)} \\ &= \mathbf{X}^{(v)}\mathbf{W}^{(v)}\mathbf{Z}^{(v)} + \mathbf{E}^{(v)}, \mathbf{W}^{(v)T}\mathbf{W}^{(v)} = \mathbf{I}. \end{aligned} \tag{10}$$

Here, $\mathbf{Z}^{(v)}$, $\mathbf{E}^{(v)}$ is the low-rank affinity matrix and error item corresponding to view $\mathbf{X}^{(v)}$. $\mathbf{X}^{(v)}\mathbf{W}^{(v)}$ represents the center of cluster of $\mathbf{X}^{(v)}$. $\mathbf{W}^{(v)T}\mathbf{W}^{(v)} = \mathbf{I}$ is a constraint to ensure the stability of the model. γ_1, γ_2 and γ_3 are penalty parameters. The parameter $\gamma^{(v)}$ is to balance the consistency of coefficient matrix between different views. The last item in (10) can help to reduce the noise propagation in low-rank affinity matrix and encourage the similarity between the representation matrices of views. Take view v as an example, the decomposition of MmCLRR is shown in Fig. 1.

We use Linearized Alternating Direction Method with Adaptive Penalty (LADMAP) [53] to solve the optimization problem in (10). In order to facilitate the solution, we introduce three auxiliary variables $\mathbf{Z}_A$, $\mathbf{Z}_B$ and $\mathbf{Z}_C$ into the objective of MmCLRR. The problem (10) is converted into

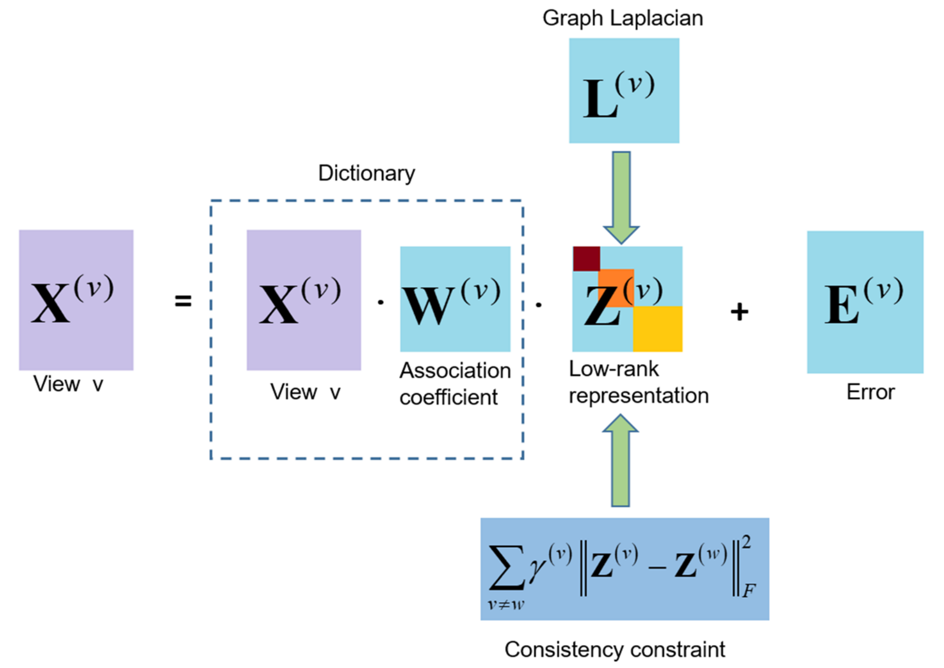


Fig. 1 The decomposition flowchart of MmCLRR on view *v*

$$\begin{aligned}
&\min \sum_{v=1}^{m_v}\left[\left\|\mathbf{Z}_A^{(v)}\right\|_* + \gamma_1\left\|\mathbf{Z}_B^{(v)}\right\|_1 + \gamma_2\left\|\mathbf{E}^{(v)}\right\|_{2,1} + \gamma_3 tr\left(\mathbf{Z}_A^{(v)}\mathbf{L}^{(v)}\mathbf{Z}_A^{(v)T}\right)\right]\\
&\quad + \sum_{1\leq v,w\leq m_v, v\neq w}\gamma^{(v)}\left\|\mathbf{Z}_C^{(v)}-\mathbf{Z}^{(w)}\right\|_F^2 \; s.t.\; \mathbf{Z}_A^{(v)}\\
&= \mathbf{Z}^{(v)}, \mathbf{Z}_B^{(v)} = \mathbf{Z}^{(v)}, \mathbf{Z}_C^{(v)} = \mathbf{Z}^{(v)}, \mathbf{H}^{(v)} = \mathbf{W}^{(v)}\mathbf{Z}^{(v)},\\
&\quad \mathbf{X}^{(v)} = \mathbf{X}^{(v)}\mathbf{H}^{(v)} + \mathbf{E}^{(v)}, \mathbf{W}^{(v)T}\mathbf{W}^{(v)} = \mathbf{I}.
\end{aligned} \tag{11}$$

Then, we draw into augmented Lagrangian method. The function (11) is recast as

$$\begin{aligned}
&\min \left\|\mathbf{Z}_A^{(v)}\right\|_* + \gamma_1\left\|\mathbf{Z}_B^{(v)}\right\|_1 + \gamma_2\left\|\mathbf{E}^{(v)}\right\|_{2,1} + \gamma_3 tr\left(\mathbf{Z}_A^{(v)}\mathbf{L}^{(v)}\mathbf{Z}_A^{(v)T}\right) + \sum_{1\leq v,w\leq m_v, v\neq w}\gamma^{(v)}\left\|\mathbf{Z}_C^{(v)}-\mathbf{Z}^{(w)}\right\|_F^2\\
&\quad + \frac{\mu_1}{2}\left\|\mathbf{X}^{(v)} - \mathbf{X}^{(v)}\mathbf{H}^{(v)} - \mathbf{E}^{(v)} + \frac{\mathbf{Y}_1}{\mu_1}\right\|_F^2 + \frac{\mu_2}{2}\left\|\mathbf{H}^{(v)} - \mathbf{W}^{(v)}\mathbf{Z}^{(v)} + \frac{\mathbf{Y}_2}{\mu_2}\right\|_F^2\\
&\quad + \frac{\mu_3}{2}\left\|\mathbf{Z}^{(v)} - \mathbf{Z}_A^{(v)} + \frac{\mathbf{Y}_3}{\mu_3}\right\|_F^2 + \frac{\mu_4}{2}\left\|\mathbf{Z}^{(v)} - \mathbf{Z}_B^{(v)} + \frac{\mathbf{Y}_4}{\mu_4}\right\|_F^2 + \frac{\mu_5}{2}\left\|\mathbf{Z}^{(v)} - \mathbf{Z}_C^{(v)} + \frac{\mathbf{Y}_5}{\mu_5}\right\|_F^2\\
&\quad \text{s.t. } \mathbf{W}^{(v)T}\mathbf{W}^{(v)} = \mathbf{I}.
\end{aligned} \tag{12}$$

Here, $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu$ are penalty parameters, $\mathbf{Y}_1 \sim \mathbf{Y}_5$ are Lagrange multipliers. Next,

The formula (12) is separated into the following sub problems with respect to $\mathbf{Z}_A^{(v)}$, $\mathbf{Z}_B^{(v)}$, $\mathbf{Z}_C^{(v)}$, $\mathbf{Z}^{(v)}$, $\mathbf{H}^{(v)}$, $\mathbf{W}^{(v)}$ and $\mathbf{E}^{(v)}$.

$$\begin{aligned} l_1\left(\mathbf{Z}^{(v)}\right) = \underset{\mathbf{Z}}{\arg\min}\, &\frac{\mu_2}{2}\left\|\mathbf{H}^{(v)} - \mathbf{W}^{(v)}\mathbf{Z}^{(v)} + \frac{\mathbf{Y}_2}{\mu_2}\right\|_F^2 + \frac{\mu_3}{2}\left\|\mathbf{Z}^{(v)} - \mathbf{Z}_A^{(v)} + \frac{\mathbf{Y}_3}{\mu_3}\right\|_F^2 \\ &+ \frac{\mu_4}{2}\left\|\mathbf{Z}^{(v)} - \mathbf{Z}_B^{(v)} + \frac{\mathbf{Y}_4}{\mu_4}\right\|_F^2 + \frac{\mu_5}{2}\left\|\mathbf{Z}^{(v)} - \mathbf{Z}_C^{(v)} + \frac{\mathbf{Y}_5}{\mu_5}\right\|_F^2. \end{aligned} \tag{13}$$

$$l_2\left(\mathbf{Z}_A^{(v)}\right) = \underset{\mathbf{Z}_A}{\arg\min} \left\|\mathbf{Z}_A^{(v)}\right\|_* + \gamma_3 tr\left(\mathbf{Z}_A^{(v)}\mathbf{L}^{(v)}\mathbf{Z}_A^{(v)T}\right) + \frac{\mu_3}{2}\left\|\mathbf{Z}^{(v)} - \mathbf{Z}_A^{(v)} + \frac{\mathbf{Y}_3}{\mu_3}\right\|_F^2. \tag{14}$$

$$l_3\left(\mathbf{Z}_B^{(v)}\right) = \underset{\mathbf{Z}_B}{\arg\min}\, \gamma_1\left\|\mathbf{Z}_B^{(v)}\right\|_1 + \frac{\mu_4}{2}\left\|\mathbf{Z}^{(v)} - \mathbf{Z}_B^{(v)} + \frac{\mathbf{Y}_4}{\mu_4}\right\|_F^2. \tag{15}$$

$$l_4\left(\mathbf{Z}_C^{(v)}\right) = \underset{\mathbf{Z}_C}{\arg\min} \sum_{1\le v,w\le m_v, v\ne w} \gamma^{(v)}\left\|\mathbf{Z}_C^{(v)} - \mathbf{Z}^{(w)}\right\|_F^2 + \frac{\mu_5}{2}\left\|\mathbf{Z}^{(v)} - \mathbf{Z}_C^{(v)} + \frac{\mathbf{Y}_5}{\mu_5}\right\|_F^2. \tag{16}$$

$$l_5\left(\mathbf{E}^{(v)}\right) = \underset{\mathbf{E}}{\arg\min}\, \gamma_2\left\|\mathbf{E}^{(v)}\right\|_{2,1} + \frac{\mu_1}{2}\left\|\mathbf{X}^{(v)} - \mathbf{X}^{(v)}\mathbf{H}^{(v)} - \mathbf{E}^{(v)} + \frac{\mathbf{Y}_1}{\mu_1}\right\|_F^2. \tag{17}$$

$$l_6\left(\mathbf{H}^{(v)}\right) = \underset{\mathbf{H}}{\arg\min} \frac{\mu_1}{2}\left\|\mathbf{X}^{(v)} - \mathbf{X}^{(v)}\mathbf{H}^{(v)} - \mathbf{E}^{(v)} + \frac{\mathbf{Y}_1}{\mu_1}\right\|_F^2 + \frac{\mu_2}{2}\left\|\mathbf{H}^{(v)} - \mathbf{W}^{(v)}\mathbf{Z}^{(v)} + \frac{\mathbf{Y}_2}{\mu_2}\right\|_F^2. \tag{18}$$

$$l_7\left(\mathbf{W}^{(v)}\right) = \underset{\mathbf{W}}{\arg\min} \frac{\mu_2}{2}\left\|\mathbf{H}^{(v)} - \mathbf{W}^{(v)}\mathbf{Z}^{(v)} + \frac{\mathbf{Y}_2}{\mu_2}\right\|_F^2 \text{ s.t. } \mathbf{W}^{(v)T}\mathbf{W}^{(v)} = \mathbf{I}. \tag{19}$$

Then, the final iterative algorithm is obtained by solving the above sub problems in turn. It is assumed that all variables after the k-th iteration are known. For example, the variable.

$\mathbf{E}^{(v)}$ in the k-th iteration is marked as $\mathbf{E}_k^{(v)}$. The iteration rules for each variable are as follows.

(1) Updating $\mathbf{Z}^{(v)}$. According to sub problem (13), we take the derivative with respect to $\mathbf{Z}^{(v)}$ and let the derivative be equal to 0. Then the iteration rule of $\mathbf{Z}^{(v)}$ is obtained as follows.

$$\begin{aligned} \mathbf{Z}^{(v)}_{(k+1)} = &\left[\mu_{2(k)}\mathbf{W}^{(v)T}_{(k)}\mathbf{W}^{(v)}_{(k)} + \left(\mu_{3(k)} + \mu_{4(k)} + \mu_{5(k)}\right)\mathbf{I}\right]^{-1} \\ &\times \left[\mu_{2(k)}\mathbf{W}^{(v)T}_{(k)}\mathbf{H}^{(v)}_{(k)} + \mu_{3(k)}\mathbf{Z}^{(v)}_{A(k)} + \mu_{4(k)}\mathbf{Z}^{(v)}_{B(k)} + \mu_{5(k)}\mathbf{Z}^{(v)}_{C(k)} \right. \\ &\left. + \mathbf{W}^{(v)T}_{(k)}\mathbf{Y}_{2(k)} - \mathbf{Y}_{3(k)} - \mathbf{Y}_{4(k)} - \mathbf{Y}_{5(k)}\right]. \end{aligned} \tag{20}$$

(2) Updating $\mathbf{Z}_A^{(v)}$. We take the derivative of the problem (14) with regard to $\mathbf{Z}_A^{(v)}$, and denote the derivative as $\nabla_{\mathbf{Z}} f\left(\mathbf{Z}^{(v)}_{A(k)}\right)$.

$$\nabla_{\mathbf{Z}} f\left(\mathbf{Z}^{(v)}_{A(k)}\right) = \gamma_3\left[\mathbf{Z}^{(v)}_{A(k)}\mathbf{L}^{(v)T} + \mathbf{Z}^{(v)}_{A(k)}\mathbf{L}^{(v)}\right] + \mu_{3(k)}\left(\mathbf{Z}^{(v)}_{A(k)} - \mathbf{Z}^{(v)}_{(k)} - \frac{\mathbf{Y}_{3(k)}}{\mu_{3(k)}}\right).$$

According to LADMAP, the solution of $\mathbf{Z}_A^{(\nu)}$ is transformed into the optimization of problem (22).

$$\min_{\mathbf{Z}_A} \left\|\mathbf{Z}_A^{(\nu)}\right\|_* + \left\langle \nabla_{\mathbf{Z}} f\left(\mathbf{Z}_{A(k)}^{(\nu)}\right), \mathbf{Z}_A^{(\nu)} - \mathbf{Z}_{A(k)}^{(\nu)} \right\rangle + \frac{\eta}{2}\left\|\mathbf{Z}_A^{(\nu)} - \mathbf{Z}_{A(k)}^{(\nu)}\right\|_F^2, \tag{22}$$

where $\eta = 2\gamma_3\left\|\mathbf{L}^{(\nu)}\right\|_2 + \mu_3\left(1 + \left\|\mathbf{X}^{(\nu)}\right\|_2^2\right)$. Then the solution to problem (14) is as follows.

$$\mathbf{Z}_{A(k+1)}^{(\nu)} = \Theta_{\frac{1}{\eta}}\left(\mathbf{Z}_{A(k)}^{(\nu)} - \nabla_{\mathbf{Z}} f\left(\mathbf{Z}_{A(k)}^{(\nu)}\right)\Big/\eta\right). \tag{23}$$

Here,$\Theta(\cdot)$ denotes skinny singular value decomposition and $\Theta_\varepsilon(\mathbf{A}) = US_\varepsilon\left(\sum\right)V^T$, where $S_\varepsilon(x) = \operatorname{sgn}(x)\max\left(|x| - \varepsilon, 0\right)$.

(3) Updating $\mathbf{Z}_B^{(\nu)}$. We find the partial derivative of problem (15) as below.

$$\frac{\partial l_3}{\partial \mathbf{Z}_{B(k)}^{(\nu)}} = \mu_{4(k)}\left(\mathbf{Z}_{B(k)}^{(\nu)} - \mathbf{Z}_{(k)}^{(\nu)} - \mathbf{Y}_{4(k)}\big/\mu_{4(k)}\right). \tag{24}$$

Let formula (24) be 0, and the expression of $\mathbf{Z}_{B(k)}^{(\nu)}$ is

$$\mathbf{Z}_{B(k)}^{(\nu)} = \mathbf{Z}_{(k)}^{(\nu)} + \mathbf{Y}_{4(k)}\big/\mu_{4(k)}. \tag{25}$$

According to literature [54], the literation rule of $\mathbf{Z}_B^{(\nu)}$ is as follows.

$$\mathbf{Z}_{B(k+1)}^{(\nu)} = S_{\frac{\gamma_1}{\mu_{4(k)}}}\left(\mathbf{Z}_{(k)}^{(\nu)} + \mathbf{Y}_{4(k)}\big/\mu_{4(k)}\right) \tag{26}$$

(4) Updating $\mathbf{Z}_C^{(\nu)}$. Similar with $\mathbf{Z}_A^{(\nu)}$, the solution of problem (16) is as bellow.

$$\mathbf{Z}_{C(k+1)}^{(\nu)} = \left[2\gamma^{(\nu)}(n_\nu - 1) + \mu_{5(k)}\right]^{-1}\left(2\gamma^{(\nu)}\sum_{1\le \nu, w\le m_\nu, \nu\ne w}^{m_\nu=3}\mathbf{Z}_{(k)}^{(w)} + \mu_{5(k)}\mathbf{Z}_{(k)}^{(\nu)} + \mathbf{Y}_{5(k)}\right). \tag{27}$$

(5) Updating $\mathbf{E}^{(\nu)}$. According to reference [34], the iterative formula of $\mathbf{E}^{(\nu)}$ is.

$$\mathbf{E}_{(k+1)}^{(\nu)}(:, i) = \begin{cases} \frac{\|\mathbf{G}(:,i)\| - \gamma_2/\mu_{1(k)}}{\|\mathbf{G}(:,i)\|}\mathbf{G}(:, i), & \lambda_2/\mu_{1(k)} < \|\mathbf{G}(:, i)\| \\ 0 & \text{otherwise} \end{cases}. \tag{28}$$

Here, $\mathbf{G} = \mathbf{X}^{(\nu)} - \mathbf{X}^{(\nu)}\mathbf{H}_{(k)}^{(\nu)} + \mathbf{Y}_{1(k)}\big/\mu_{1(k)}$.

(6) Updating $\mathbf{H}^{(\nu)}$. Similar with $\mathbf{Z}_A^{(\nu)}$ and $\mathbf{Z}_C^{(\nu)}$, the updating rule of $\mathbf{H}^{(\nu)}$ is as.

$$\begin{aligned} \mathbf{H}_{(k+1)}^{(\nu)} = &\left(\mu_{1(k)}\mathbf{X}^{(\nu)T}\mathbf{X}^{(\nu)} + \mu_{2(k)}\mathbf{I}\right)^{-1}\left(\mu_{1(k)}\mathbf{X}^{(\nu)T}\mathbf{X}^{(\nu)} - \mu_{1(k)}\mathbf{X}^{(\nu)T}\mathbf{E}_{(k)}^{(\nu)}\right. \\ &\left. + \mathbf{X}^{(\nu)T}\mathbf{Y}_{1(k)} + \mu_{2(k)}\mathbf{W}_{(k)}^{(\nu)}\mathbf{Z}_{(k)}^{(\nu)} - \mathbf{Y}_{2(k)}\right). \end{aligned} \tag{29}$$

(7) Updating $\mathbf{W}^{(\nu)}$. Referring to Theorem 1 in [55], we solve sub problem (19) and get the iteration of $\mathbf{W}^{(\nu)}$ as follows.

$$\mathbf{W}_{(k+1)}^{(\nu)} = \mathbf{U}\mathbf{V}^T. \tag{30}$$

Here, $\left(\mathbf{H}^{(v)} + \mathbf{Y}_{2(k)}/\mu_{2(k)}\right)\left(\mathbf{Z}_{(k)}^{(v)}\right)^T = \mathbf{U}\mho\mathbf{V}^T,\ \mho = \mathrm{diag}(\delta)$.

(8) Updating $\mathbf{Y}_1 \sim \mathbf{Y}_5$.

$$\mathbf{Y}_{1(k+1)} = \mathbf{Y}_{1(k)} + \mu_1\left(\mathbf{X}^{(v)} - \mathbf{X}^{(v)}\mathbf{H}_{(k)}^{(v)} - \mathbf{E}_{(k)}^{(v)}\right). \tag{31}$$

$$\mathbf{Y}_{2(k+1)} = \mathbf{Y}_{2(k)} + \mu_2\left(\mathbf{H}_{(k)}^{(v)} - \mathbf{W}_{(k)}^{(v)}\mathbf{Z}_{(k)}^{(v)}\right). \tag{32}$$

$$\mathbf{Y}_{3(k+1)} = \mathbf{Y}_{3(k)} + \mu_3\left(\mathbf{Z}_{(k)}^{(v)} - \mathbf{Z}_{A(k)}^{(v)}\right). \tag{33}$$

$$\mathbf{Y}_{4(k+1)} = \mathbf{Y}_{4(k)} + \mu_4\left(\mathbf{Z}_{(k)}^{(v)} - \mathbf{Z}_{B(k)}^{(v)}\right). \tag{34}$$

$$\mathbf{Y}_{5(k+1)} = \mathbf{Y}_{5(k)} + \mu_5\left(\mathbf{Z}_{(k)}^{(v)} - \mathbf{Z}_{C(k)}^{(v)}\right). \tag{35}$$

Finally, based on the low-rank representation matrix of each view, we calculate the fused affinity matrix $\mathbf{Z}^*$ by formula (36).

$$\mathbf{Z}^* = \frac{\mathrm{sum}\left\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \mathbf{Z}^{(3)}\right\}}{m_v}. \tag{36}$$

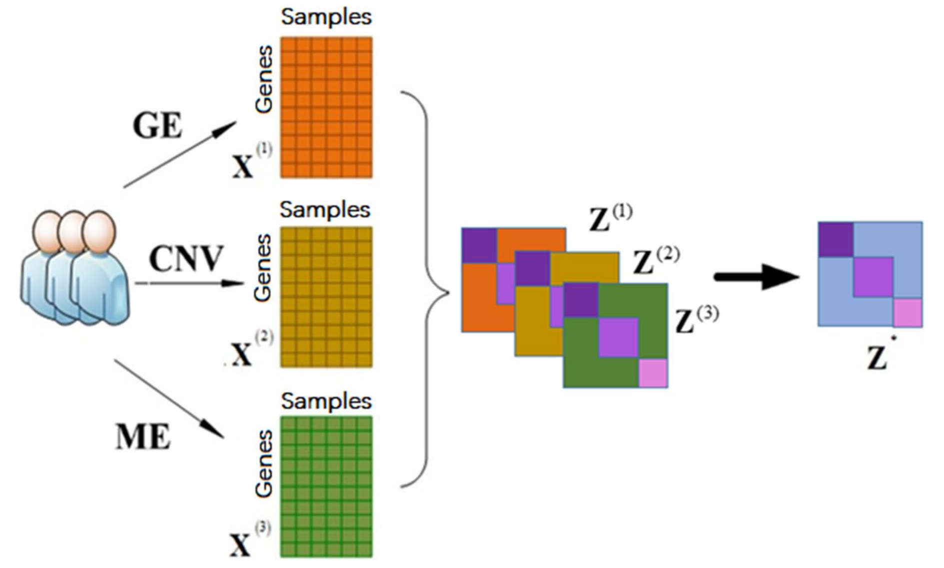


Fig. 2 The MmCLRR model on cancer multi-omics data for cancer samples clustering

```
Algorithm 1. MmCLRR Algorithm Using LADMAP

Input: Multi-view data $\left\{\mathbf{X}^{(v)}\right\}_{v=1}^{m_v}$; parameters $\gamma_1, \gamma_2, \gamma_3$ and $\left\{\gamma^{(v)}\right\}_{v=1}^{m_v}$;
Initialization: $\mathbf{Z}_{(0)}^{(v)} = \mathbf{Z}_{A(0)}^{(v)} = \mathbf{Z}_{B(0)}^{(v)} = \mathbf{Z}_{C(0)}^{(v)} = \mathbf{E}_{(0)}^{(v)} = \mathbf{H}_{(0)}^{(v)} = \mathbf{W}_{(0)}^{(v)} = 0$, $\left\{\mathbf{Y}_{i(0)} = \mathbf{0}\right\}_{i=1}^{5}$, $\mathbf{L}^{(v)}$, $\rho = 1.5$, $\mu_0 = 10$,
    $\mu_{\max} = 10^6$, $\varepsilon = 10^{-3}$.
While not converged do
    For $v = 1$ to $m_v$
        Update $\mathbf{Z}^{(v)}$ by (20);      Update $\mathbf{Z}_A^{(v)}$ by (23);
        Update $\mathbf{Z}_B^{(v)}$ by (26);    Update $\mathbf{Z}_C^{(v)}$ by (27);
        Update $\mathbf{E}^{(v)}$ by (28);      Update $\mathbf{H}^{(v)}$ by (29);
        Update $\mathbf{W}^{(v)}$ by (30);      Update $\mathbf{Y}_1 \sim \mathbf{Y}_5$ by (31)~(35).
    End
Checking convergence:
    $\max\left\{\begin{array}{l} \left\|\mathbf{Z}_{(k)}^{(v)} - \mathbf{Z}_{A(k)}^{(v)}\right\|_\infty, \left\|\mathbf{Z}_{(k)}^{(v)} - \mathbf{Z}_{B(k)}^{(v)}\right\|_\infty, \left\|\mathbf{Z}_{(k)}^{(v)} - \mathbf{Z}_{C(k)}^{(v)}\right\|_\infty, \left\|\mathbf{Z}_{(k+1)}^{(v)} - \mathbf{Z}_{(k)}^{(v)}\right\|_\infty, \\ \left\|\mathbf{X}^{(v)} - \mathbf{X}^{(v)}\mathbf{H}_{(k)}^{(v)} - \mathbf{E}_{(k)}^{(v)}\right\|_\infty, \left\|\mathbf{H}_{(k)}^{(v)} - \mathbf{W}_{(k)}^{(v)}\mathbf{Z}_{(k)}^{(v)}\right\|_\infty \end{array}\right\} \leq \varepsilon$.
    Update $\mu_{k+1}: \mu_{k+1} = \min\left(\mu_{\max}, \rho\mu_k\right)$,
End while
Calculate the matrix $\mathbf{Z}^*$ by (36)

Output: $\mathbf{Z}^*$
```

The detailed optimization process of MmCLRR method is shown in Algorithm 1.

The MmCLRR model on cancer multi-omics data

As mentioned earlier, besides gene expression data, DNA methylation and copy number variation also play important roles in the formation and development of cancer. And these omics data of cancer have been used alone or in combination with other data for cancer type research. This fully shows that these data contain the characteristic information needed in cancer type recognition. Thinking different omics data as the expression of cancer features at different levels, it is reasonable for us to regard that the feature information in these omics data can complement each other. Therefore, we intend to fuse the characteristic information of gene expression data, DNA methylation data and copy number variation data to cluster cancer samples. Here, we think of each omics data as a feature view of cancer, and use MmCLRR method to model these omics data. The schematic diagram of MmCLRR model on multi-omics data is shown in Fig. 2. In Fig. 2, gene expression data is abbreviated as GE, copy number variation is abbreviated as CNV, and DNA methylation is abbreviated as ME. $\mathbf{Z}^{(1)}$, $\mathbf{Z}^{(2)}$ and $\mathbf{Z}^{(3)}$ denote the low-rank representation matrix corresponding to GE, CNV and ME, respectively. In this model, we are not sure which omics data are more important, so we regard the proportion of each omics data in the model as the same, and use the same $\gamma^{(v)}$ for all omics data. After the decomposition of MmCLRR, we adopt NCuts clustering method to cluster cancer samples based on the fused matrix $\mathbf{Z}^*$.

Table 1 Samples and genes distribution of each omics data in the experimental datasets

Datasets	Omics data	Genes	Normal samples	Cancer samples
HNSC	Gene expression	2002	20	398
	DNA methylation	23627	20	398
	Copy number variation	21031	20	398
COAD	Gene expression	20502	19	262
	DNA methylation	23627	19	262
	Copy number variation	21031	19	262
ESCA	Gene expression	20502	9	183
	DNA methylation	23627	9	183
	Copy number variation	21031	9	183

Table 2 The clustering performance of five methods on three experimental data sets

Multi-omics data	Metrics	ioNMF (%)	SNF (%)	BLLRR (%)	MLRSSC (%)	MmCLRR (%)
HNSC	Acc	69.38	90.29	97.58	78.71	**99.52**
	RI	57.41	83.74	92.00	67.23	**99.02**
	F1	44.45	47.51	74.22	55.37	**97.24**
COAD	Acc	64.76	86.30	98.27	74.93	**98.93**
	RI	54.87	78.53	89.38	63.21	**97.88**
	F1	66.06	50.29	72.13	58.90	**96.05**
ESCA	Acc	67.32	84.31	96.88	69.53	**96.25**
	RI	55.84	77.80	**93.91**	65.71	92.83
	F1	45.18	46.97	79.44	50.03	**83.50**

Best clustering results are highlighted in bold

Results

Evaluation metrics

We use Accuracy (Acc) [56], Rand Index (RI) [57] and F1 measurment (F1) [58] as the evaluation metrics of clustering performance. The following is a brief introduction to these metrics.

Acc evaluates the clustering performance at the global level by calculating the matching degree between the experimental labels and the actual labels. It is defined as follows.

$$\mathrm{Acc} = \frac{\sum_{i=1}^{N} \delta(p_i, map(q_i))}{N} \times 100\% \,. \tag{37}$$

Here, q_i and p_i denote the experimental label and actual label of data point i, respectively. N represents the number of data points. $map(q_i)$ is a function to match the experimental labels with the actual labels, and the method called Kuhn–Munkres [59] is usually employed to implement the matching. $\delta(p_i, map(q_i))$ is a function that compares the experimental tag with the actual tag. For data point i, if the experimental label q_i is the same as the actual label p_i, the function value is assigned as 1, otherwise it is assigned as 0.

Table 3 The average clustering results of three low-rank methods

Multi-omics data	Metrics	ioNMF (%)	SNF (%)	AVG-LRSC (%)
HNSC	Acc	69.38	90.29	**91.94**
	RI	57.41	83.74	**86.08**
	F1	44.45	47.51	**75.61**
COAD	Acc	64.76	86.30	**90.71**
	RI	54.87	78.53	**83.49**
	F1	66.06	50.29	**75.69**
ESCA	Acc	67.32	84.31	**87.55**
	RI	55.84	77.80	**84.15**
	F1	45.18	46.97	**70.99**

Best clustering results are highlighted in bold

RI assesses the performance of clustering algorithm by comparing the relationship between the actual classification and the experimental classification. The following is the definition of RI.

$$\mathrm{RI} = \frac{a+b}{C^2_{n_{samples}}} \times 100\% \,. \tag{38}$$

Here, a represents the number of data point pairs belonging to the same class in the actual classification and experimental classification. And b denotes the number of data point pairs that are not in the same class. $C^2_{n_{samples}}$ is the total number of data pairs clustered or classified.

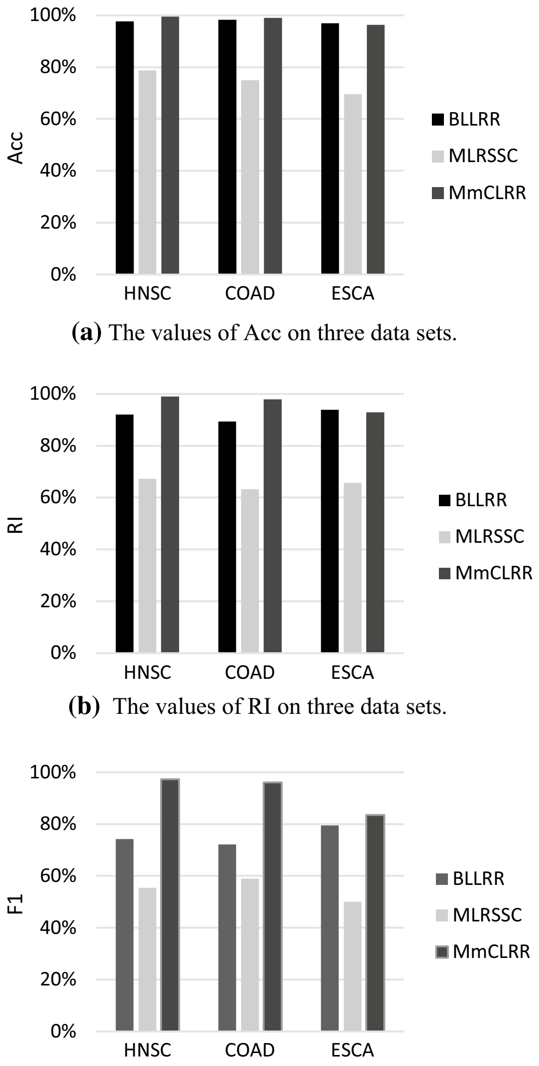
F1 is the average of precision rate and recall rate, which is defined as below.

$$\mathrm{F1} = \frac{2 * P * R}{P + R} \times 100\% \,. \tag{39}$$

Here, $P = \frac{TP}{TP+FP} \times 100\%$ and $R = \frac{TP}{TP+FN} \times 100\%$ denote precision rate and recall rate respectively, where TP means that positive samples are clustered into positive class, FP indicates that negative samples are wrongly classified into positive class, and FN means that positive samples are classified into negative classes.

Data sets

The data sets used in our study, including Head and Neck cancer (HNSC), Esophagus Cancer (ESCA) and Colon Adenocarcinoma (COAD), are downloaded from TCGA. Each data set contains three types of omics data, namely gene expression, DNA methylation data and copy number variation. And these omics data in each dataset come from the same batch of samples. Each of the three data sets includes cancer samples and normal samples. Specifically, HNSC consists of 398 cancer samples and 20 normal samples, ESCA includes 183 cancer samples and 9 normal samples, and COAD has got 262 cancer samples and 19 normal samples. The number of genes in gene expression, DNA methylation data and copy number variation data is 20502, 23,627 and 21,031, respectively. The samples and genes distribution of each omics data is shown in Table 1.



(a) The values of Acc on three data sets.

(b) The values of RI on three data sets.

(c) The values of F1 on three data sets.

Fig. 3 The clustering results of three LRSC methods

Results and analysis

In order to test and verify the performance of our method in cancer samples clustering, we compare MmCLRR with the existing multi-views analysis methods, including ioNMF [24], SNF [23], Block- constraint Laplacian regularized LRR (BLLRR) [60] and MLRSSC [32]. In order to evaluate the performance of each clustering method more objectively, the clustering experiment of each method is executed 50 times, and the average values obtained from 50 experiments are used to evaluate the clustering results. The experimental results on HNSC, COAD and ESCA are shown in Table 2. And the best results of each data set are represented in bold. From Table 2, we can see that our method outstrip all comparison methods. Next, we will compare and analyze the experimental results in detail.

Among the methods, BLLRR, MLRSSC and MmCLRR are low-rank subspace clustering (LRSC) methods. These LRSC methods mainly use the nuclear norm constraint to obtain the low-rank representation of multi-omics data, so as to explore the subspace structure of data. And, they construct the affinity matrix based on low-rank representation for cancer samples clustering. SNF is a network-based approach. It constructs similarity network for each omics data, and then integrates these networks generated by different omics data to realize samples clustering. The ioNMF approach is a NMF-based method. In ioNMF, different omics data are decomposed into a common fusion matrix and multiple independent sub matrixes at the same time, and then the common matrix is used to cluster samples. So, we firstly compare the three subspace clustering methods with ioNMF and SNF. From Table 2, we can find that the clustering results of the three subspace clustering methods are generally better than those of ioNMF and SNF. For this reason, we further calculate the mean values of BLLRR, MLRSSC and MmCLRR on each clustering metric (see Table 3). In Table 3, the average of LRSC methods is denoted as AVG-LRSC. And we also show the best results in bold. As can be seen from Table 3, the average clustering performance of these LRSC methods is significantly higher than the other two methods. The above analysis shows that LRSC method has a significant advantage in subspace learning.

Among the three LRSC methods, MLRSSC method does not take the local topology of data into account in subspace learning. Different from MLRSSC method, both BLLRR and MmCLRR methods are all committed to obtaining the global and local structures of manifold in multi-omics data by introducing manifold regularization constraint into LRR. Therefore, next, we compare MLRSSC with BLLRR and MmCLRR. For the convenience of comparison, as shown in Fig. 3, the histograms of clustering results on these three methods are given. From Fig. 3, it can be find that the values of all measures on method BLLRR and MmCLRR are higher than those on method MLRSSC. This indicates that the local geometry structure embedded in high-dimensional data is very vital to subspace segment problem. Preserving the local structure information of high-dimensional data during spatial mapping is helpful to smooth the manifold structure of the data in low-dimensional space and improve the subspace learning performance of the low-rank representation algorithm.

Thirdly, we compare MmCLRR with BLLRR. As be seen in Fig. 3, the experiment results of MmCLRR are better than BLLRR, especially on HNSC data set. First, for multi-omics analysis, the frameworks of the two methods are different. BLLRR is a method based on integrated multi-omics data. In BLLRR, the multi-omics data are integrated to form a comprehensive data matrix across omics. And the consistent low-dimensional subspace representation shared by multi-omics data is learned from the integrated data by imposing different penalty constraints on different omics data. MmCLRR is a

Table 4 The paremeter values of MmCLRR on each experimental data set

Multi-omics data	γ_1	γ_2	γ_3	γ
HNSC	10^{-1}	10^{4}	10^{-1}	10^{0}
COAD	10^{-1}	10^{-1}	10^{-1}	10^{2}
ESCA	10^{0}	10^{-1}	10^{2}	10^{2}

method of multi-views learning. In MmCLRR, it is considered that the subspace representation from different views should be consistent. MmCLRR jointly learns the representation matrix of each view by enforcing the balance constraint between different views. In addition, if BLLRR is employed to single omics data, the objective of MmCLRR is transformed into $\min \|\mathbf{Z}\|_* + \gamma_1\|\mathbf{Z}\|_1 + \gamma_2\|\mathbf{E}\|_{2,1} + \gamma_3 tr\left(\mathbf{ZLZ}^T\right)\ s.t.\ \mathbf{X} = \mathbf{XZ} + \mathbf{E}$, that is, MmCLRR method is changed into BLLRR method. Similarly, when analyzing single omics data, the objective of MmCLRR method will become $\min \|\mathbf{Z}\|_* + \gamma_1\|\mathbf{Z}\|_1 + \gamma_2\|\mathbf{E}\|_{2,1} + \gamma_3 tr\left(\mathbf{ZLZ}^T\right)\quad s.t.\ \mathbf{X} = \mathbf{XWZ} + \mathbf{E}, \mathbf{WW}^T = \mathbf{I}$. Obviously, the only difference between the two methods for single view is that the dictionary is constructed differently. BLLRR uses the original data as dictionary, which is fixed in iterative learning. And, MmCLRR applies the idea of concept factorization to construct dictionary matrix, which is constantly updated in learning. According to the above analysis, the clustering advantage of MmCLRR can be attributed to two points. One is that the multi-views learning model is more suitable for cross group analysis than the analysis model on integrated data. Another point is the successful modeling of dictionary by concept idea.

Finally, the MmCLRR approach is compared with MLRSSC. These two methods are basically consistent in the framework and main ideas for multi-view processing. There are two differences between them. On the one hand, compared with MLRSSC, manifold constraint is introduced into MmCLRR. On the other hand, the construction methods of dictionary are different. As mentioned above, MmCLRR takes the linear combination of original data as dictionary to update the dictionary matrix with the algorithm optimization, while MLRSSC uses original data as the fixed dictionary. From Fig. 3, we can see that the clustering advantage of MmCLRR method is much larger than that of BLLRR method. This fully shows that both manifold constraint and dictionary modeling make the low-rank representation matrix obtained by MmCLRR better distinguishable in subspace separation.

The setting of parameters

In MmCLRR method, there are four regularization parameters γ_1, γ_2, γ_3 and $\gamma^{(v)}(v = 1, 2, 3)$. As mentioned in the previous section, there is no prior knowledge to prove which omics data are more important in low dimensional learning. So we think that the proportion of each omics data in MmCLRR model is the same, and we use the same adjustment parameter γ for all the three omics data, i.e., $\gamma = \gamma^{(1)} = \gamma^{(2)} = \gamma^{(3)}$. In our experiment, the parameters are set by grid search, and the parameter values are shown in Table 4.

Discussion

MmCLRR is a novel multi-view integration analysis framework based low-rank decomposition. Our main contribution is to model dictionary matrix by concept factorization, which enables the dictionary matrix to update with subspace learning, thus enhancing the ability of dictionary to describe subspace. The comparative experiment of MmCLRR with other four multi-view methods is given on real multi-omics data. And the experiment results indicate that MmCLRR has a good performance in subspace clustering. In our experiment MmCLRR treats all omics data equally, so the parameter $\gamma^{(v)}$, balancing the consistency of low-rank representation of different views, is set to the same. If different views are of different importance in the analysis, the parameter $\gamma^{(v)}$ should be set to different values, which may increase the difficulty of parameter adjustment. Therefore, the increasing number of multi-view and the difference of their importance will be the main challenges for MmCLRR method.

Conclusions

In this study, we develop a multi-view low-rank subspace clustering method, named as MmCLRR, to analyze caner multi-omics data. MmCLRR aims to achieve the consistent low-rank representation from multi-view data by balancing the consistency of different views. In our method, concept factorization is adopted to model dictionary. That is, the dictionary is constructed as the combination of the original data. Furthermore, the manifold regularization is introduced into our method to grasp the local structural relationship within the data. So, MmCLRR can capture the global and local structure of submanifold shared by multi-view data more efficiently. Finally, we adopt the proposed method to cluster cancer samples based on multi-omics data from TCGA. The experimental results demonstrated that our method can outperform the state-of-the-art multi-view approaches. In the future, we will promote the application of MmCLRR in other fields of cancer research.

Abbreviations

LRR: Low-rank representation; MmCLRR: Multi-view manifold regularized compact low-rank representation; LADMAP: Linearized alternating direction method with adaptive penalty.

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 BMCBioinformatics Volume 22 Supplement 12 2021: Explainable AI methods inbiomedical data science. The full contents of the supplement are available athttps://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-22-supplement-12.

Authors' contributions

JW and CHL jointly designed and implemented the MmCLRR method. CHL carried out the experiments. JW drafted the manuscript. XZK and LYD gave computational advice and analyzed the results of the methods. SSY and XFZ participated in designing evaluation criteria and polishing the English expression. All authors read and approved the final manuscript.

Funding

Publication costs are funded by the NSFC under grant Nos. 61702299, 61902215, 61701279, and 61873117. 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 used in this study can be found in the [The Cancer Genome Atlas (TCGA)] https://cancergenome.nih.gov/.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Computer Science, Qufu Normal University, Rizhao 276826, China. [2]School of Information and Electrical Engineering, Ludong University, Yantai 264025, China.

Received: 25 May 2021 Accepted: 27 May 2021
Published online: 20 January 2022

References

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74.
2. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M. The landscape of somatic copy-number alteration across human cancers. Nature. 2010;463(7283):899–905.
3. Chen X, Wang H, Yan D: Clustering of transcriptomic data for the identification of cancer subtypes. 2018.
4. Dai W, Zheng H, Cheung A, Tang S, Ko J, Wong W, Leong M, Sham P, Cheung M, Kwong D: Whole-exome sequencing identifies MST1R as a genetic susceptibility gene in nasopharyngeal carcinoma. In: Proceedings of the National Academy of Sciences 2016.
5. Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, Ally A, Balasundaram M, Birol I, Butterfield YSN, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45:1113–20.
6. Dellinger TH, Smith DD, Ouyang C, Warden CD, Williams JC, Han ES. L1CAM is an independent predictor of poor survival in endometrial cancer—an analysis of The Cancer Genome Atlas (TCGA). Gynecol Oncol. 2016;141:336–40.
7. Han X. Improving gene expression cancer molecular pattern discovery using nonnegative principal component analysis. Genome Inf Int Conf Genome Inf. 2008;21(21):200–11.
8. Cherif W. Optimization of K-NN algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis. Proc Comput. 2018;127:293–9.
9. Chakraborty D, Maulik U. Identifying cancer biomarkers from microarray data using feature selection and semisupervised learning. IEEE J Transl Eng Health Med. 2014;2:1–11.
10. Yuan G, George C. Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics. 2005;21:3970.
11. Ye F, Xia JF, Chong YW, Zhang Y, Zheng CH: Tumor clustering using independent component analysis and adaptive affinity propagation. In: International conference on intelligent computing. 2014.
12. Zheng CH, Wang J, Ng TY, Shiu CK: Tumor clustering based on penalized matrix decomposition. 2010.
13. Nguyen DV, Rocke DM. Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics. 2002;18(9):1216–26.
14. Siegmund KD, Laird PW, Laird-Offringa IA. A comparison of cluster analysis methods using DNA methylation data. Bioinformatics. 2004;20(12):1896–904.
15. Zhuang J, Widschwendter M, Teschendorff AE. A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. BMC Bioinform. 2012;13(1):1–14.
16. Ivakhno S, Royce T, Cox AJ, Evers DJ, Cheetham RK, Tavare S. CNAseg—a novel framework for identification of copy number changes in cancer from second-generation sequencing data. Bioinformatics. 2010;26(24):3051–8.
17. Reuters T. Global variation in copy number in the human genome. Nature. 2006;444(7118):444–54.
18. Polovinkin AN, Druzhkov PN, Krylov IB, Zaikin AA, Ivanchenko MV, Zolotykh NY, Meyerov IB: Solving problems of clustering and classification of cancer diseases based on DNA methylation data. Pattern recognition and image analysis: advances in mathematical theory and applications in the USSR 2016.
19. Virmani AK, Tsou JA, Siegmund KD, Shen LYC, Laird-Offringa IA. Hierarchical clustering of lung cancer cell lines using DNA methylation markers. Cancer Epidemiol Biomark Prev. 2002;11(3):291–7.
20. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458(7239):719.
21. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, Milanesi L. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinform. 2016;17(2):S15.
22. Tianle M, Aidong Z: Affinity network fusion and semi-supervised learning for cancer patient clustering. Methods 2018:S1046202317304930.
23. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11(3):333–7.
24. Stražar M, Žitnik M, Zupan B, Ule J, Curk T. Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. Bioinformatics. 2016;32(10):1527–35.
25. Liu JX, Gao YL, Zheng CH, Xu Y, Yu J. Block-Constraint robust principal component analysis and its application to integrated analysis of TCGA data. IEEE Trans Nanobiosci. 2016;15(6):510–6.
26. 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.
27. 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: 1615–1622.
28. Tang, Kewei, Liu, Risheng, Zhang, Jie, Su, Zhixun: Structure-constrained low-rank representation. IEEE Trans Neural Netw Learn Syst. 2014.
29. 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.
30. Wang J, Lu CH, Liu JX, Dai LY, Kong XZ: Multi-cancer samples clustering via graph regularized low-rank representation method under sparse and symmetric constraints. BMC Bioinform 2019, 20(S22).
31. Wang J, Liu J-X, Kong X-Z, Yuan S-S, Dai L-Y. Laplacian regularized low-rank representation for cancer samples clustering. Comput Biol Chem. 2019;78:504–9.
32. Brbić M, Kopriva I. Multi-view low-rank sparse subspace clustering. Pattern Recogn. 2018;73:247–58.
33. Xu W, Gong Y: Document clustering by concept factorization. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. 2004: 202–209.
34. 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: 663–670.

35. Lu X, Wang Y, Yuan Y. Graph-regularized low-rank representation for destriping of hyperspectral images. IEEE Trans Geosci Remote Sens. 2013;51(7):4009–18.
36. Wang M, Yu J, Xue JH, Sun W. Denoising of hyperspectral images using group low-rank representation. IEEE J Sele Top Applied Earth Obs Remote Sens. 2016;9(9):4420–7.
37. W En J, Zhang B, Xu Y, Yang J, Han N: Adaptive weighted nonnegative low-rank representation. Pattern Recognit. 2018:326–340.
38. Chen J, Yang J. Robust subspace segmentation via low-rank representation. IEEE Trans Cybern. 2014;44(8):1432.
39. Wei L, Wu A, Yin J: Latent space robust subspace segmentation based on low-rank and locality constraints. Exp Syst Appl. 2015 42(19).
40. Wei L, Wang X, Wu A, Zhou R, Zhu C. Robust Subspace segmentation by self-representation constrained low-rank representation. Neural Process Lett. 2018;48(3):1671–91.
41. Zhou PL, et al. Shuicheng tensor low-rank representation for data recovery and clustering. IEEE Trans Pattern Anal Mach Intell. 2020. https://doi.org/10.1109/TPAMI.2019.2954874.
42. Wang C, He X, Bu J, et al. Image representation using Laplacian regularized nonnegative tensor factorization. Pattern Recognit. 2011;44(10–11):2516–26.
43. Sha L, Schonfeld D, Wang J. Graph laplacian regularization with sparse coding for image restoration and representation. IEEE Trans Circuits Syst Video Technol. 2019;PP(99):1–1.
44. Seung H. S: the manifold ways of perception. Science. 2000;290(5500):2262–9.
45. Cai D, He X, Han J. Document clustering using locality preserving indexing. IEEE Trans Knowl Data Eng. 2005;17(12):1624–37.
46. He X. Locality preserving projections. Adv Neural Inf Process Syst. 2003;16(1):186–97.
47. Chung FRK: Spectral graph theory. 2012;413–439.
48. Cai D, He X, Han J, Huang TS. Graph regularized nonnegative matrix factorization for data representation. IEEE Trans Pattern Anal Mach Intell. 2011;33(8):1548–60.
49. Xie J, Liu S, Dai H. Manifold regularization based distributed semi-supervised learning algorithm using extreme learning machine over time-varying network. Neurocomputing. 2019;355:24–34.
50. Xiao Q, Luo J, Dai J: Computational prediction of human disease- associated circRNAs based on manifold regularization learning framework. IEEE J Biomed Health Inf 2019;1–1.
51. Zhang Y, Zhang Z, Zhang Z, Zhao M, Wang M: Deep self-representative concept factorization network for representation learning. 2019.
52. Cai D, He X, Han J: Locally consistent concept factorization for document clustering: IEEE Educational Activities Department; 2011.
53. Lin Z, Liu R, Su Z: Linearized alternating direction method with adaptive penalty for low-rank representation. AdvNeural Inf Process Syst. 2011:612–620.
54. Cai JF, Candès EJ, Shen Z. A singular value thresholding algorithm for matrix completion. SIAM J Optim. 2010;20(4):1956–82.
55. Yu SX, Shi J: Multiclass spectral clustering. In: IEEE international conference on computer vision. 2003.
56. 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.
57. Steinley D, Brusco MJ: A note on the expected value of the Rand index. Br J Math Stat Psychol. 2018.
58. Hao J, Sohn LL, Huang H, Chen L. Single cell clustering based on cell-pair differentiability correlation and variance analysis. Bioinformatics. 2018;21:3684.
59. Zhu H, Zhou MC, Alkins R. Group role assignment via a kuhn-munkres algorithm-based solution. IEEE Trans Syst Man Cybern Part A Syst Hum. 2012;42(3):739–50.
60. Wang J, Liu JX, Zheng CH, Lu CH, Dai LY, Kong XZ: Block-constraint laplacian-regularized low-rank representation and its application for cancer sample clustering based on integrated TCGA data. Complexity 2020. 2020.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.