BMC Bioinformatics

RESEARCH

Open Access

Clustering of cancer data based on Stiefel manifold for multiple views

Jing Tian[1], Jianping Zhao[1*] and Chunhou Zheng[1,2]

*Correspondence:
zhaojianping@126.com
[1] College of Mathematics and System Sciences, Xinjiang University, Urumqi, China
Full list of author information is available at the end of the article

Abstract

Background: In recent years, various sequencing techniques have been used to collect biomedical omics datasets. It is usually possible to obtain multiple types of omics data from a single patient sample. Clustering of omics data plays an indispensable role in biological and medical research, and it is helpful to reveal data structures from multiple collections. Nevertheless, clustering of omics data consists of many challenges. The primary challenges in omics data analysis come from high dimension of data and small size of sample. Therefore, it is difficult to find a suitable integration method for structural analysis of multiple datasets.

Results: In this paper, a multi-view clustering based on Stiefel manifold method (MCSM) is proposed. The MCSM method comprises three core steps. Firstly, we established a binary optimization model for the simultaneous clustering problem. Secondly, we solved the optimization problem by linear search algorithm based on Stiefel manifold. Finally, we integrated the clustering results obtained from three omics by using k-nearest neighbor method. We applied this approach to four cancer datasets on TCGA. The result shows that our method is superior to several state-of-art methods, which depends on the hypothesis that the underlying omics cluster class is the same.

Conclusion: Particularly, our approach has better performance than compared approaches when the underlying clusters are inconsistent. For patients with different subtypes, both consistent and differential clusters can be identified at the same time.

Keywords: Stiefel manifold, Multi-view clustering, Cancer data, Optimization model, Linear search algorithm

Introduction

One of the challenges of cancer treatment is how to identify tumor subtypes, which can help to provide patients with specific treatment. Meanwhile, with the continuous development of all kinds of sequencing technologies, a lot of high flux data have been produced [1]. For cancer subtypes identification, integration of different types of omics data to unravel the molecular mechanism of complex diseases becomes more and more important [2]. On the one hand, multiple omics data of different subtypes of cancer provided more detailed information. On the other hand, it made data analysis more complicated. Different levels of multiple omics data often show different types, they have

different correlation structure statistical properties and expressions [3]. In addition, the same tumor specimens from different levels of data are also unlikely to be independent. Therefore, how to reasonably integrate the multiple omics data to accurately predict cancer subtypes becomes a challenging and interesting research [4].

Due to the high dimensionality of data, we usually need to take a series of dimensionality reduction measures. However, some unsupervised approaches such as KCCA 5, KPCA [6], ISOMAP [7], the projection is only optimal at preserving the variance of the data or preserving the direction of the search. The two processes of reduction and clustering are completely independent. Solving the optimization problem on Stiefel manifold, it can be found directly in the lower dimensional representation of the feasible solution. It is worth mentioning that the noise in the original data can be reduced effectively by manifold methods. The literature [8] finds that when solving the optimization problem on Stiefel manifold, it can be more simple and quickly reach an almost medium precision.

Recently, many strategies for integrating multi-omics data have emerged. Their objectives are to understand the inter-relationships between different omics, and explore the relationship between omics data and subtypes [9, 10] . For example, the methods of biclustering aim to find the internal similar structure of high-dimensional data, and can cluster samples and features simultaneously. They have good performances in many ways, but they have a high time complexity [11, 12]. Similarity Network Fusion (SNF) method [13] constructed the similarity network for each data type, and then used the iterative method to fuse them into a similar network. The final clusters are obtained by spectral clustering of fusion networks. Some multi-view clustering methods based on spectral clustering have also been proposed [9, 14, 15]. They used different integration methods to combine the spectral clustering results from a single view. The Affinity Aggregation for Spectral Clustering (AASC) algorithm [14] introduced weights in the spectral clustering of each view, and then added them together to optimize the weights in the calculation.

However, these methods were put forward based on a basic hypothesis that the underlying omics clusters are the same. In actual situation, there are inconsistent clusters [10]. In the process of integrated clustering, data clustering was carried out for each view and cluster alignment was carried out for different views, which could handle this situation [9, 15, 16]. However, the method [9] tended to obtain the local optimal as described above, and the methods [15, 16] relax excessively the original multi-view point specific tangent condition, so that the information of each viewpoint may be lost. In the paper [17], the authors proposed the Multi-View Clustering using Manifold Optimization (MVCMO) method considered the diversity of the cluster. Consistent clusters and different clusters can be identified in each group. This method can effectively identify the cluster of differences, and this theory is also used in our method.

In order to improve the algorithm stability of MVCMO [17], we introduce the "Heat Kernel" to measure similarity between patients. And we use Backtracking Line Search to find the optimal solution more accurately. In this study, we propose a Multi-view Clustering based on Stiefel Manifold (MCSM) method for multi-view clustering problems with potential clusters. Firstly, we introduce a "Heat Kernel" to measure similarity. The patient-patient similarity network is constructed using k-nearest neighbor (KNN)

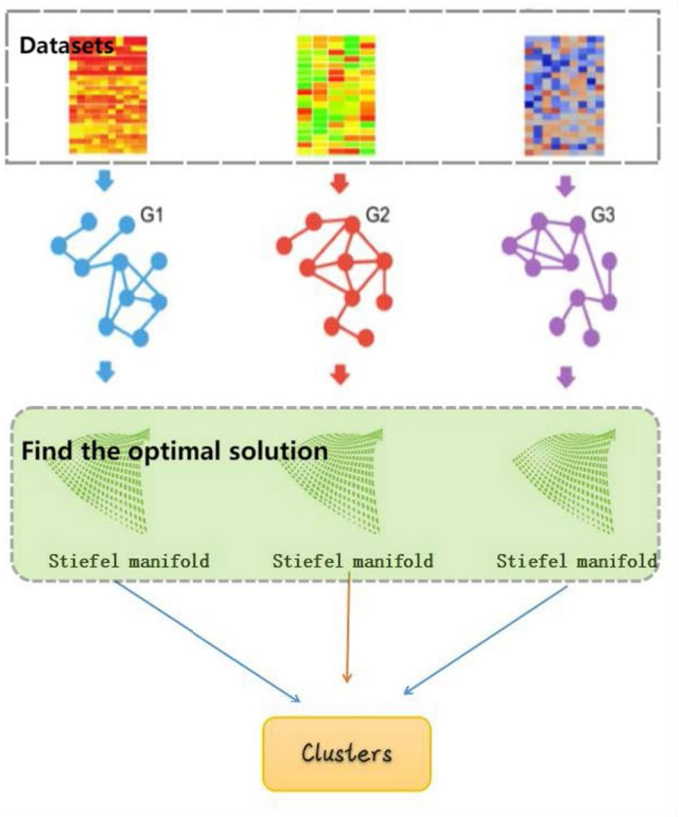

Fig. 1 The process of MCSM method

method. Then we establish a binary optimization model for the simultaneous clustering problem. The solving process of the objective optimization problem is divided into three steps. First, we project our target function onto each of Stiefel manifold's tangent vector Spaces. Second, we do Backtracking Line Search on Stiefel manifold for the objective problem. Third, we retract the found points to the manifold with singular value decomposition. Finally, the KNN method is used for integrating the obtained clusters from three omics to get the final result. The proposed MCSM method has two highlights. One is that it combines the two processes of reduction and solution optimization,which preserves as much data information of each sample as possible. The other is that it can identify the cluster effectively when the underlying clusters are different. We experiment on simulated datasets to see the algorithm's performance when there are potential clusters. The experimental results on simulated datasets and several multiple omics datasets from TCGA show that our method has better performance than state-of-art methods.

Datasets and methods

The overall design of our method is illustrated in Fig. 1.

Datasets and preprocessing

In this paper, we selected four cancer datasets in the TCGA for experiment, including gene expression data, miRNA expression data and DNA methylation data from samples of cancer patients. The cancer datasets include glioblastoma multiforme (GBM) with

215 samples, breast invasive carcinoma (BIC) with 105 samples, Skim Cutaneous Melanoma (SKCM) with 439 samples and Acute Myeloid Leukemia (AML) with 96 samples.

Firstly, if the data of a patient loses more than 20% in any data type, the patient will be deleted. Secondly, if the missing value of a feature in all patients exceeds 20%, it will be filtered out. Thirdly, the K-nearest-neighbor method is adopted to fill in missing data. We need to determine k according to the size of the sample. In our experiment, we set $k = 20$.

Fourthly, we log transform the data set to make it more stable. Finally, each feature is normalized in the constructed network to make it have a standard normal distribution. We performed the following normalization for each data type:

$$\widehat{f} = \frac{f - E(f)}{\sqrt{Var(f)}}. \tag{1}$$

where f is the characteristic of sample data, $\widehat{f}$ is the corresponding characteristic after normalization of f, E (f) and Var (f) represent the sample mean and sample variance respectively.

Construction of the patient-to-patient similarity graph

Denoted $\{X^m\}_{m=1}^M$ as multi-view data from N patient samples, which has m data type in total. Each X^m is a matrix of $p_m \times N$, then a similar network graph G^m is constructed to reflect the neighborhood relationship between the samples.

In the similar network of the type m, $G^m = (V^m, E^m, W^m)$, V^m is vertex set, E^m is edge set, and W^m is adjacency matrix. The adjacency matrix of W^m in graph G^m is a symmetric matrix.

In this paper, "Heat Kernel" is used to measure the similarity between samples [18]. The basic form is a Gaussian function with "t". It has linear complexity and robustness that is not sensitive to small changes.

$$S_{ij}^m = \exp\left(-\frac{\|x_i^m - x_j^m\|^2}{2t^2}\right), i = 1\ldots, N, j = 1\ldots, N. \tag{2}$$

Next, we construct the K-nearest neighbor graph based on the similarity matrix S^m. If the vertex has an edge between v_i and v_j, then W_{ij}^m represents the edge weight, otherwise 0.

$$W_{ij}^m = \begin{cases} S_{ij}^m, & v_j \in N_i, \\ 0, & otherwise. \end{cases} \tag{3}$$

here N_i is the neighborhood of v_i (including v_i), N_i with size k, and the number of k usually depends on the size of the sample. Essentially, we assume that local similarity is more reliable than remote similarity. This is a modest assumption, and it is widely used by other manifold learning algorithms [18].

Construction of objective optimize problem

The objective optimize problem of the spectral clustering method is:

$$\begin{aligned} &\min_{U_m \in R^{N\times k}} \operatorname{trace}\left(U_m^T L_m U_m\right) \\ &\text{s.t.} \quad U_m^T U_m = I_K. \end{aligned} \tag{4}$$

Here, the $L_m = (D_m - A_m)$. The A_m is the corresponding adjacency matrix of similar network G^m, and D_m is the diagonal matrix constructed using the degree of all the nodes in the mth network.

Then, used U_m for K-means and find its minimum k eigenvectors in order to obtain the clustering labels.

Based on the spectral clustering, [15] proposed a multi-view network clustering method. Its objective optimize problem is:

$$\begin{aligned} &\min \sum_{m=1}^{M}\sum_{k=1}^{K} \frac{\left(S_{,k}^m\right)^T (D_m - A_m)\left(S_{,k}^m\right)}{\left(S_{,k}^m\right)^T\left(S_{,k}^m\right)} - \beta \sum_{l\neq m}\sum_{k=1}^{K} \frac{\left(S_{,k}^m\right)^T\left(S_{,k}^l\right)}{S_{,k2}^m S_{,k2}^l} \\ &\text{s.t. } S_{,k}^m \in \{0,1\}, i = 1\ldots, N;\ m = 1\ldots, M;\ k = 1\ldots, k; \\ &\sum_{k=1}^{K} S_{i,k}^m = 1, \quad \text{for} \quad m = 1\ldots, M. \end{aligned} \tag{5}$$

The binary optimization problem cannot be solved in polynomial time. So, the objective function of multi-view spectral clustering can be constructed as follows:

$$\begin{aligned} &\min_{U_m \in R^{N\times k}} \operatorname{trace}\left(U^T L U\right) \\ &\text{s.t.} \quad U^T U = I_K \end{aligned} \tag{6}$$

$$\text{where } L = \begin{pmatrix} L_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & L_m \end{pmatrix} - \beta \begin{pmatrix} 0 & \cdots & I_n \\ \vdots & \ddots & \vdots \\ I_n & \cdots & 0 \end{pmatrix}, U = \begin{pmatrix} U_1 \\ U_2 \\ \\ \vdots \\ U_3 \end{pmatrix}.$$

β is used to balance the weight parameters between the network and within the network. If we have abundant prior knowledge, we can set it according to prior information. Otherwise, when building a network, we can try to establish a connection at the same level (e.g. similar connection densities) and set it directly to 1. In our experiment, we set $\beta = 1$ directly.

However, the optimization problem (6) combines the information of all networks together and will loss the information in each network. The proposed MVSM method still follows the original objective function of multi-view spectral clustering and the construction of Laplace matrix [17].

The objective optimization problems to be solved are as follows:

$$\begin{aligned} &\min_{U_m \in R^{N\times k}} \operatorname{trace}\left(U^T L U\right) \\ &\text{s.t.} \quad U_m^T U_m = I \end{aligned} \tag{7}$$

When we set $U_m = \frac{S^m}{\|S^m\|_2}$, the objective function $U_m^T U_m = I_K$ substitude as $\sum s_{i,j}^N = 1$. It transforms the constraints for each network into one equation.

The solution of objective optimize problem

To solve the objective function (7), we project it onto the Stiefel manifold and solve it by backtracking linear search. The process is roughly divided into three steps.

First, we project the target function $\mathrm{trace}\left(\mathrm{U^T LU}\right)$ onto each of Stiefel manifold's tangent vector Spaces.

The tangent vector space of M is.

$$\mathrm{TM_m} = \left\{\mathrm{U_m B} + \left(\mathrm{I} - \mathrm{U_m U_m^T}\right)\mathrm{C} : \mathrm{B} = -\mathrm{B^T}, \mathrm{C} \in \mathrm{R^{N\times k}}\right\}. \tag{8}$$

here each Stiefel manifold $\mathrm{M_m} = \mathrm{U_m} \in \mathrm{R^{N\times k}} : \mathrm{s.t.U_m^T U_m} = \mathrm{I_K}$.

So, the negative gradient of the target function $\mathrm{trace}\left(\mathrm{U^T LU}\right)$ can be expressed as:

$$\mathrm{Z} = -\nabla \mathrm{trace}\left(\mathrm{U^T LU}\right) = -\mathrm{LU} = \left(\mathrm{Z_1^T}, \mathrm{Z_2^T}, \ldots, \mathrm{Z_m^T}\right)^{\mathrm{T}} \tag{9}$$

where Z_m represents the negative derivative of the objective function on the mth omic.

Then, we search the next point along the direction η_m on each tangent vector space of the manifold. Where,

$$\eta_\mathrm{m} = \mathrm{Z_m} - \frac{1}{2}\mathrm{U^m}\left(\left(\mathrm{U^m}\right)^{\mathrm{T}}\mathrm{U^m} + \mathrm{Z_m^T U^m}\right). \tag{10}$$

Second, we do Backtracking Line Search on Stiefel manifold for problem (9–10).

The purpose of line search is to find the smallest point of the target function in the search direction. However, it is time-consuming to find the accurate minimum point. The search direction is already approximate, so we just to find the minimum point approximation at a lower cost. Backtracking Line Search (BLS) is such a Line Search algorithm. The idea of the BLS algorithm is to set an initial step size α_0 in the search direction. Then, if the step size is too large, we reduce the step size until it is appropriate.

Backtracking Line Search in the negative gradient direction of the objective function is as follow:

$$\begin{aligned} \mathrm{f(U} + \alpha\eta) &\le \mathrm{f(U)} + \alpha\mathrm{cm} \\ \mathrm{m} &= \eta^T \mathrm{Z} \end{aligned} \tag{11}$$

where η is the current search direction, α is the step size, and c is the control parameter, which needs to be manually verified according to the situation.

If the current U does not satisfy inequation (11), then a parameter τ is required to adjust the step size:

$$\alpha = \tau\alpha \tag{12}$$

where the parameter τ is controls the reduce search step size.

Third, we retract the found points to the manifold with singular value decomposition.

$$\mathrm{U} = \mathrm{W}\Sigma\mathrm{U^T}, \mathrm{U} = \mathrm{WU^T}. \tag{13}$$

After the manifold optimization process, we get the values of U. The whole process

Step 1. The negative gradient direction of the objective function is projected onto Stiefel manifold.

$$\eta_m = Z_m - \frac{1}{2}U^m((U^m)^T U^m + Z_m^T U^m);$$

$$\eta = (\eta_1^T, \eta_2^T, \cdots \eta_m^T)^T;$$

Step 2. Backtracking Linear search in tangent vector space:

$$U = U + \alpha\eta, \alpha\epsilon(0,1);$$

$$f(U + \alpha\eta) \leq f(U) + \alpha cm;$$

$$m = \eta_m{}^T Z;$$

Step 3. Retracted the points obtaining in 2 to the Stiefel manifold:

$$U = W\Sigma V^T, U = WV^T;$$

Step 4. Repeat 1- 3 until the convergence condition is satisfied:

$$((r\_err\_f > 1e-8) \,||\, (r\_err\_g > 1e-4))\&\&(iters < 1000);$$

Where r_err_f represents the relative error of objective function f, r_err_grepresents the relative error of the negative gradient of the objective function Z, and iters represents the number of iterations.

of our proposed method is summarized in Algorithm 1.

Here, we get the solution of the objective function, and then we perform k-means to cluster U and obtain the cluster labels $C_1, C_2, \ldots, C_k$. Finally, we integrate the clustering results obtained from three omics by using k-nearest neighbor method.

Remark: we set $c = 6, \tau = 0.1$ in our experiment. We will set out the reasons in Sect. 3.1.2.

Results

In this section, we selected some methods from different perspectives to compare with MVSC methods. For the methods were proposed using network structure, we chose AASC algorithm [14], SNF method [13] and MVSC [15]. In particular, AASC and MVSC method can effectively identify different clusters. For the methods based on manifold, we chose MOCMO [17] and Grassmann manifold clustering method [18]. For the state-of-the-art methods, we chose MvNE algorithm [19].

The selection of parameter

The number of clustering

When the clusters k is not known, we can select it according the value of silhouette [20] and RI coefficient. From the perspective of computational efficiency, Calinski Harabaz score [21] is the highest. So Calinski Harabaz score is more commonly used. Firstly, we did experiments with k equals 2–10. Then, we choose the clustering number corresponding to the maximum Calinski Harabaz score. To compare MVSM to other methods, we set k as a known value.

The Backtracking Line Search parameters

There are three parameters in the Backtracking Line Search parameters, η, α and c. Where, η is the current search direction, α is the step size, and c is the control parameter, which needs to be manually verified according to the situation. Firstly, we initialize $\alpha = 0.01$. During the experiment, it was found that if the value of c was too small, the step size would not be adjusted during the search process. However, if we want to adjust appropriately, then we need to set the parameter c according to the objective function value and gradient value of the initial point. Therefore, according to several data sets used in the experiment, we set $c = 6, \tau = 0.1$.

Experimental results on simulated datasets

Here, we use the simulated datasets to verify that MVSM method is suitable for datasets with uneven distribution of underlying clusters.

Since these methods (AASC, SNF and MVSC) were proposed using network tools, we simulate the network structure firstly. Then, we generate the connections within the same cluster and different clusters. The probability of connections within a given cluster is greater than the probability of connections between clusters. For M omics networks, given the number of nodes N, these nodes are assigned to K clusters with different probabilities.

In order to see the influence of the connections between clusters change, we set the following four connection probability matrices:

$$P_1 = \frac{1}{N}\begin{pmatrix} 16 & 0 & 0 \\ 0 & 18 & 0 \\ 0 & 0 & 17 \end{pmatrix}, \; P_2 = \frac{1}{N}\begin{pmatrix} 16 & 0.4 & 0.6 \\ 0.4 & 18 & 0.55 \\ 0.6 & 0.55 & 17 \end{pmatrix},$$
$$P_3 = \frac{1}{N}\begin{pmatrix} 16 & 0.8 & 1.2 \\ 0.8 & 18 & 1.1 \\ 1.2 & 1.1 & 17 \end{pmatrix}, \; P_4 = \frac{1}{N}\begin{pmatrix} 16 & 1.2 & 1.8 \\ 1.2 & 18 & 1.65 \\ 1.8 & 1.65 & 17 \end{pmatrix}.$$

The term (i,j) of the four matrices represents the connection probability between cluster i and cluster j. Each term (i, i) and each term $(i, j), i \neq j$ represent the connections within and between clusters, the larger the value the term correspond to, the tighter the connection. N represents the number of nodes.

In order to see the performance of the method, we tested two settings. For each setting, we consider that the M omics of distribution is different.

Table 1 Comparison of RI in different methods based on Setting 1

Method	P_1	P_2	P_3	P_4
AASC	0.73	0.73	0.73	0.73
SNF	0.68	0.67	0.67	0.67
MVSC	0.99	0.98	0.94	0.87
MCSM	1	0.99	0.94	0.97

Table 2 Comparison of RI in different methods based on Setting 2

Method	P_1	P_2	P_3	P_4
AASC	0.75	0.75	0.75	0.75
SNF	0.75	0.75	0.75	0.75
MVSC	0.93	0.93	0.93	0.93
MCSM	0.96	0.95	0.95	0.95

Setting 1: M = 3, N = 150, cluster distribution: (50, 50, 50); (30, 90, 30); (40, 60, 50);

Setting 2: M = 6, N = 1000, cluster distribution: (300, 300, 400); (300, 300, 400); (400, 300, 300); (300, 350, 350); (300, 400, 300); (450, 250, 300).

We use the Rand index to evaluate the clustering performance, which is defined as:

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}},$$

'TP' is defined as the number of intersection nodes in the same cluster, which are also clustered in the same cluster, and other nodes are defined similarly.

On this basis, we obtain the rand index comparison of several methods:

For each setting, we run it 50 times and take the average of the results. Tables 1 and 2 show the mean RI when the underlying clusters are different for the two settings.We can see that all four methods with an average RI is close to 1 when the cluster sizes of different groups are the same. On the one hand, because both SNF and AASC set the underlying cluster to be the same, they cannot detect the difference between the different views. So the MVSC and our method have better performance, when the size of the underlying cluster is different. On the other hand, more information of the clusters can be kept in our method by using more strict relaxation of the binary variables. Form Tables 1 and 2, when the nodes of networks are different, our method has a better performance than MVSC in both setups.

To further show the effective of our method, we also calculate the NMI coefficient on different omics. It is defined as follows:

$$\mathrm{NMI}(\mathrm{U}, \mathrm{V}) = \frac{2\mathrm{MI}(\mathrm{U}, \mathrm{V})}{\mathrm{H}(\mathrm{U}) + \mathrm{H}(\mathrm{V})}$$

where U and V represent the clusters according to clustering and real clusters, respectively. H(U), H(V) and MI(U, V) are defined as:

Table 3 Comparison of NMI in different methods based on Setting 2

Method	M_1	M_2	M_3
AASC	0.84	0.64	0.90
SNF	0.74	0.49	0.88
MVSC	0.83	0.93	0.97
MCSM	0.84	0.97	0.97

Table 4 Comparison of Cox survival p-values

Cancer type	SNF	Grassmann Cluster	MOCMO	AASC	MVSC	MvNE	MCSM
GBM(3)	0.0002	0.0043	0.0019	0.0022	0.00072	0.01113	**0.0001**
BIC(5)	0.0011	0.0002	**0.00016**	0.00015	0.0007	0.0061	0.0025
SKCM(4)	0.0001	0.19	0.00045	0.00016	0.00045	0.0098	**0.0001**
AML(5)	0.037	0.12	0.03	0.045	0.058	0.062	**0.019**

Bold values indicate the smallest Cox survival p-values on the different datasets

$$\mathrm{MI}(U,V)=\sum_{i=1}^{C}\sum_{j=1}^{C}p_{i,j}\log\left(\frac{p_{i,j}}{p_i\times p_j}\right)$$

$$H(U)=-\sum_{i=1}^{C}p_i\log p_i$$

$$H(V)=-\sum_{j=1}^{C}p_j\log p_j$$

p_i is the proportion of the number of cluster i to the total amount of sample.

For setting 2, we calculate NMI coefficients on each omics and show their contrast results in Table 3. It can be seen that the NMI coefficient of the MVSM method is higher than the contrast methods in all three omics.

Experimental results on real datasets

In order to prove the effectiveness of our method on the real datasets. We apply our method on multiple omics datasets [22], analyze and compare our results with other advanced methods. Shown are final Cox survival P values for Glioblastoma Multiforme (GBM), breast Invasive carcinoma (BIC), Skim Cutaneous Melanoma (SKCM) and Acute Myeloid Leukemia (AML) in Table 4.

Shown are Kaplan Meier plots of the overall survival of integrative clusters for GBM (a), BIC (b), SKCM (c) and AML) (d) in Fig. 2.

It can be seen from the Table 4, in three of the four cancers datasets (GBM, BIC, SKCM and AML), our method obtains more significant differences than comparison methods in survival time. For BIC dataset, the insignificant difference may be due to the small cluster difference of the data itself. Survival plots for GBM, BIC, SKCM, and AML

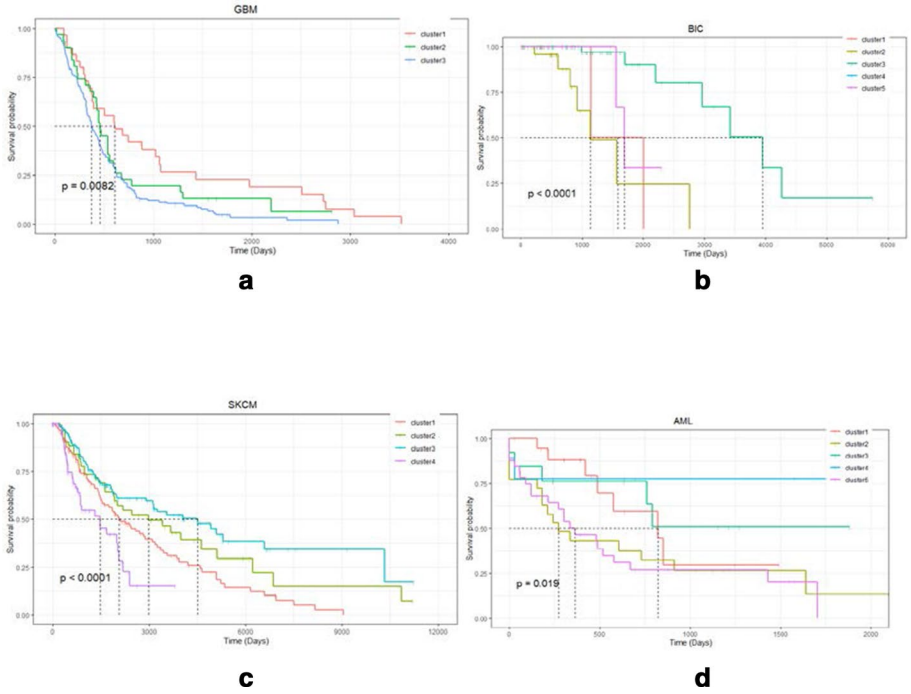

Fig. 2 Survival plots for GBM, BIC, SKCM, and AML tumors

tumors are shown in Fig. 2. We can predict survival rate in a sample according plots in Fig. 2. In the prediction task, our method performed better than other methods.

Convergence analysis

The proposed methodology can be divided into three parts, construction of Laplace matrix, process for solving optimization problems and iteration. The time complexity of each part of the algorithm is as follows:

Construction of Laplace matrix It has linear complexity.

In the process of solving the optimization problem The time complexity of matrix multiplication is $O(2Nk^2)$.

Iteration Since we need to use SVD to retract the found solution back to the Stiefel manifold, then the matrix operation complexity is $O(N^3)$.

Overall, the time complexity of MVSM method is much lower than that of MvNE method ($O(nt(ij+jk+kl+lm))$). From Fig. 3, it can be observed that for 20 iterations, there is a stable objective function value for all the datasets. It shows that our algorithm can find an appropriate solution with fewer iterations.

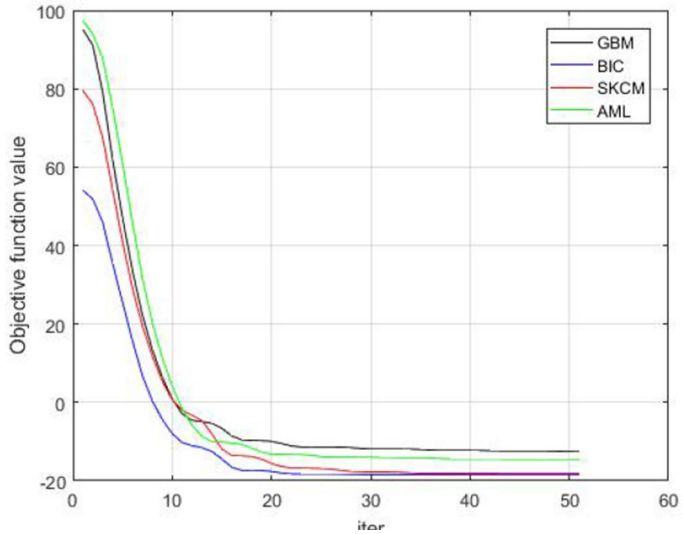

Fig. 3 Convergence of the MVSM algorithm

Table 5 Comparison of clusterings to established subtypes

Clusters	Gene expression subtypes				DNA methylation subtypes	
	Classical	Mesenchymal	Neural	Proneural	G-CIMP	Non-G-CIMP
Cluster1	0	0	1	13	16	14
Cluster2	2	25	2	2	0	31
Cluster3	55	41	31	24	2	152

A case study: comparison of established subtypes

In order to compare the results of our clustering with the established biological subtypes, we downloaded the clinical data of 215 GBMs from the cBio Cancer Genomis Portal (http://www.cbioportal.org/) at the Memorial Sloan-Kettering Cancer Center. For the GBM, there are four established subtypes defined by patients' gene expression profiles, which are Classical, Mesenchymal, Neural and Proneural [23]. According to DNA methylation clustering, they [24] divided the subtypes into Glioma-CpG island methylator phenotype (G-CIMP) and Non-G-CIMP. The results of our method are compared with the established subtypes in Table 5. It shows that the clustering of our method is not just based on one data type, it takes into account both gene expression and DNA methylation information.

For gene expression subtypes, it can be seen that cluster 1 mainly contains Proneural subtype, cluster 2 mainly contains Proneural subtype, and they have strong enrichment. However, for DNA methylation subtypes, G-CIMP subtypes are mainly distributed in cluster 1. If only the DNA methylation information is considered, cluster2 and cluster3 are likely to merge. So, we can conclude that it's important to consider both gene expression and DNA methylation information.

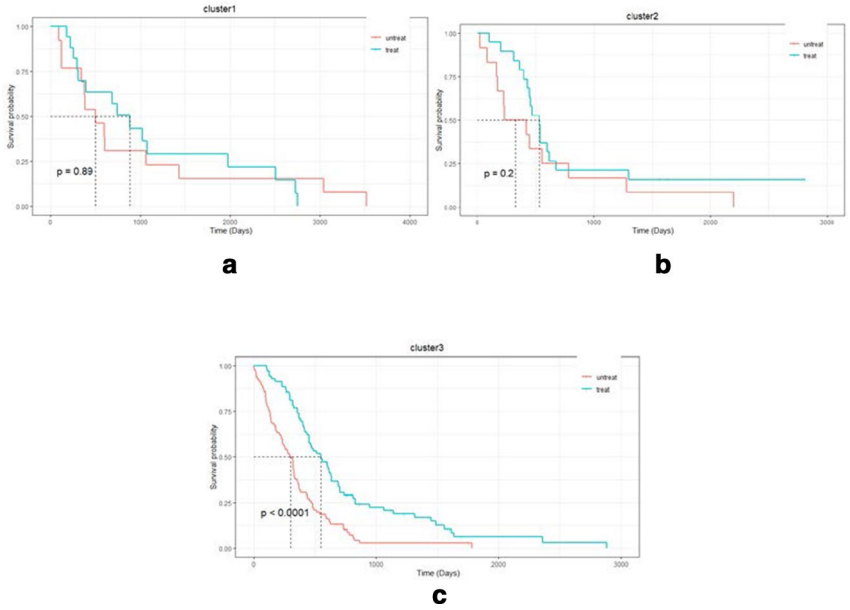

Fig. 4 Survival analysis of GBM patients for treatment with TMZ

In order to further understand the biological significance of clusters, we investigated the response to temozolomide (TMZ) treatment of the GBMs. TMZ is an alkylation agent that causes incorrect pairing of thymine during DNA replication. In the GBM dataset, 105 patients were treated with TMZ. Figure 4 indicated that the TMZ-treated samples had different drug responses compared to the samples not treated with the drug. For different clusters, the degree of drug response of TMZ was also different. Compared with Cluster 1 and Cluster 2, patients in Cluster 3 had significantly increased survival time after treatment with TMZ (P value using Cox log-rank test = 0.0001), and this medication was also more meaningful. The results show that the clusters we obtained can be used as a reference for identifying the effectiveness of drugs.

Conclusion

Multi-view data clustering is a hot topic in recent years. Recent work has focused on cases where the underlying clusters are consistent, and as we reviewed in the first section, several approaches have been proposed. When the underlying cluster is different, some methods are proposed to find different clusters. However, as we know, both consistent and differentiated clusters can exist at the same time. This leads us to study multi-view simultaneous clustering to find both consistent and different cluster data. In this paper, we propose a multi-view clustering model. On the basis of manifold optimization, the algorithm for formula optimization is proposed. Simulation results show that the performance of the proposed method is better than that of the existing algorithm under the same underlying cluster condition. We download the gene expression, miRNA expression and DNA methylation datasets of GBM, BIC, SKCM and AML from TCGA, and also carry out numerical experiments, showing that our

method is superior to several comparison methods. In the future work, the cluster difference problem is still worth researching, and we will integrate other omics information such as gene mutation data.

Acknowledgements
This work was supported by grants from the Xinjiang Autonomous Region University Research Program (No. XJEDU2019Y002), and the National Natural Science Foundation of China (No. U19A2064, 61873001).

Authors' contributions
TJ contributions are processed a multi-view clustering based on Stiefel manifold method (MCSM) and the numerical simulation results of MCSM model. ZJP contribution lies in the embellishment of the article. ZCH contribution lies in the thought guidance of the method. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
All the raw data are available at http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html (Nimrod et al., 2018), and all code scripts used are available at https://github.com/charley410/tianjing.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interest.

Author details
[1]College of Mathematics and System Sciences, Xinjiang University, Urumqi, China. [2]School of Computer Science and Technology, Anhui University, Hefei, China.

Received: 24 January 2021 Accepted: 12 May 2021
Published online: 25 May 2021

References

1. Zheng CH, Yang W, Chong YW, Xia JF. Identification of mutated driver pathways in cancer using a multi-objective optimization model. Comput Biol Med. 2016;72:22–9. https://doi.org/10.1016/j.compbiomed.2016.03.002.
2. Zhang D, Chen P, Zheng CH, Xia JF. Identification of ovarian cancer subtype-specific network modules and candidate drivers through an integrative genomics approach. Oncotarget. 2016;7(4):4298. https://doi.org/10.18632/oncotarget.6774.
3. 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. https://doi.org/10.1109/TNB.2011.2144998.
4. Bickel PJ, Chen A. A nonparametric view of network models and new managirvan and other modularities. Proc Natl Acad Sci USA. 2009;106(50):21068–73. https://doi.org/10.1073/pnas.0907096106.
5. Chaudhuri K, Kakade SM, Livescu K, Sridharan K (2009) Multi-view clustering via canonical correlation analysis. In: Proceedings of the 26th annual international conference on machine learning, pp 129–136
6. Xia R, Pan Y, Du L, Yin J. Robust multi-view spectral clustering via low-rank and sparse decomposition. In: Twenty-eighth AAAI conference on artificial intelligence (2014).
7. Kakade SM, Foster DP. Multi-view regression via canonical correlation analysis. In: International Conference on Computational Learning Theory (2007) pp. 82–96.
8. Absil PA, Mahony R, Sepulchre R (2008) Optimization algorithms on matrix manifolds. Princeton University Press, Princeton, pp 11462–11467. https://doi.org/10.1515/9781400830244
9. Kumar R, Kamdar D, Madden L, Hills C. Th1/th2 cytokine imbalance in meningioma, anaplastic astrocytoma and glioblastoma multiforme patients. Oncol Rep. 2006;15(6):1513–6. https://doi.org/10.3892/or.15.6.1513.
10. Yuan Y, Savage RS, Markowetz F. Patient-specific data fusion defines prognostic cancer subtypes. PLoS Comput Biol. 2011;7(10):e1002227. https://doi.org/10.1371/journal.pcbi.1002227.
11. Hussain SF, Bashir S. Co-clustering of multi-view datasets. Knowl Inf Syst. 2016;47(3):545–70.
12. Maran P, Shanthi S, Thenmozhi K, Hemalatha D, Nanthini K, A novel deep learning method for identification of cancer genes from gene expression dataset. In: Machine learning and deep learning in real-time applications. IGI Global, (2020), pp 129–144.
13. Wang B, Mezlini AM, Demir F, Fiume M. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11(3):333–7. https://doi.org/10.1038/nmeth.2810.
14. Huang HC, Chuang YY, Chen CS. Affinity aggregation for spectral clustering. Conf Comput Vis Pattern Recognit. 2012;2012:773–80. https://doi.org/10.1109/CVPR.2012.6247748.

15. Zhang S, Zhao H, Ng MK. Functional module analysis for gene coexpression networks with network integration. IEEE/ACM Trans Comput Biol Bioinf. 2015;12(5):1146–60. https://doi.org/10.1109/TCBB.2015.2396073.
16. Chen C, Ng MK, Zhang S. Block spectral clustering methods for multiple graphs. Numer Linear Algebra Appl. 2017;24(1):1–20. https://doi.org/10.1002/nla.2075.
17. Yu Y, Zhang LH, Zhang SQ. Simultaneous clusterin of multiview biomedical data using manifold optimization. Bioinformatics. 2019;35(20):4029–37. https://doi.org/10.1093/bioinformatics/btz217.
18. Ding H, Michael S, Wang C. Integrative cancer patient stratication via subspace merging. Bioinformatics. 2018;35(10):1653–9. https://doi.org/10.1093/bioinformatics/bty866.
19. Mitra S, Sriparna S, Mohammed H. Multi-view clustering for multi-omics data using unified embedding. Sci Rep. 2020;10(1):1–16.
20. Kaufman L, Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53.
21. Calinski T, Harabasz J. A dendrite method for cluster analysis. Commun Statist. 1974;3(1):1.
22. Nimrod R, Ron S. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucl Acids Res. 2018;46(20):10546–62.
23. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. Cancer Cell. 2010;17(1):98–110.
24. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, Cancer Genome Atlas Research Network. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell 2010; 17(5), 510–522.

Publisher's Note