Clustering of Cancer Data Based on Stiefel Manifold for Multiple Views

Jing Tian
Xinjiang University
Jianping Zhao (zhaojianping@126.com)
Xinjiang University
Chun-hou Zheng
Xinjiang University

Research Article

Keywords: Stiefel manifold, multi-view clustering, cancer data, optimization model, linear search algorithm

Posted Date: February 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-154286/v1

Clustering of Cancer Data Based on Stiefel Manifold for Multiple Views

Jing Tian[1], Jian-ping Zhao[1,*], Chun-Hou Zheng[1,2,1]

[1]College of Mathematics and System Sciences, Xinjiang University, Urumqi, China

[2]School of Computer Science and Technology, Anhui University, Hefei, China

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 these datasets has proved to be valuable for biological and medical research and helpful to reveal data structures from multiple collections. However, such data often have small sample size and high dimension. 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. Firstly, we established a binary optimization model for the simultaneous clustering problem. Secondly, the optimization problem solved 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 performs when the underlying clusters are

[1]Corresponding authors

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

1. 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 (Zheng et al. 2016). For cancer subtypes identification, integration of different types of omics data to unravel the molecular mechanism of complex diseases becomes more and more important (Zhang et al., 2016). On the one hand, multiple omics data of different subtypes of cancer provided more detailed information. On the other hand, it leaded to the new challenge for data analysis. Different levels of multiple omics data often show different types, they have different correlation structure statistical properties and expressions (Zhang et al., 2011). 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 (Bickel et al., 2009).

Recently, many strategies have emerged to integrate multi-omics data for association studies. The main objectives of these studies usually are to understand the inter-relationships among various omics data and to detect phenotype-related modules (Manko et al., 2010; Kumar et al., 2011; Yuan et al., 2011). These Multi-view data integration methods can be divided into three categories: early

integration, late integration and intermediate integration. Early integration mainly merged all data types into a single dataset and analyzed them directly (Manko et al., 2010). Late integration is the process of analyzing each data type, then combining the results for integration (Fridly et al., 2012). Intermediate integration converts each data type into an intermediate representation, which is then consolidated for analysis (Cancer Genome Atlas Research Network et al., 2012).

In recent years, many new methods are also published for clustering. For example, Similarity Network Fusion (SNF) method (Wang et al., 2014) constructed the similarity network for each data type, and then used the iterative method to fuse them into a similar network. The final clustering is obtained by spectral clustering of fusion networks. Some multi-view clustering methods based on spectral clustering have also been proposed (Huang et al., 2012; Kumar et al., 2011; Zhang et al., 2015). They used different integration methods to combine the spectral clustering results from a single view. The Affinity Aggregation for Spectral Clustering (AASC) algorithm (Huang et al., 2012) 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 cluster is the same. In actual situation, there are differences in clusters except for the same cluster (Yuan et al., 2011). 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 (Kumar et al., 2011; Zhang et al., 2015; Chen et al., 2017). However, the method (Kumar et al., 2011) tended to obtain the local optimal as described above, and the methods (Zhang et al., 2015; Chen et al., 2017) relax excessively the original multi-view point specific tangent condition, so that the information of each viewpoint may be lost. In the paper (Yu et al., 2019), the authors proposed

the Multi-View Clustering using Manifold Optimization (MVCMO) method considering 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 (Yu et al., 2019), 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) 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 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.

2.Datasets and Methods

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

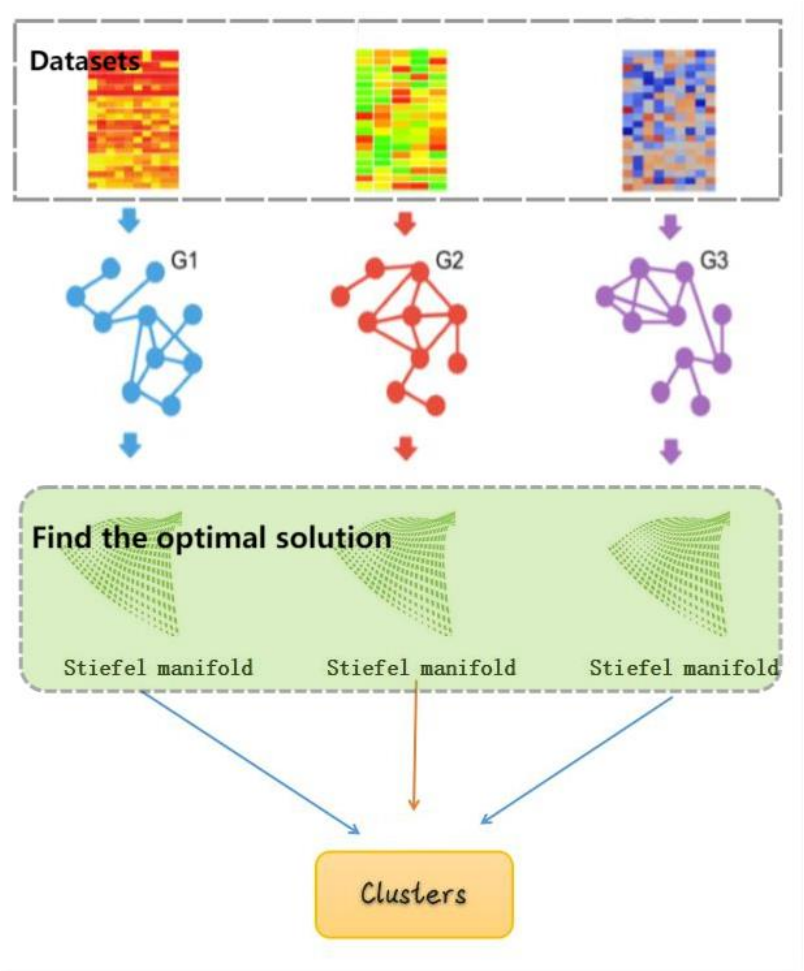


Figure 1: The process of MCSM method.

2.1 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.

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:

$$\hat{f} = \frac{f - E(f)}{\sqrt{Var(f)}}. \quad (1)$$

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

2.2 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 graphG^m is constructed to reflect the neighborhood relationship between the samples.

In the similar network of 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 (Ding et al., 2018).

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

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.

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

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 (Ding et al., 2018).

2.3 Construction of objective optimize problem

The objective optimize problem of the spectral clustering method is:

$$\min_{U_m \in R^{N \times k}} \operatorname{trace}(U_m^T L_m U_m)$$

$$s.t.\ U_m^T U_m = I_K. \tag{4}$$

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

Based on the spectral clustering, (Zhang et al., 2015) proposed a multi-view network clustering method. Its objective optimize problem is:

$$\min \sum_{m=1}^{M} \sum_{k=1}^{K} \frac{(S_{,k}^m)^T (D_m - A_m)(S_{,k}^m)}{(S_{,k}^m)^T (S_{,k}^m)} - \beta \sum_{l \neq m} \sum_{k=1}^{K} \frac{(S_{,k}^m)^T (S_{,k}^l)}{\|S_{,k}^m\|_2 \|S_{,k}^l\|_2}$$

$$s.t.\ S_{,k}^m \in \{0,1\}, i = 1 \cdots, N; m = 1 \cdots, M; k = 1 \cdots, k;$$

$$\sum_{k=1}^{K} S_{i,k}^m = 1, \quad \text{for} \quad m = 1 \cdots, M. \tag{5}$$

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

$$\min_{U_m \in R^{N\times k}} \mathrm{trace}(U^T L U)$$

$$s.t.\ U^T U = I_K \tag{6}$$

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 (Yu et al., 2019).

The objective optimization problems to be solved are as follows:

$$\min_{U_m \in R^{N\times k}} \mathrm{trace}(U^T L U)$$

$$s.t.\ {U_m}^T U_m = I \tag{7}$$

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.

2.3 The solution of objective optimize problem

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

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

The tangent vector space of M is

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

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

So, the negative gradient of the objective function can be expressed as:

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

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

$$\eta_m = Z_m - \frac{1}{2} U^m((U^m)^T U^m + Z_m^T). \quad (10)$$

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:

$$f(U + \alpha\eta) \le f(U) + \alpha c m$$

$$m = \eta^T Z \quad (11)$$

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.

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

After the manifold optimization process, we get the values of U. The whole process of our proposed method is summarized in Algorithm 1.

Algorithm 1: The linear search based on Stiefel manifold

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

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

$$\eta = (\eta_{1}^{T}, \eta_{2}^{T}, \cdots \eta_{m}^{T})^{T};$$

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:

$$U = W\Sigma U^T, U = WU^T;$$

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.

Here, we get the solution of the objective function, and then we perform k-means to cluster U and obtain the cluster labels $C_1, C_2, \cdots, C_k$. 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 **3.1.2**.

3.Results

In this section, to show the effective of our proposed method, we compare it with AASC algorithm (Huang et al., 2012), SNF method (Wang et al., 2014), MOCMO (Yu et al., 2019), MVSC (Zhang et al., 2015) and Grassmann manifold clustering method (Ding et al., 2018).

3.1The selection of parameter

3.1.1 The number of clustering

When the clusters k is not known, we can select it according the value of silhouette. Firstly, we did experiments with k equals 2 to 10. Then, we choose the number corresponding to the maximum silhouette coefficient. To compare MVSM to other methods, we set k as a known value.

3.1.2 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 $c = 6, \tau = 0.1$.

3.2 Experimental results on simulated datasets

We refer to the same method in the article (Yu et al., 2019) to verify that this method is suitable for datasets with uneven distribution of underlying clusters. Here we use a simulated dataset.

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

We used the following four connection probability matrices to define the connections within and between clusters to see the results when the connections between clusters changed:

$$P_1 = \frac{1}{N}\begin{pmatrix} 16 & 0 & 0 \\ 0 & 18 & 0 \\ 0 & 0 & 17 \end{pmatrix}, \qquad P_2 = \frac{1}{N}\begin{pmatrix} 16 & 0.4 & 0.6 \\ 0.4 & 18 & 0.55 \\ 0.6 & 0.55 & 17 \end{pmatrix},$$

$$P_3 = \frac{1}{N}\begin{pmatrix} 16 & 0.8 & 1.2 \\ 0.8 & 18 & 1.1 \\ 1.2 & 1.1 & 17 \end{pmatrix}, \qquad P_4 = \frac{1}{N}\begin{pmatrix} 16 & 1.2 & 1.8 \\ 1.2 & 18 & 1.65 \\ 1.8 & 1.65 & 17 \end{pmatrix}.$$

Where, each term (i, i) of the four matrices defines the connection probability within the cluster, and each term $(i, j), i \neq j$ represents the connection probability between cluster i and cluster j. For each setting, we consider that in all M omics, the sample size of each category is the same, and the omics of distribution is different. To see the performance of the method, we tested two settings:

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

Setting 2: N=1000, M=6, cluster size:(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:

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

with the 'TP', 'TN', 'FP', 'FN' are present the number of the true positive, true negative, false positive, and false negative. '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:

Table 1: Performance comparison of different methods. N=150, M=3, cluster size:(50,50,50); (30,90,30); (40,60,50)

Method	$\mathbf{P_1}$	$\mathbf{P_2}$	$\mathbf{P_3}$	$\mathbf{P_4}$

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: Performance comparison of different methods. N=1000, M=6, cluster size:(300,300,400), (300,300,400); (400,300,300); (300,350,350); (300,400,300); (450,250,300)

Method	$\mathbf{P_1}$	$\mathbf{P_2}$	$\mathbf{P_3}$	$\mathbf{P_4}$
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**

For each setting, we run it 50 times and take the average of the results. From Table 1 and Table 2, we can see that all four methods cluster nodes with an average Rand index is close to 1 when the cluster sizes of different groups are the same. This also illustrates the importance of data integration. When the size of the underlying cluster is different, because both AASC and SNF 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. Table 1 and Table 2 show the mean Rand index and its

standard deviation in the bracket when the underlying clusters are different for the two setups. By using more strict relaxation of the binary variables, more information of the clusters can be kept. our method out performs MVSC in both setups and when there are less networks, our method performs more stable, which achieves much less standard deviation of the Rand index.

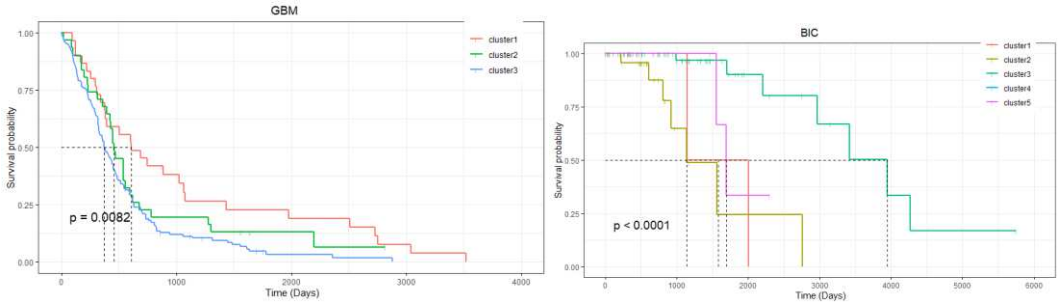
3.3 Experimental results on real datasets

In order to prove the effectiveness of our method on the real datasets. We apply our method on multiple datasets analyzed by (Wang et al., 2014), and compare our results with those of SNF fused by similarity network. The final Cox survival P values are listed in Table 3.

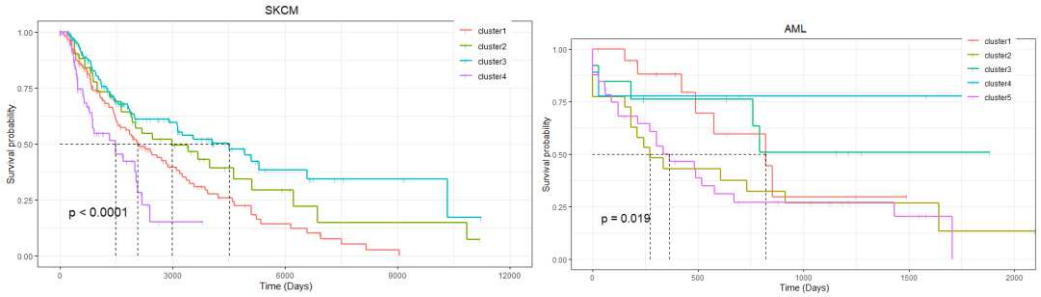
Table 3: Comparison of Cox survival p-values.

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

Integrative clustering discovers clinically significant subtypes of cancer. Shown are Kaplan Meier plots of the overall survival of integrative clusters for Glioblastoma Multiforme (GBM) (a), breast Invasive carcinoma (BIC) (b), Skim Cutaneous Melanoma (SKCM) (c) and Acute Myeloid Leukemia (AML) (d) in figure 2. P-values are computed from the log rank test.



(a) (b)



(c) (d)

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

It can be seen from the Table 3, 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 tumors are shown in Figure 2. Our proposed method can also be extended to improve the survival rate prediction task.

3.4 A case study: Comparison of clusterings to 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

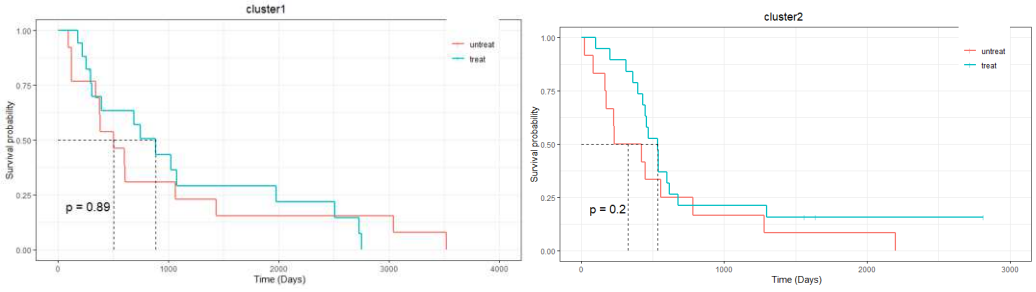
exist four established subtypes determined by patients' gene expression profiles, which are Classical, Mesenchymal, Neural and Proneural (Verhaak et al., 2010), and a subtype called Glioma-CpG island methylator phenotype (G-CIMP) generated by DNA methylation clustering (Noushmehr et al., 2010). Comparison of our GBM clustering with these existing subtypes (Table 4) shows that our method not only reflects evidence from one data type, but also finds a cluster that considers both gene expression and DNA methylation information.

Table 4: Comparison of clusterings to established subtypes

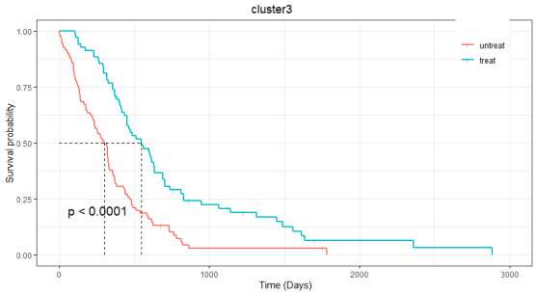
	Gene expression subtypes (Verhaak et al., 2010)				**DNA methylation subtypes** (Noushmehr et al., 2010)	
Clusters	Classical	Mesenchymal	Neural	Proneural	G-CIMP	Non-G-CIMP
Cluster1	0	0	1	13	16	14
Cluster2	2	25	2	2	0	31
Cluster3	55	41	31	24	2	152

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.

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 3 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.



(a) (b)



(c)

Figure 3: Survival analysis of GBM patients for treatment with Temozolomide in the different clusterings.

4.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.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials:

All data generated or analysed during this study are included in this published article [and its suppl ementary information files]

Competing interests: Not applicable.

Funding: Not applicable.

Authors' contributions: T J’ contributions are processed a multi-view clustering based on Stiefel

manifold method (MCSM) and the numerical simulation results of MCSM model. Z JP' contribution lies in the embellishment of the article. Z CH' contribution lies in the thought guidance of the method.

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).

References

[1] Absil PA, Mahony R and Sepulchre R. Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton, NJ (2008).11462-11467. doi: 10.1515/9781400830244

[2] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, and Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America (1999) 96(12): 6745-6750. doi: 10.1073 / pnas.96.12.6745

[3] Ashburner M, Ball CA, Botstein D and Blake JA. Gene ontology: tool for the unification of biology. Nature Genetics (2000) 25(1): 25-29. doi: 10.1038/75556

[4] Bendor A, Shamir R and Yakhini Z. Clustering gene expression patterns. Journal of Computational Biology (1999) 6: 281-297. doi: 10.1145/299432.299448

[5] Bickel PJ and Chen A. A nonparametric view of network models and new managirvan and other modularities. Proceedings of the National Academy of Sciences of the United States of America (2009)106(50): 21068-21073. doi: 10.1073/pnas.0907096106

342 [6] Cancer Genome Atlas Research Network. Comprehensive genomic characterization of
343 squamous cell lung cancers. Nature (2012) 489: 519-525. doi: 10.1016/j.ypat.2012.11.048

344 [7] Chang HY, Nuyten DSA, Sneddon JB, Hastie Tibshirani R, Dai HY and He YD. Robustness,
345 scalability and integration of a wound-response gene expression signature in predicting breast
346 cancer survival. Proceedings of the National Academy of Sciences of the United States of
347 America (2005) 102(10): 3738-3743. doi:10.1073/pnas.0409462102

348 [8] Chalise P, Koestler DC, Bimall M, Yu Q. Integrative clustering methods for high dimensional
349 molecular data. Translational cancer research (2014) 3(3), 202–216. doi: 10.3978/j.issn.2218-
350 676X.2014.06.03

351 [9] Chen C, Ng MK, and Zhang S. Block spectral clustering methods for multiple graphs.
352 Numerical Linear Algebra with Applications (2017) 24(1): 1-20. doi: 10.1002/nla.2075

353 [10] Dai LY, Feng CM, Liu JX, Zheng CH, Yu J. Robust graph regularized discriminative
354 nonnegative matrix factorization for characteristic gene selection. IEEE International
355 Conference on Bioinformatics and Biomedicine (2016) 1253-1258. doi:
356 10.1109/BIBM.2016.7822698

357 [11] Ding H, Michael S, Wang C. Integrative cancer patient stratication via subspace merging.
358 Bioinformatics (2018) 35(10): 1653-1659. doi: 10.1093/bioinformatics/bty866

359 [12] Fridley BL, Lund SP, Jenkins GD, Wang LW. A Bayesian integrative genomic model for
360 pathway analysis of complex traits. Genetic epidemiology (2012) 36:352-359. doi:
361 10.1002/gepi.21628

362 [13] Ge SG, Xia J, Sha W, Zheng CH. Cancer subtype discovery based on integrative model of
363 multigenomic data. IEEE/ACM transactions on computational biology and bioinformatics
364 (2017) 14(5): 1115-1121. doi: 10.1109/TCBB.2016.2621769

365 [14] Ha MJ, Baladandayuthapani V and Do KA. Dingo: Differential network analysis in genomics.
366 Bioinformatics (2015) 31(21): 3413-3420. doi: 10.1093/bioinformatics/btv406

367 [15] Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP and Cancer
368 Genome Atlas Research Network. Identification of a CpG island methylator phenotype that
369 defines a distinct subgroup of glioma. Cancer cell (2010) 17(5): 510-522.

370 [16] Huang HC, Chuang YY and Chen CS. Affinity aggregation for spectral clustering. Conference
371 on Computer Vision and Pattern Recognition (2012) 2012: 773-780. doi:
372 10.1109/CVPR.2012.6247748

373 [17] Huang BY, Zhang HB, Gu LJ, Ye BX and Xiong X. Advances in immunotherapy for
374 glioblastoma multiforme. Clinical Developmental Immunology (2017) 2017(3): 1-11. doi:
375 10.1155/2017/3597613

376 [18] Huang E, Cheng SH, Dressman H and Pittman J. Gene expression predictors of breast cancer
377 outcomes. The Lancet (2003) 361(9369): 1590-1596. doi: 10.1016/S0140-6736(03)13308-9

378 [19] Kanehisa M, Goto S, Furumichi M, Tanabe M and Hirakawa M. Kegg for representation and
379 analysis of molecular networks involving diseases and drugs. Nucleic Acids Research (2010)
380 38(suppl_1): D355-D360. doi: 10.1142/9781848165632_0020

381 [20] Kumar R, Kamdar D, Madden L, Hills C. Th1/th2 cytokine imbalance in meningioma,

anaplastic astrocytoma and glioblastoma multiforme patients. Oncology Reports (2006) 15(6): 1513-1516. doi: 10.3892/or.15.6.1513

[21] Shen R, Olshen AB and Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics (2010) 25(22): 2906-2912. doi: 10.1093/bioinformatics/btp659

[22] Shen R, Mo Q, Schultz N and Seshan VE. Integrative subtype discovery in glioblastoma using icluster. Plosone (2012) 7(4): e35236. doi: 10.1371/journal.pone.0035236

[23] Speicher NK and Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery Bioinformatics (2015) 31(12): i268-i275. doi: 10.1093/bioinformatics/btv244

[24] Stefansson OA, Moran S, Gomez A and Sayols S. A DNA methylation-based definition of biologically distinct breast cancer subtypes. Molecular Oncology (2015) 9(3): 555-568. doi: 10.1016/j.molonc.2014.10.012

[25] Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, and 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.

[26] West M, Blanchette C, Dressman H and Huang E. Predicting the clinical status of human breast cancer by using gene expression profiles. Proceedings of the National Academy of Sciences of the United States of America (2001) 98(20): 11462-11467. doi: 10.1073/pnas.201162998

[27] Wang B, Mezlini AM, Demir F and Fiume M. Similarity network fusion for aggregating data

types on a genomic scale. Nature Methods (2014) 11(3): 333-337. doi: 10.1038/nmeth.2810

[28] Yuan Y, Savage RS, Markowetz F. Patient-specific data fusion defines prognostic cancer subtypes. Plos Computational Biology (2011) 7(10): e1002227. doi: 10.1371/journal.pcbi.1002227

[29] Yu Y, Zhang LH and Zhang SQ. Simultaneous clusterin of multiview biomedical data using manifold optimization, Bioinformatics (2019) 35(20): 4029-4037. doi: 10.1093/bioinformatics/btz217

[30] Zhang D, Chen P, Zheng CH, and Xia JF. Identification of ovarian cancer subtype-specific network modules and candidate drivers through an integrative genomics approach. Oncotarget (2016) 7(4): 4298. doi: 10.18632/oncotarget.6774

[31] Zhang S, Zhao H and Ng MK. Functional module analysis for gene coexpression networks with network integration. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2015) 12(5): 1146-1160. doi: 10.1109/TCBB.2015.2396073

[32] Zheng CH, Ng TY, Zhang L, Shiu, C. K., & Wang, H. Q. Tumor classification based on non-negative matrix factorization using gene expression data. IEEE transactions on nanobioscience (2011) 10(2): 86-93. doi: 10.1109/TNB.2011.2144998

[33] Zheng CH, Yang W, Chong YW, Xia JF. Identification of mutated driver pathways in cancer using a multi-objective optimization model. Computers in Biology and Medicine (2016) 72: 22-29. doi: 10.1016/j.compbiomed.2016.03.002

Figures

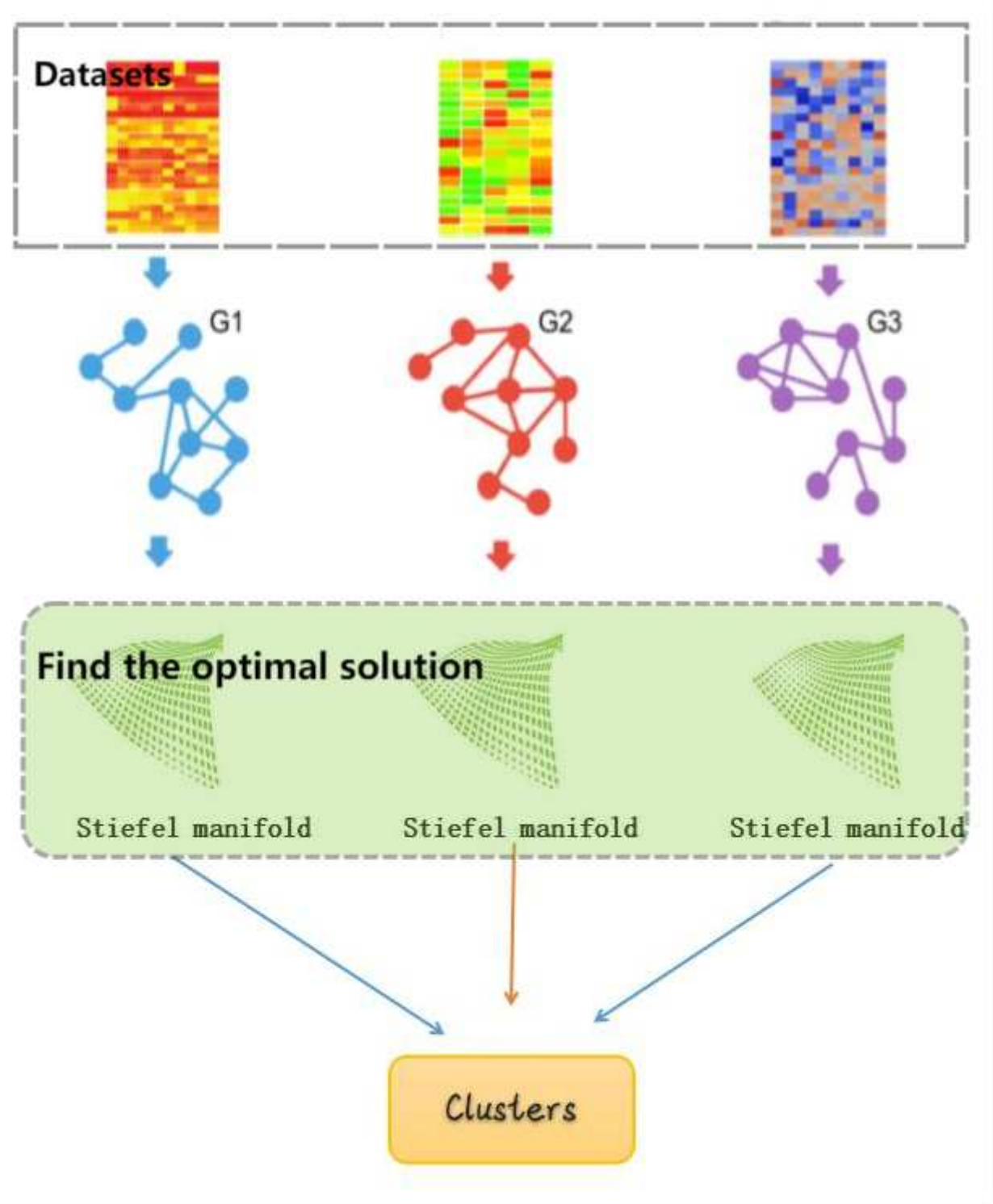


Figure 1

The process of MCSM method.

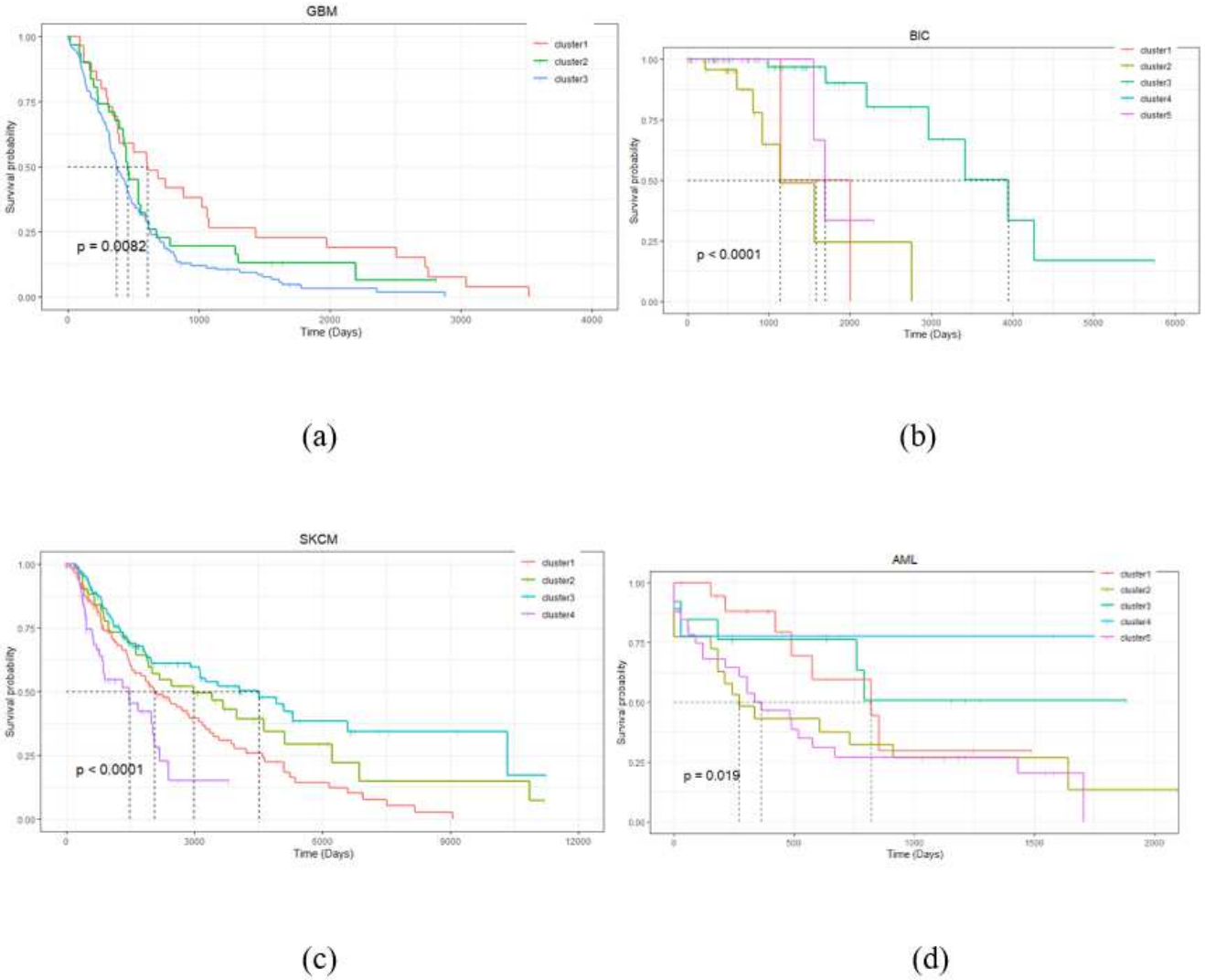


Figure 2

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

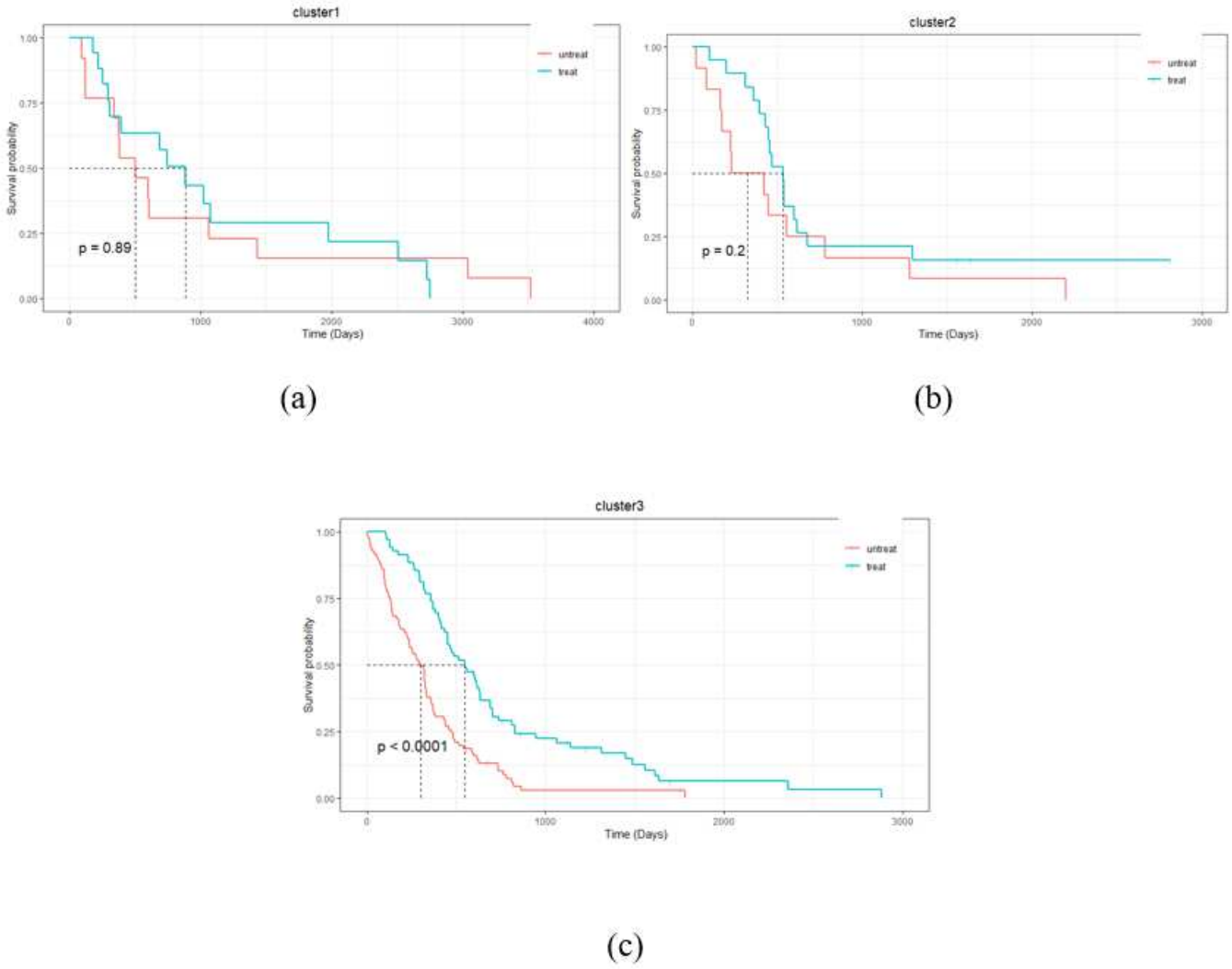


Figure 3

Survival analysis of GBM patients for treatment with Temozolomide in the different clusterings.