Network Modeling Analysis in Health Informatics and Bioinformatics (2022) 11:20
https://doi.org/10.1007/s13721-022-00362-6

ORIGINAL ARTICLE

A new clustering algorithm for genes with multiple cancer diseases by self-consistent field iteration method

Ye Liu[1] · **Michael K. Ng**[2]

Received: 13 September 2021 / Revised: 25 February 2022 / Accepted: 8 March 2022

Abstract
An increasing body of literature shows that predicting gene clusters related to human cancer disease using biological networks is significant in bioinformation, it would help to understand disease mechanisms, and benefit the development of diagnostics and therapeutics. However, due to noise and preprocessing of data, a single network or graph generated from one cancer disease is insufficient to cluster genes. As some cancer diseases are correlated with each other in practice, by integrating several gene expression networks generated from those associated cancer diseases, more accurate and robust partition of genes can be obtained. In this paper, we propose a multiple graph spectral clustering method with graph association, it helps us to discover functional modules in each cancer disease more accurately and comprehensively, meanwhile the degree of association among cancer diseases can also be determined. Our idea is to construct a block adjacency matrix to integrate the adjacency matrix of each graph and the degree of association among multiple graphs together, then spectral clustering would be employed to calculate clusters for each graph. The proposed algorithm is based on a self-consistent field iteration such that both the degree of association and gene clusters can be identified during iterations. Moreover, we establish the condition under which convergence of the proposed algorithm is guaranteed with some assumptions. Experimental results on two datasets of human cancer diseases are presented, which demonstrate that the proposed method can not only identify gene functional modules, but also calculate the degree of association among different cancer diseases accurately.

Keywords Gene clusters · Spectral clustering · Gene coexpression data · Nonlinear eigenvalue problem · Self-consistent field iteration

1 Introduction

The completion of Human Genome Project (HGP) and the development of advanced high-throughput technologies have introduced large quantities of gene expression datasets from different cancer diseases. Such datasets have motivated lots of research works to investigate the relationship among genes. Particularly, numerous research works have been done to predict gene clusters related to human cancer diseases (Chen et al. 2014; Gill et al. 2014; Hu et al. 2018; Li et al. 2014; Sun et al. 2011). Usually, genes in the same cluster are functionally related, so disease-related gene clusters can support the existence of distinct disease-specific functional modules (Goh et al. 2007; Oti et al. 2006; Wu et al. 2010). The identification of gene clusters can not just help to better understand cancer disease mechanisms, but provide new and exact diagnostics and therapeutics.

An undirected graph is often used to represent the gene expression dataset, where interesting genes are vertices and the pairwise relationships among genes are edges. Usually, one single graph is insufficient to analyze due to noise and uncertainty in data generation and preprocess. Nowadays, datasets over the same set of genes with different kinds of the relationship generated from different sources are available. For example, different tissues and diseases may share some common functional modules when analyzing multiple gene expression datasets obtained from different tissues or diseases. Quantities of works about such multiple relational

✉ Ye Liu
yliu03@scut.edu.cn

Michael K. Ng
mng@maths.hku.hk

[1] School of Future Technology, South China University of Technology, Guangzhou, Guangdong, China

[2] Department of Mathematics, The University of Hong Kong, Pokfulam, Hong Kong

Published online: 07 April 2022

datasets have been done, refer to (Hu et al. 2005; Shen et al. 2015; Zhang and Ng 2016; Zhang et al. 2015).

According to literatures, multi-view clustering has been well developed to discover the underlying clustering structures of such datasets, see for instance (Chen et al. 2017; Kumar and Daumé 2011; Tang et al. 2009; Xie and Sun 2013). However, most existing methods assume that all graphs contribute the same when exploiting partitions of objects across graphs. While, in reality, some graphs may be strongly or weakly associated with the other graphs. For example, when we integrate gene expression datasets from different cancers, some cancers may be the same subtypes from the same tissue, which may share the same features compared with those from different tissues. Better clustering results can be obtained with the consideration of association among graphs.

In this paper, we propose to solve multiple graphs clustering by self-consistent field (SCF) iteration. Our idea is to construct a block adjacency matrix to integrate multiple graphs, diagonal blocks refer to the relationship within each graph, the off-diagonal blocks refer to the degree of association among different graphs. During each iteration, spectral clustering is performed, clusters in each graph and degree of association among graphs would be calculated by utilizing the corresponding eigenvectors. The whole process can be formulated as a nonlinear eigenvalue problem, which can be solved by SCF iteration, the condition under which the algorithm converges is analyzed. Experimental results on two gene expression datasets are given to illustrate its superior clustering performance and excellent ability in measuring the degree of association among cancers.

The rest of this paper is organized as follows. In Sect. 2, we briefly review some related work. In Sect. 3, the proposed model and algorithm are introduced. In Sect. 4, two gene expression datasets are used to illustrate the performance of our method. Conclusion is given in Sect. 5. In "Appendix", we show the convergence analysis of our proposed model and identify the condition under which the convergence of the algorithm is guaranteed.

2 Related work

We will show the related works about graph clustering from three aspects.

2.1 Clustering for single graph

Spectral clustering is one of the most popular clustering algorithms when dealing with one single graph. Given a graph $G = (V, E)$, V represent the set of N vertices, which refers to the interested objects for clustering. And E is the set of edges which connect vertices. The corresponding adjacency matrix $\boldsymbol{A}$ can be constructed for an undirected graph G, where $\boldsymbol{A}_{ij}$ represents the connection between ith object and jth object. If the ith object and the jth object have a link, then $\boldsymbol{A}_{ij} = 1$, otherwise $\boldsymbol{A}_{ij} = 0$. If there are K clusters, we denote a N-by-K matrix $\boldsymbol{U}$ as an object-cluster-indicator matrix. When the ith object is assigned to the jth cluster, then $\boldsymbol{U}_{ij} = 1$, otherwise it is equal to 0. From the graph cut point of view, spectral clustering aims to find a partition of a graph such that the edges between different groups have very low weights and the edges within groups have high weights. The optimization problem is formulated as:

$$\max_{\mathbf{U}} \ \mathrm{Tr}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}) \quad \text{subject to} \quad \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}_K \tag{1}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. The optimal solution $\boldsymbol{U}$ in Eq. (1) is K eigenvectors corresponding to the first K largest eigenvalues of $\boldsymbol{A}$. The clustering result is obtained by clustering the embeddings of data points $\boldsymbol{U}$, which are the K eigenvectors corresponding to the first K largest eigenvalues of $\boldsymbol{A}$. The detailed theoretical and practical introduction about spectral clustering is shown in Von Luxburg (2007).

2.2 Clustering for multiple graphs

According to the literature, more accurate clustering results can be obtained by integrating information from multiple graphs. In this section, we will briefly review some related works. A multi-objective method with the pareto frontier is proposed in Wang et al. (2013) to search the joint numerical range of multiple graphs. While choosing preferred cuts from alternative clusters brings high computational costs. Xie and Sun (2013) extended clustering ensembles to multi-view clustering. In Liu et al. (2012), the authors presented a tensor-based framework to solve multi-domain spectral clustering problem, this formulation is solved by high-order singular value decomposition. A probabilistic generative model is proposed by Shiga and Mamitsuka (2010), based on variational Bayesian estimation, a robust learning scheme is used to solve. Dong et al. (2012) combined the spectrum of multiple graphs to cluster vertices. Linked matrix factorization (LMF) method is proposed in Tang et al. (2009). For each graph, a specific factor and a common factor for all graphs can be calculated by applying matrix factorization. The common factor provides features used to cluster all vertices. In Zhou and Burges (2007), a method based on spectral clustering is developed for multi-view data, which generalizes the normalized cut from a single view to multiple views. Kumar and Daumé (2011) proposed to combine multi-view spectral clustering with co-training. For each graph, the structure would be updated by using clustering results from other graphs, such a process would repeat iteratively. In Kumar et al. (2011), Kumar et al. presented a co-regularization framework by a cost function containing two

parts, one measures disagreement between clusters of two views, another is the spectral clustering objective function of each individual view.

There are two drawbacks to the above methods. One is that all methods seek the common clusters across all graphs, they assume that the same objects of different graphs should be in the same cluster. Another is that they do not take the association among different graphs into consideration when integrating information from multiple graphs, clustering performance may be degraded when some graphs are less relevant to the other graphs.

2.3 Multiple graphs association

In multiple graphs clustering, more accurate clustering results would be obtained by identifying relevant and irrelevant graphs for each graph. In Cheng et al. (2016), co-regularized graph clustering (CGC) method based on non-negative matrix factorization was proposed, it makes use of relationships across multiple graphs to enhance clustering result and reevaluate the consistency across graphs. Ni et al. (2015) developed NoNClus method, it models the similarity of graphs as a graph, which can be utilized to regularize the clustering structures in different networks. While the calculation of NMF in each iteration of CGC and NoNClus causes high computational cost. Sun et al. (2013) proposed a probabilistic approach to integrate meta-path selection with user-guided clustering, the weight of each meta path would be learned and clusters with the updated weight would be generated. Recently, Chen et al. (2017) combined the Laplacian matrix of all graphs into an block Laplacian matrix and used a trade-off parameter to control the association among graphs, then apply spectral clustering on this block Laplacian matrix.

The disadvantages of the above-mentioned methods are that their optimization process is computationally intensive and costly, and some of them seek consensus partition across all graphs.

3 The proposed model

In this section, we propose a multiple graphs spectral clustering method with graph association.

3.1 Degree of association calculation

It is significant to design an appropriate method to calculate the degree of association among graphs such that accurate clustering results would be achieved by data integration. Assume that $\boldsymbol{X}_j$ is a N-by-K matrix containing the clustering information of the jth graph. The degree of association between two graphs should be higher when they have similar clustering structure. In Liu et al. (2018), the degree of association between ith domain and jth domain is calculated using cosine similarity $\frac{\max\{\boldsymbol{x}_i^T\boldsymbol{x}_j,-\boldsymbol{x}_i^T\boldsymbol{x}_j\}}{\|\boldsymbol{x}_i\|_2\|\boldsymbol{x}_j\|_2}$. Normalization is used here in order to control the scale of association and make it comparable across graphs, thus maximum is to avoid the case that the signs of two vectors may be different. Inspired by it, instead of using one eigenvector in Liu et al. (2018), here, we calculate the degree of association $\boldsymbol{B}(i,j)$ between ith graph and jth graph using K eigenvectors according to

$$\boldsymbol{B}(i,j)=\frac{|\mathrm{Tr}(\boldsymbol{X}_i\boldsymbol{X}_j^{\mathrm{T}})|}{\|\boldsymbol{X}_i\|_F\|\boldsymbol{X}_j^{\mathrm{T}}\|_F}\quad(i\neq j) \tag{2}$$

Here, Tr() is the trace of a matrix. |.| means taking the absolute value instead of maximum. By this way, when two graphs have similar clustering structures, a large $\boldsymbol{B}(i,j)$ would be achieved, which reflects a large correlation between the ith graph and the jth graph.

3.2 Block adjacency matrix with domain association

When given M undirected graphs $G_1, G_2, \ldots, G_M$ over the same set of N vertices. $\boldsymbol{A}_m \in \mathbb{R}^{N\times N}$ denotes the adjacency matrix of graph G_m. For the graph G_m, if the ith object and the jth object are connected with a link, then $\boldsymbol{A}_m(i,j)=1$. Otherwise there is no link. Then we can construct a M-by-M block adjacency matrix $\boldsymbol{A}(\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_M)$ to integrate the relationship within each graph and degree of association among multiple graphs together:

$$\begin{aligned}&\boldsymbol{A}(\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_M)\\ =&\mathbf{A}+\alpha\mathbf{B}\otimes\boldsymbol{I}_N\\ =&\begin{bmatrix}\boldsymbol{A}_1 & \frac{\alpha|\mathrm{Tr}(\boldsymbol{X}_1\boldsymbol{X}_2^{\mathrm{T}})|}{\|\boldsymbol{X}_1\|_F\|\boldsymbol{X}_2\|_F}\boldsymbol{I}_N & \cdots & \frac{\alpha|\mathrm{Tr}(\boldsymbol{X}_1\boldsymbol{X}_M^{\mathrm{T}})|}{\|\boldsymbol{X}_1\|_F\|\boldsymbol{X}_M\|_F}\boldsymbol{I}_N\\ \frac{\alpha|\mathrm{Tr}(\boldsymbol{X}_2\boldsymbol{X}_1^{\mathrm{T}})|}{\|\boldsymbol{X}_2\|_F\|\boldsymbol{X}_1\|_F}\boldsymbol{I}_N & \boldsymbol{A}_2 & \cdots & \frac{\alpha|\mathrm{Tr}(\boldsymbol{X}_2\boldsymbol{X}_M^{\mathrm{T}})|}{\|\boldsymbol{X}_2\|_F\|\boldsymbol{X}_M\|_F}\boldsymbol{I}_N\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\alpha|\mathrm{Tr}(\boldsymbol{X}_M\boldsymbol{X}_1^{\mathrm{T}})|}{\|\boldsymbol{X}_M\|_F\|\boldsymbol{X}_1\|_F}\boldsymbol{I}_N & \frac{\alpha|\mathrm{Tr}(\boldsymbol{X}_M\boldsymbol{X}_2^{\mathrm{T}})|}{\|\boldsymbol{X}_M\|_F\|\boldsymbol{X}_2\|_F}\boldsymbol{I}_N & \cdots & \boldsymbol{A}_M\end{bmatrix}\end{aligned} \tag{3}$$

$\boldsymbol{B}$ in Eq. (4) is a matrix describing the degree of association among graphs, a larger value in $\boldsymbol{B}(i,j)$ illustrate that ith graph and jth graph has a larger degree of association.

$$\boldsymbol{B}=\begin{bmatrix}0 & \frac{|\mathrm{Tr}(\boldsymbol{X}_1\boldsymbol{X}_2^{\mathrm{T}})|}{\|\boldsymbol{X}_1\|_F\|\boldsymbol{X}_2^{\mathrm{T}}\|_F} & \cdots & \frac{|\mathrm{Tr}(\boldsymbol{X}_1\boldsymbol{X}_M^{\mathrm{T}})|}{\|\boldsymbol{X}_1\|_F\|\boldsymbol{X}_M^{\mathrm{T}}\|_F}\\ \frac{|\mathrm{Tr}(\boldsymbol{X}_2\boldsymbol{X}_1^{\mathrm{T}})|}{\|\boldsymbol{X}_2\|_F\|\boldsymbol{X}_1^{\mathrm{T}}\|_F} & 0 & \cdots & \frac{|\mathrm{Tr}(\boldsymbol{X}_2\boldsymbol{X}_M^{\mathrm{T}})|}{\|\boldsymbol{X}_2\|_F\|\boldsymbol{X}_M^{\mathrm{T}}\|_F}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{|\mathrm{Tr}(\boldsymbol{X}_M\boldsymbol{X}_1^{\mathrm{T}})|}{\|\boldsymbol{X}_M\|_F\|\boldsymbol{X}_1^{\mathrm{T}}\|_F} & \frac{|\mathrm{Tr}(\boldsymbol{X}_M\boldsymbol{X}_2^{\mathrm{T}})|}{\|\boldsymbol{X}_M\|_F\|\boldsymbol{X}_2^{\mathrm{T}}\|_F} & \cdots & 0\end{bmatrix} \tag{4}$$

Moreover, α is a known constant to balance the inner-relation information within each graph and the association

among graphs. $\boldsymbol{A}$ contains the relationship among interest of objects for each graph.

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_2 & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{A}_M \end{bmatrix} \tag{5}$$

And $\boldsymbol{A}_i$ is a symmetric matrix describing the similarity among objects in the ith graph. It is easy to verify that both $\boldsymbol{A}(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_M)$ and $\boldsymbol{B}$ are symmetric matrices.

3.3 Multiple graphs clustering by self-consistent field iteration (SCF) method

In order to obtain better classification performance and precise degree of association simultaneously, we propose to solve the following self-consistent field (SCF) iteration problem:

$$\boldsymbol{A}(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_M)\boldsymbol{X} = \boldsymbol{\Lambda}_k \boldsymbol{X} \quad \text{s.t. } \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{I}_K \tag{6}$$

where $\boldsymbol{X} = [\boldsymbol{X}_1^{\mathrm{T}}, \boldsymbol{X}_2^{\mathrm{T}}, \dots, \boldsymbol{X}_M^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{MN \times K}$, it contains the clustering information of each graph, and $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{I}_K$. $\boldsymbol{A}(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_M) \in \mathbb{R}^{MN \times MN}$ is defined in Eq. (3),it is a M-by-M block matrix, and each block is a N-by-N matrix. Assume that each graph contains K clusters, then $\boldsymbol{X}$ contains K eigenvectors corresponding to the first K largest eigenvalue $\boldsymbol{\Lambda}_K$ of $\boldsymbol{A}(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_M)$. Further we would use this $\boldsymbol{X}$ to construct a new block adjacency matrix $\boldsymbol{A}(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_M)$ according to Eq. (3). In each iteration, clusters for each graph are determined with the largest separability by considering both correlation between objects within each graph and correlation among graphs. Then the new updated vectors $\boldsymbol{X}$ containing clustering information would be used to calculate new degree of association. The SCF procedure will be terminated until the difference between block matrix $\boldsymbol{A}(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_M)$ in two successive iterations is negligible, then stable partition of objects and finalized accurate association among graphs would be achieved. The detail of algorithm is shown in Algorithm 1. The convergence of Algorithm 1 also be provided in "Appendix". We remark that (1) we only use information within each graph to calculate the degree of association in the beginning of algorithm. (2) By using Eq. (4), we have $\boldsymbol{B}_{i,j} = \boldsymbol{B}_{j,i}$. When two different graphs are similar (or not similar) clustering structures, their association $\boldsymbol{B}_{i,j}$ would be large (or small).

Algorithm 1 Multiple graphs clustering by self-consistent iteration algorithm

Input: M-by-M block symmetric adjacency matrix $\mathbf{A}$.

Step 1: $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M$ are the first K largest eigenvectors of $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_M$ separately, let $\mathbf{X} = [\mathbf{X}_1^T, \mathbf{X}_2^T, \dots, \mathbf{X}_M^T]^T$, use Eq. (3) to update $\mathbf{A}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M)$.

repeat

Step 2: Compute the first K largest eigenvectors $\mathbf{X}^i$ of $\mathbf{A}^i(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M)$ by optimization of Eq. (6);

Step 3: Apdate $\mathbf{A}^{(i+1)}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M)$ according to the Eq. (3) using $\mathbf{X}^i$

until $\|\mathbf{A}^{(i+1)}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M) - \mathbf{A}^i(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M)\| < tol$ ($tol = 10^{-5}$ in the experiment).

Step 4: Calculate corresponding Laplacian matrix of $\mathbf{A}^{(i+1)}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M)$, then apply spectral clustering on it.

4 Experimental results

We used two gene expression datasets to evaluate the performance of the proposed algorithm in clustering and its ability in the degree of association, thus we compared it with the BLSC method (Chen et al. 2017).

4.1 Gene expression data: LUAD-LUSC-OV-UCEC

The gene expression data for four cancer diseases is downloaded from The Cancer Genome Atlas (TCGA), these four cancer diseases are: lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), and uterine corpus endometrial carcinoma (UCEC). All cancer disease datasets are generated with Affy-metrix HT_HG-U133A by the Broad Institute. There are 32 samples, 154 samples, 597 samples and 54 samples in LUAD, LUSC, OV and UCEC respectively, 17,825 common genes are involved in total. Firstly, we calculated the variance of all the genes across samples, and selected the first 1500 genes with the largest variance for each cancer. Then we took the union of genes from four cancers, finally there are 2831 common genes selected as the interested objects for further analysis.

To construct the gene coexpression networks, first, the similarity between two genes is calculated according to the Pearson correlation coefficient, then use hard thresholding to create the adjacency matrix. It means if the value of similarity between two genes is larger than a given value, then an edge with weight 1 is assigned between them; if not, they do not connect.

We tried different thresholds and compute the linear regression coefficient between the frequency of degree $d(\log10(f(d)))$ and log10 transformed degree $d(\log10(d))$ to see whether these networks have approximately scale-free property (Zhang and Horvath 2005). We finally set the hard threshold for LUAD, LUSC, OV and UCEC as 0.68, 0.51, 0.74 and 0.59 respectively. Then we constructed the corresponding adjacency matrix A_j for four cancer diseases. So here $M = 4$ and $N = 2831$.

In terms of the number of clusters K, we firstly calculated the eigenvalues of corresponding Laplacian matrix of the adjacency matrix A_j for each cancer, then observed the first big jump $(\lambda_{i+1}-\lambda_i)/(\lambda_i-\lambda_{i-1})$ in the relative change of successive eigenvalues, the first big jump of four cancers are shown in Fig. 2. The first big jump of eigenvalue for LUAD occurs in 533, that of LUSC occurs in 550, that of OV appears in 1306, and that of UCEC is 545. So in the experiment, we considered 533, 550, 545 and 1306 as the value of K, and calculated the consistent clusters for each cancer with different values of K.

Figure 1 shows the average calculated degree of association among four cancers with the change of α when $K = 533$, it can be seen that the average degree of association close to zero when α is smaller than 1, and it is close to 1 when α close to 10. So to control the final convergent value of $\boldsymbol{B}_{ij}$ to be reasonable, α is set as 2 . After applying Algorithm 1, we calculate the average degree of association when K is set as 533, 550, 545 and 1306, it is displayed in Table 1. Then we calculated the common clusters among four different cancers and deleted those clusters whose size was smaller than 5 to simplify the analysis. Thus, we show the common genes shared by clusters among four cancer diseases in Table 2. It is clear that the number of common genes is consistent with the value of $\boldsymbol{B}_{ij}$, which illustrates that our method can determine an accurate degree of association among cancer diseases.

4.1.1 Clusters analysis

For OV and UCEC, there are 28 common clusters, the detailed information is shown in Table 3. We further performed enrichment analysis for Gene Ontology (GO, biological process) and KEGG pathways for these clusters with DAVID (Huang et al. 2009). In Table 3, size means the number of genes in the cluster. N_{GO} or N_{KEGG} refer to the number of enriched GO terms or KEGG pathways. 20 gene clusters enrich GO terms, and 14 gene clusters enrich KEGG pathways. The enrichment results with P value less than 1E-6 for GO categories are shown in Table 4.

In Cluster #1, six among 7 genes are involved in the regionalization and pattern specification process (Zhang 2018). All six genes belong to the gene family HOX, they are HOXB2, HOXB3, HOXB5, HOXB6, HOXB8, HOXB9. According to existing results, the HOX gene is highly dysregulated in cancer. What's more, according to current evidence, some HOX genes play a general supportive role in malignancy at the cellular level and tumor level. For example, it promotes proliferation and blocking apoptosis (Morgan et al. 2007) at the cellular level. Moreover, at tumor level, where they have been shown to variously induce angiogenesis (Kachgal et al. 2012), drive metastasis (Hong et al. 2015), facilitate drug (Jin and Sukumar 2016; Li et al. 2015).

For Cluster #9, it is mainly related to responses, including immune response and inflammatory response with P value being 3.6E−45 and 4.05E−29. Immune response is recognized to have the potential to destroy cancer cells and inhibit tumor growth. A lot of evidences suggest that immunotherapeutic approaches will become the next major therapeutic advance for cancer, including lung cancer and ovarian cancer (Carbone et al. 2015; Domagala-Kulawik 2015; Domagala-Kulawik et al. 2014).For example, the immune system is recognized as an important mediator of ovarian carcinogenesis (Charbonneau et al. 2013). Inflammatory is considered as a key factor to promote tumor progression. Cancer-related inflammation influences many

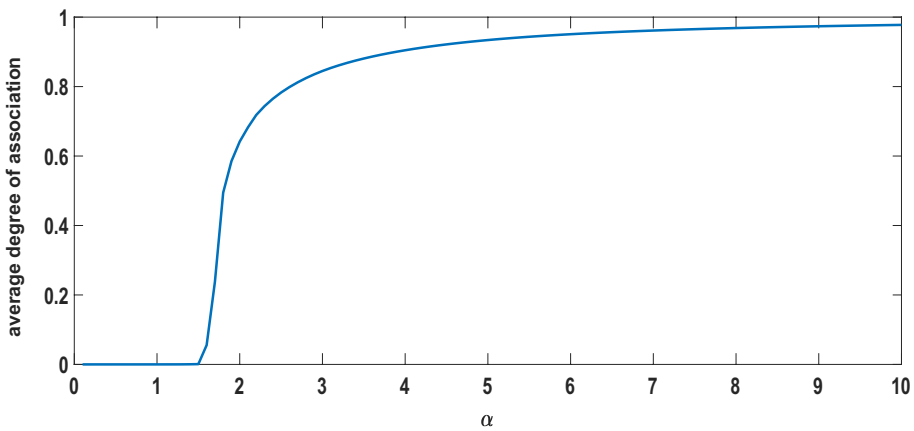

Fig. 1 The average degree of association among four cancers $\boldsymbol{B}_{ij}$ with the change of α when $K = 533$

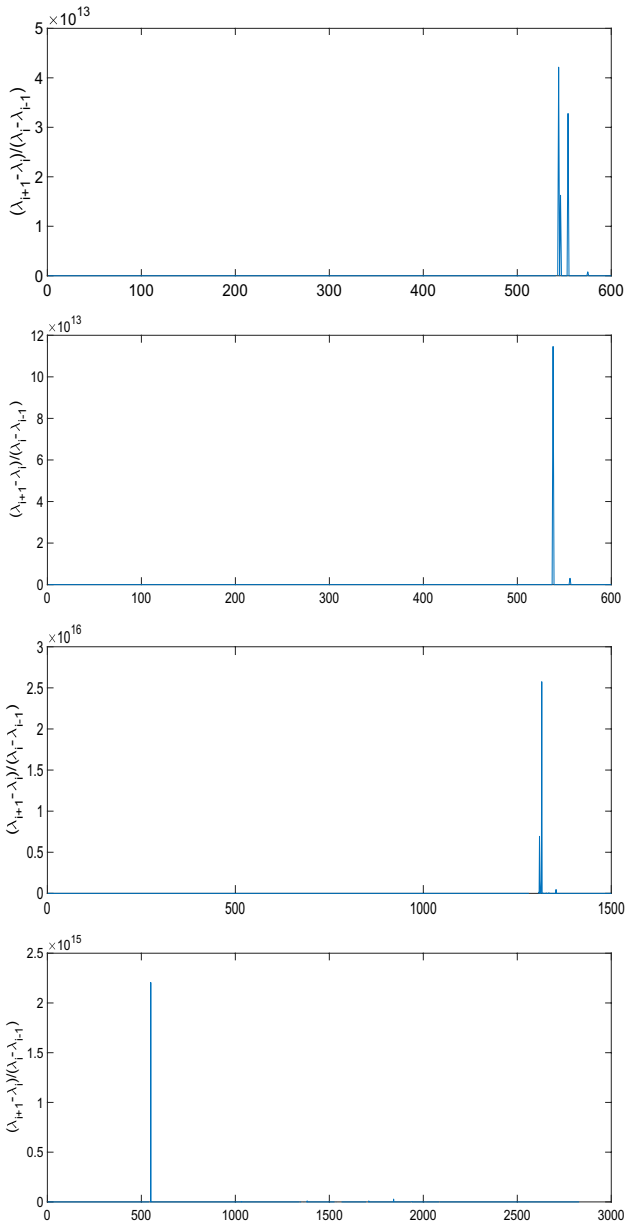

Fig. 2 The first big jump of eigenvalue for LUAD, LUSC, OV, UCEC appear near the 533th, 550th, 1306th, 545th eigenvalue

Table 1 The convergent values of $\boldsymbol{B}_{ij}$ for LUAD–LUSC–OV–UCEC

	LUAD	LUSC	OV	UCEC
LUAD	0	0.5243	0.7519	0.5857
LUSC	0.5243	0	0.7084	0.5235
OV	0.7519	0.7084	0	0.7559
UCEC	0.5857	0.5235	0.7559	0

Table 2 Overlapped genes shared by clusters among four cancers

	LUAD	LUSC	OV	UCEC
LUAD	0	1212	1237	1148
LUSC	1212	0	1283	1201
OV	1237	1283	0	1221
UCEC	1148	1201	1221	0

Table 3 Common clusters identified for OV and UCEC by our method

# Cluster	Size	N_{GO}	N_{KEGG}
1	7	5	0
2	23	25	1
3	15	9	2
4	11	0	0
5	23	4	0
6	15	1	0
7	7	0	0
8	8	4	1
9	214	148	40
10	18	7	2
11	6	0	0
12	8	0	0
13	12	2	0
14	12	6	1
15	119	107	9
16	6	0	0
17	11	1	0
18	11	6	1
19	54	11	2
20	67	23	12
21	13	1	0
22	20	11	1
23	116	15	1
24	6	0	0
25	6	0	0
26	10	0	0
27	41	27	10
28	119	44	2

aspects of malignancy, including the proliferation and survival of malignant cells, angiogenesis, tumor metastasis and tumor response to chemotherapeutic drugs and hormones (Gomes et al. 2014; Ness and Cottreau 1999). More comprehensive understanding towards inflammation will provide new insights to the treatment of cancer.

Cluster #10 is also related to immune responses, evidences already show that Type I interferons are protective in acute viral infections, which can activate the adaptive immune system, then boosting the development of high-affinity antigen-specific T and B cell responses and immunological memory (Ivashkiv and Donlin 2014). Moreover, Type I interferons have become key regulators that modulate tumor cell growth, proliferation, migration, apoptosis

Table 4 Gene ontology enrichment of the common clusters for OV and UCEC by our method

Cluster	Enriched GO terms	%	*P* value
Cluster 1	GO:0009952 anterior/posterior pattern specification	56.02	1.29E−11
	GO:0048704 embryonic skeletal system morphogenesis	46.69	3.71E−10
Cluster 9	GO:0006955 immune response	26.62	3.60E−45
	GO:0006954 inflammatory response	19.61	4.05E−29
	GO:0070098 chemokine-mediated signaling pathway	9.34	1.26E−21
	GO:0006935 chemotaxis	9.34	7.17E−17
	GO:0071346 cellular response to interferon-gamma	6.07	1.03E−12
	GO:0042102 positive regulation of T cell proliferation	6.07	1.97E−12
	GO:0002548 monocyte chemotaxis	5.14	2.24E−11
	GO:0050729 positive regulation of inflammatory response	6.07	2.29E−11
	GO:0060326 cell chemotaxis	5.60	1.15E−10
	GO:0050776 regulation of immune response	7.94	1.26E−10
	GO:0030593 neutrophil chemotaxis	5.60	1.37E−10
	GO:0007166 cell surface receptor signaling pathway	9.34	1.96E−10
	GO:0007186 G-protein coupled receptor signaling pathway	16.35	2.23E−10
	GO:0060333 interferon-gamma-mediated signaling pathway	5.60	3.11E−10
	GO:0006968 cellular defense response	5.14	1.35E−9
	GO:0007267 cell–cell signaling	8.41	3.17E−9
	GO:0007165 signal transduction	17.28	1.27E−8
	GO:0071347 cellular response to interleukin-1	4.67	8.14E−8
	GO:0031295 T cell costimulation 10	4.67	1.87E−7
	GO:0032496 response to lipopolysaccharide	6.07	2.75E−7
	GO:0071356 cellular response to tumor necrosis factor	5.14	3.74E−7
	GO:0048247 lymphocyte chemotaxis	3.27	5.08E−7
	GO:0070374 positive regulation of ERK1 and ERK2 cascade	6.07	5.53E−7
	GO:0090026 positive regulation of monocyte chemotaxis	2.80	6.09E−7
	GO:0002504 antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	2.80	8.55E−7
Cluster 10	GO:0060337 type I interferon signaling pathway	61.61	1.05E−18
	GO:0009615 response to virus	49.29	4.65E−12
	GO:0051607 defense response to virus	49.29	8.27E−11
	GO:0045071 negative regulation of viral genome replication	30.81	4.92E−8
Cluster 15	GO:0030198 extracellular matrix organization	12.98	1.83E−20
	GO:0030199 collagen fibril organization	7.08	4.28E−16
	GO:0030574 collagen catabolic process	7.67	4.19E−15
	GO:0007155 cell adhesion	12.39	5.52E−12
Cluster 27	GO:0055114 oxidation-reduction process	38.08	6.01E−13
	GO:0044597 daunorubicin metabolic process	10.15	3.17E−7
	GO:0044598 doxorubicin metabolic process	10.15	3.17E−7
Cluster 28	GO:0008544 epidermis development	13.96	5.60E−19
	GO:0031424 keratinization	6.98	7.78E−9
	GO:0030216 keratinocyte differentiation	7.85	9.03E−9

and so on. More research on it would improve the treatment of cancer (Lu et al. 2019).

For Cluster #15, four significantly enriched GO terms contribute to the extracellular matrix, which is a non-cellular component of tissue. Rather than its structural role, it serves for cell–cell communication, cell adhesion and cell proliferation (Walker et al. 2018). Collagen is the most significant component and most abundant fibrous protein in the

extracellular matrix, it makes up to 30% of total protein in multicellular animals (Daley et al. 2008). The primary functional properties of the matrix are dominated by collagen, changes in the deposition or degradation will result in the loss of extracellular matrix homeostasis, which would have great influences on the proliferation of cancer cells (Fang et al. 2014; Provenzano et al. 2006). An overall understanding toward extracellular matrix has been intensely researched to investigate its potential markers for better diagnosis and prognosis of various cancers, including ovarian cancer and lung cancer (Cho et al. 2015; Lim et al. 2017; Nadiarnykh et al. 2010).

In Cluster #27, three enriched GO terms are all about metabolic process, both doxorubicin and daunorubicin are anthracycline drugs which used in the treatment of many cancers, including lung, ovarian and breast (Thorn et al. 2011). Oxidation-reduction process is a type of redox reaction, which has been shown to control carcinogenesis, cancer cell proliferation, migration, invasion, metastasis and cancer vascularization (Hegedűs et al. 2018). A lot of researches have been provided to show that cell death in cancer can be induced by targeting redox-mediated signaling cascade (Acharya et al. 2010; Nagane et al. 1998; Pegram et al. 2000; Tsai et al. 1996). Therefore, further study about this cluster would contribute to new drug targets and novel drug developments of cancer.

For Cluster #28, all enriched GO terms are about epidermis and corresponding cell junctions, which are essential components for mamals (Liu et al. 2013). Epidermis is the outermost of three layers that make up the skin, it is primarily composed of keratinocytes, both keratinization and keratinocyte differentiation are a normal physiological process about keratinocytes. There are a lot of researches which show that epidermal growth factor receptor inhibition may be an effective strategy to enhance the effectiveness of treatment in ovarian cancer and endometrial cancer (Alper et al. 2001; Dong et al. 2013; Granados et al. 2015; Hudson et al. 2009; Nishimura et al. 2015; Siwak et al. 2010).

4.1.2 Comparison with BLSC method

To illustrate the performance of our methods better, we conducted an experiment on the LUAD-LUSC-OV-UCEC dataset using BLSC method (Chen et al. 2017), which assumes the association among four cancers is the same. We kept the same number of clusters K and α, and did the same process as what we did in the last subsection. We also conducted enrichment analysis for Gene Ontology and KEGG pathway on the common clusters of OV and UCEC using DAVID. The final result is shown in Tables 5 and 6.

Table 5 Common clusters identified for OV and UCEC by BLSC

# Cluster	Size	N_{GO}	N_{KEGG}
A	248	160	39
B	187	134	15
C	29	26	13
D	114	49	3
E	7	5	0

For cluster #A and cluster #B in Table 5, they enriched similar GO terms with closed P value compared to cluster #9 and cluster #15 in Table 3. Compared with cluster #C and cluster #D in Table 5, cluster #27 and cluster #28 in Table 3 enriches the same three GO terms with much smaller P value, because some genes are not discovered by BLSC. Cluster #E in Table 5 is the same with Cluster #1 in Table 3.

In conclusion, compared with BLSC, GO terms enriched by clusters discovered by our methods have lower P value, which indicates that better clustering results for multiple cancer diseases would be obtained with taking associations among cancer diseases into consideration.

4.2 Gene co-expression data: COAD-KIPAN-KIRC

In this data set, we did research on the gene expression profiles of three different cancer diseases, they are kidney renal clear cell carcinoma (KIRC), Colon adenocarcinoma (COAD) and pan-kidney cohorts (KIPAN). These three datasets are also produced with Affymetrix HT_HG-U133A by the Broad Institute. This dataset contains 17185 common genes expressed in 172 COAD samples, 72 KIRC samples, 88 KIPAN samples. Similar to the last gene expression data, the variance of all the genes across samples for each cancer was calculated. 1500 genes with the largest variance were selected for each cancer. Finally, 2513 common genes were determined for data integration analysis by taking the union from three cancers. So $M = 3$ and $N = 2513$. Similar as the previous subsection, adjacency matrices for three gene co-expression networks were constructed. The threshold for COAD, KIPAN and KIRC are 0.58, 0.46 and 0.61 respectively. Moreover, we determined the number of clusters K following the way in the last subsection. The first big jump of eigenvalue for COAD occurs in 524, that of KIPAN appears in 117, that of KIRC is 340. So we consider the value of K as 117, 340 and 524, then consistent clusters for each cancer are computed with three different values of K.

Figure 3 shows the average degree of association among three cancers calculated by Algorithm 1 with the change of

Table 6 Gene Ontology Enrichment of the common clusters for OV and UCEC by BLSC

Cluster	Enriched GO terms	%	*P* value
Cluster A	GO:0006955 immune response	25.19	4.42E−45
	GO:0006954 inflammatory response	17.64	1.45E−26
	GO:0070098 chemokine-mediated signaling pathway	8.40	1.92E−20
	GO:0009615 response to virus	7.98	2.47E-15
	GO:0006935 chemotaxis	7.98	1.65E−14
	GO:0051607 defense response to virus	8.82	2.62E−14
	GO:0060337 type I interferon signaling pathway	6.30	4.98E−14
	GO:0050776 regulation of immune response	8.40	1.21E−12
	GO:0060333 interferon-gamma-mediated signaling pathway	5.88	4.64E−12
	GO:0071346 cellular response to interferon-gamma	5.46	5.45E−12
	GO:0042102 positive regulation of T cell proliferation	5.46	1.04E−11
	GO:0002548 monocyte chemotaxis	4.62	9.01E−11
	GO:0050729 positive regulation of inflammatory response	5.46	1.19E−10
	GO:0007166 cell surface receptor signaling pathway	8.82	3.24E−10
	GO:0060326 cell chemotaxis	5.04	5.19E−10
	GO:0007267 cell–cell signaling	8.40	6.06E−10
	GO:0030593 neutrophil chemotaxis	5.04	6.16E−10
	GO:0006968 cellular defense response	4.62	5.28E−9
	GO:0045071 negative regulation of viral genome replication	3.78	3.19E−8
	GO:0007165 signal transduction	16.38	5.79E−8
	GO:0007186 G-protein coupled receptor signaling pathway	13.87	1.16E−7
	GO:0071347 cellular response to interleukin-1	4.20	2.71E−7
	GO:0031295 T cell costimulation	4.20	6.14E−7
Cluster B	GO:0030198 extracellular matrix organization	9.68	3.07E−20
	GO:0030574 collagen catabolic process	5.81	6.73E−16
	GO:0030199 collagen fibril organization	5.03	9.01E−16
	GO:0007155 cell adhesion	10.07	1.05E−12
Cluster C	GO:0055114 oxidation-reduction process	46.39	4.73E−11
	GO:0044597 daunorubicin metabolic process	15.46	1.09E−7
	GO:0044598 doxorubicin metabolic process	15.46	1.09E−7
Cluster D	GO:0008544 epidermis development	12.91	1.10E−17
	GO:0031424 keratinization	6.89	5.28E−9
Cluster E	GO:0009952 anterior/posterior pattern specification	56.02	1.29E−11
	GO:0048704 embryonic skeletal system morphogenesis	46.69	3.71E−10

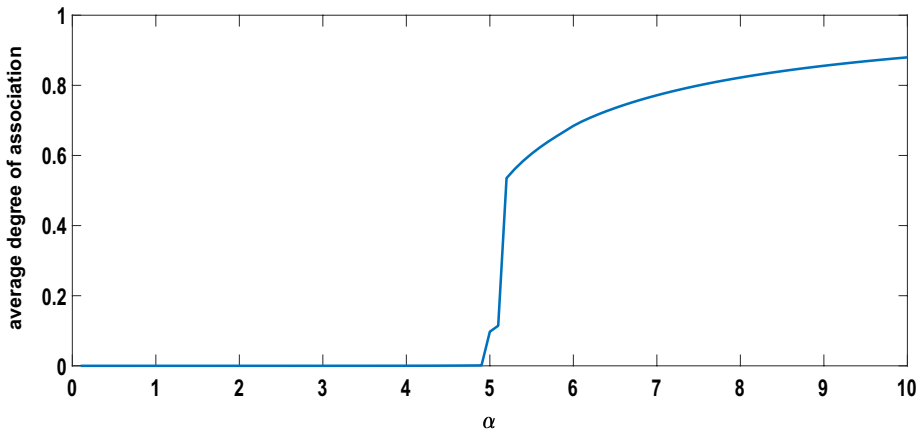

Fig. 3 The average degree of association among three cancers $\boldsymbol{B}_{ij}$ with the change of α when $K = 524$

Table 7 The convergent values of $\boldsymbol{B}_{ij}$ for COAD-KIPAN-KIRC

	COAD	KIPAN	KIRC
COAD	0	0.6362	0.5855
KIPAN	0.6362	0	0.5706
KIRC	0.5855	0.5706	0

Table 8 Overlapped genes shared by clusters among four cancers

	COAD	KIPAN	KIRC
COAD	0	1810	1798
KIPAN	1810	0	1799
KIRC	1798	1799	0

Table 9 Common clusters identified for COAD and KIPAN by our method

# Cluster	Size	N_{GO}	N_{KEGG}
1	28	4	4
2	634	109	17
3	8	3	0
4	10	3	0
5	105	37	8
6	94	86	18
7	7	0	0
8	8	0	0
9	16	1	1
10	18	2	1
11	6	0	0
12	11	0	1
13	11	3	1
14	71	10	3
15	25	0	0
16	647	82	33
17	22	1	1
18	52	21	0
19	8	0	0
20	24	7	4
21	6	0	0
22	21	6	1
23	16	4	4

α when $K = 524$. Similarly, when α close to 10, the degree of association increases. So we set α to be 10 in our experiment. The average value of $\boldsymbol{B}_{ij}$ with three different numbers of clusters K are shown in Table 7. Then the overlapped clusters with size larger than 5 among three different cancers are calculated for simplified analysis. What's more, the number of overlapped genes among three cancers are shown in Table 8. The consistency between the degree of association and the number of overlapped genes illustrates that the association identified by our method is reasonable.

4.2.1 Clusters analysis

For COAD and KIPAN, 23 clusters are shown in Table 9 are discoveried. Enrichment analysis for Gene Ontology (GO, biological process) and KEGG pathways are conducted for these clusters with DAVID (Huang et al. 2009). Within 23 clusters, 16 clusters enriched GO terms, 14 clusters enriched KEGG pathway. The enrichment results with P value less than $1E-6$ for GO categories are displayed in Table 10. It is apparent that Cluster #4, #6, and #16 significantly enriched GO terms.

Cluster #4 enriches one GO term which is related to acute-phase response, an acute inflammatory response that is mediated by several acute phase proteins whose concentrations in the plasma increase or decrease when people are infected or injured. Until now, evidence suggests acute phase proteins may have a profound impact on cancer growth, and it has tremendous potential as cancer biomarkers (Orr et al. 2011).

For Cluster #6, six GO terms are enriched significantly. Cell adhesion is indispensable to assemble individual cells into the three-dimensional tissue of animals, and it is an important unit for the intercellular communication in all kinds of tissues (Gumbiner 1996; Missler et al. 2012). Besides, the other three GO terms are also related to collagen and matrix, we already show that collagen is the dominant component in the extracellular matrix organization, and loss of extracellular matrix homeostasis would result in the proliferation of cancer cells.

Cluster #16 is related to immune response and many other responses belonging to the immune system. It seems that there are many responses occurring in the human body everyday to prevent the deterioration of cancer diseases. For instance, the generation of immune response would lead to the proliferation of antigen-specific lymphocytes, then it would result in antibodies up-regulation so that carcinogenesis can be better controlled (Adam et al. 2003).

4.2.2 Comparison with BLSC method

For comparison, we also conducted an experiment on COAD-KIPAN-KIRC using the BLSC method with the same K and α as our method. Enrichment analysis for Gene Ontology and KEGG pathways on the clusters of COAD and KIPAN are also conducted using DAVID, clusters enriched significantly are shown in Tables 11 and 12.

Table 10 Gene ontology enrichment of the clusters for COAD and KIPAN by our method

Cluster	Enriched GO terms	%	*P* value
Cluster 4	GO:0006953 acute-phase response	43.573	9.64E−7
Cluster 6	GO:0007155 cell adhesion	15.43	3.09E−12
	GO:0030574 collagen catabolic process	7.31	5.97E−10
	GO:0030198 extracellular matrix organization	9.75	1.87E−9
	GO:0001501 skeletal system development	8.12	1.56E−8
	GO:0006954 inflammatory response	11.37	2.21E−8
	GO:0030199 collagen fibril organization	5.68	2.58E−8
Cluster 16	GO:0006955 immune response	5.84	8.75E−14
	GO:0006954 inflammatory response	4.97	4.62E−11
	GO:0070098 chemokine-mediated signaling pathway	1.86	2.91E−8
	GO:0007204 positive regulation of cytosolic calcium ion concentration	2.49	3.06E−8
	GO:0007166 cell surface receptor signaling pathway	3.60	3.10E-8
	GO:0007165 signal transduction	8.70	1.12E−7
	GO:0008015 blood circulation	1.37	8.79E−7

Table 11 Common clusters identified for COAD and KIPAN by BLSC

#Cluster	Size	N_{GO}	N_{KEGG}	Minimal *P* value
A	589	197	28	1.14E−12
B	42	51	12	2.40E−10
C	18	3	1	5.54E−7
D	10	5	0	9.64E−7

Cluster #A and cluster #16 in Table 9 have 528 common genes, cluster #B and cluster #16 have 22 common genes, and genes in cluster #C are all contained in cluster #16. However, only a few GO terms are significantly enriched by clusters discovered by BLSC. Cluster #D enriched the same GO terms with the same *P* value as cluster #4 discovered by our method. So we can also conclude that the association among diseases would help us to obtain more accurate clustering results for multiple cancer diseases.

5 Concluding remarks

In this paper, we proposed a clustering method for multiple graphs with graphs association using a self-consistent field iteration method. In each iteration of our method, the consistency among graphs are updated based on first K eigenvectors containing the clustering structure information of each graph, then construct a new adjacency matrix. In terms of the experiment's results, the clustering performance of our methods is better than the BLSC method using two gene expression datasets. Besides, results on two gene expression datasets also demonstrate better clustering performance with the degree of association among graphs. Moreover, the convergence of a nonlinear eigenvalue problem in which the consistency weight depends on eigenvectors of the block adjacency matrix, has been studied in "Appendix". This research also opens up several avenues for future work. In our framework, all graphs should contain the same size and

Table 12 Gene ontology enrichment of the clusters for COAD and KIPAN by BLSC

Cluster	Enriched GO terms	%	*P* value
Cluster A	GO:0006955 immune response	5.77	1.14E−12
	GO:0006954 inflammatory response	5.37	2.90E−12
	GO:0007155 cell adhesion	5.63	7.16E−11
	GO:0007166 cell surface receptor signaling pathway	4.02	1.02E−9
	GO:0007204 positive regulation of cytosolic calcium ion concentration	2.82	1.14E−9
	GO:0070098 chemokine-mediated signaling pathway	1.74	6.29E−7
Cluster B	GO:0030198 extracellular matrix organization	15.61	2.40E−10
Cluster C	GO:0045926 negative regulation of growth	20.52	5.54E−7
	GO:0071294 cellular response to zinc ion	20.52	5.54E−7
Cluster D	GO:0006953 acute-phase response	42.87	9.64E−7

types of objects. However, in many applications, different types of objects with different size are connected with different types of relationships, handling such multiple graphs with different topological structures and semantic meanings is an important and challenging topic of future research.

Appendix: Convergence analysis

As we all know, self-consistent iteration (SCF) may not converge (Kouteckỳ and Bonačić 1971). According to Algorithm 1, in order to analyze convergence more easily, we are concerned about the following problem in which $\boldsymbol{B}$ is replaced by $\bar{\boldsymbol{B}}$.

$$\boldsymbol{A}(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M)\boldsymbol{X} = \boldsymbol{\Lambda}_K \boldsymbol{X} \quad \text{s.t.} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{I}_k \tag{7}$$

where $\boldsymbol{X} \in \mathbb{R}^{MN\times K}, \boldsymbol{A}(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M) \in \mathbb{R}^{MN\times MN}$ is a M-by-M block matrix, and the size of each block is N-by-N. $\boldsymbol{X}$ is K eigenvectors corresponding to the first K largest eigenvalues $\boldsymbol{\Lambda}_k$ of $\boldsymbol{A}(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M)$, and $\boldsymbol{X} = [\boldsymbol{X}_1^{\mathrm{T}}, \boldsymbol{X}_2^{\mathrm{T}}, \ldots, \boldsymbol{X}_M^{\mathrm{T}}]^{\mathrm{T}}$. The matrix $\boldsymbol{A}(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M)$ depends on eigenvectors $\boldsymbol{X}$. It is defined as:

$$\boldsymbol{A}(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M) = \mathbf{A} + \alpha\bar{\mathbf{B}} \otimes \boldsymbol{I}_N \tag{8}$$

where α is a known constant. And

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A}_2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{A}_M \end{bmatrix} \tag{9}$$

is an block diagonal matrix, and $\boldsymbol{A}_i$ is symmetric matrix. The dependency is expressed through a matrix $\bar{\boldsymbol{B}}$, defined as:

$$\bar{\boldsymbol{B}} = \bar{\boldsymbol{W}} \odot \begin{bmatrix} 0 & |\mathrm{Tr}(\boldsymbol{X}_1\boldsymbol{X}_2^{\mathrm{T}})| & \ldots & |\mathrm{Tr}(\boldsymbol{X}_1\boldsymbol{X}_M^{\mathrm{T}})| \\ |\mathrm{Tr}(\boldsymbol{X}_2\boldsymbol{X}_1^{\mathrm{T}})| & 0 & \ldots & |\mathrm{Tr}(\boldsymbol{X}_2\boldsymbol{X}_M^{\mathrm{T}})| \\ \vdots & \vdots & \ddots & \vdots \\ |\mathrm{Tr}(\boldsymbol{X}_M\boldsymbol{X}_1^{\mathrm{T}})| & |\mathrm{Tr}(\boldsymbol{X}_M\boldsymbol{X}_2^{\mathrm{T}})| & \ldots & 0 \end{bmatrix} \tag{10}$$

$\bar{\mathbf{B}}$ is a M-by-M matrix, and $\bar{\mathbf{B}}_{ij} = \bar{w}_{ij}|\mathrm{Tr}(\boldsymbol{X}_i\boldsymbol{X}_j^{\mathrm{T}})|$ for $i \neq j$, $\bar{\boldsymbol{W}}$ is:

$$\bar{\boldsymbol{W}} = \bar{\boldsymbol{W}}^{\mathrm{T}} = \begin{bmatrix} 0 & \bar{w}_{12} & \ldots & \bar{w}_{1M} \\ \bar{w}_{21} & 0 & \ldots & \bar{w}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{w}_{M1} & \bar{w}_{M2} & \ldots & 0 \end{bmatrix} \tag{11}$$

Assume the M-by-M block matrix $\boldsymbol{A}(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M)$ is:

$$\boldsymbol{A}(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M) = \begin{bmatrix} \boldsymbol{A}_1 & \ldots & \alpha|\mathrm{Tr}(\boldsymbol{X}_1\boldsymbol{X}_M^{\mathrm{T}})|\bar{w}_{1M}\boldsymbol{I}_N \\ \alpha|\mathrm{Tr}(\boldsymbol{X}_2\boldsymbol{X}_1^{\mathrm{T}})|\bar{w}_{21}\boldsymbol{I}_N & \ldots & \alpha|\mathrm{Tr}(\boldsymbol{X}_1\boldsymbol{X}_M^{\mathrm{T}})|\bar{w}_{2M}\boldsymbol{I}_N \\ \vdots & \ddots & \vdots \\ \alpha|\mathrm{Tr}(\boldsymbol{X}_M\boldsymbol{X}_1^{\mathrm{T}})|\bar{w}_{M1}\boldsymbol{I}_N & \ldots & \boldsymbol{A}_M \end{bmatrix}$$

Here notation Tr(.) means trace of a matrix, and |.| means taking the absolute value.

When given $\boldsymbol{X}$ with constrain $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{I}_K$, then the eigenvectors of $\boldsymbol{A}(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M)$ corresponding to the first K largest eigenvalues is $\boldsymbol{Y} = [\boldsymbol{Y}_1^{\mathrm{T}}, \boldsymbol{Y}_2^{\mathrm{T}}, \ldots, \boldsymbol{Y}_M^{\mathrm{T}}]^{\mathrm{T}}, \boldsymbol{Y} \in \mathbb{R}^{MN\times k}$, $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_M \in \mathbb{R}^{N\times k}$. Then these new eigenvectors would be used to construct a new block adjacency matrix. In the following analysis, $\boldsymbol{A}(\boldsymbol{X})$ is used to represent $\boldsymbol{A}(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M)$. The changing term of $\boldsymbol{A}(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M)$ during iterations is

$$\bar{\boldsymbol{\Gamma}} = \begin{bmatrix} 0 & |\mathrm{Tr}(\boldsymbol{X}_1\boldsymbol{X}_2^{\mathrm{T}})| & \ldots & |\mathrm{Tr}(\boldsymbol{X}_1\boldsymbol{X}_M^{\mathrm{T}})| \\ |\mathrm{Tr}(\boldsymbol{X}_2\boldsymbol{X}_1^{\mathrm{T}})| & 0 & \ldots & |\mathrm{Tr}(\boldsymbol{X}_2\boldsymbol{X}_M^{\mathrm{T}})| \\ \vdots & \vdots & \ddots & \vdots \\ |\mathrm{Tr}(\boldsymbol{X}_M\boldsymbol{X}_1^{\mathrm{T}})| & |\mathrm{Tr}(\boldsymbol{X}_M\boldsymbol{X}_2^{\mathrm{T}})| & \ldots & 0 \end{bmatrix} \tag{12}$$

Suppose the eigenvalues of $\boldsymbol{A}(\boldsymbol{X})$ are

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > \lambda_{k+1} \geq \cdots \geq \lambda_{MN}$$

Since $\lambda_k > \lambda_{k+1}$, a filter function $\bar{\phi}(\lambda)$ that satisfies

$$\bar{\phi}(\lambda) = \begin{cases} 0 & for\ \lambda = \lambda_{k+1}, \ldots, \lambda_{MN} \\ 1 & for\ \lambda = \lambda_1, \lambda_2, \ldots, \lambda_{k-1}, \lambda_k \end{cases} \tag{13}$$

can be constructed, then

$$\bar{\phi}(\boldsymbol{A}(\boldsymbol{X})) = \boldsymbol{Y}\bar{\phi}(\boldsymbol{\Lambda})\boldsymbol{Y}^{\mathrm{T}} = \boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{Y}_1\boldsymbol{Y}_1^{\mathrm{T}} & \boldsymbol{Y}_1\boldsymbol{Y}_2^{\mathrm{T}} & \ldots & \boldsymbol{Y}_1\boldsymbol{Y}_M^{\mathrm{T}} \\ \boldsymbol{Y}_2\boldsymbol{Y}_1^{\mathrm{T}} & \boldsymbol{Y}_2\boldsymbol{Y}_2^{\mathrm{T}} & \ldots & \boldsymbol{Y}_2\boldsymbol{Y}_M^{\mathrm{T}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{Y}_M\boldsymbol{Y}_1^{\mathrm{T}} & \boldsymbol{Y}_M\boldsymbol{Y}_2^{\mathrm{T}} & \ldots & \boldsymbol{Y}_M\boldsymbol{Y}_M^{\mathrm{T}} \end{bmatrix}$$

Here $\boldsymbol{\Lambda}$ denotes the eigenvalue matrix of $\boldsymbol{A}(\boldsymbol{X})$, then the changing term can be represented as

$$\bar{\boldsymbol{\Gamma}} = \mathrm{abs}(\mathrm{Tr}_2(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}})) \odot \boldsymbol{M}$$

Here abs(.) means taking the absolute value on each element of the matrix. $\mathrm{Tr}_2(.)$ means to take a trace for each block in a block matrix. For example, if $\boldsymbol{T}$ is a M-by-M block matrix, each block is a N-by-N matrix.

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{T}_{11} & \boldsymbol{T}_{12} & \ldots & \boldsymbol{T}_{1M} \\ \boldsymbol{T}_{21} & \boldsymbol{T}_{22} & \ldots & \boldsymbol{T}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{T}_{M1} & \boldsymbol{T}_{M2} & \ldots & \boldsymbol{T}_{MM} \end{bmatrix}$$

Then

$$\mathrm{Tr}_2(\boldsymbol{T}) = \begin{bmatrix} \mathrm{Tr}(\boldsymbol{T}_{11}) & \mathrm{Tr}(\boldsymbol{T}_{12}) & \dots & \mathrm{Tr}(\boldsymbol{T}_{1M}) \\ \mathrm{Tr}(\boldsymbol{T}_{21}) & \mathrm{Tr}(\boldsymbol{T}_{22}) & \dots & \mathrm{Tr}(\boldsymbol{T}_{2M}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Tr}(\boldsymbol{T}_{M1}) & \mathrm{Tr}(\boldsymbol{T}_{M2}) & \dots & \mathrm{Tr}(\boldsymbol{T}_{MM}) \end{bmatrix}$$

So $\boldsymbol{A}(\boldsymbol{X})$ can be constructed from $\bar{\boldsymbol{\Gamma}}$. It is

$$\boldsymbol{A}(\bar{\boldsymbol{\Gamma}}) = \boldsymbol{A} + \alpha\bar{\boldsymbol{\Gamma}} \odot \boldsymbol{W} \otimes \boldsymbol{I}_N$$

If $\boldsymbol{A}(\boldsymbol{X})$ is constructed from the $\bar{\boldsymbol{\Gamma}}_{in}$, then $\bar{\boldsymbol{\Gamma}}_{out}$ can be represented as a function of $\bar{\boldsymbol{\Gamma}}_{in}$.

$$\bar{\boldsymbol{\Gamma}}_{out} = \mathrm{abs}(\mathrm{Tr}_2(\bar{\phi}(\boldsymbol{A}(\bar{\boldsymbol{\Gamma}}_{in})))) \odot \boldsymbol{M}$$

So there is a mapping F from $\bar{\boldsymbol{\Gamma}}_{in}$ to $\bar{\boldsymbol{\Gamma}}_{out}$. In order to show the convergence of (7), we need to seek a condition such that F becomes a contraction, it is

$$\|F(\bar{\boldsymbol{\Gamma}}_1) - F(\bar{\boldsymbol{\Gamma}}_2)\|_1 < \rho\|\bar{\boldsymbol{\Gamma}}_1 - \bar{\boldsymbol{\Gamma}}_2\|_1 \tag{14}$$

For function $\bar{\phi}(t)$, we define $\bar{\phi}(t)$ as Fermi-Dirac distribution, it is

$$\bar{\phi}(t) = \bar{f}_\mu(t) = \frac{1}{1 + e^{\beta(t-\mu)}}$$

Here μ is decided by the input matrix of $\bar{\phi}(t)$. $\beta > 0$ is a positive constant. μ is the solution of the following problem

$$\mathrm{trace}(\bar{\phi}(\boldsymbol{A})) = \mathrm{trace}(\bar{f}_\mu(\boldsymbol{A})) = K \tag{15}$$

According to Yang et al. (2009), because $\sum_{i=1}^{n} f_\mu(\lambda_i)$ is monotonic with respect to μ for a fixed β, the solution to Eq. (15) is unique for any choice of β and $\boldsymbol{A}$. Moreover, Yang et al. (2009) already shows that a larger β results in a sharper drop-off of $\bar{\phi}(t)$ from 1 to 0 and a large enough constant β can make Eq. (13) satisfied in finite precision arithmetic if the UWP condition holds.

Lemma 1 *Let* $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{M\times M}$, *then*

$$\|\mathrm{abs}(\boldsymbol{B}) - \mathrm{abs}(\boldsymbol{C})\|_1 \le \|\boldsymbol{B} - \boldsymbol{C}\|_1$$

hold.

Proof If $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{M\times M}$, then

$$\mathrm{abs}(\boldsymbol{B}) = \begin{bmatrix} |b_{11}| & |b_{12}| & \dots & |b_{1M}| \\ |b_{21}| & |b_{22}| & \dots & |b_{2M}| \\ \vdots & \vdots & \ddots & \vdots \\ |b_{M1}| & |b_{M2}| & \dots & |b_{MM}| \end{bmatrix}$$

$$\mathrm{abs}(\boldsymbol{C}) = \begin{bmatrix} |c_{11}| & |c_{12}| & \dots & |c_{1M}| \\ |c_{21}| & |c_{22}| & \dots & |c_{2M}| \\ \vdots & \vdots & \ddots & \vdots \\ |c_{M1}| & |c_{M2}| & \dots & |c_{MM}| \end{bmatrix}$$

furthermore,

$$\begin{aligned} &\|\mathrm{abs}(\boldsymbol{B}) - \mathrm{abs}(\boldsymbol{C})\|_1 \\ =&\max_j(|\,|b_{1j}| - |c_{1j}|\,| + |\,|b_{2j}| - |c_{2j}|\,| + \dots \\ &+ |\,|b_{Mj}| - |c_{Mj}|\,|) \\ \le&\max_j(|\,b_{1j} - c_{1j}\,| + |\,b_{2j} - c_{2j}\,| + \dots \\ &+ |\,b_{Mj} - c_{Mj}\,|) \\ =&\|\boldsymbol{B} - \boldsymbol{C}\|_1 \end{aligned}$$

so the Lemma 1 holds. □

Assume that $\boldsymbol{A_1}$ and $\boldsymbol{A_2}$ are symmetric matrix created from the changing term $\bar{\boldsymbol{\Gamma}}_1$ and $\bar{\boldsymbol{\Gamma}}_2$ respectively. If $\mu_1 > \mu_2$, then $\bar{f}_{\mu_1}(t) \ge \bar{f}_{\mu_2}(t)$ for any t. According to Lemma 1. We can show that

$$\begin{aligned} &\|F(\bar{\boldsymbol{\Gamma}}_1) - F(\bar{\boldsymbol{\Gamma}}_2)\|_1 \\ =&\|\mathrm{abs}(\mathrm{Tr}_2(\bar{f}_{\mu_1}(\boldsymbol{A_1}))) \odot \boldsymbol{M} \\ &-\mathrm{abs}(\mathrm{Tr}_2(\bar{f}_{\mu_2}(\boldsymbol{A_2}))) \odot \boldsymbol{M}\|_1 \\ \le&(M-1)\|\mathrm{abs}(\mathrm{Tr}_2(\bar{f}_{\mu_1}(\boldsymbol{A_1}))) \\ &-\mathrm{abs}(\mathrm{Tr}_2(\bar{f}_{\mu_2}(\boldsymbol{A_2})))\|_1 \\ \le&(M-1)\|\mathrm{Tr}_2(\bar{f}_{\mu_1}(\boldsymbol{A_1})) - \mathrm{Tr}_2(\bar{f}_{\mu_2}(\boldsymbol{A_2}))\|_1 \\ \le&(M-1)[\|\mathrm{Tr}_2(\bar{f}_{\mu_1}(\boldsymbol{A_1})) - \mathrm{Tr}_2(\bar{f}_{\mu_2}(\boldsymbol{A_1}))\|_1 \\ &+\|\mathrm{Tr}_2(\bar{f}_{\mu_2}(\boldsymbol{A_1})) - \mathrm{Tr}_2(\bar{f}_{\mu_2}(\boldsymbol{A_2}))\|_1] \end{aligned} \tag{16}$$

Lemma 2 *Let* $\bar{\mathbf{X}} \in \mathbb{R}^{M\times M}$ *be a symmetric positive semidefinite matrix, then the inequality*

$$\|\bar{\mathbf{X}}\|_1 \le \sqrt{M}\mathrm{Tr}(\bar{\mathbf{X}})$$

holds.

Proof If $\bar{\mathbf{X}} \in \mathbb{R}^{M\times M}$ is a symmetric positive semidefinite matrix, define the Schatten p-norm of matrix $\bar{\mathbf{X}}$ as $\|\bar{\mathbf{X}}\|_{S_p}$. It is easy to know that

$$\begin{aligned} \|\bar{\mathbf{X}}\|_1 \le&\sqrt{M}\|\bar{\mathbf{X}}\|_2 \le \sqrt{M}\|\bar{\mathbf{X}}\|_F \\ =&\sqrt{M}\|\bar{\mathbf{X}}\|_{S_2} \le \sqrt{M}\|\bar{\mathbf{X}}\|_{S_1} = \sqrt{M}\mathrm{Tr}(\bar{\mathbf{X}}) \end{aligned}$$

So the Lemma 2 can be proved. □

Lemma 3 *If* $\bar{\boldsymbol{A}}$ *is a symmetric positive semidefinite matrix, then* $\mathrm{Tr}_2(\bar{\boldsymbol{A}})$ *is also a symmetric positive semidefinite matrix.*

Proof $\bar{\boldsymbol{A}}$ is a symmetric positive semidefinite matrix. So it can be decomposed as $\bar{\boldsymbol{A}} = \boldsymbol{U}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{U}_{\boldsymbol{A}}^{\mathbf{T}}$ with nonnegative eigenvalue $\boldsymbol{\Lambda}_{\boldsymbol{A}}$ and its corresponding eigenvectors $\boldsymbol{U}_{\boldsymbol{A}}$. $\boldsymbol{U}_{\boldsymbol{A}} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_{MN}] \in \mathbb{R}^{MN\times MN}$, $\boldsymbol{u}_j\,(1 \le j \le MN) \in \mathbb{R}^{MN\times 1}$, $\boldsymbol{\Lambda}_{\boldsymbol{A}} = \mathrm{Diag}([\lambda_1, \lambda_2, \dots, \lambda_{MN}])$, $\lambda_1, \lambda_2, \dots, \lambda_{MN} \ge 0$, $Diag(x)$

means a diagonal matrix with x on its diagonal, so $\bar{\boldsymbol{A}} = \lambda_1 \boldsymbol{u}_1 \boldsymbol{u}_1^{\mathrm{T}} + \cdots + \lambda_{MN} \boldsymbol{u}_{MN} \boldsymbol{u}_{MN}^{\mathrm{T}}$.

$$\begin{aligned}\mathrm{Tr}_2(\bar{\boldsymbol{A}}) &= \mathrm{Tr}_2(\lambda_1 \boldsymbol{u}_1 \boldsymbol{u}_1^{\mathrm{T}} + \cdots + \lambda_{MN} \boldsymbol{u}_{MN} \boldsymbol{u}_{MN}^{\mathrm{T}}) \\ &= \lambda_1 \mathrm{Tr}_2(\boldsymbol{u}_1 \boldsymbol{u}_1^{\mathrm{T}}) + \lambda_2 \mathrm{Tr}_2(\boldsymbol{u}_2 \boldsymbol{u}_2^{\mathrm{T}}) \\ &\quad + \cdots + \lambda_{MN} \mathrm{Tr}_2(\boldsymbol{u}_{MN} \boldsymbol{u}_{MN}^{\mathrm{T}})\end{aligned}$$

For any vector $\boldsymbol{v} = [v_1, v_2, \ldots, v_M]^{\mathrm{T}} \in \mathbb{R}^M$, then

$$\begin{aligned}\boldsymbol{v}^{\mathrm{T}} \mathrm{Tr}_2(\bar{\boldsymbol{A}}) \boldsymbol{v} =& \lambda_1 \boldsymbol{v}^{\mathrm{T}} \mathrm{Tr}_2(\boldsymbol{u}_1 \boldsymbol{u}_1^{\mathrm{T}}) \boldsymbol{v} + \lambda_2 \boldsymbol{v}^{\mathrm{T}} \mathrm{Tr}_2(\boldsymbol{u}_2 \boldsymbol{u}_2^{\mathrm{T}}) \boldsymbol{v} \\ &+ \cdots + \lambda_{MN} \boldsymbol{v}^{\mathrm{T}} \mathrm{Tr}_2(\boldsymbol{u}_{MN} \boldsymbol{u}_{MN}^{\mathrm{T}}) \boldsymbol{v}\end{aligned}$$

for $1 \le j \le MN$, let $\boldsymbol{u}_j = [(\boldsymbol{u}_j^1)^{\mathrm{T}}, (\boldsymbol{u}_j^2)^{\mathrm{T}}, \ldots, (\boldsymbol{u}_j^M)^{\mathrm{T}}]^{\mathrm{T}}$, $\boldsymbol{u}_j^i \in \mathbb{R}^{N \times 1}$, then

$$\begin{aligned}\boldsymbol{v}^{\mathrm{T}} \mathrm{Tr}_2(\boldsymbol{u}_j \boldsymbol{u}_j^{\mathrm{T}}) \boldsymbol{v} &= \sum_{i=1}^{M} \sum_{t=1}^{M} \boldsymbol{v}_i \mathrm{Tr}(\boldsymbol{u}_j^i (\boldsymbol{u}_j^t)^{\mathrm{T}}) \boldsymbol{v}_t \\ &= \sum_{i=1}^{M} \sum_{t=1}^{M} \boldsymbol{v}_i (\boldsymbol{u}_j^i)^{\mathrm{T}} \boldsymbol{u}_j^{\mathrm{T}} \boldsymbol{v}_j \\ &= (\boldsymbol{v}_1 \boldsymbol{u}_j^1 + \boldsymbol{v}_2 \boldsymbol{u}_j^2 + \cdots + \boldsymbol{v}_M \boldsymbol{u}_j^M)^{\mathrm{T}} \\ &\quad (\boldsymbol{v}_1 \boldsymbol{u}_j^1 + \boldsymbol{v}_2 \boldsymbol{u}_j^2 + \cdots + \boldsymbol{v}_M \boldsymbol{u}_j^M) \\ &\ge 0\end{aligned}$$

thus $\lambda_1, \lambda_2, \ldots, \lambda_{MN} \ge 0$, so $\boldsymbol{v}^{\mathrm{T}} \mathrm{Tr}_2(\bar{\boldsymbol{A}}) \boldsymbol{v} \ge 0$ for any $\boldsymbol{v} \in \mathbb{R}^M$. Therefore, $\mathrm{Tr}_2(\bar{\boldsymbol{A}})$ is a symmetric positive semidefinite matrix. The Lemma holds. □

According to Lemmas 2 and 3:

$$\begin{aligned}&\|F(\bar{\boldsymbol{\Gamma}}_1) - F(\bar{\boldsymbol{\Gamma}}_2)\|_1 \\ \le& (M-1)MN\sqrt{M} \|\bar{f}_{\mu_2}(\boldsymbol{A_1})) - \bar{f}_{\mu_2}(\boldsymbol{A_2})\|_1 \\ &+ N \|\bar{f}_{\mu_2}(\boldsymbol{A_1})) - \bar{f}_{\mu_2}(\boldsymbol{A_2})\|_1\end{aligned} \tag{17}$$

According to LEMMA 4 in Yang et al. (2009), suppose that $\boldsymbol{A_1} = \boldsymbol{X_1} \boldsymbol{\Lambda_1} \boldsymbol{X_1}^{\mathrm{T}}$ and $\boldsymbol{A_2} = \boldsymbol{X_2} \boldsymbol{\Lambda_2} \boldsymbol{X_2}^{\mathrm{T}}$ are the spectral decompositions of $\boldsymbol{A_1}$ and $\boldsymbol{A_2}$ respectively, then

$$\begin{aligned}&\|\bar{f}_{\mu_2}(\boldsymbol{A_2}) - \bar{f}_{\mu_2}(\boldsymbol{A_1})\|_1 \\ \le& \alpha M^2 N^2 \|\hat{\boldsymbol{C}}\|_1 \|\boldsymbol{W}\|_1 \|\bar{\boldsymbol{\Gamma}}_2 - \bar{\boldsymbol{\Gamma}}_1\|_1\end{aligned} \tag{18}$$

Using the mean value theorem and the fact that

$$|f'_{\mu}(t)| \le \frac{\beta}{4}$$

we can show that

$$\max_{j,k} |\hat{\boldsymbol{C}_{j,k}}| \le \frac{\beta}{4}$$

Therefore

$$\|\hat{\boldsymbol{C}}\|_1 \le \frac{MN\beta}{4} \tag{19}$$

Combining Eq. (17), Eq. (18) and Eq. (19). we obtain

$$\begin{aligned}&\|F(\boldsymbol{\Gamma}_1) - F(\boldsymbol{\Gamma}_2)\|_1 \\ \le& \frac{\alpha M^3 N^4 (M-1)(M\sqrt{M}+1)\beta \|\boldsymbol{W}\|_1 \|\boldsymbol{\Gamma}_2 - \boldsymbol{\Gamma}_1\|_1}{4}\end{aligned}$$

If α satisfies

$$\alpha \le \frac{4}{M^3 N^4 (M-1)(M\sqrt{M}+1)\beta \|\boldsymbol{W}\|_1} \tag{20}$$

We can conclude that F is a contraction, then the convergence of problem Eq. (7) is guaranteed.

Acknowledgements M. Ng's work was supported by HKRGC GRF 12300218, 12300519, 17201020 and 17300021.

References

Acharya A, Das I, Chandhok D et al (2010) Redox regulation in cancer: a double-edged sword with therapeutic potential. Oxidative Med Cell Longev 3(1):23–34. https://doi.org/10.4161/oxim.3.1.10095

Adam JK, Odhav B, Bhoola KD (2003) Immune responses in cancer. Pharmacol Ther 99(1):113–132. https://doi.org/10.1016/s0163-7258(03)00056-1

Alper O, Bergmann-Leitner ES, Bennett TA et al (2001) Epidermal growth factor receptor signaling and the invasive phenotype of ovarian carcinoma cells. J Natl Cancer Inst 93(18):1375–1384. https://doi.org/10.1093/jnci/93.18.1375

Carbone DP, Gandara DR, Antonia SJ et al (2015) Non-small-cell lung cancer: role of the immune system and potential for immunotherapy. J Thorac Oncol 10(7):974–984. https://doi.org/10.1097/JTO.0000000000000551

Charbonneau B, Goode EL, Kalli KR et al (2013) The immune system in the pathogenesis of ovarian cancer. Crit Rev Immunol. https://doi.org/10.1615/critrevimmunol.2013006813

Chen B, Wang J, Li M et al (2014) Identifying disease genes by integrating multiple data sources. BMC Med Genom 7(2):1–12. https://doi.org/10.1186/1755-8794-7-S2-S2

Chen C, Ng MK, Zhang S (2017) Block spectral clustering methods for multiple graphs. Numer Linear Algebra Appl 24(1):e2075. https://doi.org/10.1002/nla.2075

Cheng W, Guo Z, Zhang X et al (2016) Cgc: a flexible and robust approach to integrating co-regularized multi-domain graph for clustering. ACM Trans Knowl Discov Data (TKDD) 10(4):1–27. https://doi.org/10.1145/2903147

Cho A, Howell VM, Colvin EK (2015) The extracellular matrix in epithelial ovarian cancer: a piece of a puzzle. Front Oncol 5:245. https://doi.org/10.3389/fonc.2015.00245

Daley WP, Peters SB, Larsen M (2008) Extracellular matrix dynamics in development and regenerative medicine. J Cell Sci 121(3):255–264. https://doi.org/10.1242/jcs.006064

Domagala-Kulawik J (2015) The role of the immune system in non-small cell lung carcinoma and potential for therapeutic intervention. Transl Lung Cancer Res 4(2):177. https://doi.org/10.3978/j.issn.2218-6751.2015.01.11

Domagala-Kulawik J, Osinska I, Hoser G (2014) Mechanisms of immune response regulation in lung cancer. Transl Lung Cancer Res 3(1):15. https://doi.org/10.3978/j.issn.2218-6751.2013.11.03

Dong X, Frossard P, Vandergheynst P et al (2012) Clustering with multi-layer graphs: a spectral perspective. IEEE Trans Signal Process 60(11):5820–5831. https://doi.org/10.1109/TSP.2012.2212886

Dong P, Kaneuchi M, Konno Y et al (2013) Emerging therapeutic biomarkers in endometrial cancer. BioMed Res Int. https://doi.org/10.1155/2013/130362

Fang M, Yuan J, Peng C et al (2014) Collagen as a double-edged sword in tumor progression. Tumor Biol 35(4):2871–2882. https://doi.org/10.1007/s13277-013-1511-7

Gill N, Singh S, Aseri TC (2014) Computational disease gene prioritization: an appraisal. J Comput Biol 21(6):456–465. https://doi.org/10.1089/cmb.2013.0158

Goh KI, Cusick ME, Valle D et al (2007) The human disease network. Proc Natl Acad Sci 104(21):8685–8690. https://doi.org/10.1073/pnas.0701361104

Gomes M, Teixeira AL, Coelho A et al (2014) The role of inflammation in lung cancer. Inflamm Cancer. https://doi.org/10.1007/978-3-0348-0837-8_1

Granados ML, Hudson LG, Samudio-Ruiz SL (2015) Contributions of the epidermal growth factor receptor to acquisition of platinum resistance in ovarian cancer cells. PLoS One 10(9):e0136893. https://doi.org/10.1371/journal.pone.0136893

Gumbiner BM (1996) Cell adhesion: the molecular basis of tissue architecture and morphogenesis. Cell 84(3):345–357. https://doi.org/10.1016/s0092-8674(00)81279-9

Hegedűs C, Kovács K, Polgár Z et al (2018) Redox control of cancer cell destruction. Redox Biol 16:59–74. https://doi.org/10.1016/j.redox.2018.01.015

Hong CS, Jeong O, Piao Z et al (2015) Hoxb5 induces invasion and migration through direct transcriptional up-regulation of β-catenin in human gastric carcinoma. Biochem J 472(3):393–403. https://doi.org/10.1042/BJ20150213

Hu K, Hu JB (2018) Tang L et al Predicting disease-related genes by path structure and community structure in protein–protein networks. J Stat Mech Theory Exp 10:100001. https://doi.org/10.1088/1742-5468/aae02b

Hu H, Yan X, Huang Y et al (2005) Mining coherent dense subgraphs across massive biological networks for functional discovery. Bioinformatics 21(suppl 1):i213–i221. https://doi.org/10.1093/bioinformatics/bti1049

Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37(1):1–13. https://doi.org/10.1093/nar/gkn923

Hudson LG, Zeineldin R, Silberberg M et al (2009) Activated epidermal growth factor receptor in ovarian cancer. Ovarian Cancer. https://doi.org/10.1016/j.canep.2012.06.005

Ivashkiv LB, Donlin LT (2014) Regulation of type i interferon responses. Nat Rev Immunol 14(1):36–49. https://doi.org/10.1038/nri3581

Jin K, Sukumar (2016) Hox genes: major actors in resistance to selective endocrine response modifiers. Biochimica et Biophysica Acta (BBA) Rev Cancer 1865(2):105–110. https://doi.org/10.1016/j.bbcan.2016.01.003

Kachgal S, Mace KA, Boudreau NJ (2012) The dual roles of homeobox genes in vascularization and wound healing. Cell Adhes Migr 6(6):457–470. https://doi.org/10.4161/cam.22164

Kouteckỳ J, Bonačić V (1971) On convergence difficulties in the iterative hartree-fock procedure. J Chem Phys 55(5):2408–2413. https://doi.org/10.1063/1.1676424

Kumar A, Rai P, Daume H (2011) Co-regularized multi-view spectral clustering. Adv Neural Inf Process Syst 24:1413–1421. https://doi.org/10.5555/2986459.2986617

Kumar A, Daumé H (2011) A co-training approach for multi-view spectral clustering. In: Proceedings of the 28th international conference on machine learning (ICML-11), Citeseer, pp 393–400. https://doi.org/10.5555/3104482.3104532

Li M, Zhang J, Liu Q et al (2014) Prediction of disease-related genes based on weighted tissue-specific networks by using DNA methylation. BMC Med Genom 7(2):1–8. https://doi.org/10.1186/1755-8794-7-S2-S4

Li N, Jia X, Wang J et al (2015) Knockdown of homeobox a5 by small hairpin RNA inhibits proliferation and enhances cytarabine chemosensitivity of acute myeloid leukemia cells. Mol Med Rep 12(5):6861–6866. https://doi.org/10.3892/mmr.2015.4331

Lim SB, Tan SJ, Wan-Teck L et al (2017) An extracellular matrix-related prognostic and predictive indicator for early-stage non-small cell lung cancer. Nat Commun 8(1):1–11. https://doi.org/10.1038/s41467-017-01430-6

Liu X, Ji S, Glänzel W et al (2012) Multiview partitioning via tensor methods. IEEE Trans Knowl Data Eng 25(5):1056–1069. https://doi.org/10.1109/TKDE.2012.95

Liu S, Zhang H, Duan E (2013) Epidermal development in mammals: key regulators, signals from beneath, and stem cells. Int J Mol Sci 14(6):10869–10895. https://doi.org/10.3390/ijms140610869

Liu Y, Ng MK, Wu S (2018) Multi-domain networks association for biological data using block signed graph clustering. IEEE/ACM Trans Comput Biol Bioinform 17(2):435–448. https://doi.org/10.1109/TCBB.2018.2848904

Lu C, Klement JD, Ibrahim ML et al (2019) Type i interferon suppresses tumor growth through activating the stat3-granzyme b pathway in tumor-infiltrating cytotoxic t lymphocytes. J Immunother Cancer 7(1):1–11. https://doi.org/10.1186/s40425-019-0635-8

Missler M, Südhof TC, Biederer T (2012) Synaptic cell adhesion. Cold Spring Harbor Perspect Biol 4(4):a005694. https://doi.org/10.1101/cshperspect.a005694

Morgan R, Pirard PM, Shears L et al (2007) Antagonism of hox/pbx dimer formation blocks the in vivo proliferation of melanoma. Cancer Res 67(12):5806–5813. https://doi.org/10.1158/0008-5472.CAN-06-4231

Nadiarnykh O, LaComb RB, Brewer MA et al (2010) Alterations of the extracellular matrix in ovarian cancer studied by second harmonic generation imaging microscopy. BMC Cancer 10(1):1–14. https://doi.org/10.1186/1471-2407-10-94

Nagane M, Levitzki A, Gazit A et al (1998) Drug resistance of human glioblastoma cells conferred by a tumor-specific mutant epidermal growth factor receptor through modulation of bcl-xl and caspase-3-like proteases. Proc Natl Acad Sci 95(10):5724–5729. https://doi.org/10.1073/pnas.95.10.5724

Ness RB, Cottreau C (1999) Possible role of ovarian epithelial inflammation in ovarian cancer. J Natl Cancer Inst 91(17):1459–1467. https://doi.org/10.1093/jnci/91.17.1459

Nishimura T, Nakamura K, Yamashita S et al (2015) Effect of the molecular targeted drug, erlotinib, against endometrial cancer expressing high levels of epidermal growth factor receptor. BMC Cancer 15(1):1–11. https://doi.org/10.1186/s12885-015-1975-5

Ni J, Tong H, Fan W, et al (2015) Flexible and robust multi-network clustering. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp 835–844. https://doi.org/10.1145/2783258.2783262

Orr WS, Malkas LH, Hickey RJ, et al (2011) Acute phase proteins as cancer biomarkers. Acute phase proteins as early non-specific

biomarkers of human and veterinary diseases, 408. https://doi.org/10.5772/25181
Oti M, Snel B, Huynen MA et al (2006) Predicting disease genes using protein–protein interactions. J Med Genet 43(8):691–698. https://doi.org/10.1136/jmg.2006.041376
Pegram MD, Konecny G, Slamon DJ (2000) The molecular and cellular biology of her2/neu gene amplification/overexpression and the clinical development of herceptin (trastuzumab) therapy for breast cancer. Adv Breast Cancer Manag. https://doi.org/10.1007/978-1-4757-3147-7_4
Provenzano PP, Eliceiri KW, Campbell JM et al (2006) Collagen reorganization at the tumor-stromal interface facilitates local invasion. BMC Med 4(1):1–15. https://doi.org/10.1186/1741-7015-4-38
Shen C, Pan J, Zhang S et al (2015) Multiple networks modules identification by a multi-dimensional Markov chain method. Netw Model Anal Health Inform Bioinform 4(1):1–13. https://doi.org/10.1007/s13721-015-0106-1
Shiga M, Mamitsuka H (2010) A variational bayesian framework for clustering with multiple graphs. IEEE Trans Knowl Data Eng 24(4):577–590. https://doi.org/10.1109/TKDE.2010.272
Siwak DR, Carey M, Hennessy BT et al (2010) Targeting the epidermal growth factor receptor in epithelial ovarian cancer: current knowledge and future challenges. J Oncol. https://doi.org/10.1155/2010/568938
Sun PG, Gao L, Han S (2011) Prediction of human disease-related gene clusters by clustering analysis. Int J Biol Sci 7(1):61. https://doi.org/10.7150/ijbs.7.61
Sun Y, Norick B, Han J et al (2013) Pathselclus: integrating meta-path selection with user-guided object clustering in heterogeneous information networks. ACM Trans Knowl Discov Data (TKDD) 7(3):1–23. https://doi.org/10.1145/2500492
Tang W, Lu Z, Dhillon IS (2009) Clustering with multiple graphs. In: 2009 Ninth IEEE international conference on data mining. IEEE, pp 1016–1021. https://doi.org/10.1109/ICDM.2009.125
Thorn CF, Oshiro C, Marsh S et al (2011) Doxorubicin pathways: pharmacodynamics and adverse effects. Pharmacogenet Genom 21(7):440. https://doi.org/10.1097/FPC.0b013e32833ffb56
Tsai CM, Levitzki A, Wu LH et al (1996) Enhancement of chemosensitivity by tyrphostin ag825 in high-p185neu expressing non-small cell lung cancer cells. Cancer Res 56(5):1068–1074
Von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416. https://doi.org/10.1007/s11222-007-9033-z
Walker C, Mojares E, del Río Hernández A (2018) Role of extracellular matrix in development and cancer progression. Int J Mol Sci 19(10):3028. https://doi.org/10.3390/ijms19103028
Wang X, Qian B, Ye J, et al (2013) Multi-objective multi-view spectral clustering via pareto optimization. In: Proceedings of the 2013 SIAM international conference on data mining. SIAM, pp 234–242. https://doi.org/10.1137/1.9781611972832.26
Wu H, Xiong WC, Mei L (2010) To build a synapse: signaling pathways in neuromuscular junction assembly. Development 137(7):1017–1033. https://doi.org/10.1242/dev.038711
Xie X, Sun S (2013) Multi-view clustering ensembles. In: 2013 international conference on machine learning and cybernetics. IEEE, pp 51–56. https://doi.org/10.1109/ICMLC.2013.6890443
Yang C, Gao W, Meza JC (2009) On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems. SIAM J Matrix Anal Appl 30(4):1773–1788. https://doi.org/10.1137/080716293
Zhang S (2018) Comparisons of gene coexpression network modules in breast cancer and ovarian cancer. BMC Syst Biol 12(1):75–87. https://doi.org/10.1186/s12918-018-0530-9
Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. https://doi.org/10.2202/1544-6115.1128
Zhang S, Ng MK (2016) Gene-microrna network module analysis for ovarian cancer. BMC Syst Biol 10(4):445–455. https://doi.org/10.1186/s12918-016-0357-1
Zhang S, Zhao H, Ng MK (2015) Functional module analysis for gene coexpression networks with network integration. IEEE/ACM Trans Comput Biol Bioinform 12(5):1146–1160. https://doi.org/10.1109/TCBB.2015.2396073
Zhou D, Burges CJ (2007) Spectral clustering and transductive learning with multiple views. In: Proceedings of the 24th international conference on machine learning. pp 1159–1166. https://doi.org/10.1145/1273496.1273642