Robust Principal Component Analysis Based On Hypergraph Regularization for Sample Clustering and Co-Characteristic Gene Selection

Ying-Lian Gao, Ming-Juan Wu, Jin-Xing Liu, Chun-Hou Zheng, and Juan Wang

Abstract—Extracting genes involved in cancer lesions from gene expression data is critical for cancer research and drug development. The method of feature selection has attracted much attention in the field of bioinformatics. Principal Component Analysis (PCA) is a widely used method for learning low-dimensional representation. Some variants of PCA have been proposed to improve the robustness and sparsity of the algorithm. However, the existing methods ignore the high-order relationships between data. In this paper, a new model named Robust Principal Component Analysis via Hypergraph Regularization (HRPCA) is proposed. In detail, HRPCA utilizes L2,1-norm to reduce the effect of outliers and make data sufficiently row-sparse. And the hypergraph regularization is introduced to consider the complex relationship among data. Important information hidden in the data are mined, and this method ensures the accuracy of the resulting data relationship information. Extensive experiments on multi-view biological data demonstrate that the feasible and effective of the proposed approach.

Index Terms—Robust principal component analysis, sample clustering, co-characteristic gene selection, L2,1-norm, hypergraph regularization

1 Introduction

In modern molecular biology, the study of cancer has become increasingly widespread. Cancer (malignant tumors) has become the top issue threatening human health [1]. Experimental studies show that the occurrence and development of cancer are always accompanied by changes in the genome and mutations in genes [2]. Changes in genes can be detected by sequencing techniques, such as gene chip technology and gene expression profiles [3]. The rapid development of sequencing technology has allowed sequencing of complete genes. Additionally, the direction and structure of recombinant genes can be determined, which provides technical support for disease research and development. The main information and features of cancer are hidden in gene expression data obtained using these technologies. Some judgments about cancer can be made using gene expression data, and relevant information has been found for different cancers [4], [5]. However, these data are characterized as high dimension low sample size; that is, the number of variables (number of genes) far exceeds the number of samples [6], [7]. Therefore, researchers have difficulty carrying out bioinformatics studies.

As an effective data processing method, principal component analysis (PCA) has attracted the attention of researchers. This method is widely used for face recognition, image clustering, biology, and other applications [8]. With the widespread application of PCA and its outstanding results, some shortcomings of the original PCA method have been gradually revealed. For example, each new principal component (PC) in low-dimensional subspace is a linear combination of the original data in high-dimensional space. Thus, PCA is often affected by the fact that PCs are dense [9], and new obtained PCs are difficult to explain. Because of this issue, many versions of PCA have been proposed to improve the interpretation of PCs. Therefore, this issue is not discussed in this paper. In addition, the covariance matrix of PCA is derived from L_2-norm, which is very sensitive to outliers and noise. Therefore, for some practical problems, PCA fails to deal with outliers and noise, and many related PCA methods have been proposed to reduce these effects. For example, PCA algorithms were proposed by Feng *et al.* to enhance the robustness of the algorithm by introducing the L_p-norm [10]. Shi *et al.* proposed a novel robust PCA method via an optimal mean by joint use of $L_{2,1}$ and schatten p-norms [11]. Robust principal component analysis (RPCA) overcomes the lack of robustness of the PCA method for the actual data [12]. Finally, Jiang *et al.* introduced $L_{2,1}$-norm and manifold learning into PCA to reduce the impact of data outliers and consider the internal geometry of the data, respectively [13].

The squared loss function of PCA is sensitive to outliers. Some outliers are also included in the cancer gene expression data used in this paper, and their existence is a major factor

- *Ying-Lian Gao is with the Qufu Normal University Library, Qufu Normal University, Rizhao, Shandong 276826, China. E-mail: yingliangao@126.com.*
- *Ming-Juan Wu, Jin-Xing Liu, Chun-Hou Zheng, and Juan Wang are with the School of Computer Science, Qufu Normal University, Rizhao, Shandong 276826, China. E-mail: {mingjuansw, wangjuansdu}@163.com, {sdcavell, zhengch99}@126.com.*

Manuscript received 19 Sept. 2019; revised 21 July 2020; accepted 27 Nov. 2020. Date of publication 10 Mar. 2021; date of current version 8 Aug. 2022. (Corresponding author: Juan Wang.)
Digital Object Identifier no. 10.1109/TCBB.2021.3065054

underlying the inaccuracy of cancer gene expression experiments [14], [15]. In recent years, researchers have proposed many relevant PCA methods to improve the robustness to outliers. The basic idea of this kind of method is to add various robust constraints on the loss term. Recently, the $L_{2,1}$-norm has been proven to be more robust to outliers [16]. Therefore, in this paper, we take advantage of the $L_{2,1}$-norm to reduce the impact of data outliers.

However, the above methods ignore the complex relationships among data, and these relationships are likely to hide key information about cancer lesions. Therefore, the geometry among the data cannot be ignored. The proposed manifold learning method establishes a relationship between data points, and it can well reflect the similarity between two cancer samples. However, the relationships among data are not simple relationships between two samples but also involve more complex high-order relationships. If multiple complex relationships are simply compressed into a relationship between the two, much useful information will inevitably be lost, which will have a certain impact on the accuracy of the learning algorithm. With the wide application of hypergraph [17], [18], [19], its advantages have gradually emerged. It can well capture the high-order relationship among data and reflect the sample geometry. Therefore, in this paper, in order to capture important information among data, we will take the advantages of hypergraph to mine complex and variable high-order relationships.

Based on the above ideas, $L_{2,1}$-norm and hypergraph regularization are jointly introduced into the PCA model. A new robust PCA method based on hypergraph regularization (HRPCA) is proposed. Different from the previous PCA methods, the highlights of the HRPCA are as follows:

(a) A new model named HRPCA is proposed for sample clustering and co-characteristic gene selection. In detail, the $L_{2,1}$-norm is applied to decrease the outliers and ensure that the data are sufficiently row-sparse, which avoid the impact of the square approximation error on the model. At the same time, hypergraph regularization is used to consider the complex relationships among data, which ensures the accuracy of the learning algorithm.
(b) The augmented Lagrange multiplier (ALM) is applied to solve the optimization problem. In addition, a series of theoretical analysis of HRPCA that includes convergence and computational complexity analyses are provided to validate the feasibility and effectiveness of HRPCA.
(c) The experimental results from multi-view datasets demonstrate the advancement of our approach. This method provides a good basis to explore potential relationships among different diseases. New oncogenes can be discovered from co-characteristic genes. The study of these genes is contributing to the early diagnosis and treatment of cancer.

The rest of the article is organized as follows: Section 2 introduces the related work, including PCA and hypergraph regularization. In Section 3, we introduce the new method HRPCA. Experiments and discussion are presented in Section 4. Section 5 is a summary of this paper and future work.

2 Related Work

2.1 Principal Component Analysis

PCA is one of the main tools for dimensionality reduction and is widely used in face recognition, image clustering, and biology. Assume that the $\mathbf{X} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{m \times n}$ is the data matrix with m variables on the rows, and n samples on the columns. In gene expression data, each column of $\mathbf{X}$ represents the transcriptional expression of a sample in m genes, and each row of $\mathbf{X}$ represents the expression level of a gene in n samples. PCA can find the optimal low dimensional subspace defined by the principal directions $\mathbf{M} = (m_1, m_2, \ldots, m_k) \in \mathbb{R}^{m \times k}$ and the projected samples in the new subspace $\mathbf{Q} = (q_1, q_2, \ldots, q_k) \in \mathbb{R}^{n \times k}$. Therefore, PCA can be expressed in the following problem:

$$\min_{\mathbf{M},\mathbf{Q}} \left\| \mathbf{X} - \mathbf{M}\mathbf{Q}^T \right\|_F^2 \; s.t. \; \mathbf{Q}^T\mathbf{Q} = \mathbf{I}. \tag{1}$$

PCA learns a low dimensional representation of data lying on a linear structure.

2.2 Hypergraph Regularization

As a generalization of a graph, a hypergraph is an important tool for data representation. The main difference between a graph and hypergraph is that the edges of a hypergraph can include multiple vertices, whereas the edges of a graph only include two vertices [20]. When the number of vertices in the hyperedge is equal to two, the hypergraph is equivalent to the common graph. For many practical problems, compression of multiple complex relationships into a relationship between two samples will result in the inevitable loss of a lot of useful information.

This loss will have some impact on the accuracy of the learning algorithm. Due to the characteristics of the above hyperedges, a hypergraph has greater flexibility to depict higher-order relationships [20]. To better illustrate the structure of the hypergraph, some experiments are performed on the gene expression data of PAAD-ESCA-CHOL dataset, as shown in Fig. 1. For example, some sample data of the dataset are selected, and different samples are described by different marks, as shown in Fig. 1a. Fig. 1b is a hypergraph constructed in accordance with the principles. Here, some basic hypergraph concepts and symbols are given.

Given a hypergraph $G = (V, E, \mathbf{W})$ consisting of a vertex set V and a hyperedge set E where $\mathbf{W}$ is the weight matrix of the hyperedge, the weight of each hyperedge is defined as $w(e)$. The weight $w(e)$ indicates the degree of necessity of retaining the corresponding sample relationship during the hypergraph partition. In the hypergraph, each vertex corresponds to a sample, and each hyperedge is an arbitrary subset of V that groups the samples based on a relationship. We use $\mathbf{H}$ to represent an incidence matrix of hypergraph G; then, $\mathbf{H}$ can be defined as:

$$\mathbf{H}(v, e) = \begin{cases} if \; v \in e, \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

Based on matrix $\mathbf{H}$, the degree of each vertex is defined as:

$$d(v) = \sum_{e \in E} w(e)\mathbf{H}(v, e). \tag{3}$$

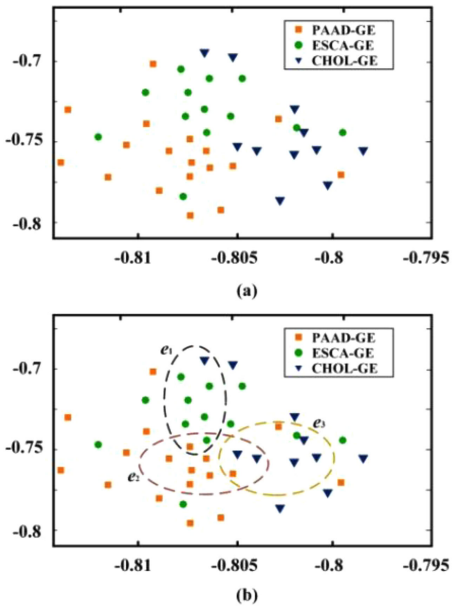


Fig. 1. Hypergraph constructed gene expression data. (a) Sample category. (b) Construct hypergraph edges (e1-e3).

This equation can be used to measure the degree of importance of vertex v for hypergraph G. Similarly, the degree of a hyperedge is defined as:

$$\delta(e) = \sum_{v} \mathbf{H}(v, e). \tag{4}$$

If we use $\mathbf{D}_v$, $\mathbf{W}_e$, and $\mathbf{D}_e$ to represent diagonal matrixes corresponding to the vertical degree, edge weight, and hyperedge degree, respectively [21], then the hypergraph Laplacian matrix can be defined as follows:

$$\mathbf{L}_h = \mathbf{D}_v - \mathbf{H}\mathbf{W}_e(\mathbf{D}_e)^{-1}\mathbf{H}^T. \tag{5}$$

3 Methodology

In this section, the objective function and optimal solution of HRPCA are presented. In addition, we provide a theoretical analysis of HRPCA that includes convergence and computational complexity analyses.

3.1 HRPCA Method

Adding the $L_{2,1}$-norm to HRPCA makes it more accurate to find co-characteristic genes. When extracting co-characteristic genes, it is in-evitable to take into account outlines and noise. This part of the noise information is F-norm cannot be avoided, which leads to the selected genes not what we want to obtain. Therefore, considering the robustness of the $L_{2,1}$-norm to outliers and noise, we introduce the $L_{2,1}$-norm into the objective function. In addition, the potential relationships among data makes sample clustering more accurate, because the relationships among the same types of data are more closely related. At the same time, the relationships among data are not simple relationships between two data points Therefore, in order to mine potential high-order relationships in the data, hypergraph regularization are introduced into the objective function.

The paper proposed HRPCA method, by adding $L_{2,1}$-norm and hypergraph regularization into PCA model mentioned above. The objective function of HRPCA is calculated as follows:

$$\min_{\mathbf{M},\mathbf{Q}} \left\|\mathbf{X} - \mathbf{M}\mathbf{Q}^T\right\|_{2,1} + \alpha \mathrm{Tr}\left(\mathbf{Q}^T\mathbf{L}_h\mathbf{Q}\right) \; s.t. \; \mathbf{Q}^T\mathbf{Q} = \mathbf{I}, \tag{6}$$

where α is a parameter to balance the contributions of the two parts. $\mathbf{L}_h$ is a hypergraph regularization matrix, and the $L_{2,1}$-norm is defined as:

$$\|\mathbf{X}\|_{2,1} = \sum\nolimits_{j=1}^{n} \sqrt{\sum\nolimits_{i=1}^{m} x_{ij}^2}. \tag{7}$$

3.2 Algorithmic Solution Strategy

The ALM method is used to update the algorithm. ALM is an important method for solving constrained optimization problems that is widely used in various fields. The basic idea is to introduce the Lagrange multiplier and penalty factor into the objective function, which makes finding the optimal solution for the transformed problem easier [22], [23].

An auxiliary variable is introduced to rewrite Eq. (6) as:

$$\begin{aligned} &\min_{\mathbf{M},\mathbf{Q},\mathbf{S}} \|\mathbf{S}\|_{2,1} + \alpha \mathrm{Tr}\left(\mathbf{Q}^T\mathbf{L}_h\mathbf{Q}\right), \\ &s.t. \; \mathbf{Q}^T\mathbf{Q} = \mathbf{I}, \; \mathbf{S} = \mathbf{X} - \mathbf{M}\mathbf{Q}^T. \end{aligned} \tag{8}$$

The augmented Lagrange function of Eq. (8) is defined as:

$$\begin{aligned} L(\mathbf{M}, \mathbf{Q}, \mathbf{S}, C) = \|\mathbf{S}\|_{2,1} + \alpha \mathrm{Tr}\left(\mathbf{Q}^T\mathbf{L}_h\mathbf{Q}\right) + \langle C, \mathbf{S} - \mathbf{X} + \mathbf{M}\mathbf{Q}^T\rangle \\ + \frac{\mu}{2}\left\|\mathbf{S} - \mathbf{X} + \mathbf{M}\mathbf{Q}^T\right\|_F^2, \end{aligned} \tag{9}$$

where C is the Lagrange multiplier matrix and μ is the penalty parameter. For the inner product of matrix $\mathbf{A}$ and matrix $\mathbf{B}$, denoted by $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{Tr}(\mathbf{A}^T\mathbf{B})$. Specifically, the optimal solution of Eq. (8) can be solved via the following three steps.

Computing $\mathbf{M}_{r+1}$. Fixing variables $\mathbf{Q}_r$, and $\mathbf{S}_r$, update $\mathbf{M}_{r+1}$ by:

$$\begin{aligned} \mathbf{M}_{r+1} &= \arg\min_{\mathbf{M}} L(\mathbf{M}, \mathbf{Q}_r, \mathbf{S}_r, C_r) \\ &= \arg\min_{\mathbf{M}} \frac{\mu}{2}\left\|\mathbf{S} - \mathbf{X} + \mathbf{M}_r\mathbf{Q}_r^T + \frac{C_r}{\mu}\right\|_F^2. \end{aligned} \tag{10}$$

We set $\mathbf{E}_r = \mathbf{X} - \mathbf{S}_r - C_r/\mu$. By seeking partial derivatives of $\mathbf{M}$, we can get:

$$\mathbf{M}_{r+1} = \mathbf{E}_r\mathbf{Q}_r. \tag{11}$$

Computing $\mathbf{Q}_{r+1}$. Similar, let $\mathbf{E}_r = \mathbf{X} - \mathbf{S}_r - C_r/\mu$. Fix others to calculate $\mathbf{Q}_{r+1}$ as follows:

$$\begin{aligned} \mathbf{Q}_{r+1} &= \arg\min_{\mathbf{Q}} L(\mathbf{M}_{r+1}, \mathbf{Q}, \mathbf{S}_r, C_r) \\ &= \arg\min_{\mathbf{Q}} \frac{\mu}{2}\left\|\mathbf{E}_r - \mathbf{M}_{r+1}\mathbf{Q}^T\right\|_F^2 + \alpha \mathrm{Tr}\left(\mathbf{Q}^T\mathbf{L}_h\mathbf{Q}\right). \end{aligned} \tag{12}$$

Substituting $\mathbf{M}$ into the Eq. (12), By some algebra, we have:

$$\begin{aligned}\mathbf{Q}_{r+1} &= \arg\min_{\mathbf{Q}} \left\|\mathbf{E}-\mathbf{E}_r\mathbf{Q}\mathbf{Q}^T\right\|_F^2 + \frac{2\alpha}{\mu}\mathrm{Tr}\left(\mathbf{Q}^T\mathbf{L}_h\mathbf{Q}\right)\\ &= \arg\min_{\mathbf{Q}} \mathrm{Tr}\left[\left(\mathbf{E}_r-\mathbf{E}_r\mathbf{Q}\mathbf{Q}^T\right)^T\left(\mathbf{E}_r-\mathbf{E}_r\mathbf{Q}\mathbf{Q}^T\right)\right]\\ &\quad + \frac{2\alpha}{\mu}\mathrm{Tr}\left(\mathbf{Q}^T\mathbf{L}_h\mathbf{Q}\right)\\ &= \min_{\mathbf{Q}} \mathrm{Tr}\left[\mathbf{Q}^T\left(-\mathbf{E}_r^T\mathbf{E}_r + \frac{2\alpha}{\mu}\mathbf{L}_h\right)\mathbf{Q}\right].\end{aligned} \tag{13}$$

Therefore, the Eq. (12) is equal to the following:

$$\mathbf{Q}_{r+1} = \arg\min_{\mathbf{Q}} \mathrm{Tr}\mathbf{Q}^T\left(-\mathbf{E}_r^T\mathbf{E}_r + \frac{2\alpha}{\mu}\mathbf{L}_h\right)\mathbf{Q}. \tag{14}$$

The solution of $\mathbf{Q}_{r+1}$ can be obtained by the eigenvectors corresponding to the first k smallest eigenvalues of the matrix P_α:

$$P_\alpha = -\mathbf{E}_r^T\mathbf{E}_r + \frac{2\alpha}{\mu}\mathbf{L}_h. \tag{15}$$

Computing $\mathbf{S}_{r+1}$. When computing $\mathbf{S}_{r+1}$, fix $\mathbf{M}_{r+1}$ and $\mathbf{Q}_{r+1}$, and minimize L for $\mathbf{S}_{r+1}$ as follows:

$$\begin{aligned}S_{r+1} &= \arg\min_{\mathbf{S}} L(\mathbf{M}_{r+1}, \mathbf{Q}_{r+1}, \mathbf{S}, C_r)\\ &= \arg\min_{\mathbf{S}} \|\mathbf{S}\|_{2,1} + \frac{\mu}{2}\|\mathbf{S}-\mathbf{D}\|_F^2,\end{aligned} \tag{16}$$

where $\mathbf{D} = \mathbf{X} - \mathbf{M}_{r+1}\mathbf{Q}_{r+1}^T - C_r/\mu$. Eq. (16) can be decomposed into the form of n independent problems:

$$\min_{s_i} \|s_i\|_{2,1} + \frac{\mu}{2}\|s_i - d_i\|_F^2, \tag{17}$$

where s_i and d_i are the i-th column of matrixes $\mathbf{S}$ and $\mathbf{D}$, respectively. The solution of the Eq. (17) can be obtained by [24].

$$s_i = \max\left(1 - \frac{1}{\mu\|d_i\|}, 0\right)d_i. \tag{18}$$

At the end of each iteration, the Lagrange multiplier matrix C and the parameter μ are updated to:

$$\begin{aligned}C_{r+1} &= C_r + \mu\left(\mathbf{S}_{r+1} - \mathbf{X} + \mathbf{M}_{r+1}\mathbf{Q}_{r+1}^T\right),\\ \mu &= \rho\mu,\end{aligned} \tag{19}$$

where $\rho > 1$. After a large number of experiments, it is found that the best results are obtained when$\rho \in [1.1, 1.6]$. The alternating updating algorithm of HRPCA is summarized in Algorithm 1.

3.3 Convergence Analysis

Before analyzing the convergence of the algorithm, a lemma is given.

Lemma 1. *For any non-zero vectors b and $s \in \mathbb{R}^m$, we have:*

$$\|b\|_2 - \frac{\|b\|_2^2}{2\|s\|_2} \le \|s\|_2 - \frac{\|s\|_2^2}{2\|s\|_2}. \tag{20}$$

The convergence analysis of HRPCA is summarized by Theorem 1.

Algorithm 1. HRPCA

Input: Data matrix:$\mathbf{X} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{m\times n}$
Weight matrix:$\mathbf{D}_w$
Parameters: α, k
Output: $\mathbf{M}_{r+1}$, $\mathbf{Q}_{r+1}$, $\mathbf{S}_{r+1}$

Repeat
Update $\mathbf{M}_{r+1}$ by (11);
Update $\mathbf{Q}_{r+1}$ by the eigenvectors of the matrix P_α corresponding to the first k smallest eigenvalues;
Update $\mathbf{S}_{r+1}$ by (18);
Update C_{r+1}, μ by (19)
Until convergence

Theorem 1. *The objective function value of each iteration is monotonically decreasing and converges to a local optimal solution.*

Proof. *The objective function in Eq. (6) is denoted as $J(M, Q)$. When the algorithm performs the $t+1-$th iteration, $\mathbf{M}^t$ and $\mathbf{Q}^t$ are obtained. We have:*

$$\begin{aligned}J\left(M^{t+1}, Q^{t+1}\right) = &\mathrm{Tr}\left(\mathbf{X} - \mathbf{M}^t\left(\mathbf{Q}^t\right)^T\right)^T\mathbf{D}^t\left(\mathbf{X} - \mathbf{M}^t\left(\mathbf{Q}^t\right)^T\right)\\ &+ \alpha\mathrm{Tr}\left(\left(\mathbf{Q}^t\right)^T\mathbf{L}_h\mathbf{Q}^t\right).\end{aligned} \tag{21}$$

First, since the eigenvalue decomposition gives the optimal $\mathbf{Q}^t$, and we get:

$$J\left(M^{t+1}, Q^{t+1}\right) \le J\left(M^{t+1}, Q^t\right). \tag{22}$$

Then, the following formulation can be obtained:

$$\begin{aligned}&\mathrm{Tr}\left(\mathbf{X} - \mathbf{M}^{t+1}\left(\mathbf{Q}^{t+1}\right)^T\right)^T\mathbf{D}^t\left(\mathbf{X} - \mathbf{M}^{t+1}\left(\mathbf{Q}^{t+1}\right)^T\right)\\ &+ \alpha\mathrm{Tr}\left(\left(\mathbf{Q}^{t+1}\right)^T\mathbf{L}_h\mathbf{Q}^{t+1}\right)\\ &\le \mathrm{Tr}\left(\mathbf{X} - \mathbf{M}^{t+1}\left(\mathbf{Q}^{t+1}\right)^T\right)^T\mathbf{D}^t\left(\mathbf{X} - \mathbf{M}^{t+1}\left(\mathbf{Q}^{t+1}\right)^T\right)\\ &+ \alpha\mathrm{Tr}\left(\left(\mathbf{Q}^t\right)^T\mathbf{L}_h\mathbf{Q}^t\right).\end{aligned} \tag{23}$$

In addition, since $\|\mathbf{L}\|_{2,1} = \sum_{i=1}^n \|l_i\|_2$, according to lemma 1 we have:

$$\begin{aligned}&\sum\nolimits_{i=1}^n\left(\left\|x - m_i^{t+1}\left(q_i^{t+1}\right)^T\right\|_2 - \frac{\left\|x - m_i^{t+1}\left(q_i^{t+1}\right)^T\right\|_2}{\left\|x - m_i^t\left(q_i^t\right)^T\right\|_2}\right)\\ &\le \sum\nolimits_{i=1}^n\left(\left\|x - m_i^t\left(q_i^t\right)^T\right\|_2 - \frac{\left\|x - m_i^t\left(q_i^t\right)^T\right\|_2}{\left\|x - m_i^t\left(q_i^t\right)^T\right\|_2}\right).\end{aligned} \tag{24}$$

TABLE 1
Summary of the Two Multi-View Datasets

Datasets	Samples	Genes
COAD	281 (262, 19)	20502
ESCA	192 (183, 9)	20502
PAAD	180 (176, 4)	20502
CHOL	45 (36, 9)	20502
PAAD-ESCA-CHOL	395	20502
PAAD-CHOL-COAD	474	20502

Finally, by combining Eqs. (23) and (24), we have:

$$\begin{aligned} &\mathrm{Tr}\left(\mathbf{X}-\mathbf{M}^{t+1}\left(\mathbf{Q}^{t+1}\right)^{T}\right)^{T}\mathbf{D}^{t}\left(\mathbf{X}-\mathbf{M}^{t+1}\left(\mathbf{Q}^{t+1}\right)^{T}\right)\\ &+\alpha\mathrm{Tr}\left(\left(\mathbf{Q}^{t+1}\right)^{T}\mathbf{L}_{h}\mathbf{Q}^{t+1}\right)\\ &\leq \mathrm{Tr}\left(\mathbf{X}-\mathbf{M}^{t}\left(\mathbf{Q}^{t}\right)^{T}\right)^{T}\mathbf{D}^{t}\left(\mathbf{X}-\mathbf{M}^{t}\left(\mathbf{Q}^{t}\right)^{T}\right)\\ &+\alpha\mathrm{Tr}\left(\left(\mathbf{Q}^{t}\right)^{T}\mathbf{L}_{h}\mathbf{Q}^{t}\right). \end{aligned} \tag{25}$$

That is,

$$J\left(M^{t+1},Q^{t+1}\right)\leq J\left(M^{t},Q^{t}\right). \tag{26}$$

Therefore, the convergence of the algorithm is proved.

3.4 Computational Complexity Analysis

The computational complexity of HRPCA is divided into two steps. The first step is to compute the Laplacian matrix of the hypergraph with $O(n^2)$. The second is to compute the solution of $\mathbf{M}$ and $\mathbf{Q}$, whose computational complexity is $O(t(mn^2))$. t is the number of iterations. Therefore, the total computational complexity is $O(t(mn^2))$.

4 Experimental Results and Analysis

In this section, to verify the effectiveness of the HRPCA algorithm, sample clustering and co-characteristic gene selection experiments are performed with multi-view biological data. We compare these results with those obtained using six other methods: Laplacian embedding (LE) [25], PCA, Non-negative Matrix Factorization (NMF) [26], Low-Rank Representation (LRR) [27], graph-Laplacian PCA (gLPCA) [13], and robust gLPCA (RgLPCA) [13]. Since the datasets used in this paper are multi-view datasets, the genes selected by each method are defined as co-characteristic genes. First, sample clustering can be used to evaluate the effects of hypergraph regularization. Secondly, the joint effects of the $L_{2,1}$-norm and hypergraph regularization in the model are evaluated by identifying co-characteristic genes and exploring the potential associations of the genes with disease. New oncogenes associated with disease can be found among the co-characteristic genes. Then, the pathogenic principles and functions of the new oncogenes are identified, which can help uncover more relevant information about cancer. Details of the datasets and experimental results used in the sample clustering and co-characteristic gene selection are summarized in the following sections.

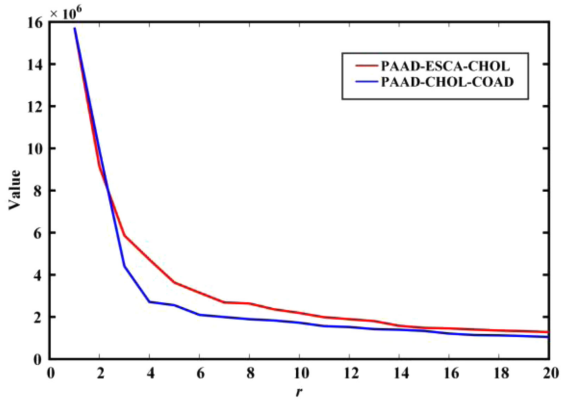


Fig. 2. The singular values of different data matrixes.

4.1 Composition of Datasets

Since completion of the Human Genome Project, cancer research has entered the era of genomics [28]. The development of multi-view data plays an important role in exploring the relationships between diseases and genes. Gene expression data from different diseases are combined to form multi-view datasets, which can help find oncogenes that cause disease and capture links between different diseases [5]. In this paper, two multi-view datasets (PAAD-ESCA-CHOL and PAAD-CHOL-COAD) are used to verify the effectiveness of the algorithm. In detail, each multi-view dataset consists of three gene expression datasets downloaded from TCGA database (The Cancer Genome Atlas, http://cancergenome.nih.gov/). The PAAD-ESCA-CHOL multi-view dataset contains pancreatic cancer (PAAD), oesophageal cancer (ESCA), and cholangiocarcinoma (CHOL) gene expression data. The PAAD-CHOL-COAD multi-view dataset consists of pancreatic cancer (PAAD), cholangiocarcinoma (CHOL), and colon adenocarcinoma (COAD) gene expression data. Sample types are divided into normal samples and diseased samples. To maintain sample balance, during data preprocessing, for the gene expression data of different cancers, we only keep the diseased samples with a larger proportion in the data set, and remove the normal samples with a small total sample size. For example, the COAD dataset sample contains 262 and 191 diseased and normal samples, respectively. Table 1 lists the details of the two multi-view datasets.

4.2 Experimental Settings

In this paper, the parameters k and α need to be adjusted. To be fair, the optimal parameter values are selected within a reasonable range. The best results are reported using the optimal parameters for the comparison method during the experiment. Singular value decomposition (SVD) is used to determine the number of dimensionality reductions k. First, we perform SVD on the

original matrix $\mathbf{X}$ when conducting the experiments. The results are shown in Fig. 2. The horizontal axis in Fig. 2 represents the numbers of singular values, and the vertical axis represents the values of the singular values. Different datasets are represented by different colored lines. As shown in Fig. 2, the values of the singular values are gradually reduced. Singular values often correspond to important information implied in the matrix, and the importance of this information is positively correlated with the sizes of the

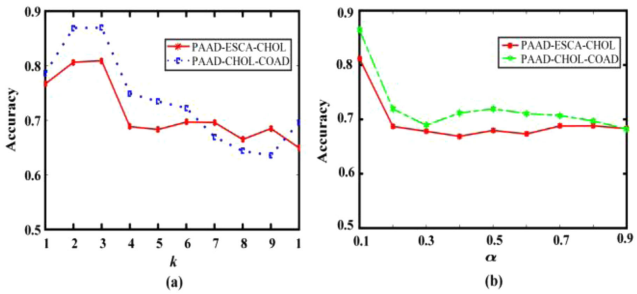


Fig. 3. Parameter selection. (a) is the clustering performance of HRPCA about k. (b) is the clustering performance of IHPCA about α.

singular values. Larger singular values correspond to the main information in the matrix. Therefore, large singular values are preserved during the experiment, which not only reduces the data complexity but also retains useful information within the data. In this paper, the three largest singular value points are retained, and the inflection point $r = 3$ is selected as the number of dimensionality reductions k.

To further verify the k obtained by SVD decomposition can achieve ideal experimental results. Fig. 3a shows the clustering accuracy at different k. It can be seen that the clustering accuracy on PAAD-ESCA-CHOL and PAAD-CHOL-COAD datasets are relatively high when k=3. The regularization parameter α is chosen form $\{0.1, 0.2, \ldots, 0.9\}$. To avoid randomness, we run ten times with different initializations and report the average results. Fig. 3b shows the sensitivity of accuracy under different α.

4.3 Sample Clustering

The new algorithm introduces hypergraph regularization to consider the high-order relationships among data. Therefore, the effectiveness of this method is validated by sample clustering. Sample clustering experiments are performed on matrix $\mathbf{Q}$. In this paper, various measures are adopted as evaluation criteria, including accuracy (ACC), recall, precision, and F-measure. ACC, Recall, precision, and F-measure are defined as follows:

$$ACC = \frac{\sum_{i=1}^{n} \delta(y_i, map(p_i))}{n}, \tag{27}$$

$$recall = \frac{TP}{TP + FN}, \tag{28}$$

$$precision = \frac{TP}{TP + FP}, \tag{29}$$

$$F - measure = \frac{2 \times precision \times recall}{recision + recall}, \tag{30}$$

where n is the number of all samples in the dataset, y_i is the true label of x_i , and p_i is the clustering label obtained by the clustering algorithm. $map(p_i)$ is the best matching function that can map clustering labels to real data labels. $\delta(x, y)$ is a function in which $\delta(x, y)$=1 if $x = y$ and $\delta(x, y)$= 0 otherwise. This metric can find a one-to-one relationship between a cluster and real categories [29]. A high ACC represents a better clustering performance. TP, FP, FN are the numbers of true positives, false positives, and false negatives.

Finally, this method is compared with six other methods (LE, PCA, NMF, LRR, gLPCA, and RgLPCA), since these methods are closely related to HRPCA. From Table 2, we can conclude the following:

(a) The clustering performance of gLPCA is better than that of LE and PCA for the two multi-view datasets. The possible reason is that the latter two methods do not consider the internal geometry of the data, which leads to loss of useful information and affects the clustering results.

(b) For the PAAD-CHOL-COAD dataset, when all of the evaluation indicators are combined, the clustering performance of RgLPCA is better than that of gLPCA. The difference between the two methods is that the $L_{2,1}$-norm in RgLPCA is used to reduce the influence of outliers and noise on the loss term. The adverse effects of outliers on the clustering results are reflected, which indicates that the existence of outliers must be considered in multi-view data.

(c) Taking Table 2 as a whole, the proposed method has a good experimental effect. In terms of the ACC, F-measure, Precision and recall, HRPCA outperformed the other methods by approximately 6.4, 16.3, 14.78, and 11.2 percent, respectively. In summary, the HRPCA method obtains more satisfactory clustering results, which indicates that the hypergraph structure has a good effect on data clustering. On the other hand, the potential links between different data are mined, and the complex relationship between genes and diseases can be reflected through the data, which lays a good foundation for the prevention and treatment of disease.

4.4 Embedding Evaluation

The sample clustering experiment is shown in the previous section. To visualize the low-dimensional embedding effects of the five methods, we performed related experiments on matrix $\mathbf{Q}$. The experimental results will be

TABLE 2
The Clustering Results on Two Multi-View Datasets

Methods	PAAD-ESCA-CHOL (%)				PAAD-CHOL-COAD (%)			
	Recall	Precision	F-measure	ACC	Recall	Precision	F-measure	ACC
LE	46.30	74.38	40.40	61.51	51.54	74.20	52.09	65.70
PCA	47.20	75.54	49.50	72.53	46.96	74.03	50.05	74.18
NMF	46.51	76.33	47.61	71.65	47.84	76.81	51.64	74.98
LRR	50.34	79.63	56.98	78.83	48.11	79.43	48.62	78.64
gLPCA	51.24	73.36	55.03	70.56	43.25	77.77	45.56	79.85
RgLPCA	49.73	74.65	50.65	76.36	43.28	73.34	50.05	80.22
HRPCA	**67.87**	**87.12**	**66.21**	**81.27**	**52.81**	**92.55**	**52.27**	**86.40**

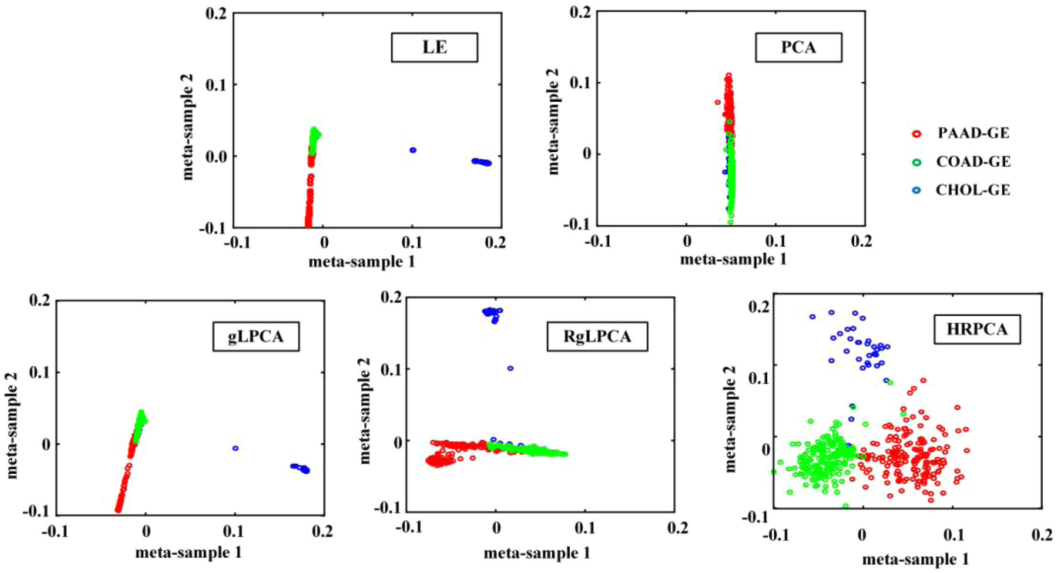


Fig. 4. A visual comparison of low-dimensional embedding by LE, PCA gLPCA, RgLPCA, and HRPCA on PAAD-CHOL-COAD, respectively, where different colors indicate different classes of samples.

presented in a two-dimensional space. Samples from the PAAD-CHOL-COAD dataset can be divided into three categories. If the method can clearly distinguish whether a sample belongs to one class or that all samples can be divided into three categories, this outcome will indicate the superiority of the sample embeddings.

Fig. 4 shows the low-dimensional embeddings obtained using LE, PCA, gLPCA, RgLPCA, and our HRPCA for PAAD-CHOL-COAD. From Fig. 4, we have the following observations:

(a) The low-dimensional embedding effect of PCA is not ideal. The three types of samples are linearly distributed, and the low-dimensional embeddings of samples from different classes by PCA are mixed together. The sample categories are not well differentiated. The sample distributions of LE and gLPCA are similar. However, gLPCA can correctly distinguish the number of samples better than LE, which shows that the graph regularization in the model has a good effect. The use of the manifold learning structure can increase the discriminative ability of PCA.

(b) The sample distributions of RgLPCA and HRPCA are as follows. Obviously, HRPCA can separate the embeddings of samples in different classes to a large extent. The samples can be roughly divided into three categories by HRPCA, which represent the three types of disease samples in the multi-view dataset. The experimental effect of RgLPCA is not as good as that of HRPCA. The difference between the two methods is that RgLPCA considers the internal geometry of the data through a Laplacian graph, whereas HRPCA introduces a hypergraph to capture the higher-order geometry of the data, which further demonstrates that introduction of hypergraph regularization into the model can guarantee the accuracy of the learning algorithm.

(c) When all of the information in Fig. 4 is taken into consideration, the low-dimensional embeddings by HRPCA are more discriminable, which validates that the joint effect of $L_{2,1}$-norm and hypergraph regularization in the model can achieve the desired results.

4.5 The Selection of Co-Characteristic Genes

In this subsection, co-characteristic gene selection is applied to verify the advancement of our method. Co-characteristic gene selection experiments are performed on matrix $\mathbf{M}$. Practice proved that this approach was the simplest method to identify differentially expressed genes, which could effectively improve the experimental performance. First, a matrix decomposed by algorithms for co-characteristic gene selection is processed. In detail, the absolute values of the matrix columns are summed to obtain a new vector. Then, the absolute values of the new vectors are sorted in descending order. Next, the first 500 values in the vector are selected, and the corresponding genes are identified for analysis. Without loss of generality, if the element in the vector has the higher ranking, then the gene expression is more differential. Second, the ToppFun (https://toppgene.cchmc.org/enrichment.jsp) and GeneCards (https://www.genecards.org/) tools are used to perform the gene enrichment analysis and explore potential links between genes and diseases, respectively. The specific experimental results will be discussed in the following subsection.

4.5.1 *Experiments on the PAAD-ESCA-CHOL Multi-View Dataset*

HRPCA is applied to discover co-characteristic genes and is compared with the LE, PCA, NMF, LRR, gLPCA, and RgLPCA methods. The genes selected by each method are placed into ToppFun, which is a type of GO term finder, for the co-characteristic gene enrichment analysis and to obtain their P-values and hit counts. The functions are classified into three categories: molecular function, cellular component, and biological process. The main role of the programme is to discover the commonalities in a large gene

TABLE 3
The P-Values and Hit Counts on PAAD-ESCA-CHOL Dataset

ID	Name	LE		PCA		NMF		gLPCA	
		P-values	Hits	P-values	Hits	P-values	Hits	P-values	Hits
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	5.27E-26	31	9.27E-46	42	6.31E-50	45	5.64E-72	60
GO:0070972	protein localization	4.14E-24	33	2.61E-31	36	3.151E-33	38	5.60E-68	64
GO:0005198	structural molecule activity	6.60E-49	107	1.36E-49	112	2.66E-54	120	3.18E-72	131
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	1.32E-24	33	1.55E-27	43	2.95E-37	49	3.10E-64	61
GO:0019083	viral transcription	3.73E-18	32	1.14E-44	55	2.53E-50	57	6.11E-52	61
GO:0006364	rRNA processing	1.97E-12	31	1.27E-40	61	4.65E-42	64	2.04E-38	59
GO:0015934	large ribosomal subunit	2.16E-11	19	9.91E-34	38	8.19E-38	41	2.27E-32	37
GO:0019843	rRNA binding	1.06E-11	16	1.69E-21	24	3.98E-23	26	1.33E-21	24
GO:0006508	proteolysis	1.48E-19	110	9.29E-14	97	7.10E-16	101	1.15E-18	108
GO:0008233	peptidase activity	6.17E-09	45	1.02E-06	40	5.14E-07	42	4.78E-09	45

ID	Name	LRR		RgLPCA		HRPCA	
		P-values	Hits	P-values	Hits	P-values	Hits
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	4.64E-56	51	2.00E-59	53	8.40E-78	63
GO:0070972	protein localization	6.18E-53	53	7.78E-57	57	7.87E-75	68
GO:0005198	structural molecule activity	4.23E-61	121	7.91E-68	127	4.01E-73	132
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	2.70E-56	57	9.42E-55	55	2.51E-69	64
GO:0019083	viral transcription	4.54E-52	63	8.16E-43	54	8.75E-55	63
GO:0006364	rRNA processing	1.98E-48	60	4.27E-32	53	7.51E-42	62
GO:0015934	large ribosomal subunit	3.03E-29	36	7.90E-26	32	3.82E-35	39
GO:0019843	rRNA binding	2.56E-21	24	3.32E-20	23	2.43E-24	26
GO:0006508	proteolysis	1.50E-16	105	8.62E-18	106	2.89E-20	111
GO:0008233	peptidase activity	5.06E-08	43	1.94E-09	46	5.55E-10	47

expression dataset. The analysis provides critical information for the co-characteristic gene extraction experiment.

The P-values and hit counts of the top 10 items obtained using these different methods are listed in Table 3. The former indicates the significance of the gene enrichment analysis using GO terms. A smaller P-values indicates more important GO terms. Additionally, a lower P-values indicates that the algorithm is less affected by noise and outliers and has higher efficiency. The degree of gene enrichment will be reduced if the algorithm is heavily influenced by noise and outliers. The hit counts is the number of genes from the input, and the P-values is affected by the number of input genes and other factors [30]. Table 3 shows that the P-values of the co-characteristic genes selected by the HRPCA method are obviously smaller than those obtained with the other methods and that the hit counts are greater than those obtained with the other methods. Good experimental results are shown in bold. In detail, GO:0070972 is protein localization to the endoplasmic reticulum. The hit counts for this item in the genome is 125, including 68 hits found by our method and more hits found using the other methods. GO:0000184 is the nuclear-transcribed mRNA catabolic process, nonsense-mediated decay. Nonsense-mediated mRNA decay (NMD) mainly maintains normal physiological functions by regulating the intracellular environment, which is closely related to the occurrence and development of tumours and changes in expression of corresponding genes. The corresponding name of GO:0006614 is SRP-dependent protein targeting to the membrane. The P-values for this GO term based on the HRPCA method is 8.40E-78, which is much smaller than the values obtained with the other four methods. Moreover, 94 genes were contained in this genetic term category,

of which 63 genes can be identified by our method, and the number of genes identified by the other four methods was 31, 42, 60, and 53 respectively. The experiment demonstrates that the degree of gene enrichment selected by HRPCA is better than that of the other methods, which further confirms the effectiveness of this method.

Moreover, the genes selected by each method are screened. Some genes unique to the HRPCA method are selected. Then, these unique genes are subjected to functional analysis to summarize their related pathogenic genes and relevance scores. Information for these unique genes is listed in Table 4. These genes cannot be ignored when studying the relationships among the three diseases in each dataset. Therefore, these genes need to be studied further in the future. The SLPI gene encodes a secreted inhibitor that has affinity for trypsin, leukocyte elastase, and cathepsin G [31]. This gene has the highest relevance score in the PAAD

TABLE 4
The Name, Functions, and Relevance Scores of Two Genes

Genes	Name	Summary of function	Relevance scores
SLPI	Secretory Leukocyte Peptidase Inhibitor	This gene encodes a secreted inhibitor that protects epithelial tissues from serine proteases. Its inhibitory effect contributes to the immune response by protecting epithelial surfaces from attack by endogenous proteolytic enzymes.	10.52, 4.33, 5.76
AZGP1	Alpha-2-Glycoprotein 1, Zinc-Binding	This gene induces a concentration-dependent increase in UCP-1 expression in primary cultures. This effect is attenuated by the β3-adrenergic receptor (β3-AR) antagonist SR59230A, which may play an important role in lipid utilization during cancer cachexia.	14.24, 10.91, 2.75

TABLE 5
Summary of the Top 7 Genes in the PAAD-ESCA-CHOL Multi-View Dataset

Genes	Official name	GO annotations	Related diseases	Relevance scores
EGFR	Epidermal Growth Factor Receptor	Identical protein binding and protein kinase activity	Inflammatory skin and bowel disease, neonatal 2, and lung cancer	76.40,87.92,18.57
CTNNB1	Catenin Beta 1	DNA binding transcription factor activity and binding	Mental retardation, autosomal dominant 19, and pilomatrixoma	66.25,75.84,13.70
MUC1	Mucin 1, Cell Surface Associated	RNA polymerase II proximal promoter sequence-specific DNA binding and p53 binding	Medullary cystic kidney disease 1 and secretory meningioma	38.50,42.72,18.22
KRT7	Keratin 7	Structural molecule activity	Cystadenoma and adenosquamous carcinoma	36.26,37.02,17.48
KRT19	Keratin 19	Structural molecule activity and structural constituent of cytoskeleton	Breast cancer, type I	28.13,26.98,18.93
CD44	CD44 Molecule (Indian Blood Group)	Transmembrane signalling receptor activity and cytokine receptor activity	Superficial keratitis and lichen sclerosus	26.09,33.61,13.00
CEACAM5	Carcinoembryonic Antigen Related Cell Adhesion Molecule 5	Protein homodimerization activity and GPI anchor binding	Lung cancer and rectal neoplasm	31.33,28.79,11.45

dataset, which indicates that it is most relevant to this disease. In the future, we should pay more attention to the relationship between this gene and pancreatic cancer. AZGP1 has a high relevance score for the PAAD and ESCA datasets. Many studies have shown that this gene is involved in the development of PAAD and ESCA diseases [32], [33]. The superiority of HRPCA can be further illustrated through the study of unique genes.

Finally, the co-characteristic genes are ranked in descending order according to their relevance scores, and the top 7 genes are selected for detailed analysis. Table 5 summarizes the official name, GO annotations, related diseases, and relevance scores of these genes. Studying the relationships among these three diseases is critical, because these genes are highly expressed in all three cancers, especially in the PAAD dataset. Mutations in one gene are likely to have an effect on all three cancers. Therefore, the relationship between genes and diseases is a cause for concern. EGFR is found in many cancers as a growth factor receptor for cancer therapeutic targets, such as non-small cell lung cancer, squamous cell carcinoma of the head and neck, colorectal cancer, and pancreatic cancer [34]. The MUC1 oncoprotein is over-expressed in most human carcinomas and blocks the induction of apoptosis by genotoxic agents. The literature indicates that MUC1 promotes the underlying mechanism of tumourigenesis by regulating β-catenin localization and the cytoskeleton [35], [36]. Therefore, more medical research should focus on the relationship between MUC1 and various diseases. Both KRT7 and CEACAM5 are found in oesophageal cancer and have a high correlation score, which indicates that the genes are likely to be causative genes for the disease. Therefore, the links between these two genes and ESCA should be investigated. MMP2 is involved in the pathological changes of cancer [37]. Currently, people are paying more attention to its transfer-promoting properties. The pathogenic genes of cancer can be further explored through a detailed introduction of genes, which provides a good foundation for cancer prevention and treatment.

5 Conclusion

In this paper, a new method named HRPCA is proposed. This method improves the robustness to data outliers by introducing $L_{2,1}$-norm in the loss term. In addition, the higher-order geometric structure inside the multi-view data is considered, and the hypergraph regularization is introduced into the objective function. The HRPCA makes full use of the outstanding characteristics of hyper-edge to capture the modular relationship among different disease data, which not only improves the accuracy of the algorithm, but also provides new research ideas and directions for disease data processing. The sample clustering and co-characteristic gene selection experiments are performed using multi-view datasets. The effectiveness and progress of the proposed method are well documented by the experimental results.

Our current method has some limitations. First, although HRPCA method enhances the robustness to outliers, it still needs to be verified to adapt to large-scale biological data.

Second, we can find more ways to construct hyperedge to better find higher-order relationships among data.

In the future, we will further explore these problems and study better methods for identification of differentially expressed genes.

ACKNOWLEDGMENTS

This work was supported in part by the NSFC under Grants 61872220, and 61572284.

REFERENCES

[1] C. Feng, Y. Xu, J. Liu, Y. Gao, and C. Zheng, "Supervised discriminative sparse PCA for com-characteristic gene selection and tumor classification on multiview biological data," *IEEE Trans. neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 2926–2937, Oct. 2019.

[2] L. Angus, M. Smid, S. M. Wilting, J. v. Riet, and J. W. M. Martens, "The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies," *Nat. Genet.*, vol. 51, no. 10, pp. 1–9, 2019.

[3] M. Kunz, "DNA microarray technology," *Seminars Cutan. Med. Surg.*, vol. 27, no. 1, pp. 16–24, 2008.

[4] X. Zhao *et al.*, "Identifying cancer-related microRNAs based on gene expression data," *Bioinformatics*, vol. 31, no. 8, pp. 1226–1234, 2015.

[5] F. He, G. Zhu, Y. Wang, X. Zhao, and D. Huang, "PCID: A novel approach for predicting disease comorbidity by integrating multiscale data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 3, pp. 678–686, May/Jun. 2017.

[6] Z.-Y. Yang, X.-Y. Liu, J. Shu, H. Zhang, and Y. Liang, "Multi-view based integrative analysis of gene expression data for identifying biomarkers," *Sci. Rep.*, vol. 9, no. 1, pp. 1–15, 2019.

[7] J. H. Lee *et al.*, "Integrative analysis of mutational and transcriptional profiles reveals driver mutations of metastatic breast cancers," *Cell Discov.*, vol. 2, 2016, pp. Art. no.16025.

[8] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev.: Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

[9] S. Yi, Z. Lai, Z. He, Y. Cheung, and Y. Liu, "Joint sparse principal component analysis," *Pattern Recognit.*, vol. 61, pp. 524–536, 2017.

[10] C. Feng, Y. L. Gao, J. X. Liu, C. H. Zheng, and J. Yu, "PCA based on graph laplacian regularization and P-norm for gene selection and clustering," *IEEE Trans. Nanobiosci.*, vol. 16, no. 4, pp. 257–265, Jun. 2017.

[11] X. S. Shi, F. P. Nie, Z. H. Lai, and Z. H. Guo, "Robust principal component analysis via optimal mean by joint $\ell 2$, 1 and schatten p-norms minimization," *Neurocomputing*, vol. 283, pp. 205–213, 2018.

[12] J. X. Liu, Y. T. Wang, C. H. Zheng, W. Sha, J. X. Mi, and Y. Xu, "Robust PCA based method for discovering differentially expressed genes," *BMC Bioinf.*, vol. 14, no. S8, 2013, pp. Art. no. S3.

[13] B. Jiang, C. Ding, B. Luo, and J. Tang, "Graph-Laplacian PCA: Closed-form solution and robustness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3492–3498.

[14] C. Yan, G. Chen, and Y. Shen, "Outlier analysis for gene expression data," *J. Comput. Sci. Technol.*, vol. 19, no. 1, pp. 13–21, 2004.

[15] M. Alshalalfa, T. A. Bismar, and R. Alhajj, "Detecting cancer outlier genes with potential rearrangement using gene expression data and biological networks," *Adv. Bioinf.*, vol. 2012, pp. 373506–373506, vol. 2012.

[16] D. G. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using L21-norm," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 673–682.

[17] D. Saxton and A. Thomason, "Hypergraph containers," *Inventiones Mathematicae*, vol. 201, no. 3, pp. 925–992, 2015.

[18] C. Shangguan, Y. Zhang, and G. Ge, "Centralized coded caching schemes: A hypergraph theoretical approach," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5755–5766, Aug. 2018.

[19] G. C. Shang, Y. W. Zhang, and G. Ge, "Centralized coded caching schemes: A hypergraph theoretical approach," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5755–5766, Aug. 2018.

[20] S. Huang, H. Wang, Y. Ge, L. Huangfu, X. Zhang, and D. Yang, "Improved hypergraph regularized nonnegative matrix factorization with sparse representation," *Pattern Recognit. Lett.*, vol. 102, no. 15, pp. 8–14, 2018.

[21] T. Jin, J. Yu, J. You, K. Zeng, C. Li, and Z. Yu, "Low-rank matrix factorization with multiple hypergraph regularizer," *Pattern Recognit.*, vol. 48, no. 3, pp. 1011–1022, 2015.

[22] M. V. Dolgopolik, "Existence of augmented lagrange multipliers: Reduction to exact penalty functions and localization principle," *Math. Prog.*, vol. 166, no. 1, pp. 297–326, 2017.

[23] F. Meng, X. Yang, and C. Zhou, "The augmented lagrange multipliers method for matrix completion from corrupted samplings with application to mixed Gaussian-impulse noise removal," *Plos One*, vol. 9, no. 9, 2014, pp. Art. no. e108125.

[24] Y. Lou and M. Yan, "Fast L1–L2 minimization via a proximal operator," *J. Sci. Comput.*, vol. 74, no. 2, pp. 767–785, 2018.

[25] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[26] C. Zheng, D. Huang, L. Zhang, and X. Kong, "Tumor clustering using nonnegative matrix factorization with gene selection," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 4, pp. 599–607, Jul. 2009.

[27] Y. Cui, C. H. Zheng, J. Yang, and P. Paolo, "Identifying subspace gene clusters from microarray data using low-rank representation," *PLoS One*, vol. 8, no. 3, 2013, pp. Art. no. e59377.

[28] A. E. Kwitek *et al.*, "Automated construction of high-density comparative maps between rat, human, and mouse," *Genome Res.*, vol. 11, no. 11, pp. 1935–1943, 2001.

[29] C. Hou, F. Nie, D. Yi, and D. Tao, "Discriminative embedded clustering: A framework for grouping high-dimensional data," *IEEE Trans. Neural Netw.*, vol. 26, no. 6, pp. 1287–1299, Jun. 2015.

[30] C. M. Feng, Y. L. Gao, J. X. Liu, J. Wang, D. Q. Wang, and C. G. Wen, "Joint L1/2-norm constraint and graph-laplacian PCA method for feature extraction," *Biomed Res. Int.*, vol. 2017, pp. 1–14, 2017.

[31] D. Zhang, R. C. M. Simmen, F. J. Michel, G. Zhao, D. Vale-Cruz, and F. A. Simmen, "Secretory leukocyte protease inhibitor mediates proliferation of human endometrial epithelial cells by positive and negative regulation of growth-associated genes," *J. Biol. Chem.*, vol. 277, no. 33, pp. 29999–30009, 2002.

[32] H. Tang *et al.*, "Reduction of AZGP1 predicts poor prognosis in esophageal squamous cell carcinoma patients in northern china," *OncoTargets Ther.*, vol. 10, pp. 85–94, 2016.

[33] K. Moore, Z. J. Bryant, G. S. Ghatnekar, U. P. Singh, R. G. Gourdie, and J. D. Potts, "A synthetic connexin 43 mimetic peptide augments corneal wound healing," *Exp. Eye Res.*, vol. 115, pp. 178–188, 2013.

[34] F. Ciardiello and G. Tortora, "EGFR antagonists in cancer treatment," *New Engl. J. Med.*, vol. 358, no. 11, pp. 1160–1174, 2008.

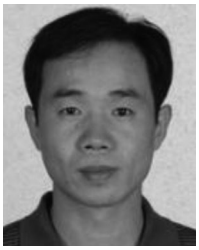
[35] D. Raina *et al.*, "MUC1 oncoprotein blocks nuclear targeting of c-Abl in the apoptotic response to DNA damage," *EMBO J.*, vol. 25, no. 16, pp. 3774–3783, 2006.

[36] J. A. Schroeder, M. C. Adriance, M. C. Thompson, T. D. Camenisch, and S. J. Gendler, "MUC1 alters β-catenin-dependent tumor formation and promotes cellular invasion," *Oncogene*, vol. 22, no. 9, pp. 1324–1332, 2003.

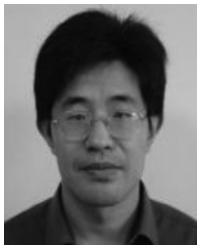
[37] W. Shao *et al.*, "Prognostic impact of MMP-2 and MMP-9 expression in pathologic stage IA non-small cell lung cancer," *J. Surg. Oncol.*, vol. 104, no. 7, pp. 841–846, 2011.

Ying-Lian Gao received the BS degree in chemical education and the MS degree in chemical engineering from Qufu Normal University, China, in 1997 and 2000, respectively. Currently she is a lecturer with the Qufu Normal University. Her research interests include pattern recognition, information management and data mining.

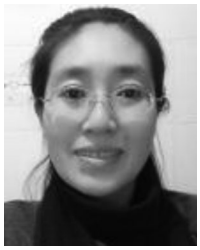
Ming-Juan Wu received the BS degree in school of information science and engineering from QuFu Normal University, China, in 2017, the MS degree in computer science and technology from QuFu Normal University, China, in 2020. Her research interests include feature selection, principal component analysis, pattern recognition, and bioinformatics.



Jin-Xing Liu (Member, IEEE) received the BS degree in electronic information and electrical engineering from Shandong University, China, in 1997, the MS degree in control theory and control engineering from QuFu Normal University, China, in 2003, and the PhD degree in computer simulation and control from the South China University of Technology, China, in 2008. From June 2011 to December 2015, he worked with Shenzhen graduate school, Harbin Institute of Technology as a postdoctoral research fellow. He is a professor with the School of Computer Science, Qufu Normal University, Rizhao, China. His research interests include pattern recognition, machine learning, and bioinformatics.



Chun-Hou Zheng (Member, IEEE) received the BS degree in physics education and the MS degree in control theory and control engineering from QuFu Normal University, China, in 1995 and 2001, respectively, and the PhD degree in pattern recognition and intelligent system from the University of Science and Technology of China, in 2006. He is currently with the School of Computer Science, Qufu Normal University, Rizhao China. His research interests include pattern recognition and bioinformatics.



Juan Wang received the BS degree in applied electronic technology from QuFu Normal University, China, in 2000, the MS degree in circuits and systems from Shandong University, China, in 2003. She is an associate professor with the School of Computer Science, Qufu Normal University, Rizhao, China. Her research interests include pattern recognition and bioinformatics.