Hindawi
BioMed Research International
Volume 2017, Article ID 5073427, 14 pages
https://doi.org/10.1155/2017/5073427

Research Article

Joint $L_{1/2}$-Norm Constraint and Graph-Laplacian PCA Method for Feature Extraction

Chun-Mei Feng,[1] Ying-Lian Gao,[2] Jin-Xing Liu,[1] Juan Wang,[1] Dong-Qin Wang,[1] and Chang-Gang Wen[1]

[1] *School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China*
[2] *Library of Qufu Normal University, Qufu Normal University, Rizhao 276826, China*

Correspondence should be addressed to Ying-Lian Gao; yinliangao@126.com

Received 30 December 2016; Revised 12 February 2017; Accepted 1 March 2017; Published 2 April 2017

Academic Editor: Jialiang Yang

Principal Component Analysis (PCA) as a tool for dimensionality reduction is widely used in many areas. In the area of bioinformatics, each involved variable corresponds to a specific gene. In order to improve the robustness of PCA-based method, this paper proposes a novel graph-Laplacian PCA algorithm by adopting $L_{1/2}$ constraint ($L_{1/2}$ gLPCA) on error function for feature (gene) extraction. The error function based on $L_{1/2}$-norm helps to reduce the influence of outliers and noise. Augmented Lagrange Multipliers (ALM) method is applied to solve the subproblem. This method gets better results in feature extraction than other state-of-the-art PCA-based methods. Extensive experimental results on simulation data and gene expression data sets demonstrate that our method can get higher identification accuracies than others.

1. Introduction

With the rapid development of gene-chip and deep-sequencing technologies, a lot of gene expression data have been generated. It is possible for biologists to monitor the expression of thousands of genes with the maturation of the sequencing technology [1–3]. It is reported that a growing body of research has been used to select the feature genes from gene expression data [4–6]. Feature extraction is a typical application of gene expression data. Cancer has become a threat to human health. Modern medicine has proved all cancers are directly or indirectly related to genes. How to identify what is believed to be related to cancer has become a hotspot in the field of bioinformatics. The major bottleneck of the development of bioinformatics is how to build an effective approach to integrate and analyze the expression data [7].

One striking feature of gene expression data is the case that the number of genes is far greater than the number of samples, commonly called the high-dimension-small-sample-size problem [8]. Typically this means that expression data are always with more than thousands of genes, while the size of samples is generally less than 100. The huge expression data make them hard to analyze, but only a small size of genes can control the gene expression. More attention has been attached to the importance of feature genes by modern biologists. Correspondingly, it is especially important how to discover these genes effectively, so many dimensionality reduction approaches are proposed.

Traditional dimensionality reduction methods have been widely used. For example, Principal Component Analysis (PCA) recombines the original data which have a certain relevance into a new set of independent indicators [9–11]. However, because of the sparsity of gene regulation, the weaknesses of traditional approaches in the field of feature extraction become increasingly evident [12, 13]. With the development of deep-sequencing technique, the inadequacy of conventional methods is emerging. Within the process of feature selection on biological data, the principal components of PCA are dense, which makes it difficult to give an objective and reasonable explanation on the significance of biology. PCA-based methods have achieved good results in the application of feature extraction [3, 12]. Although this method shows the significance of sparsity in the aspect of handling high dimensional data, there are still a lot of shortcomings in the algorithm.

(1) The high dimensionality of data poses a great challenge to the research, which is called data disaster.

(2) Facing with millions of data points, it is reasonable to consider the internal geometric structure of the data.

(3) Gene expression data usually contain a lot of outliers and noise, but the above methods cannot effectively deal with these problems.

With the development of graph theory [14] and manifold learning theory [15], the embedded structure problem has been effectively resolved. Laplacian embedding as a classical method of manifold learning has been used in machine learning and pattern recognition, whose essential idea is recovery of low dimensional manifold structure from high dimensional sampled data. The performance of feature extraction will be improved remarkably after joining Laplacian in gene expression data. In the case of maintaining the local adjacency relationship of the graph, the graph can be drawn from the high dimensional space to a low dimensional space (drawing graph). However, graph-Laplacian cannot dispose outliers.

In the field of dimensionality reduction, L_p $(0 < p < 1)$-norm was getting more and more popular to replace L_1, which was first proposed by Nie et al. [16]. Research shows that a proper value of p can achieve a more exact result for dimensionality reduction [17]. Furthermore, Xu et al. developed an simple iterative thresholding representation theory for $L_{1/2}$-norm [18], which was similar to the notable iterative soft thresholding algorithm for the solution of L_0 [19] and L_1-norm [20]. Xu et al. have shown that L_p-norm generates more better solution than L_1-norm [21]. Besides, among all regularization with p in $(0, 1/2]$, there is no obvious difference. However, when $p \in [1/2, 1)$, the smaller p is, the more effective result will be [17]. This provides a motivation to introduce $L_{1/2}$-norm constraint into original method. Since the error of each data point is calculated in the form of the square. It will also cause a lot of errors while the data contains some tiny abnormal values.

In order to solve the above problems, we propose a novel method based on $L_{1/2}$-norm constraint, graph-Laplacian PCA ($L_{1/2}$ gLPCA) which provides a good performance. In summary, the main work of this paper is as follows. (1) The error function based on $L_{1/2}$-norm is used to reduce the influence of outliers and noise. (2) Graph-Laplacian is introduced to recover low dimensional manifold structure from high dimensional sampled data.

The remainder of the paper is organized as follows. Section 2 provides some related work. We present our formulation and algorithm for $L_{1/2}$-norm constraint graph-Laplacian PCA in Section 3. We evaluate our algorithm on both simulation data and real gene expression data in Section 4. The correlations between the identified genes and cancer data are also included. The paper is concluded in Section 5.

2. Related Work

2.1. Principal Component Analysis. In the field of bioinformatics, the principal components (PCs) of PCA are used to select feature genes. Assume $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbf{R}^{m\times n}$ is the input data matrix, which contains the collection of n data column vectors and m dimension space. Traditional PCA approaches recombine the original data which have a certain relevance into a new set of independent indicators [9]. More specifically, this method reduces the input data to k-dim ($k < n$) subspace by minimizing:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & \left\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\right\|_F^2 \\ \text{s.t.} \quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}, \end{aligned} \tag{1}$$

where each column of $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_k) \in \mathbf{R}^{m\times k}$ is the principal directions and $\mathbf{V}^T = (\mathbf{v}_1, \dots, \mathbf{v}_n) \in \mathbf{R}^{k\times n}$ is the projected data points in the new subspace.

2.2. Graph-Laplacian PCA. Since the traditional PCA has not taken into account the intrinsic geometrical structure within input data, the mutual influences among data may be missed during a research project [9]. With the increasing popularity of the manifold learning theory, people are becoming aware that the intrinsic geometrical structure is essential for modeling input data [15]. It is a well-known fact that graph-Laplacian is the fastest approach in the manifold learning method [14]. The essential idea of graph-Laplacian is to recover low dimensional manifold structure from high dimensional sampled data. PCA closely relates to K-means clustering [22]. The principal components V are also the continuous solution of the cluster indicators in the K-means clustering method. Thus, it provides a motivation to embed Laplacian to PCA whose primary purpose is clustering [23, 24]. Let symmetric weight matrix $\mathbf{W} \in \mathbf{R}^{n\times n}$ be the nearest neighbor graph where $\mathbf{W}_{ij}$ is the weight of the edge connecting vertices i and j. The value of $\mathbf{W}_{ij}$ is set as follows:

$$\mathbf{W}_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \in \mathbf{N}_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathbf{N}_k(\mathbf{x}_i), \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathbf{N}_k(\mathbf{x}_i)$ is the set of k nearest neighbors of $\mathbf{x}_i$. $\mathbf{V}^T = (\mathbf{v}_1, \dots, \mathbf{v}_n) \in R^{k\times n}$ is supposed as the embedding coordinates of the data and $\mathbf{D} = \operatorname{diag}(\mathbf{d}_1, \dots, \mathbf{d}_n)$ is defined as a diagonal matrix and $\mathbf{d}_i = \sum_j \mathbf{W}_{ij}$. $\mathbf{V}$ can be obtained by minimizing:

$$\begin{aligned} \min_{\mathbf{V}} \quad & \sum_{i,j=1}^{n} \left\|\mathbf{v}_i - \mathbf{v}_j\right\|^2 \mathbf{W}_{ij} = \operatorname{tr}\left(\mathbf{V}^T(\mathbf{D} - \mathbf{W})\mathbf{V}\right) \\ & = \operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right) \\ \text{s.t.} \quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}, \end{aligned} \tag{3}$$

where $\mathbf{d}_i$ is the column or row sums of $\mathbf{W}$ and $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is named as Laplacian matrix. Simply put, in the case of maintaining the local adjacency relationship of the graph, the graph can be drawn from the high dimensional space to a low dimensional space (drawing graph). In the view of the function of graph-Laplacian, Jiang et al. proposed a model named graph-Laplacian PCA (gLPCA), which incorporates graph structure encoded in $\mathbf{W}$ [23]. This model can be considered as follows:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & J = \left\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\right\|_F^2 + \alpha \operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right) \\ \text{s.t.} \quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}, \end{aligned} \tag{4}$$

where $\alpha \geq 0$ is a parameter adjusting the contribution of the two parts. This model has three aspects. (a) It is a data representation, where $\mathbf{X} \simeq \mathbf{U}\mathbf{V}^T$. (b) It uses $\mathbf{V}$ to embed manifold learning. (c) This model is a nonconvex problem but has a closed-form solution and can be efficient to work out.

In (4), from the perspective of data point, it can be rewritten as follows:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & J = \sum_{j=1}^{n}\left(\left\|\mathbf{X}_n - \mathbf{U}\mathbf{v}_n{}^T\right\|_F^2 + \alpha \operatorname{tr}\left(\mathbf{v}_n{}^T\mathbf{L}\mathbf{v}_n\right)\right) \\ \text{s.t.} \quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}. \end{aligned} \tag{5}$$

In this formula, the error of each data point is calculated in the form of the square. It will also cause a lot of errors while the data contains some tiny abnormal values. Thus, the author formulates a robust version using $L_{2,1}$-norm as follows:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & \left\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\right\|_{2,1} + \alpha \operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right) \\ \text{s.t.} \quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}, \end{aligned} \tag{6}$$

but the major contribution of $L_{2,1}$-norm is to generate sparse on rows, in which the effect is not so obvious [3, 25].

3. Proposed Algorithm

Research shows that a proper value of p can achieve a more exact result for dimensionality reduction [17]. When $p \in [1/2, 1)$, the smaller p is, the more effective result will be [17]. Then, Xu et al. developed a simple iterative thresholding representation theory for $L_{1/2}$-norm and obtained the desired results [18]. Thus, motivated by former theory, it is reasonable and necessary to introduce $L_{1/2}$-norm on error function to reduce the impact of outliers on the data. Based on the half thresholding theory, we propose a novel method using $L_{1/2}$-norm on error function by minimizing the following problem:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & \left\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\right\|_{1/2}^{1/2} + \alpha \operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right) \\ \text{s.t.} \quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}, \end{aligned} \tag{7}$$

where $L_{1/2}$-norm is defined as $\|\mathbf{A}\|_{1/2}^{1/2} = \sum_j^n \sum_j^m |\mathbf{a}_{ij}|^{1/2}$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbf{R}^{m\times n}$ is the input data matrix, and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_k) \in \mathbf{R}^{m\times k}$ and $\mathbf{V}^T = (\mathbf{v}_1, \ldots, \mathbf{v}_n) \in \mathbf{R}^{k\times n}$ are the principal directions and the subspace of projected data, respectively. We call this model graph-Laplacian PCA based on $L_{1/2}$-norm constraint ($L_{1/2}$ gLPCA).

At first, the subproblems are solved by using the Augmented Lagrange Multipliers (ALM) method. Then, an efficient updating algorithm is presented to solve this optimization problem.

3.1. Solving the Subproblems

3.1. Solving the Subproblems. ALM is used to solve the subproblem. Firstly, an auxiliary variable is introduced to rewrite the formulation (4) as follows:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V},\mathbf{S}} \quad & \|\mathbf{S}\|_{1/2}^{1/2} + \alpha \operatorname{tr} \mathbf{V}^T(\mathbf{D} - \mathbf{W})\mathbf{V}, \\ \text{s.t.} \quad & \mathbf{S} = \mathbf{X} - \mathbf{U}\mathbf{V}^T, \ \mathbf{V}^T\mathbf{V} = \mathbf{I}. \end{aligned} \tag{8}$$

The augmented Lagrangian function of (8) is defined as follows:

$$\begin{aligned} L_\mu(\mathbf{S}, \mathbf{U}, \mathbf{V}, \mathbf{\Lambda}) = {} & \|\mathbf{S}\|_{1/2}^{1/2} + \operatorname{tr} \mathbf{\Lambda}^T\left(\mathbf{S} - \mathbf{X} + \mathbf{U}\mathbf{V}^T\right) \\ & + \frac{\mu}{2}\left\|\mathbf{S} - \mathbf{X} + \mathbf{U}\mathbf{V}^T\right\|_F^2 \\ & + \alpha \operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right), \\ \text{s.t.} \quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}, \end{aligned} \tag{9}$$

where $\mathbf{\Lambda}$ is Lagrangian multipliers and μ is the step size of update. By mathematical deduction, the function of (9) can be rewritten as

$$\begin{aligned} L_\mu(\mathbf{S}, \mathbf{U}, \mathbf{V}, \mathbf{\Lambda}) = {} & \|\mathbf{S}\|_{1/2}^{1/2} + \frac{\mu}{2}\left\|\mathbf{S} - \mathbf{X} + \mathbf{U}\mathbf{V}^T + \frac{\mathbf{\Lambda}}{\mu}\right\|_F^2 \\ & + \alpha \operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right), \\ \text{s.t.} \quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}. \end{aligned} \tag{10}$$

The general approach of (10) consists of the following iterations:

$$\begin{aligned} \mathbf{S}^{k+1} &= \arg\min_{\mathbf{S}} L_\mu\left(\mathbf{S}, \mathbf{U}^k, \mathbf{V}^k, \mathbf{\Lambda}^k\right), \\ \mathbf{V}^{k+1} &= \left(\mathbf{v}_1, \ldots, \mathbf{v}_k\right), \\ \mathbf{U}^{k+1} &= \mathbf{M}\mathbf{V}^k, \\ \mathbf{\Lambda}^{k+1} &= \mathbf{\Lambda}^k + \mu\left(\mathbf{S}^{k+1} - \mathbf{X} + \mathbf{U}^k\mathbf{V}^{T^k}\right), \\ \mu^{k+1} &= \rho\mu^k. \end{aligned} \tag{11}$$

Then, the details to update each variable in (11) are given as follows.

Updating $\mathbf{S}$. At first, we solve $\mathbf{S}$ while fixing $\mathbf{U}$ and $\mathbf{V}$. The update of $\mathbf{S}$ relates the following issue:

$$\mathbf{S}^{k+1} = \arg\min_{\mathbf{S}} \|\mathbf{S}\|_{1/2}^{1/2} + \frac{\mu}{2}\left\|\mathbf{S} - \mathbf{X} + \mathbf{U}^k\mathbf{V}^{T^k} + \frac{\mathbf{\Lambda}^k}{\mu}\right\|_F^2, \tag{12}$$

which is the proximal operator of $L_{1/2}$-norm. Since this formulation is a nonconvex, nonsmooth, non-Lipschitz, and complex optimization problem; an iterative half thresholding approach is used for fast solution of $L_{1/2}$-norm and summarizes according to the following lemma [18].

Lemma 1. *The proximal operator of $L_{1/2}$-norm minimizes the following problem:*

$$\min_{\mathbf{X}\in\mathbf{R}^{m\times n}} \|\mathbf{X}-\mathbf{A}\|_F^2 + \lambda \|\mathbf{X}\|_{1/2}^{1/2}, \tag{13}$$

which is given by

$$\mathbf{X}^* = \mathbf{H}_\lambda(\mathbf{A}) = \mathbf{U}\operatorname{diag}\left(\mathbf{H}_\lambda(\sigma)\right)\mathbf{V}^T, \tag{14}$$

where $\mathbf{H}_\lambda(\sigma) := (h_\lambda(\sigma_1), h_\lambda(\sigma_2), \ldots, h_\lambda(\sigma_n))^T$ and $h_\lambda(\sigma_i)$ is the half threshold operator and defined as follows:

$$h_\lambda(\sigma_i) = \begin{cases} \dfrac{2}{3}\sigma_i\left(1+\cos\left(\dfrac{2\pi}{3}-\dfrac{2}{3}\psi\lambda(\sigma_i)\right)\right), & \text{if } |\sigma_i| > \dfrac{\sqrt[3]{54}}{4}\lambda^{2/3} \\ 0, & \text{otherwise,} \end{cases} \tag{15}$$

where $\psi\lambda(\sigma_i) = \arccos((\lambda/8)(|\sigma_i|/3)^{-2/3})$.

Solving **U** *and* **V**. Here, we solve **U** while fixing others. The update of **U** amounts to solving

$$\mathbf{U}^{k+1} = \arg\min_{\mathbf{U}} \frac{\mu}{2}\left\|\mathbf{S}^k - \mathbf{X} + \mathbf{U}^k\mathbf{V}^{T^k} + \frac{\mathbf{\Lambda}^k}{\mu}\right\|_F^2. \tag{16}$$

Letting $\mathbf{X}-\mathbf{S}-\mathbf{\Lambda}/\mu = \mathbf{M}$, (16) becomes $\mathbf{U}^{k+1} = \arg\min_{\mathbf{U}}(\mu/2)\|\mathbf{M}-\mathbf{U}\mathbf{V}^{T^k}\|_F^2$, taking partial derivatives of **U** as follows:

$$\frac{\partial J}{\partial \mathbf{U}} = -\mu\left(\mathbf{M}-\mathbf{U}\mathbf{V}^{T^k}\right)\mathbf{V}. \tag{17}$$

Setting the partial derivatives to 0, we have

$$\mathbf{U}^{k+1} = \mathbf{M}\mathbf{V}^k. \tag{18}$$

Then, we solve **V** while fixing others. Similarly, letting $\mathbf{X}-\mathbf{S}-\mathbf{\Lambda}/\mu = \mathbf{M}$, $\mathbf{U} = \mathbf{M}\mathbf{V}$, the update of **V** can be listed as follows:

$$\begin{aligned} \mathbf{V}^{k+1} &= \arg\min_{\mathbf{V}} \frac{\mu}{2}\left\|\mathbf{M}-\mathbf{M}\mathbf{V}\mathbf{V}^T\right\|_F^2 + \alpha\operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right), \\ \text{s.t.}\quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}. \end{aligned} \tag{19}$$

By some algebra, we have

$$\begin{aligned} \mathbf{V}^{k+1} &= \arg\min_{\mathbf{V}} \left\|\mathbf{M}-\mathbf{M}\mathbf{V}\mathbf{V}^T\right\|_F^2 + \frac{2\alpha}{\mu}\operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right) \\ &= \arg\min_{\mathbf{V}} \operatorname{tr}\left(\mathbf{M}\mathbf{M}^T\right) - 2\left(\sqrt{\operatorname{tr}\left(\mathbf{M}\mathbf{M}^T\right)}\right)^2 \\ &\quad + \frac{2\alpha}{\mu}\operatorname{tr}\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right) \\ &= \arg\min_{\mathbf{V}} \operatorname{tr}\mathbf{V}^T\left(-\mathbf{M}^T\mathbf{M}+\frac{2\alpha}{\mu}\mathbf{L}\right)\mathbf{V}. \end{aligned} \tag{20}$$

Therefore, (19) can be rewritten as follows:

$$\begin{aligned} \mathbf{V}^{k+1} &= \arg\min_{\mathbf{V}} \operatorname{tr}\mathbf{V}^T\left(-\mathbf{M}^T\mathbf{M}+\frac{2\alpha}{\mu}\mathbf{L}\right)\mathbf{V}, \\ \text{s.t.}\quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}. \end{aligned} \tag{21}$$

Thus, the optimal $\mathbf{V}^{k+1}$ can be obtained by calculating eigenvectors

$$\mathbf{V}^{k+1} = \left(\mathbf{v}_1, \ldots, \mathbf{v}_k\right), \tag{22}$$

which corresponds to the first k smallest eigenvalues of the matrix $G_\alpha = -\mathbf{M}^T\mathbf{M} + 2\alpha\mathbf{L}/\mu$.

Updating $\mathbf{\Lambda}$ *and* μ. The update of $\mathbf{\Lambda}$ and μ is standard:

$$\begin{aligned} \mathbf{\Lambda}^{k+1} &= \mathbf{\Lambda}^k + \mu\left(\mathbf{S}^{k+1} - \mathbf{X} + \mathbf{U}^k\mathbf{V}^{T^k}\right), \\ \mu^{k+1} &= \rho\mu^k, \end{aligned} \tag{23}$$

where $\rho > 1$ is used to update the parameter μ. Since the value of ρ is usually bigger than 1, and over a large number of experiments, we find $\rho = 1.1\sim1.5$ are good choice. We selected $\rho = 1.2$ in such practice conditions.

The complete procedure is summarized in Algorithm 1.

3.2. Properties of Algorithm. We set $\rho = 1.2$ through all our gene expression data experiments. Whereas we introduce σ_m, σ_l is the largest eigenvalue of matrix $\mathbf{M}^T\mathbf{M}$ and **L** to normalize them, respectively. Setting

$$\frac{2\alpha}{\mu} = \frac{\beta}{1-\beta}\frac{\sigma_m}{\sigma_l}, \tag{24}$$

where β is the parameter to substitute for α, (20) can be rewritten as

$$\begin{aligned} \mathbf{V} &= \arg\min_{\mathbf{V}} \operatorname{tr}\mathbf{V}^T\left[(1-\beta)\left(\mathbf{I}-\frac{\mathbf{M}^T\mathbf{M}}{\sigma_m}\right)+\frac{2\beta}{\mu}\frac{\mathbf{L}}{\sigma_l}\right]\mathbf{V}, \\ s.t.\quad & \mathbf{V}^T\mathbf{V} = \mathbf{I}. \end{aligned} \tag{25}$$

Therefore, the solution of **V** can be expressed by the eigenvectors of G_β:

$$G_\beta = (1-\beta)\left(\mathbf{I}-\frac{\mathbf{M}^T\mathbf{M}}{\sigma_m}\right)+\frac{2\beta}{\mu}\frac{\mathbf{L}}{\sigma_l}. \tag{26}$$

It is easy to see that β should be in the range $0 \le \beta \le 1$. Without $L_{1/2}$-norm, there will be standard PCA if $\beta = 0$. Similarly, when $\beta = 1$, it reduces to Laplacian embedding.

Furthermore, we rewrite the matrix G_β as follows:

$$G_\beta = (1-\beta)\left(\mathbf{I}-\frac{\mathbf{M}^T\mathbf{M}}{\sigma_m}\right)+\frac{2\beta}{\mu}\left(\frac{\mathbf{L}}{\sigma_l}+\frac{\mathbf{e}\mathbf{e}^T}{n}\right), \tag{27}$$

where $\mathbf{e} = (1\cdots1)^T$ is an eigenvector of G_β: $G_\beta\mathbf{e} = (1-\beta)\mathbf{e}$. We have $\mathbf{M}\mathbf{e} = 0$, because **X** is centered and it is easy to

```
Input: Data matrix $\mathbf{X} \in \mathbf{R}^{m\times n}$;
       weight matrix: $\mathbf{W} \in R^{m\times n}$;
       parameters: $\beta$, $\rho$, $k$, $\mu$.
Output: Optimized matrix: $\mathbf{U}$, $\mathbf{V}$.
Initialization: $\mathbf{S} = \mathbf{\Lambda} = 0$, $\mathbf{U} = 0$, $\mathbf{V} = 0$.
repeat
Step 1. Update S and fix the others by $\mathbf{S}^{k+1} = \arg\min_{\mathbf{s}} \|\mathbf{S}\|_{1/2}^{1/2} + (\mu/2)\|\mathbf{S} - \mathbf{X} + \mathbf{U}\mathbf{V}^T + \mathbf{\Lambda}/\mu\|_F^2$.
Step 2. Update U and fix the others by $\mathbf{U}^{k+1} = \mathbf{M}\mathbf{V}^k$.
Step 3. Update V and fix the others by $\mathbf{V}^{k+1} = (\mathbf{v}_1, \dots, \mathbf{v}_r)$.
Step 4. Update Λ, μ and fix the others by
$\mathbf{\Lambda}^{k+1} = \mathbf{\Lambda}^k + \mu(\mathbf{S}^{k+1} - \mathbf{X} + \mathbf{U}^k\mathbf{V}^{T^k})$
$\mu^{k+1} = \rho\mu^k$
until converge
```

Algorithm 1: Procedure of $L_{1/2}$ gLPCA.

see that $\mathbf{M} = \mathbf{X} - \mathbf{S} - \mathbf{\Lambda}/\mu$ is centered. G_β is semipositive definite, because σ_m is the biggest eigenvalue of $\mathbf{M}^T\mathbf{M}$; thus $\mathbf{I} - \mathbf{M}^T\mathbf{M}/\sigma_m$ is semipositive definite. Meanwhile, it is easy to see that $\mathbf{L}$ is semipositive definite. Since G_β is a symmetric real matrix that eigenvectors are mutually orthogonal, thus $\mathbf{e}$ is orthogonal to others. Although we apply $\mathbf{e}\mathbf{e}^T/n$ in the Laplacian matrix part, the eigenvectors and eigenvalues do not change, which guarantees that the lowest k eigenvectors do not include $\mathbf{e}$.

4. Experiments

In this section, we compare our algorithm with Laplacian embedding (LE) [26], PCA [9], L_0 PCA, L_1 PCA [12], gLPCA, and RgLPCA [23] on simulation data and real gene expression data, respectively, to verify the performance of our algorithm. Among them, PCA and LE are obtained by adjusting the parameters of gLPCA $\beta = 0$ and $\beta = 1$, respectively. Since our algorithm is not sensitive to parameter mu in practice. In the first subsection, we provide the source of simulation data and experimental comparison results. The experimental results and the function of selected genes on real gene expression data with different methods are compared in the next two subsections.

4.1. Results on Simulation Data

4.1.1. Data Source. Here, we describe a method to produce simulation data. Supposing we generate the data matrix $\mathbf{A} \in \mathbf{R}^{k\times j}$, where $k = 2000$ and $j = 10$ are the number of genes and samples, respectively, the simulation data are generated as $\mathbf{A} \sim (0, \Sigma_4)$. Let $\tilde{\mathbf{v}}_1 \sim \tilde{\mathbf{v}}_4$ be four 2000-dimensional vectors; for instance, $\tilde{\mathbf{v}}_{1k} = 1$, $k = 1, \dots, 50$, and $\tilde{\mathbf{v}}_{1k} = 0$, $k = 51, \dots, 2000$; $\tilde{\mathbf{v}}_{2k} = 1$, $k = 51, \dots, 100$, and $\tilde{\mathbf{v}}_{2k} = 0$, $k \neq 51, \dots, 100$; $\tilde{\mathbf{v}}_{3k} = 1$, $k = 101, \dots, 150$, and $\tilde{\mathbf{v}}_{3k} = 0$, $k \neq 101, \dots, 150$; $\tilde{\mathbf{v}}_{4k} = 1$, $k = 151, \dots, 200$, and $\tilde{\mathbf{v}}_{4k} = 1$, $k \neq 151, \dots, 200$. Given a matrix $\mathbf{E} \sim N(0, 1)$ as a noise matrix with 2000-dimension and different Signal-to-Noise Ratio (SNR), which is added into $\tilde{v}$, the four eigenvectors of Σ_4 can be expressed as $\tilde{\mathbf{v}}_k = \tilde{\mathbf{v}}_k/\|\tilde{\mathbf{v}}_k\|$, $k = 1, 2, 3, 4$. Let the four eigenvectors dominate; the eigenvalues of $\mathbf{A}$ can be denoted as $c_1 = 400$, $c_2 = 300$, $c_3 = 200$, $c_4 = 100$, and $c_k = 1$ for $k = 5, \dots, 2000$.

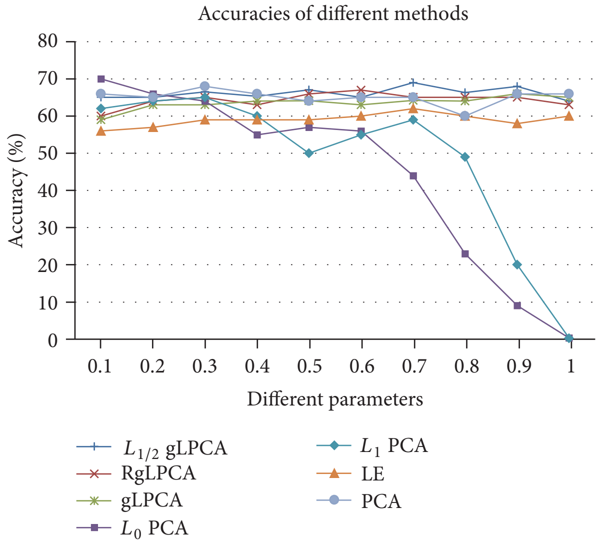

Figure 1: The accuracy of different methods on simulation data with different parameters.

4.1.2. Detailed Results on Simulation Data. In order to give more accurate experiment results, the average values of the results of 30 times are adopted. For fairness and uniformity, 200 genes are selected by the five methods with their unique parameters. Here, we show the accuracy (%) of these methods. In Figure 1, two factors as two different axes are in the figure. In Figure 2, x-axis is the number of samples. x-axis is the value of parameter μ. The accuracy is defined as follows:

$$\text{Accuracy} = \frac{1}{t}\sum_{i=1}^{t}\text{Acc}_i \times 100\%, \qquad (28)$$

TABLE 1: The average accuracy and variance of different methods on simulation data with different parameters.

Methods	$L_{1/2}$ gLPCA	RgLPCA	gLPCA	L_0 PCA	L_1 PCA	PCA	LE
Average accuracy (%)	66.12	65.47	63.53	44.43	48.43	59.00	65.10
Variance	1.48	1.62	1.76	23.60	20.30	1.61	1.97

TABLE 2: The average accuracy and variance of different methods on simulation data with different numbers of samples.

Methods	$L_{1/2}$ gLPCA	RgLPCA	gLPCA	L_0 PCA	L_1 PCA	PCA	LE
Average accuracy (%)	70.25	68.25	67.90	67.30	69.20	58.62	69.60
Variance	2.58	3.84	4.41	3.52	2.23	1.79	2.50

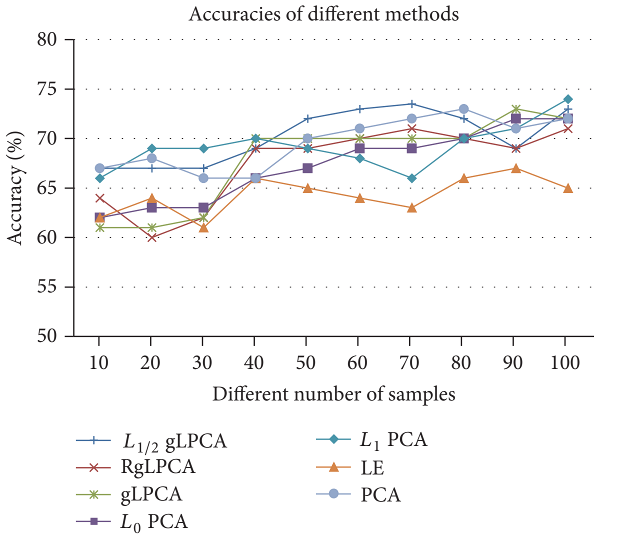

FIGURE 2: The accuracy of different methods on simulation data with different numbers of samples.

where t is the iterative times and Acc_i is the identification accuracy of the ith time. We define Acc as follows:

$$\mathrm{Acc} = \frac{1}{r}\sum_{j=1}^{r}\delta\left(I_j, \mathrm{map}\left(I_j\right)\right), \qquad (29)$$

where r denotes the number of genes, $\delta(m,n)$ is a function that equals to 0 if $m \neq n$ and equals to 1 if $m = n$. We use the function map(I) to map the identification of labels. In Figure 1, we show the average accuracies of the seven methods with different sparse parameters while the simulation data is 2000 × 10 and the average accuracy with all parameters is listed in Table 1. In general, if the algorithm is more sensitive to noise and outliers, the deviation will be greater and the accuracy will be greatly reduced. It is worthy to notice that $L_{1/2}$ gLPCA works better than other six methods with higher identification accuracies. This means that our algorithm has lower sensitivity to noise and outliers. This table clearly displays the detail of the identification accuracies in different sparse parameters; our method indicates the superiority when the parameter is larger than 0.4 and the curve is more stable. The accuracy of L_0 PCA and L_1 PCA starts a precipitous decline when the parameter is larger than 0.7 and 0.8. Compared with L_0 PCA and L_1 PCA, the methods of $L_{1/2}$ gLPCA, RgLPCA, gLPCA, PCA, and LE are not sensitive to the parameter, so there is no substantial change. The stability and average accuracy of various methods can be seen from Table 1.

Furthermore, the number of samples in real gene expression data has a significant influence on the identification accuracy when we select feature gene. Based on this theory, we test different numbers of samples with their best parameters and the average values of the results of 30 times. From the results of Figure 1, we select 0.8 as the parameters of $L_{1/2}$ gLPCA, gLPCA, RgLPCA, PCA, and LE. For L_0 PCA and L_1 PCA, we do not change its parameters, since it can obtain the best result from the author's description. The details of average identification accuracies which use seven methods with different sample numbers can be seen from Figure 2. As seen in Figure 2, the accuracy of $L_{1/2}$ gLPCA is generally better than other methods and increases with the increase of the number of samples. Besides, Table 2 shows the average accuracy and variance of seven different methods on simulation data with different number of samples. From Table 2, our approach performs better than other methods, even though, in the case of a small number of samples, the accuracy is still high.

4.2. Results on Gene Expression Data. In this subsection, the features (genes) are selected by these methods and sent to ToppFun to detect the gene-set enrichment analysis, which is a type of GOTermFinder [27]. The primary role of GOTermFinder is to discover the common of large amounts of gene expression data. The analysis of GOTermFinder provides critical information for the experiment of feature extraction. It is available publicly at https://toppgene.cchmc.org/enrichment.jsp. We set P value cutoff to 0.01 through all the experiment. For fair comparison, about $L_{1/2}$ gLPCA, RgLPCA, and gLPCA, we both set $\beta = 0.5$ to control the degree of Laplacian embedding through all experiments in this paper. When $\beta = 0$, $\beta = 1$, it results in standard PCA and LE, respectively. Since our algorithm is not sensitive to parameter μ mu in practice, we set $\mu = 0.3$ through our experiment.

4.2.1. Results on ALLAML Data. The data of ALLAML as a matrix includes 38 samples and 5000 features (genes), which are publicly available at https://sites.google.com/site/feipingnie/file. It is made up of 11 types of acute myelogenous leukemia (AML) and 27 types of acute lymphoblastic leukemia (ALL) [28]. This data contains the difference between AML and ALL, and ALL is divided into T and B cell subtypes. In this experiment, 300 genes are selected and sent to ToppFun. A series of enrichment analyses are conducted on the extracted top 500 genes corresponding to different methods. The complete experimental data have been listed as supplementary data. The P value and hit count of top nine terms about molecular function, biological process, and cellular component of ALLAML data by different methods are listed in Table 3. The P value is significance for these genes enrichment analysis in these GO terms; the smaller the P value is, the more significant these GO terms are. In this Table, the number of hits is the number of genes from input, and the P value was influenced by the number of genes from input and so on. Thus, the difference in number of hits is smaller than the difference in P value. It shows clearly that our method performs better than compared methods in 8 terms. The lower P value shows that the algorithm is less affected by noise and outliers and thus has high efficiency. If the algorithm is affected by noise and outliers significantly, the degree of gene enrichment will be reduced. Nevertheless, LE has the lowest P value in term GO: 0098552. From this table, we can see that there are 93 genes in the item of "immune response" which are selected by our method. This item can be considered as the most probable enrichment item, since it has the lowest P value. And many researches were focused on the immune status of leukemia [29–32]. Besides, 210 genes associated with leukemia are listed in an article, and 26 out of top 30 genes selected by our method can be found in this article [33]. And 30 genes selected by our method can be found in another published article [34]. The high overlap rate of these genes selected by our method with this published literature approved the effectiveness of our method.

4.2.2. Pathway Search Result on ALLAML Data. For the sake of the correlations between the selected genes and ALLAML data, the genes selected by $L_{1/2}$ gLPCA are proved based on gene-set enrichment analysis (GSEA) that is publicly available at http://software.broadinstitute.org/gsea/msigdb/annotate.jsp. We make analysis by GSEA to compute overlaps for selected genes. Figure 3 displays the pathway of hematopoietic cell lineage that has highest gene overlaps in this experiment. From Figure 3, 15 genes from our experiment are contained. Among them, HLA-DR occurs seven times. Hematopoietic cell lineage belongs to organismal systems and immune system. On the subject of acute myeloid leukemia (AML), there is consensus about the target cell within the hematopoietic stem cell hierarchy that is sensitive to leukemic transformation, or about the mechanism, that is, basic phenotypic, genotypic, and clinical heterogeneity [35]. Hematopoietic stem cell (HSC) developing from the blood-cell can undergo self-renewal and differentiate into a multilineage committed progenitor cell: one is a common lymphoid progenitor (CLP) and the other is called a common myeloid progenitor (CMP) [36]. A CLP causes the lymphoid lineage of white blood cells or leukocytes, the natural killer (NK) cells and the T and B lymphocytes. A CMP causes the myeloid lineage, which comprises the rest of the leukocytes, the erythrocytes (red blood cells), and the megakaryocytes that produce platelets important in blood clotting. Cells express a stage- and lineage-specific set of surface markers in the differentiation process. So the specific expression pattern of these genes is one way to identify the cellular stages. Related diseases include hemophilia, Bernard-Soulier syndrome, and castleman disease. In medicine, leukemia is a kind of malignant clonal disease of hematopoietic stem cells. Bone marrow transplantation is a magic weapon for the cure of leukemia, by recreating the hematopoietic system to cure leukemia. Generally speaking, when a person has problem in hematopoietic system, it might be related to leukemia [37].

4.2.3. Results on TCGA with PAAD-GE Data. As the largest public database of cancer gene information, The Cancer Genome Atlas (TCGA, https://tcgadata.nci.nih.gov/tcga/) has been producing multimodal genomics, epigenomics, and proteomics data for thousands of tumor samples across over 30 types of cancer. At the same time, as a multidimensional combination of data, five levels of data are involved, such as gene expression (GE), Protein Expression (PE), DNA Methylation (ME), DNA Copy Number (CN), and microRNA Expression (miRExp). Two disease data sets are downloaded from TCGA to be analyzed in the following two experiments. Pancreatic cancer is a type of disease that threatens human health. In this experiment, pancreatic cancer gene expression data (PAAD-GE) is analyzed by these methods. The data of PAAD-GE data as a matrix includes 180 samples and 20502 features (genes). In this subsection, we extract PAAD-GE data to complete this set of comparative experiments and 500 genes are selected and sent to ToppFun. We select top nine terms from molecular function, biological process, and cellular component by $L_{1/2}$ gLPCA and compare with other methods. The P value and hit count of these terms are listed in Table 4. It is indicated clearly in Table 4 that our method is more stable than other methods, which has lower P value in 7 terms. But in terms GO:0045047 and GO:0072599, PCA performs better than other methods. Nevertheless, $L_{1/2}$ gLPCA has the same P value with gLPCA in terms GO:0045047 and GO:0072599. 196 genes in the item of "extracellular space" are selected by our method.

4.2.4. Pathway Search Result on PAAD-GE Data. Similarly as the last experiment, we send our result to GSEA and list the highest genes overlap pathway map in Figure 4. In 1982, Ohhashi reported 4 cases with unique clinical pathological features and is different from normal pancreatic cancer cases, and these 4 cases belong to a completely new clinical type, known as "mucus production type carcinoma (mucin-producing carcinoma, M-pC)." Focal adhesion belongs to cellular processes and cellular community. More specifically, cell-matrix adhesions play important roles in biological processes including cell motility, cell proliferation, cell differentiation, regulation of gene expression, and cell survival. At the cell-extracellular matrix contact points, specialized structures are created and termed focal adhesions, where bundles of actin filaments are fixed to transmembrane receptors of the integrin family through a multimolecular complex of junctional plaque proteins. Integrin signaling is dependent on the nonreceptor tyrosine kinase activities of the FAK

Table 3: Enrichment analysis of the top 500 genes in the ALLAML data corresponding to different methods.

ID	Name	$L_{1/2}$ gLPCA		RgLPCA		gLPCA		L_0 PCA		L_1 PCA		PCA		LE	
		P-value	Hit	P-value	Hit	P-value	Hit	P-value	Hit	P-value	Hit	P-value	Hit	P-value	Hit
GO:0006955	Immune response	**1.34E − 36**	93	2.51E − 34	91	1.20E − 34	91	2.45E − 31	87	5.14E − 32	89	4.05E − 35	91	1.98E − 35	91
GO:0002684	Positive regulation of immune system process	**2.44E − 29**	67	2.17E − 25	63	1.24E − 26	64	3.60E − 28	66	1.56E − 27	66	8.98E − 28	65	3.45E − 28	66
GO:0098552	Side of membrane	3.80E − 25	46	5.19E − 34	45	2.70E − 22	43	2.23E − 20	41	7.25E − 21	42	2.01E − 23	44	**4.24E − 26**	47
GO:0009897	External side of plasma membrane	**1.83E − 17**	30	6.34E − 16	29	9.51E − 14	26	1.14E − 13	26	1.41E − 12	25	1.31E − 16	29	1.83E − 14	26
GO:0005615	Extracellular space	**2.01E − 17**	63	8.37E − 15	60	2.39E − 14	58	6.12E − 16	61	3.52E − 13	57	2.27E − 16	61	4.74E − 16	61
GO:0005764	Lysosome	**3.49E − 17**	38	7.43E − 16	37	5.46E − 14	34	1.20E − 14	35	9.22E − 11	30	1.08E − 15	36	3.49E − 16	37
GO:0009986	Cell surface	**3.58E − 17**	48	4.82E − 15	45	4.68E − 13	42	6.13E − 13	42	5.58E − 12	41	6.58E − 16	46	3.58E − 16	46
GO:0042277	Peptide binding	**5.03E − 14**	25	5.92E − 13	24	2.85E − 11	22	3.33E − 11	22	7.54E − 08	18	3.09E − 12	23	1.80E − 10	21
GO:0033218	Amide binding	**7.37E − 14**	26	4.34E − 12	24	3.44E − 11	23	4.04E − 11	23	7.36E − 08	19	3.95E − 12	24	2.04E − 10	22

Table 4: Enrichment analysis of the top 500 genes in the PAAD-GE data corresponding to different methods.

ID	Name	$L_{1/2}$ gLPCA		RgLPCA		gLPCA		L_0 PCA		L_1 PCA		PCA		LE	
		P value	Hit	*P* value	Hit	*P* value	Hit	*P* value	Hit	*P* value	Hit	*P* value	Hit	*P* value	Hit
GO:0005615	Extracellular space	**3.20E − 93**	196	3.56*E* − 80	183	2.18*E* − 72	173	2.742*E*−61	160	7.82*E* − 61	161	1.44*E* − 58	157	3.20*E* − 89	191
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	**2.79E − 86**	67	6.82*E* − 75	56	1.37*E* − 73	64	3.45*E* − 51	48	8.17*E* − 56	51	7.45*E* − 82	63	2.76*E* − 56	51
GO:0070972	Protein localization to endoplasmic reticulum	**1.01E − 83**	73	2.42*E* − 72	69	6.37*E* − 71	68	5.31*E* − 48	51	4.63*E* − 52	54	2.88*E* − 76	71	4.70*E* − 51	53
GO:0006613	Cotranslational protein targeting to membrane	**1.86E − 82**	67	3.48*E* − 79	65	7.58*E* − 73	64	3.27*E* − 49	48	1.19*E* − 53	51	2.04*E* − 80	66	4.04*E* − 54	51
GO:0045047	Protein targeting to ER	**5.01E − 82**	67	5.19*E* − 74	65	2.00*E* − 71	64	6.04*E* − 49	48	2.33*E* − 53	51	**5.80E − 82**	67	7.90*E* − 54	51
GO:0022626	Cytosolic ribosome	**1.34E − 81**	68	2.13*E* − 75	64	1.44*E* − 70	62	8.30*E* − 44	47	8.45*E* − 48	50	6.34*E* − 74	66	4.01*E* − 51	52
GO:0072599	Establishment of protein localization to endoplasmic reticulum	**2.77E − 80**	67	8.15*E* − 78	66	4.44*E* − 70	64	6.44*E* − 48	48	3.09*E* − 52	51	**3.20E − 80**	67	1.05*E* − 52	51
GO:0005198	Structural molecule activity	**1.82E − 68**	126	2.46*E* − 65	124	6.14*E* − 63	121	3.62*E* − 52	110	5.16*E* − 54	113	3.32*E* − 65	124	1.03*E* − 54	113
GO:0044391	Ribosomal subunit	**5.14E − 64**	69	5.18*E* − 60	65	3.22*E* − 56	63	2.86*E* − 37	49	1.42*E* − 40	52	1.58*E* − 63	68	3.70*E* − 42	53

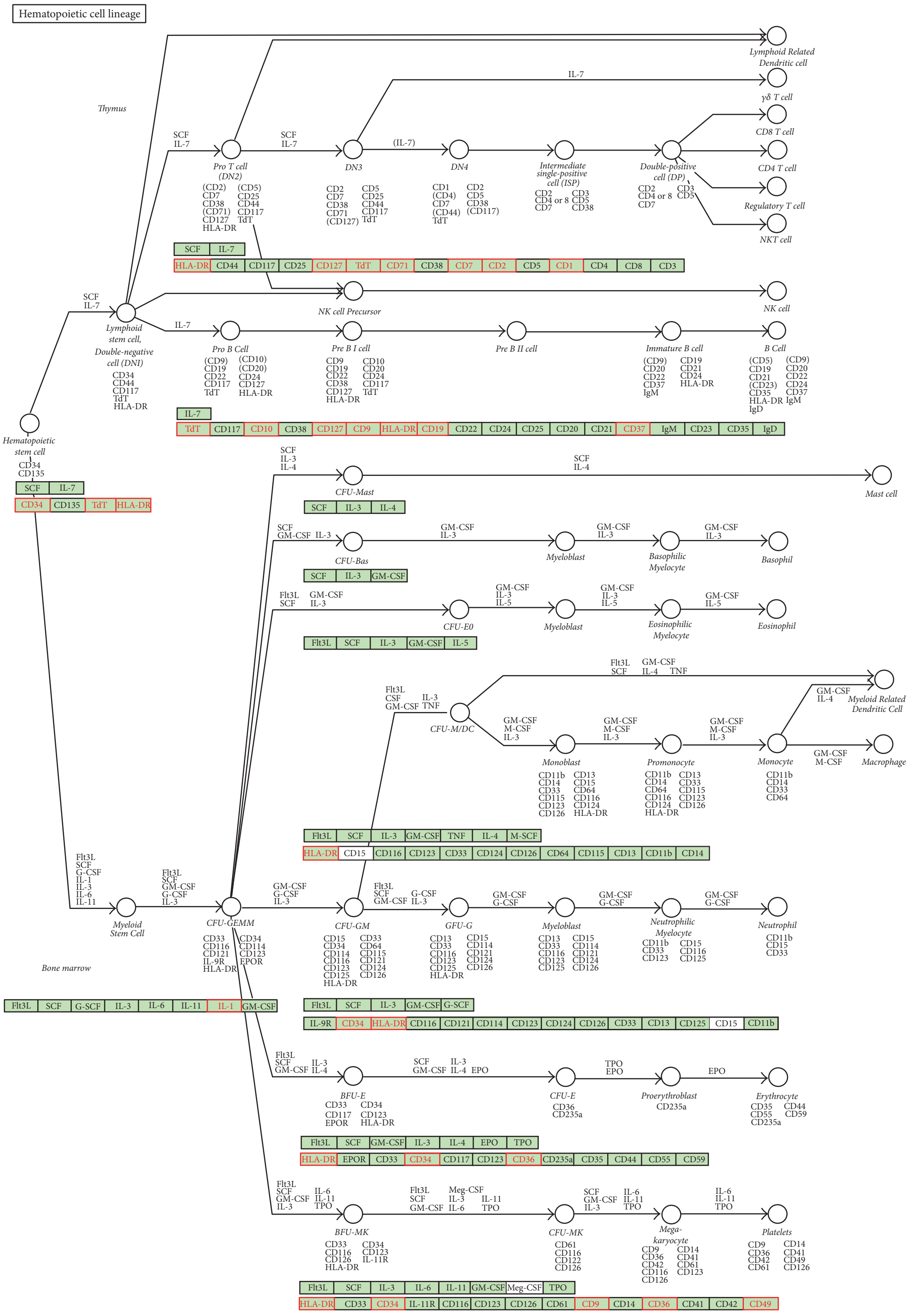
Hematopoietic cell lineage
Thymus
Lymphoid Related Dendritic cell
IL-7
γδ T cell
CD8 T cell
CD4 T cell
Regulatory T cell
NKT cell
SCF IL-7
Pro T cell (DN2)
(CD2) (CD5) CD7 CD25 CD38 CD44 (CD71) CD117 CD127 TdT HLA-DR
SCF IL-7
DN3
CD2 CD5 CD7 CD25 CD38 CD44 CD71 CD117 (CD127) TdT
(IL-7)
DN4
CD1 CD2 (CD4) CD5 CD7 CD38 (CD44) (CD117) TdT
Intermediate single-positive cell (ISP)
CD2 CD3 CD4 or 8 CD5 CD7 CD38
Double-positive cell (DP)
CD2 CD3 CD4 or 8 CD5 CD7
SCF IL-7
HLA-DR CD44 CD117 CD25 CD127 TdT CD71 CD38 CD7 CD2 CD5 CD1 CD4 CD8 CD3
SCF IL-7
Lymphoid stem cell, Double-negative cell (DNI)
CD34 CD44 CD117 TdT HLA-DR
NK cell Precursor
NK cell
IL-7
Pro B Cell
(CD9) (CD10) CD19 (CD20) CD22 CD24 CD117 CD127 TdT HLA-DR
Pre B I cell
CD9 CD10 CD19 CD20 CD22 CD24 CD38 CD117 CD127 TdT HLA-DR
Pre B II cell
Immature B cell
(CD9) CD19 CD20 CD21 CD22 CD24 CD37 HLA-DR IgM
B Cell
(CD5) (CD9) CD19 CD20 CD21 CD22 (CD23) CD24 CD35 CD37 HLA-DR IgM IgD
IL-7
TdT CD117 CD10 CD38 CD127 CD9 HLA-DR CD19 CD22 CD24 CD25 CD20 CD21 CD37 IgM CD23 CD35 IgD
Hematopoietic stem cell
CD34 CD135
SCF IL-7
CD34 CD135 TdT HLA-DR
SCF IL-3 IL-4
CFU-Mast
SCF IL-4
Mast cell
SCF IL-3 IL-4
SCF GM-CSF IL-3
CFU-Bas
GM-CSF IL-3
Myeloblast
GM-CSF IL-3
Basophilic Myelocyte
GM-CSF IL-3
Basophil
SCF IL-3 GM-CSF
Flt3L SCF
GM-CSF IL-3
CFU-E0
GM-CSF IL-3 IL-5
Myeloblast
GM-CSF IL-3 IL-5
Eosinophilic Myelocyte
GM-CSF IL-5
Eosinophil
Flt3L SCF IL-3 GM-CSF IL-5
Flt3L SCF GM-CSF IL-4 TNF
Myeloid Related Dendritic Cell
GM-CSF IL-4
Flt3L CSF GM-CSF IL-3 TNF
CFU-M/DC
GM-CSF M-CSF IL-3
Monoblast
CD11b CD13 CD14 CD15 CD33 CD64 CD115 CD116 CD123 CD124 CD126 HLA-DR
GM-CSF M-CSF IL-3
Promonocyte
CD11b CD13 CD14 CD33 CD64 CD115 CD116 CD123 CD124 CD126 HLA-DR
GM-CSF M-CSF IL-3
Monocyte
CD11b CD14 CD33 CD64
GM-CSF M-CSF
Macrophage
Flt3L SCF IL-3 GM-CSF TNF IL-4 M-SCF
HLA-DR CD15 CD116 CD123 CD33 CD124 CD126 CD64 CD115 CD13 CD11b CD14
Flt3L SCF G-CSF IL-1 IL-3 IL-6 IL-11
Myeloid Stem Cell
Flt3L SCF GM-CSF G-CSF IL-3
CFU-GEMM
CD33 CD34 CD116 CD114 CD121 CD123 IL-9R EPOR HLA-DR
Bone marrow
Flt3L SCF G-SCF IL-3 IL-6 IL-11 IL-1 GM-CSF
GM-CSF G-CSF IL-3
CFU-GM
CD15 CD33 CD34 CD64 CD114 CD115 CD116 CD121 CD123 CD124 CD125 CD126 HLA-DR
Flt3L SCF G-CSF GM-CSF IL-3
GFU-G
CD13 CD15 CD33 CD114 CD116 CD121 CD123 CD124 CD125 CD126 HLA-DR
GM-CSF G-CSF
Myeloblast
CD13 CD15 CD33 CD114 CD116 CD121 CD123 CD124 CD125 CD126
GM-CSF G-CSF
Neutrophilic Myelocyte
CD11b CD15 CD33 CD116 CD123 CD125
GM-CSF G-CSF
Neutrophil
CD11b CD15 CD33
Flt3L SCF IL-3 GM-CSF G-SCF
IL-9R CD34 HLA-DR CD116 CD121 CD114 CD123 CD124 CD126 CD33 CD13 CD125 CD15 CD11b
Flt3L SCF GM-CSF IL-3 IL-4
BFU-E
CD33 CD34 CD117 CD123 EPOR HLA-DR
SCF IL-3 GM-CSF IL-4 EPO
CFU-E
CD36 CD235a
TPO EPO
Proerythroblast
CD235a
EPO
Erythrocyte
CD35 CD44 CD55 CD59 CD235a
Flt3L SCF GM-CSF IL-3 IL-4 EPO TPO
HLA-DR EPOR CD33 CD34 CD117 CD123 CD36 CD235a CD35 CD44 CD55 CD59
Flt3L SCF GM-CSF IL-3 IL-6 IL-11 TPO
BFU-MK
CD33 CD34 CD116 CD123 CD126 IL-11R HLA-DR
Flt3L SCF GM-CSF Meg-CSF IL-3 IL-6 IL-11 TPO
CFU-MK
CD61 CD116 CD122 CD126
SCF GM-CSF IL-3 IL-6 IL-11 TPO
Mega-karyocyte
CD9 CD14 CD36 CD41 CD42 CD61 CD116 CD123 CD126
IL-6 IL-11 TPO
Platelets
CD9 CD14 CD36 CD41 CD42 CD49 CD61 CD126
Flt3L SCF IL-3 IL-6 IL-11 GM-CSF Meg-CSF TPO
HLA-DR CD33 CD34 IL-11R CD116 CD123 CD126 CD61 CD9 CD14 CD36 CD41 CD42 CD49

Figure 3: The pathway of hematopoietic cell lineage.

and src proteins as well as the adaptor protein functions of FAK, src and Shc to start downstream signaling events. Similar morphological alterations and modulation of gene expression are started by the binding of growth factors to their respective receptors, underling the considerable crosstalk between adhesion- and growth factor-mediated signaling. The early literatures have shown that there is a certain relationship between the pancreatic cancer and focal adhesion [38]. Activation of focal adhesion kinase enhances the adhesion and invasion of pancreatic cancer cells. Besides,

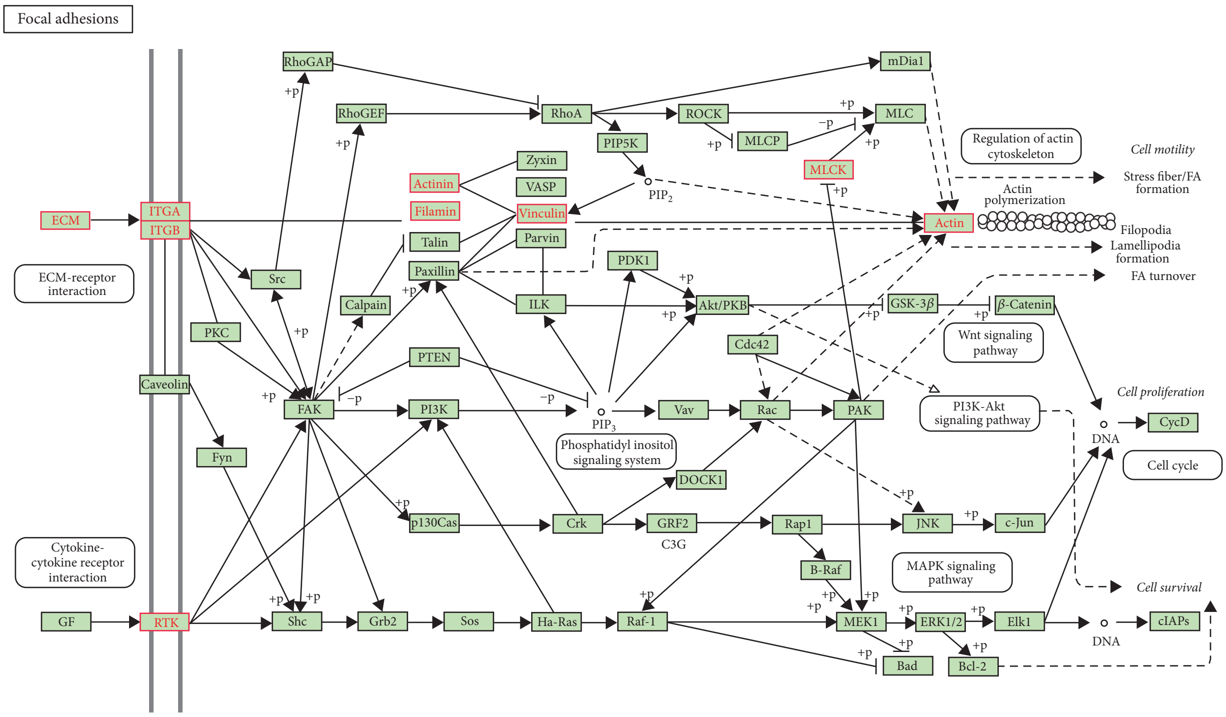

Figure 4: The pathway of focal adhesion.

Table 5: The function of top 7 extraction genes.

Gene ID	Gene name	Related GO annotations	Related diseases	Paralogous genes
5644	PRSS1	Serine-type endopeptidase activity	Trypsinogen deficiency and prss1-related hereditary pancreatitis	KLK12
5406	PNLIP	Carboxylic ester hydrolase activity and triglyceride lipase activity	Pancreatic colipase deficiency and pancreatic lipase deficiency	LPL
1357	CPA1	Metallocarboxypeptidase activity and exopeptidase activity	Borna disease and pancreatitis, hereditary	CPA3
1360	CPB1	Metallocarboxypeptidase activity and carboxypeptidase activity	Acute pancreatitis and tricuspid valve insufficiency	CPA3
63036	CELA2A	Serine-type endopeptidase activity and serine hydrolase activity	Pancreatitis, hereditary	CELA2B
5967	REG1A	Carbohydrate binding and growth factor activity	Acinar cell carcinoma and tropical calcific pancreatitis	REG3G
1056	CEL	Hydrolase activity and carboxylic ester hydrolase activity	Maturity-onset diabetes of the young, Type VIII and maturity-onset diabetes of the young	CES2

Type II diabetes mellitus is another important pathway and is widely believed to be associated with pancreatic cancer; a meta-analysis has examined this association [39].

4.2.5. Correlations between the Selected Genes and PAAD-GE Data. The function of top 7 genes selected by $L_{1/2}$ gLPCA is listed in Table 5 based on literatures and GeneCards (http://www.genecards.org/). As can be clearly seen from the table, most of these genetic lesions would likely incur pancreas-related diseases. The etiology of pancreatic cancer is not very clear; it is noted that there is a certain relationship between the incidence of chronic pancreatitis and pancreatic cancer, and we find a significant increase in the proportion of chronic pancreatitis patients with pancreatic cancer. This view is consistent with our experimental result. The clinical observation shows that abdominal pain is the most obvious symptom in the early stage of pancreatic cancer. Some literature on these genes also made a further research as follows. The gene PRSS1 variant likely affects disease susceptibility by altering expression of the primary trypsinogen gene [40]. The pancreatic lipase gene (PNLIP) is located within the genomic region of a bovine marbling quantitative trait locus. PNLIP is a positional and functional candidate for the marbling gene [41].

Table 6: The Acc and highest relevance score of these methods.

Dataset	$L_{1/2}$ gLPCA		RgLPCA		gLPCA		L_0 PCA		L_1 PCA		PCA		LE	
	Acc (%)	Relevance score	Acc (%)	Relevance score	Acc (%)	Relevance score	Acc (%)	Relevance score	Acc (%)	Relevance score	Acc (%)	Relevance score	Acc (%)	Relevance score
AMLALL	**51.33**	**55.37**	49.88	46.11	48.67	46.11	40.00	38.15	52.00	46.11	49.00	46.11	49.60	46.11
PAAD-GE	**61.60**	**85.56**	60.51	61.01	59.40	61.01	43.80	54.77	47.20	54.77	57.20	82.20	61.40	82.20

4.3. The Accuracy and Highest Relevance Score. Because ALLAML and PPAD are human disease data sets, we can find them directly from GeneCards and they are publicly available at http://www.genecards.org/.

In order to summarize the experiments on gene expression data, we compute the accuracy and highest relevance score of these methods from GeneCards and list the details in Table 6. The accuracy in Table 6 indicates the proportion of genes which are real associated with the disease in all of the genes selected by these methods. From Table 6, we observe the following. (1) Both PCA and LE commonly provide better accuracy results than L_0 PCA and L_1 PCA, demonstrating the usefulness of PCA and LE. (2) gLPCA has a good performance in some conditions and is unstable. Thus, it is necessary to reduce the effects of outliers and noise. (3) $L_{1/2}$ gLPCA and RgLPCA consistently perform better than other methods, but $L_{1/2}$ gLPCA has the highest relevance score and highest accuracy.

5. Conclusions

This paper investigates a new method of graph-Laplacian PCA ($L_{1/2}$ gLPCA) by applying $L_{1/2}$-norm constraint on the former method. $L_{1/2}$-norm constraint is applied on error function to improve the robustness of the PCA-based method. Augmented Lagrange Multipliers (ALM) method is applied to solve the optimization problem. Extensive experiments on both simulation and real gene expression data have been performed. Results on these two kinds of data show that our proposed method performs better than compared methods. Based on our proposed method, many genes have been extracted to analyze. The identified genes are demonstrated that they are closely related to the corresponding cancer data set.

In future, we will modify the model to improve sparse and robustness of the structure at the same time.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the NSFC under Grants nos. 61572284 and 61502272.

References

[1] M. J. Heller, "DNA microarray technology: devices, systems, and applications," *Annual Review of Biomedical Engineering*, vol. 4, pp. 129–153, 2002.

[2] C. K. Sarmah and S. Samarasinghe, "Microarray gene expression: a study of between-platform association of Affymetrix and cDNA arrays," *Computers in Biology and Medicine*, vol. 41, no. 10, pp. 980–986, 2011.

[3] J. Liu, D. Wang, Y. Gao et al., "A joint-L2,1-norm-constraint-based semi-supervised feature extraction for RNA-Seq data analysis," *Neurocomputing*, vol. 228, pp. 263–269, 2017.

[4] J.-X. Liu, Y. Xu, Y.-L. Gao, C.-H. Zheng, D. Wang, and Q. Zhu, "A class-information-based sparse component analysis method to identify differentially expressed genes on RNA-Seq data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 2, pp. 392–398, 2016.

[5] E. Levine, Z. Zhang, T. Kuhlman, and T. Hwa, "Quantitative characteristics of gene regulation by small RNA," *PLoS Biology*, vol. 5, no. 9, article e229, 2007.

[6] J. Liu, Y. Gao, C. Zheng, Y. Xu, and J. Yu, "Block-constraint robust principal component analysis and its application to integrated analysis of TCGA data," *IEEE Transactions on NanoBioscience*, vol. 15, no. 6, pp. 510–516, 2016.

[7] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[8] K.-J. Kim and S.-B. Cho, "Meta-classifiers for high-dimensional, small sample classification for gene expression analysis," *Pattern Analysis and Applications*, vol. 18, no. 3, pp. 553–569, 2015.

[9] B. E. Jolli, *Principal Component Analysis*, Springer, 2nd edition, 2012.

[10] D. Meng, Q. Zhao, and Z. Xu, "Improve robustness of sparse PCA by L_1-norm maximization," *Pattern Recognition*, vol. 45, no. 1, pp. 487–497, 2012.

[11] D. Meng, H. Cui, Z. Xu, and K. Jing, "Following the entire solution path of sparse principal component analysis by coordinate-pairwise algorithm," *Data and Knowledge Engineering*, vol. 88, pp. 25–36, 2013.

[12] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *Journal of Machine Learning Research*, vol. 11, pp. 517–553, 2010.

[13] Y. Zheng, B. Jeon, D. Xu, Q. M. J. Wu, and H. Zhang, "Image segmentation by generalized hierarchical fuzzy C-means algorithm," *Journal of Intelligent and Fuzzy Systems*, vol. 28, no. 2, pp. 961–973, 2015.

[14] F. R. Chung, *Spectral Graph Theory*, vol. 92, American Mathematical Society, Providence, RI, USA, 1997.

[15] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proceedings of the 15th Annual Neural Information Processing Systems Conference (NIPS '01)*, pp. 585–591, December 2001.

[16] F. Nie, H. Huang, and C. H. Ding, "Low-rank matrix recovery via efficient schatten p-norm minimization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2012.

[17] S. Jia, X. Zhang, and Q. Li, "Spectral-spatial hyperspectral image classification using regularized low-rank representation and sparse representation-based graph cuts," *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing*, vol. 8, pp. 2473–2484, 2015.

[18] Z. Xu, X. Chang, F. Xu, and H. Zhang, "L1/2 regularization: a thresholding representation theory and a fast solver," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 7, pp. 1013–1027, 2012.

[19] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *The Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 629–654, 2008.

[20] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[21] Z.-B. Xu, H.-L. Guo, Y. Wang, and H. Zhang, "Representative of $L_{1/2}$ regularization among L_q $(0 < q \le 1)$ regularizations: an experimental study based on phase diagram," *Acta Automatica Sinica*, vol. 38, no. 7, pp. 1225–1228, 2012.

[22] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*, p. 29, ACM, July 2004.

[23] B. Jiang, C. Ding, B. Luo, and J. Tang, "Graph-laplacian PCA: closed-form solution and robustness," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 3492–3498, June 2013.

[24] C. M. Feng, J. X. Liu, Y. L. Gao, J. Wang, D. Q. Wang, and Y. Du, "A graph-Laplacian PCA based on L1/2-norm constraint for characteristic gene selection," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '16)*, pp. 1795–1799, Shenzhen, China, 2016.

[25] S. Yang, C. Hou, F. Nie, and Y. Wu, "Unsupervised maximum margin feature selection via $L_{2,1}$-norm minimization," *Neural Computing and Applications*, vol. 21, no. 7, pp. 1791–1799, 2012.

[26] X. Xin, Z. Li, and A. K. Katsaggelos, "Laplacian embedding and key points topology verification for large scale mobile visual identification," *Signal Processing: Image Communication*, vol. 28, no. 4, pp. 323–333, 2013.

[27] J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga, "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization," *Nucleic Acids Research*, vol. 37, no. 2, pp. W305–W311, 2009.

[28] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 12, pp. 4164–4169, 2004.

[29] T. K. Richmond, E. Tili, M. Brown, M. Chiabai, D. Palmieri, and R. Cui, "Abstract LB-289: interaction between miR-155 and Quaking in the innate immune response and leukemia," *Cancer Research*, vol. 75, pp. 534–538, 2015.

[30] M. Essex, A. Sliski, W. D. Hardy Jr., and S. M. Cotter, "Immune response to leukemia virus and tumor-associated antigens in cats," *Cancer Research*, vol. 36, no. 2, part 2, pp. 640–645, 1976.

[31] X. Zhang, Y. Su, H. Song, Z. Yu, B. Zhang, and H. Chen, "Attenuated A20 expression of acute myeloid leukemia-derived dendritic cells increased the anti-leukemia immune response of autologous cytolytic T cells," *Leukemia Research*, vol. 38, no. 6, pp. 673–681, 2014.

[32] N. A. Gillet, M. Hamaidia, A. de Brogniez et al., "The bovine leukemia virus microRNAs permit escape from innate immune response and contribute to viral replication in the natural host," *Retrovirology*, vol. 12, supplement 1, pp. 1190–1195, 2015.

[33] M.-Y. Wu, D.-Q. Dai, X.-F. Zhang, and Y. Zhu, "Cancer subtype discovery and biomarker identification via a new robust network clustering algorithm," *PLoS ONE*, vol. 8, no. 6, Article ID e66256, 2013.

[34] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–527, 1999.

[35] D. Bonnet and J. E. Dick, "Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell," *Nature Medicine*, vol. 3, no. 7, pp. 730–737, 1997.

[36] D. Metcalf, "On hematopoietic stem cell fate," *Immunity*, vol. 26, no. 6, pp. 669–673, 2007.

[37] F. Ciceri, M. Labopin, F. Aversa et al., "A survey of fully haploidentical hematopoietic stem cell transplantation in adults with high-risk acute leukemia: a risk factor analysis of outcomes for patients in remission at transplantation," *Blood*, vol. 112, no. 9, pp. 3574–3581, 2008.

[38] H. Sawai, Y. Okada, H. Funahashi et al., "Activation of focal adhesion kinase enhances the adhesion and invasion of pancreatic cancer cells via extracellular signal-regulated kinase-1/2 signaling pathway activation," *Molecular Cancer*, vol. 4, article 37, 12 pages, 2005.

[39] R. Huxley, A. Ansary-Moghaddam, A. Berrington De González, F. Barzi, and M. Woodward, "Type-II diabetes and pancreatic cancer: a meta-analysis of 36 studies," *British Journal of Cancer*, vol. 92, no. 11, pp. 2076–2083, 2005.

[40] D. C. Whitcomb, J. LaRusch, A. M. Krasinskas et al., "Common genetic variants in the CLDN2 and PRSS1-PRSS2 loci alter risk for alcohol-related and sporadic pancreatitis," *Nature*, vol. 44, no. 12, pp. 1349–1354, 2012.

[41] Y. Muramatsu, H. Tanomura, T. Ohta, H. Kose, and T. Yamada, "Allele frequency distribution in PNLIP promoter SNP is different between high-marbled and low-marbled Japanese Black Beef Cattle," *Open Journal of Animal Sciences*, vol. 6, p. 137, 2016.