Hindawi
Complexity
Volume 2017, Article ID 4216797, 11 pages
https://doi.org/10.1155/2017/4216797

Research Article

Robust Nonnegative Matrix Factorization via Joint Graph Laplacian and Discriminative Information for Identifying Differentially Expressed Genes

Ling-Yun Dai, Chun-Mei Feng, Jin-Xing Liu, Chun-Hou Zheng, Jiguo Yu, and Mi-Xiao Hou

School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China

Correspondence should be addressed to Jin-Xing Liu; sdcavell@126.com and Chun-Hou Zheng; zhengch99@126.com

Received 17 January 2017; Accepted 6 March 2017; Published 6 April 2017

Academic Editor: Fang X. Wu

Differential expression plays an important role in cancer diagnosis and classification. In recent years, many methods have been used to identify differentially expressed genes. However, the recognition rate and reliability of gene selection still need to be improved. In this paper, a novel constrained method named robust nonnegative matrix factorization via joint graph Laplacian and discriminative information (GLD-RNMF) is proposed for identifying differentially expressed genes, in which manifold learning and the discriminative label information are incorporated into the traditional nonnegative matrix factorization model to train the objective matrix. Specifically, $L_{2,1}$-norm minimization is enforced on both the error function and the regularization term which is robust to outliers and noise in gene data. Furthermore, the multiplicative update rules and the details of convergence proof are shown for the new model. The experimental results on two publicly available cancer datasets demonstrate that GLD-RNMF is an effective method for identifying differentially expressed genes.

1. Introduction

Cancer is one of the most serious diseases that endanger the health of human being. Millions of people die of cancer every year. With the development of gene sequencing technology and other gene detection technologies, huge gene data have been generated [1, 2]. Therefore, it is important and challenging for scientists to find pathogenic genes from a large number of gene expression data. Microarray datasets on each chip usually contain many gene expression data, and the number of samples is far less than that of genes, which makes the identification of differentially expressed genes difficult [3]. In addition, irrelevant or noisy variables may reduce the accuracy of the results. In recent years, many effective mathematical methods have been applied to identify differentially expressed genes. For example, principal component analysis (PCA) [4, 5] and penalized matrix decomposition (PMD) [6] have been used to analyze gene expression data. Liu et al. used robust principal component analysis (RPCA) to discover differentially expressed genes [7]. Zheng et al. employed nonnegative matrix factorization (NMF) on the selection of tumor genes [8]. Cai et al. proposed an algorithm named graph regularized nonnegative matrix factorization (GNMF) for data representation [9]. Wang et al. used robust graph regularized nonnegative matrix factorization (RGNMF) for identifying differentially expressed genes [10]. A CIPMD (Class-Information-Based Penalized Matrix Decomposition) algorithm was proposed to identify the differentially expressed genes on RNA-Seq data, which introduced the class information via a total scatter matrix [11]. The Consensus Clustering methodology was proposed for microarray data analysis by Giancarlo and Utro [12].

However, two characteristics of gene expression data pose a serious challenge to the existing methods. Firstly, a large number of researchers hold that gene expression data probably reside in a low dimensional manifold embedded in a high dimensional ambient space. Therefore it is critical to consider the geometrical structure in the original gene expression data. Manifold learning is clearly an effective method to preserve the data geometric structure embedded in the original gene expression data [13, 14]. Cai et al. proposed GNMF [9], in which the geometrical structure of data

was constructed by an affinity graph. Another variant of NMF called manifold regularized discriminative nonnegative matrix factorization (MD-NMF) was also introduced [15]. MD-NMF considered both the local geometry of data and the discriminative information of different classes simultaneously. Long et al. proposed a method called graph regularized discriminative nonnegative matrix factorization (GDNMF) [16], in which both the geometrical structure and discriminative label information were considered in the objective function. Secondly, gene expression data often contain a lot of outliers and noise. However, existing methods cannot effectively eliminate outliers and noise. For example, least squares methods are sensitive to outliers and noise. In recent years, many researchers have been devoted to improving the robustness to outliers and noise. Zheng et al. proposed an algorithm named generalized hierarchical fuzzy C-means [17], which is robust to noise and outliers. Wang et al. used $L_{2,1}$-norm to reduce the effect of outliers and noise [10].

A novel algorithm, which we call robust nonnegative matrix factorization via joint graph Laplacian and discriminative information (GLD-RNMF), is proposed to overcome the aforementioned problems together. The proposed algorithm preserves the geometric structure of data space by constructing an affinity graph and improves the discriminative ability by the supervised label information. To do so, a new matrix decomposition objective function by integrating the geometric structure and label information is constructed. In addition, we employ $L_{2,1}$-norm instead of L_2-norm on the error function and the regularization term to reduce the influence of outliers and noise. For completeness, we present that the convergence proof of our iterative scheme is also shown in the Appendix. Experimental results indicate that the GLD-RNMF algorithm has better results than other existing algorithms for identifying differentially expressed genes.

The remainder of the paper is arranged as follows. In Section 2, we briefly introduce some relevant mathematical foundation and propose the GLD-RNMF algorithm in detail. In Section 3, the results of differentially expressed gene selection using our GLD-RNMF method and the other four methods (GNMF, NMFSC, RGNMF, and GDNMF) are shown for comparison. Finally, we conclude this paper in Section 4.

2. Materials and Methods

2.1. Mathematical Definition of $L_{2,1}$. The mathematical definition of $L_{2,1}$-norm [18] is

$$\|\mathbf{Y}\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{s} \mathbf{y}_{ij}^2} = \sum_{i=1}^{n} \|\mathbf{y}_i\|_2, \tag{1}$$

where $\mathbf{y}_i$ is the ith row of $\mathbf{Y}$ and $\mathbf{Y}$ is $n \times s$ matrix. $L_{2,1}$-norm is interpreted as follows. Firstly, we compute L_2-norm of the vector $\mathbf{y}_i$ and then compute L_1-norm of vector $\mathbf{p}(\mathbf{Y}) = (\|\mathbf{Y}_1\|_2, \|\mathbf{Y}_2\|_2, \ldots, \|\mathbf{Y}_s\|_2)$. The value of the elements of vector $\mathbf{p}(\mathbf{Y})$ represents the importance of each dimension. $L_{2,1}$-norm enables the vector $\mathbf{p}(\mathbf{Y})$ sparse to achieve the purpose of dimension reduction.

2.2. Manifold Learning. The purpose of this work is to get the best approximation of the original data. We also hope that the new representation can respect the intrinsic Riemannian structure. Recently, many researchers hold that high dimensional data often reside on a much lower dimensional manifold. The "manifold assumption" could be that data points nearby in the intrinsic geometry structure are also close under the new basis. Therefore, they usually have similar characteristics and can be categorized into the same class. In this paper, we employ manifold learning to achieve the aforementioned goal.

For a graph with N vertices, each vertex corresponds to a data point. For each data point, we can find its k nearest neighbors and connect it with the neighbors. There are many ways to define the weight matrix $\mathbf{W}$ on the graph, for example, 0-1 weighting, heat kernel weighting, and dot-product weighting. Considering that 0-1 weighting is the simplest and easy to compute, we choose 0-1 weighting as the measure in this paper.

0-1 Weight. $\mathbf{W}_{ij} = 1$, if and only if two nodes i and j are connected by an edge. That is,

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

where $\mathbf{N}_k(\mathbf{x}_j)$ consists of k nearest neighbors of $\mathbf{x}_j$ and the neighbors have the same label with $\mathbf{x}_j$.

Therefore, the smoothness of the dimensional representation can be measured as follows:

$$\begin{aligned} R &= \frac{1}{2}\sum_{i,j=1}^{n} \left\|\mathbf{s}_i - \mathbf{s}_j\right\|^2 \mathbf{W}_{ij} = \sum_{i=1}^{n} \mathbf{s}_i^T \mathbf{s}_i \mathbf{B}_{ii} - \sum_{i,j=1}^{n} \mathbf{s}_i^T \mathbf{s}_j \mathbf{W}_{ij} \\ &= \operatorname{tr}\left(\mathbf{G}^T \mathbf{B}\mathbf{G}\right) - \operatorname{tr}\left(\mathbf{G}^T \mathbf{W}\mathbf{G}\right) = \operatorname{tr}\left(\mathbf{G}^T \mathbf{L}\mathbf{G}\right), \end{aligned} \tag{3}$$

where $\operatorname{tr}(\cdot)$ represents the trace of a matrix. $\mathbf{B}$ is a diagonal matrix and $\mathbf{B}_{ii}$ is the row sum (or column, because $\mathbf{W}$ is symmetric, $\mathbf{W} \in R^{n\times n}$) of $\mathbf{W}$; that is, $\mathbf{B}_{ii} = \sum_{j=1}^{n} \mathbf{W}_{ij}$. $\mathbf{L} = \mathbf{B} - \mathbf{W}$ is graph Laplacian matrix and $\mathbf{L} \in R^{n\times n}$. We measure the distance of two points in the low dimensional space by the Euclidean distance $R(\mathbf{s}_i, \mathbf{s}_j) = \|\mathbf{s}_i - \mathbf{s}_j\|^2$.

2.3. Nonnegative Matrix Factorization (NMF). We review the standard NMF in this section. Although the algorithm has been widely used in many aspects, there are still many shortcomings.

Given n nonnegative samples $[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ in R^m, arranged in columns of a matrix $\mathbf{X} \in R^{m\times n}$, in this paper, each row of $\mathbf{X}$ represents the transcriptional response of the n genes in one sample and each column of $\mathbf{X}$ represents the expression level of a gene across all samples. Letting matrices $\mathbf{X} \in R^{m\times n}$, $\mathbf{V} \in R^{m\times k}$, and $\mathbf{H} \in R^{k\times n}$, NMF decomposes $\mathbf{X}$ into the product of $\mathbf{V}$ and $\mathbf{H}$; that is, $\mathbf{X} \approx \mathbf{V}\mathbf{H}$.

To ensure an approximate factorization $\mathbf{X} \approx \mathbf{V}\mathbf{H}$, two update rules are introduced [19]. One of the objective functions is constructed by minimizing the square of the

Euclidean distance between $\mathbf{X}$ and $\mathbf{VH}$. The optimization problem is described as follows:

$$\begin{aligned} \min_{\mathbf{V},\mathbf{H}} \quad & \|\mathbf{X} - \mathbf{VH}\|_F^2 \\ \text{s.t.} \quad & \mathbf{V} \geq 0, \\ & \mathbf{H} \geq 0, \end{aligned} \tag{4}$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm. The corresponding optimization rules are as follows:

$$\begin{aligned} \mathbf{H}_{qj} &\longleftarrow \mathbf{H}_{qj} \frac{\left(\mathbf{V}^T\mathbf{X}\right)_{qj}}{\left(\mathbf{V}^T\mathbf{VH}\right)_{qj}}, \\ \mathbf{V}_{iq} &\longleftarrow \mathbf{V}_{iq} \frac{\left(\mathbf{XH}^T\right)_{iq}}{\left(\mathbf{VHH}^T\right)_{iq}}. \end{aligned} \tag{5}$$

The convergence of the above optimization rules has been proven [19].

2.4. Graph Regularized Discriminative Nonnegative Matrix Factorization (GDNMF). Supervised label information is added to the objective function of GNMF [16]. The definition and iterative rules of GDNMF are presented below.

Class indicator matrix $\mathbf{S} \in R^{c\times n}$ is defined as follows:

$$\mathbf{S}_{ij} = \begin{cases} 1, & \text{if } \mathbf{y}_j = i,\ j = 1, 2, \ldots, n;\ i = 1, 2, \ldots, c, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where $\mathbf{y}_j \in \{1, 2, \ldots, c\}$ is the class label of $\mathbf{x}_j$ and c is the total number of classes in $\mathbf{X}$.

The objective function of GDNMF is formulated as follows:

$$\begin{aligned} \min_{\mathbf{V},\mathbf{H}} \quad & \|\mathbf{X} - \mathbf{VH}\|_F^2 + \beta \operatorname{tr}\left(\mathbf{HLH}^T\right) + \alpha \|\mathbf{S} - \mathbf{AH}\|_F^2 \\ \text{s.t.} \quad & \mathbf{V} \geq 0, \\ & \mathbf{H} \geq 0, \\ & \mathbf{A} \geq 0. \end{aligned} \tag{7}$$

The corresponding optimization rules are as follows:

$$\begin{aligned} \mathbf{H}_{qj} &\longleftarrow \mathbf{H}_{qj} \frac{\left(\mathbf{V}^T\mathbf{X} + \alpha\mathbf{A}^T\mathbf{S} + \beta\mathbf{HW}\right)_{qj}}{\left(\mathbf{V}^T\mathbf{VH} + \alpha\mathbf{A}^T\mathbf{AH} + \beta\mathbf{HB}\right)_{qj}}, \\ \mathbf{V}_{iq} &\longleftarrow \mathbf{V}_{iq} \frac{\left(\mathbf{XH}^T\right)_{iq}}{\left(\mathbf{VHH}^T\right)_{iq}}, \end{aligned} \tag{8}$$

where $\mathbf{A} \in R^{c\times k}$ is initialized to a random nonnegative matrix in the algorithm. α and β are nonnegative regularization parameters, respectively. Essentially, GDNMF incorporates the graph Laplacian and supervised label information into the objective function of NMF, which ensures the algorithm to keep consistent with the intuitive geometric structure of the data and improves the discriminative power of different classes.

2.5. Robust Nonnegative Matrix Factorization via Joint Graph Laplacian and Discriminative Information (GLD-RNMF)

2.5.1. The Objective Function. For the purpose of dimension reduction, NMF represents original data $\mathbf{X} \in R^{m\times n}$ by a product of a nonnegative matrix $\mathbf{V} \in R^{m\times k}$ and coefficient matrix $\mathbf{H} \in R^{k\times n}$. The approximation error is calculated according to the squared residuals; that is, $\|\mathbf{X} - \mathbf{VH}\|_F^2 = \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{Vh}_i\|^2$. Due to the squared term in the objective function, smaller outliers can lead to larger errors. In this paper, we enforce $L_{2,1}$-norm constraint on the objective function to reduce the impact of outliers and noise.

By employing $L_{2,1}$-norm on GDNMF model, we can formulate the objective function of GLD-RNMF as follows:

$$\begin{aligned} \min_{\mathbf{V},\mathbf{H}} \quad & \|\mathbf{X} - \mathbf{VH}\|_{2,1} + \beta \operatorname{tr}\left(\mathbf{HLH}^T\right) + \alpha \|\mathbf{S} - \mathbf{AH}\|_{2,1} \\ \text{s.t.} \quad & \mathbf{V} \geq 0, \\ & \mathbf{H} \geq 0, \\ & \mathbf{A} \geq 0. \end{aligned} \tag{9}$$

This objective function can solve high dimensional, negative, noisy and sparse data simultaneously, keep consistent with the intuitive geometric structure of data, and improve the discriminative power of different classes.

2.5.2. The Multiplication Update Rules of GLD-RNMF. Although the objective function is not convex jointly about $(\mathbf{V}, \mathbf{H}, \mathbf{A})$, it is convex in regard to one of variables in $(\mathbf{V}, \mathbf{H}, \mathbf{A})$ when the others are fixed. The objective function can be expanded as follows:

$$\begin{aligned} J &= \operatorname{tr}\left((\mathbf{X} - \mathbf{VH})\,\mathbf{Q}\,(\mathbf{X} - \mathbf{VH})^T\right) + \beta \operatorname{tr}\left(\mathbf{HLH}^T\right) \\ &\quad + \alpha \operatorname{tr}\left((\mathbf{S} - \mathbf{AH})\,\mathbf{Q}\,(\mathbf{S} - \mathbf{AH})^T\right) \\ &= \operatorname{tr}\left(\mathbf{XQX}^T\right) - 2\operatorname{tr}\left(\mathbf{XQH}^T\mathbf{V}^T\right) + \operatorname{tr}\left(\mathbf{VHQH}^T\mathbf{V}^T\right) \\ &\quad + \beta\left(\mathbf{HLH}^T\right) + \alpha \operatorname{tr}\left(\mathbf{SGS}^T\right) - 2\alpha \operatorname{tr}\left(\mathbf{SGH}^T\mathbf{A}^T\right) \\ &\quad + \alpha \operatorname{tr}\left(\mathbf{AHGH}^T\mathbf{A}^T\right), \end{aligned} \tag{10}$$

where $\mathbf{Q}$ and $\mathbf{G}$ both are diagonal matrices and the diagonal elements are as follows:

$$\mathbf{Q}_{jj} = \frac{1}{\sqrt{\sum_{i=1}^m (\mathbf{X} - \mathbf{VH})_{ij} + \varepsilon}}, \tag{11}$$

$$\mathbf{G}_{jj} = \frac{1}{\sqrt{\sum_{i=1}^m (\mathbf{S} - \mathbf{AH})_{ij} + \varepsilon}}, \tag{12}$$

in which ε is an infinitesimal positive number.

In order to solve the optimization problem in (9), we introduce the Lagrange multipliers Φ, Ψ, and Ω for $\mathbf{V}$, $\mathbf{H}$, and

A, respectively. Firstly, we formulate the Lagrange function of GLD-RNMF as follows:

$$\begin{aligned} L_J = {} & \operatorname{tr}\left(\mathbf{XQX}^T\right) - 2\operatorname{tr}\left(\mathbf{XQH}^T\mathbf{V}^T\right) \\ & + \operatorname{tr}\left(\mathbf{VHQH}^T\mathbf{V}^T\right) + \beta\left(\mathbf{HLH}^T\right) + \alpha\operatorname{tr}\left(\mathbf{SGS}^T\right) \\ & - 2\alpha\operatorname{tr}\left(\mathbf{SGH}^T\mathbf{A}^T\right) + \alpha\operatorname{tr}\left(\mathbf{AHGH}^T\mathbf{A}^T\right) \\ & + \operatorname{tr}(\Phi\mathbf{V}) + \operatorname{tr}(\Psi\mathbf{H}) + \operatorname{tr}(\Omega\mathbf{A}). \end{aligned} \tag{13}$$

Taking the partial derivatives of L_J with respect to **V**, **A**, and **H** and setting them to zero and in view of $\operatorname{tr}(\mathbf{XY}) = \operatorname{tr}(\mathbf{YX})$ and $\operatorname{tr}(\mathbf{X}^T) = \operatorname{tr}(\mathbf{X})$, we get

$$\begin{aligned} \frac{\partial L_J}{\partial \mathbf{V}} &= -2\mathbf{XQH}^T + 2\mathbf{VHQH}^T + \Phi, \\ \frac{\partial L_J}{\partial \mathbf{A}} &= -2\alpha\mathbf{SGH}^T + 2\alpha\mathbf{AHGH}^T + \Omega, \\ \frac{\partial L_J}{\partial \mathbf{H}} &= -2\mathbf{V}^T\mathbf{XQ} + 2\mathbf{V}^T\mathbf{VHQ} + 2\beta\mathbf{HL} \\ &\quad + \alpha\left[-2\mathbf{A}^T\mathbf{SG} + 2\mathbf{A}^T\mathbf{AHG}\right] + \Psi. \end{aligned} \tag{14}$$

According to the KKT (Karush-Kuhn-Tucker) conditions [20], that is, $\Phi_{iq}\mathbf{V}_{iq} = 0$, $\Omega_{kq}\mathbf{A}_{kq} = 0$, and $\Psi_{qj}\mathbf{H}_{qj} = 0$, we can obtain the following equations:

$$\begin{aligned} & \left[-2\mathbf{XQH}^T + 2\mathbf{VHQH}^T\right]_{iq}\mathbf{V}_{iq} + \Phi_{iq}\mathbf{V}_{iq} = 0, \\ & \left[-2\alpha\mathbf{SGH}^T + 2\alpha\mathbf{AHGH}^T\right]_{kq}\mathbf{A}_{kq} + \Omega_{kq}\mathbf{A}_{kq} = 0, \\ & \left[-2\mathbf{V}^T\mathbf{XQ} + 2\mathbf{V}^T\mathbf{VHQ} + 2\beta\mathbf{HL}\right. \\ & \quad \left. + \alpha\left(-2\mathbf{A}^T\mathbf{SG} + 2\mathbf{A}^T\mathbf{AHG}\right)\right]_{qj}\mathbf{H}_{qj} + \Psi_{qj}\mathbf{H}_{qj} = 0. \end{aligned} \tag{15}$$

Then we can get the multivariate updating rules as follows:

$$\mathbf{V}_{iq} \longleftarrow \mathbf{V}_{iq}\frac{\left(\mathbf{XQH}^T\right)_{iq}}{\left(\mathbf{VHQH}^T\right)_{iq}}, \tag{16}$$

$$\mathbf{A}_{kq} \longleftarrow \mathbf{A}_{kq}\frac{\left(\mathbf{SGH}^T\right)_{kq}}{\left(\mathbf{AHGH}^T\right)_{kq}}, \tag{17}$$

$$\mathbf{H}_{qj} \longleftarrow \mathbf{H}_{qj}\frac{\left(\mathbf{V}^T\mathbf{XQ} + \alpha\mathbf{A}^T\mathbf{SG} + \beta\mathbf{HW}\right)_{qj}}{\left(\mathbf{V}^T\mathbf{VHQ} + \alpha\mathbf{A}^T\mathbf{AHG} + \beta\mathbf{HB}\right)_{qj}}. \tag{18}$$

The details of our method are described in Algorithm 1. The iterative procedure is performed until the algorithm converges.

Considering the three update rules above, we ensure the convergence of the algorithm by the following theorem.

Theorem 1. *The objective function* $\mathrm{O} = \|\mathbf{X} - \mathbf{VH}\|_{2,1} + \beta\operatorname{tr}(\mathbf{HLH}^T) + \alpha\|\mathbf{S} - \mathbf{AH}\|_{2,1}$ *is nonincreasing under the iterative rules in (16), (17), and (18).*

The detailed proof of the theorem is shown in the Appendix.

3. Results and Discussion

In order to verify the effectiveness of GLD-RNMF algorithm for identifying differentially expressed genes, we perform experiments on real gene expression datasets to compare our algorithm with the other four feature extraction algorithms: (a) GNMF algorithm (Cai et al. [9]); (b) NMFSC algorithm (Hoyer [21]); (c) RGNMF algorithm (Wang et al. [10]); (d) GDNMF algorithm (Long et al. [16]). We conduct these experiments on two publicly available cancer datasets: pancreatic cancer dataset (PAAD) and cholangiocarcinoma dataset (CHOL).

3.1. Identifying Differentially Expressed Genes by GLD-RNMF. In this section, we use GLD-RNMF to identify differentially expressed genes. The matrix **X** with size $m \times n$ is the original gene expression data. Each row of **X** indicates the transcriptional response of n genes in a sample. Each column of **X** indicates the expression level of a gene in all samples. Therefore, **X** can be written as follows:

$$\mathbf{X} \approx \mathbf{VH}, \tag{19}$$

where **V** is the basis matrix with size $m \times k$ and **H** is the coefficient matrix with size $k \times n$ and $k \ll \min(m, n)$. Since the matrix **V** contains all of the genes, the differentially expressed genes can be identified from the matrix **V** [10]. By GLD-RNMF, the evaluating vector $\overline{\mathbf{V}}$ is obtained in which elements are sorted in descending order:

$$\overline{\mathbf{V}} = \left[\sum_{i=1}^{k}\left|\mathbf{V}_{1i}\right|, \ldots, \sum_{i=1}^{k}\left|\mathbf{V}_{ni}\right|\right]^T. \tag{20}$$

Generally, the larger the entry in $\overline{\mathbf{V}}$ is, the more differential this gene is. Therefore, the differentially expressed genes can be obtained by the first num (num $\leq j$) largest elements in $\overline{\mathbf{V}}$.

The objective of the experiment is to identify the differentially expressed genes by GLD-RNMF algorithm. The identifying process is described below.

(1) Obtain the nonnegative matrix **X** according to the genomic dataset.

(2) Construct the label matrix **S** and the diagonal matrixes **G** and **Q**.

(3) Gain the Laplacian matrix **L** and basis matrix **V** via GLD-RNMF algorithm.

(4) Identify differentially expressed genes through the vector $\overline{\mathbf{V}}$.

(5) Check differentially expressed genes by gene ontology tool.

3.2. Parameters Selection. We assign parameters in our GLD-RNMF algorithm following the same way proposed by Long et al. [16]. Distinguishingly, there are two parameters, that is, β and α, in GLD-RNMF method.

Fortunately, if β and α are set in a reasonable range they have little effect on the performance of the algorithm [15, 16].

Input: Data matrix $\mathbf{X} \in R^{m\times n}$, indicator matrix $\mathbf{S} \in R^{c\times n}$, parameters α, β, k.
Output: Matrices $\mathbf{V} \in R^{m\times k}$, $\mathbf{H} \in R^{k\times n}$ and $\mathbf{A} \in R^{c\times k}$.
(1) Initialization: Randomly initialize three nonnegative matrices $\mathbf{V}_0 \in R^{m\times k}$, $\mathbf{H}_0 \in R^{k\times n}$ and $\mathbf{A}_0 \in R^{c\times k}$, initialize $\mathbf{Q}_0 \in R^{n\times n}$, $\mathbf{G}_0 \in R^{n\times n}$ to be identity matrix. Set $r = 0$
(2) Repeat
Update $\mathbf{V}_{r+1}$,$\mathbf{A}_{r+1}$ and $\mathbf{H}_{r+1}$ separately by

$$\mathbf{V}_{ir+1} \longleftarrow \mathbf{V}_r \frac{\left(\mathbf{X}\mathbf{Q}_r\mathbf{H}_r^T\right)}{\left(\mathbf{V}_r\mathbf{H}_r\mathbf{Q}_r\mathbf{H}_r^T\right)}$$

$$\mathbf{A}_{r+1} \longleftarrow \mathbf{A}_r \frac{\left(\mathbf{S}\mathbf{G}_r{\mathbf{H}_r}^T\right)}{\left(\mathbf{A}_r\mathbf{H}_r\mathbf{G}_r{\mathbf{H}_r}^T\right)}$$

$$\mathbf{H}_{r+1} \longleftarrow \mathbf{H}_r \frac{\left({\mathbf{V}_{r+1}}^T\mathbf{X}\mathbf{Q}_{r+1} + \alpha{\mathbf{A}_{r+1}}^T\mathbf{S}\mathbf{G}_{r+1} + \beta\mathbf{H}_r\mathbf{W}\right)}{\left({\mathbf{V}_{r+1}}^T\mathbf{V}\mathbf{H}_r\mathbf{Q}_{r+1} + \alpha{\mathbf{A}_{r+1}}^T\mathbf{A}\mathbf{H}_r\mathbf{G}_{r+1} + \beta\mathbf{H}_r\mathbf{B}\right)}$$

Calculate the diagonal matrices $\mathbf{Q}_{r+1}$ and $\mathbf{G}_{r+1}$ by (11) and (12), separately.
Until convergence.

Algorithm 1: GLD-RNMF.

Table 1: Comparison of p values of different methods on PAAD.

Gene ID	Gene name	NMFSC	GNMF	RGNMF	GDNMF	GLD-RNMF
GO:0030198	Extracellular matrix organization	$1.36E-15$	$9.04E-17$	$1.36E-15$	$1.36E-15$	**2.99E − 42**
GO:0043062	Extracellular structure organization	$1.44E-15$	$9.57E-17$	$1.44E-15$	$1.44E-15$	**3.35E − 42**
GO:0031012	Extracellular matrix	$5.64E-15$	$3.23E-17$	$5.64E-15$	$5.64E-15$	**3.53E − 37**
GO:0005615	Extracellular space	$3.09E-27$	$3.09E-27$	$3.09E-27$	$3.09E-27$	**9.70E − 37**
GO:0005578	Proteinaceous extracellular matrix	$6.74E-12$	$4.59E-14$	$6.74E-12$	$6.74E-12$	**2.00E − 29**
GO:0044420	Extracellular matrix component	$1.55E-10$	$7.70E-12$	$1.55E-10$	$1.55E-10$	**4.96E − 27**
GO:0030574	Collagen catabolic process	$3.93E-14$	$8.45E-16$	$5.64E-15$	$3.93E-14$	**8.03E − 27**
GO:0044243	Multicellular organism catabolic process	$1.24E-13$	$3.00E-15$	$1.24E-13$	$1.24E-13$	**6.06E − 26**
GO:0032963	Collagen metabolic process	$3.41E-11$	$1.45E-12$	$3.41E-11$	$3.41E-11$	**2.37E − 23**
GO:0098644	Complex of collagen trimers	$1.35E-09$	$1.56E-11$	$1.35E-09$	$1.35E-09$	**3.53E − 20**

In our experiments, we set $\beta = 0.9$ and $\alpha = 0.5$ in the GLD-RNMF algorithm. Another important parameter in our GLD-RNMF algorithm is k which is used to construct a k-nearest graph. Empirically, we set $k = 5$ and adopt the mode as the heat kernel in LPP [22]. Besides, we set the reduced dimension to 5 for all the methods. All the parameters in the other methods keep in line with those described in their paper [10, 16, 21, 23].

3.3. Gene Ontology Analysis. The gene ontology (GO) tool can interpret the genes that are input and discover the functions that these genes may have in common. As a web-based tool [24], GO Enrichment Analysis can find important GO items from a large number of genes and provide important information for the biological interpretation of high-throughput experiments. Another online tool that we use is ToppFun. It usually is used to interpret the differentially expressed genes.

To be fair, we extract 100 genes from the gene expression data by GNMF, NMFSC, RGNMF, GDNMF, and GLD-RNMF methods. Threshold parameters of ToppFun are set as follows: the maximum p value is set to 0.01 and the minimum number of gene products is set to 2.

3.4. Pancreatic Cancer Dataset. Pancreatic cancer is a tumor with high malignancy, which is difficult to diagnose and treat. Early diagnosis of pancreatic cancer is not difficult and the mortality rate is high. The cause of pancreatic cancer is still not clear until now. In the experiment, there are 20502 genes in 180 samples contained in the dataset.

The top 10 GO items extracted and the p values of the five methods are listed in Table 1. In this table, "ID" and "Name" represent items and their names associated with the GO in the whole genome and the lowest p value of the five methods has been marked in bold font. As we find from Table 1, the p values generated by GLD-RNMF are much smaller than those by the other four methods for the PAAD. Therefore, GLD-RNMF method is more superior than the other four methods for the PAAD. The name of "GO:0030198" is extracellular matrix organization. It contains DPT (Dermatopontin), POSTN (Periostin), sec24d (Sec24-related protein d), and

Table 2: Pancreatic cancer genes extracted by GLD-RNMF.

Gene ID	Gene name	Gene annotations	Relative diseases
4313	MMP2	Serine-type endopeptidase activity and metallopeptidase activity	Arthropathy and multicentric osteolysis of Torg
3486	IGFBP3	Fibronectin binding and insulin-like growth factor I binding.	Insulin-like growth factor I and acromegaly
3630	INS	Identical protein binding and protease binding	Diabetes mellitus
4316	MMP7	Peptidase activity and metallopeptidase activity	Spastic entropion and focal myositis
3880	KRT19	Structural molecule activity and structural constituent of cytoskeleton	Thyroid cancer and lung cancer
7057	THBS1	Calcium ion binding and heparin binding	Posterior uveal melanoma and thrombotic thrombocytopenic purpura
3320	HSP90AA1	Poly(A) RNA binding and identical protein binding	Lobular neoplasia and candidiasis
1508	CTSB	Peptidase activity and cysteine-type peptidase activity	Occlusion of gallbladder and ileum cancer
7076	TIMP1	Cytokine activity and protease binding.	Oral submucous fibrosis and lung giant cell carcinoma
2335	FN1	Heparin binding and protease binding	Glomerulopathy and plasma fibronectin deficiency

other genes which are related to pancreatic cancer [25–27]. For example, POSTN can create a tumor-supportive microenvironment in the pancreas [28]. Genes and gene products associated with extracellular structure organization (GO:0043062) can be found by GO tool, in which, DPT, POSTN, sec24d, uxs1 (UDP-Glucuronate Decarboxylase 1), fkrp (Fukutin Related Protein), and other genes have been illustrated to be associated with pancreatic cancer [29, 30]. For example, Sec24d is ubiquitously expressed but exhibits predominant expression in heart, placenta, liver, and pancreas. The other GO items can also be proven to be related to pancreatic cancer by some relevant literature material. Clearly, GLD-RNMF method is an effective method for identifying differentially expressed genes.

Comparing 100 genes extracted by GLD-RNMF with what we obtain from Gene Cards (http://www.genecards.org/) about pancreatic cancer, 82 of the 100 genes are associated with pancreatic cancer. Many genes, which were previously thought to be unrelated to clinical outcomes, are identified. We present the top 10 of 82 genes with higher relevance scores in Table 2, including their gene ID, names, function, and related diseases. Among the identified differentially expressed genes, MMP2, MMP7, IGFBP3, INS, and the other genes have been demonstrated to be related to pancreatic cancer [31–34]. For example, the effect of MMP-2 and its activators MT1-MMP, MT2-MMP, and MT3-MMP in pancreatic tumor cell invasion and the development of the desmoplastic reaction characteristic of pancreatic cancer tissues have been discussed [35]. Akihisa Fukuda et al. demonstrated that serum MMP7 level in human pancreatic ductal adenocarcinoma patients is correlated with metastatic disease and survival. Conditioned medium from Capan-1 pancreatic cancer cells which contains abundant IGFBP-3 has been mentioned [36]. Other genes identified by GLD-RNMF have been illustrated to be related to pancreatic cancer by some relevant literature materials as well.

On the other hand, we use Kyoto Encyclopedia of Genes and Genomes (KEGG) online analysis tool to analyze the differentially expressed genes identified by GLD-RNMF. In this experiment, putting the identified 100 genes into KEGG, we can obtain the corresponding disease pathway. Figure 1 is the pathway of pancreatic cancer. The genes that have been found by GLD-RNMF are marked with red. The chromosomal instability pathway records the disease progression from normal duct to pancreatic cancer. Infiltrating ductal adenocarcinoma is the most common malignancy of the pancreas. Normal duct epithelium progresses to the stage of infiltrating cancer through a series of histologically defined precursors. These differentially expressed genes contain two oncogenes: K-Ras and HER2/neu. p16, p53, BRCA2, and Smad4 are tumor suppressors. Slebos et al. assessed that K-ras oncogene mutations and p53 protein accumulation are associated with known or postulated risk factors for pancreatic cancer [37]. Gu et al. found the expression of Smad4 and p16 is significantly lower in pancreatic cancer tissue compared with normal tissue. The lower expression of the proteins may impact the development of pancreatic cancer [38]. From Figure 1, we can see that the pancreatic cancer pathway map contains six other pathways: PI3K-Akt signaling pathway, ErbB signaling pathway, MAPK signaling pathway, VEGF signaling, Jak-STAT signaling pathway, p53 signaling pathway, and TGF-β signaling pathway.

PI3K-Akt signaling pathway is presented in Figure 2. From Figure 2, we can find four proteinases from the differentially expressed genes identified by GLD-RNMF. Therefore, our algorithm achieves better results.

3.5. Cholangiocarcinoma Dataset. Cholangiocarcinoma is diagnosed in 12,000 patients in the US each year, but only 10 percent are discovered early enough to allow for successful surgical treatment. In our experiments, we apply the five methods on CHOL which contains 20502 genes on 45 samples. The 8 GO items closely related to cholangiocarcinoma

Table 3: Comparison of p values of different methods on CHOL.

Gene ID	Gene name	NMFSC	GNMF	RGNMF	GDNMF	GLD-RNMF
GO:0072562	Blood microparticle	$1.73E-20$	$1.73E-20$	$1.85E-17$	$1.73E-20$	**$1.25E-50$**
GO:0060205	Cytoplasmic membrane-bounded vesicle lumen	$4.19E-23$	$4.19E-23$	$3.90E-18$	$4.19E-23$	**$1.34E-41$**
GO:0031983	Vesicle lumen	$8.66E-23$	$8.66E-23$	$7.05E-18$	$8.66E-23$	**$4.47E-41$**
GO:0005615	Extracellular space	$3.21E-19$	$3.21E-19$	$2.12E-17$	$3.21E-19$	**$2.66E-37$**
GO:0034774	Secretory granule lumen	$1.59E-19$	$1.59E-19$	$1.35E-14$	$1.59E-19$	**$3.33E-34$**
GO:0004857	Enzyme inhibitor activity	$5.71E-06$	$5.71E-06$	$5.71E-06$	$5.71E-06$	**$3.93E-17$**
GO:0004866	Endopeptidase inhibitor activity	$2.92E-05$	$2.92E-05$	$2.92E-05$	$2.92E-05$	**$3.74E-16$**
GO:0061135	Endopeptidase regulator activity	$3.64E-05$	$3.64E-05$	$3.64E-05$	$3.64E-05$	**$6.50E-16$**

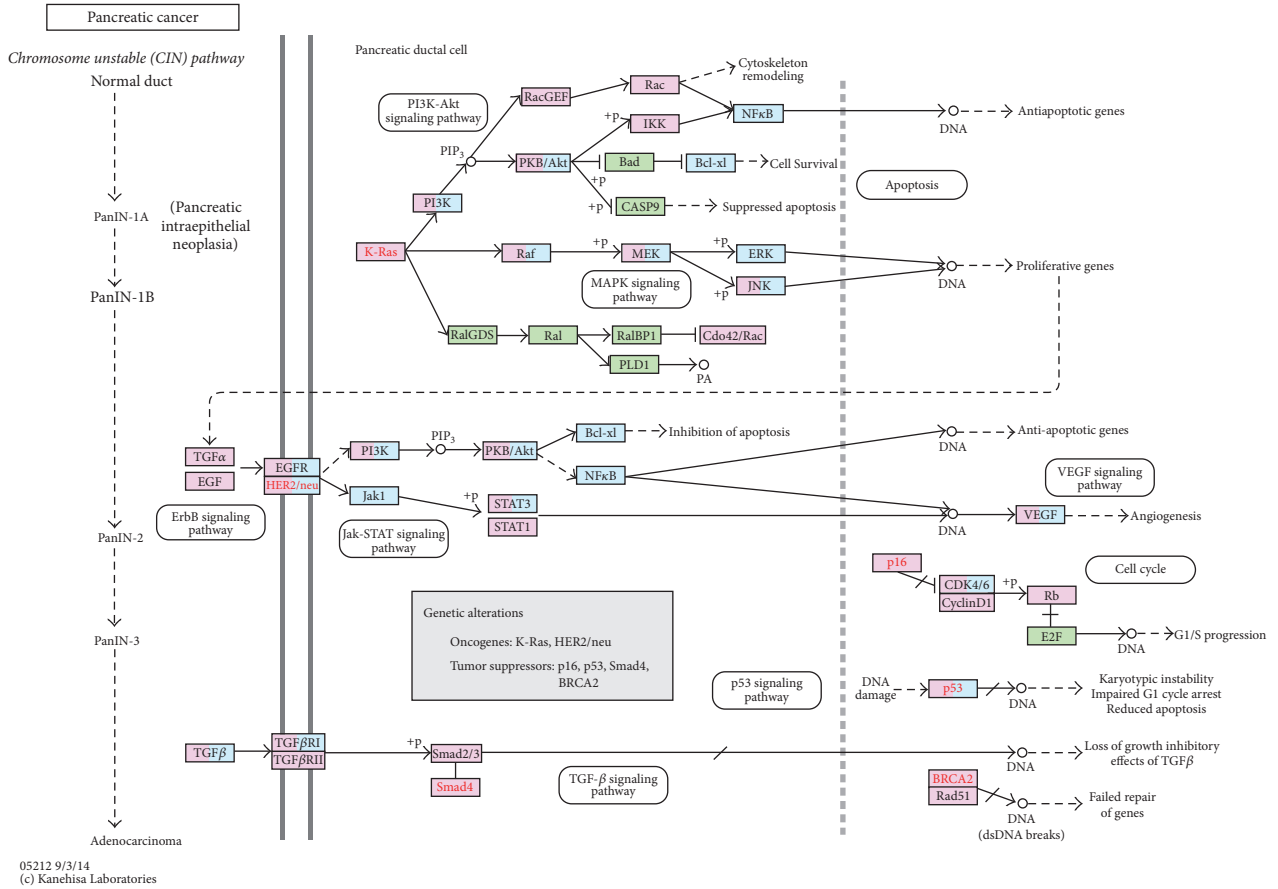

Figure 1: Pathway of pancreatic cancer, where pink background represents disease genes, light blue represents drug target genes, and light green represents human genes. Genes found by GLD-RNMF are marked with red.

and the p values of the five methods are listed in Table 3. In this table, "ID" and "Name" represent items and their names associated with the GO in the whole genome. The lowest p value of the five methods has been marked in bold font. We can see from the table that GLD-RNMF is much better than the other four methods. Genes and gene products associated with blood microparticle (GO:0072562) can be found by the GO tool, in which APOA4 (Apolipoprotein A4), kng1 (Kininogen 1), and other genes have been illustrated to be associated with cholangiocarcinoma [39–41]. The name of "GO:0060205" is cytoplasmic vesicle lumen. It contains ada (Adenosine Deaminase), DBH (Dopamine Beta-Hydroxylase), and other genes which are related to cholangiocarcinoma [42, 43]. The top 8 genes identified by GLD-RNMF are listed in Table 4 including the gene ID, names, gene annotations, and related diseases. Consistent with the previous study, ALB (Albumin), HP (Haptoglobin), SERPINC1 (Serpin Family C Member 1), C3 (Complement C3),

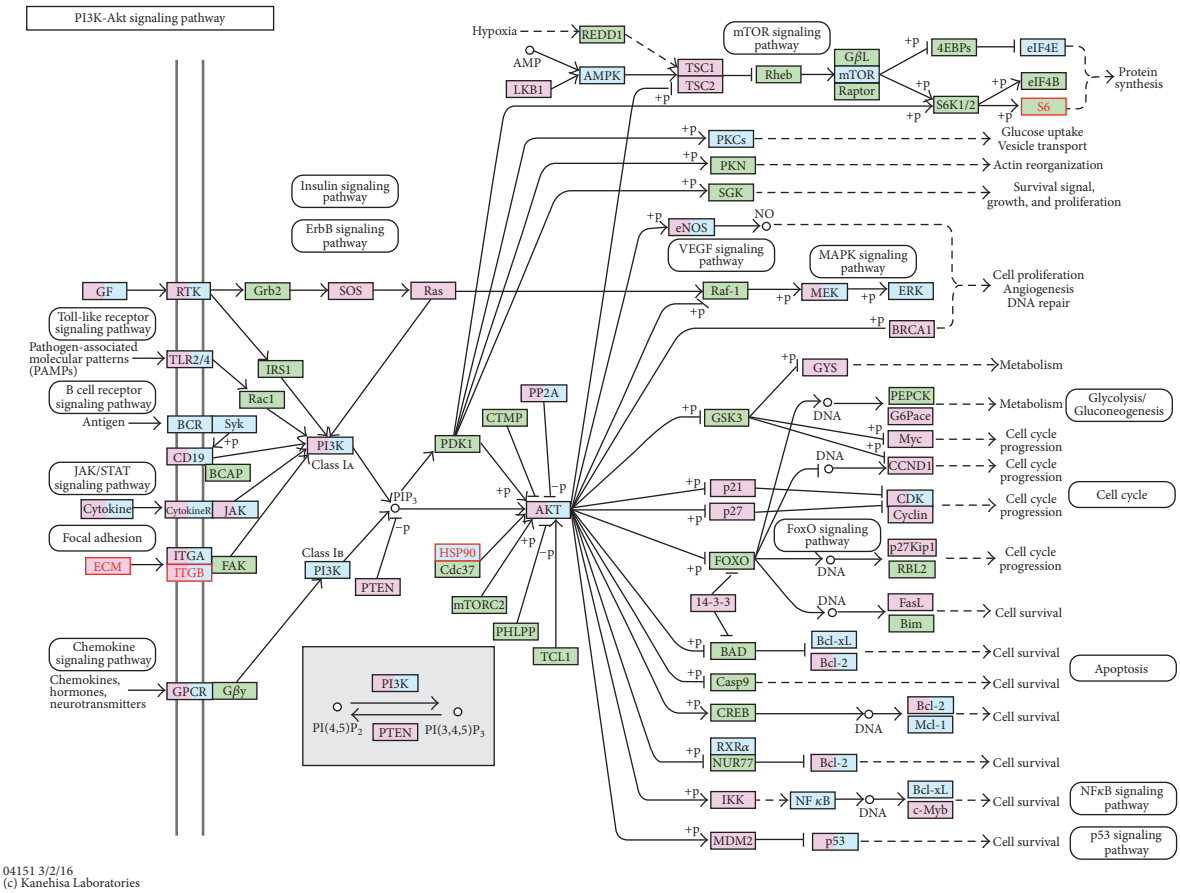

FIGURE 2: PI3K-AKT signaling pathway, where pink background represents disease genes, light blue represents drug target genes, and light green represents human genes. Genes found by GLD-RNMF are marked with red.

and other genes are successfully identified which represent potential biomarkers for cholangiocarcinoma and potential targets for clarifying the molecular mechanisms associated with cholangiocarcinoma. For example, HP was proposed as pronucleating proteins because they were highly expressed in the fast nucleating bile of patients with cholesterol stones [44]. Waghray et al. predicted survival in patients with hilar cholangiocarcinoma by serum albumin [45]. C3 and HP were identified as more abundant in cholangiocarcinoma [46]. The relative documents can illustrate that the other genes identified by GLD-RNMF are associated with cholangiocarcinoma.

4. Conclusions

By introducing $L_{2,1}$-norm, manifold graph and discriminative label information, we propose an efficient algorithm named robust nonnegative matrix factorization via joint graph Laplacian and discriminative information (GLD-RNMF) in this paper. $L_{2,1}$-norm can reduce the influence of outliers and noise, manifold graph can find the low dimensional manifold in high dimensional data space and the intrinsic law of the observed data, and the label information can increase the discriminative power of different classes. Nonnegative matrix factorization avoids the problems of high dimension and nonnegative data. As a result, GLD-RNMF can handle nonnegative, high dimension, outliers, and noise and improve the discriminative power of different classes. Experimental results on two datasets show that GLD-RNMF is superior to the state-of-the-art methods for identifying differentially expressed gene.

Appendix

Detailed Proof of Theorem

To prove Theorem, we need to show that the objective function in (9) is nonincreasing under the iterative rules in (16), (17), and (18). For the objective function O, we need to fix $\mathbf{H}$ and $\mathbf{A}$ when we update $\mathbf{V}$. Similarly, we need to fix $\mathbf{H}$ and $\mathbf{V}$ when we update $\mathbf{A}$, and we need to fix $\mathbf{A}$ and $\mathbf{V}$ when we update $\mathbf{H}$. For the reason that we have similar update rules for $\mathbf{A}$ and $\mathbf{V}$ in GLD-RNMF with those in NMF, the detailed proof can be found [27]. Hence, we just need to prove that O is nonincreasing under the iterative

Table 4: Cholangiocarcinoma genes extracted by GLD-RNMF.

Gene ID	Gene name	Gene annotations	Relative diseases
213	ALB	Enzyme binding and chaperone binding.	Analbuminemia and congenital analbuminemia
3240	HP	Serine-type endopeptidase activity and hemoglobin binding	Anhaptoglobinemia and plasmodium falciparum malaria
5265	SERPINA1	Identical protein binding and protease binding	Emphysema due to Aat deficiency and alpha 1-antitrypsin deficiency
718	C3	Receptor binding and C5L2 anaphylatoxin chemotactic receptor binding	Macular degeneration, age-related, 9, and C3 deficiency
2243	FGA	Receptor binding and protein binding, bridging	Dysfibrinogenemia, congenital, and afibrinogenemia, congenital
7448	VTN	Heparin binding and scavenger receptor activity	Glanzmann thrombasthenia and camptodactyly-arthropathy-coxa vara-pericarditis syndrome.
7018	TF	Ubiquitin protein ligase binding and ferric iron transmembrane transporter activity	Atransferrinemia and iron overload In Africa
335	APOA1	Identical protein binding and lipid binding	Amyloidosis, familial visceral, and hypoalphalipoproteinemia

rules in (18). We use an auxiliary function similar to what is used in the Expectation-Maximization algorithm [47]. In the demonstration, we present the definition of auxiliary function [16]. Definition G(h, h') is an auxiliary function of $F(h)$ if the following conditions are satisfied.

$$\begin{aligned} G\left(h, h'\right) &\geq F\left(h\right), \\ G\left(h, h\right) &= F\left(h\right). \end{aligned} \tag{A.1}$$

The auxiliary function is vital due to the following lemmas.

Lemma A.1. *If G is an auxiliary function of F, then F is nonincreasing under the update rule:*

$$h^{(t+1)} = \arg\min_{h} \quad G\left(h, h^{(t)}\right). \tag{A.2}$$

Proof. Obviously

$$\begin{aligned} F\left(h^{(t+1)}\right) &\leq G\left(h^{(t+1)}, h^{(t)}\right) \leq G\left(h^{(t)}, h^{(t)}\right) \\ &= F\left(h^{(t)}\right); \end{aligned} \tag{A.3}$$

now, we will present the update step for $\mathbf{H}$ in (18) is exactly the update in (A.2) with an appropriate auxiliary function

$$\begin{aligned} F'_{ij} &= \left(\frac{\partial O}{\partial \mathbf{H}}\right)_{ij} = \left(-2\mathbf{V}^T\mathbf{XQ} + 2\mathbf{V}^T\mathbf{VHQ} + 2\beta\mathbf{HL}\right. \\ &\quad \left. - 2\alpha\mathbf{A}^T\mathbf{SG} + 2\alpha\mathbf{A}^T\mathbf{AHG}\right)_{ij}, \\ F''_{ij} &= \left(2\mathbf{V}^T\mathbf{VQ}\right)_{ii} + 2\beta\mathbf{L}_{jj} + \left(2\alpha\mathbf{A}^T\mathbf{AG}\right)_{ii}. \end{aligned} \tag{A.4}$$

It is enough to prove that each F_{ij} is nonincreasing under the update rules because our update is essentially wised. Therefore, we present the following lemma. □

Lemma A.2. *Function,*

$$\begin{aligned} G\left(h, h_{ij}^{(t)}\right) &= F_{ij}\left(h_{ij}^{(t)}\right) + F'_{ij}\left(h_{ij}^{(t)}\right)\left(h - h_{ij}^{(t)}\right) \\ &\quad + \frac{\left(\mathbf{V}^T\mathbf{VHQ}\right)_{ij} + \beta\left(\mathbf{HB}\right)_{ij} + \alpha\left(\mathbf{A}^T\mathbf{AHG}\right)_{ij}}{h_{ij}^{(t)}}\left(h\right. \\ &\quad \left. - h_{ij}^{(t)}\right)^2, \end{aligned} \tag{A.5}$$

is an auxiliary function of F_{ij}.

Proof. We only need to demonstrate that $G(h, h_{ij}^{(t)}) \geq F_{ij}(h)$, because $G(h, h) = F_{ij}(h)$ is obvious. Consequently, comparing the Taylor series expansion of $F_{ij}(h)$,

$$\begin{aligned} &F_{ij}\left(h\right) \\ &= F_{ij}\left(h_{ij}^{(t)}\right) + F'_{ij}\left(h_{ij}^{(t)}\right)\left(h - h_{ij}^{(t)}\right) \\ &\quad + \left[\left(\mathbf{V}^T\mathbf{VQ}\right)_{ii} + \beta\mathbf{L}_{jj} + \left(\alpha\mathbf{A}^T\mathbf{AG}\right)_{ii}\right]\left(h - h_{ij}^{(t)}\right)^2, \end{aligned} \tag{A.6}$$

with (A.2), we can find that $G(h, h_{ij}^{(t)}) \geq F_{ij}(h)$ is equivalent to

$$\begin{aligned} &\frac{\left(\mathbf{V}^T\mathbf{VHQ}\right)_{ij} + \beta\left(\mathbf{HB}\right)_{ij} + \alpha\left(\mathbf{A}^T\mathbf{AHG}\right)_{ij}}{h_{ij}^{(t)}} \\ &\quad \geq \left(\mathbf{V}^T\mathbf{VQ}\right)_{ii} + \beta\mathbf{L}_{jj} + \left(\alpha\mathbf{A}^T\mathbf{AG}\right)_{ii}. \end{aligned} \tag{A.7}$$

Actually, we have

$$\left(\mathbf{V}^T\mathbf{VHQ}\right)_{ij} = \sum_{l=1}^{k} h_{il}^{(t)} \left(\mathbf{V}^T\mathbf{VQ}\right)_{lj} \geq \left(\mathbf{V}^T\mathbf{VQ}\right)_{ii} h_{ij}^{(t)},$$

$$\left(\mathbf{A}^T\mathbf{AHG}\right)_{ij} = \sum_{l=1}^{k} h_{il}^{(t)} \left(\mathbf{A}^T\mathbf{AG}\right)_{lj} \geq \left(\mathbf{A}^T\mathbf{AG}\right)_{ii} h_{ij}^{(t)}. \tag{A.8}$$

$$\begin{aligned}\beta\left(\mathbf{HB}\right)_{ij} &= \beta\sum_{l=1}^{n} h_{il}^{(t)} B_{lj} \geq \beta h_{ij}^{(t)} B_{jj} \\ &\geq \beta h_{ij}^{(t)} \left(B - W\right)_{jj} = \beta h_{ij}^{(t)} L_{jj}.\end{aligned}$$

□

Therefore, (A.7) holds and $G(h, h_{ij}^{(t)}) \geq F_{ij}(h)$. Now we can prove the convergence of theorem.

Proof of Theorem. Replacing $G(h, h_{ij}^{(t)})$ in (A.2) by (A.5), we can obtain the following update rules:

$$\begin{aligned}h_{ij}^{(t+1)} &= h_{ij}^{(t)} \\ &\quad - h_{ij}^{(t)} \frac{F_{ij}'\left(h_{ij}^{(t)}\right)}{\left(2\mathbf{V}^T\mathbf{VHQ}\right)_{ij} + 2\beta\left(\mathbf{HB}\right)_{ij} + \left(2\alpha\mathbf{A}^T\mathbf{AHG}\right)_{ij}} \\ &= h_{ij}^{(t)} \left(\frac{\mathbf{V}^T\mathbf{XQ} + \alpha\mathbf{A}^T\mathbf{SG} + \beta\mathbf{HW}}{\mathbf{V}^T\mathbf{VHQ} + \beta\mathbf{HB} + \alpha\mathbf{A}^T\mathbf{AHG}}\right)_{ij}.\end{aligned} \tag{A.9}$$

Since $G(h, h_{ij}^{(t)})$ is an auxiliary function, F_{ij} is nonincreasing under the update rule. □

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported in part by the grants of the National Science Foundation of China, nos. 61572284, 61502272, 61373027, and 61672321.

References

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

[2] Y. Li and Z. Zhang, "Computational biology in microRNA," *Wiley Interdisciplinary Reviews: RNA*, vol. 6, no. 4, pp. 435–452, 2015.

[3] A. B. Tchagang and A. H. Tewfik, "DNA microarray data analysis: a novel biclustering algorithm approach," *Eurasip Journal on Applied Signal Processing*, vol. 2006, Article ID 59809, 12 pages, 2006.

[4] A. Wang and E. A. Gehan, "Gene selection for microarray data analysis using principal component analysis," *Statistics in Medicine*, vol. 24, no. 13, pp. 2069–2087, 2005.

[5] R. Luss and A. D'Aspremont, "Clustering and feature selection using sparse principal component analysis," *Optimization and Engineering*, vol. 11, no. 1, pp. 145–157, 2010.

[6] C. H. Zheng, L. Zhang, T. Y. Ng, K. S. Chi, and S. L. Wang, "Inferring the transcriptional modules using penalized matrix decomposition," in *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence: 6th International Conference on Intelligent Computing, ICIC 2010, Changsha, China, August 18–21, 2010. Proceedings*, vol. 6216 of *Lecture Notes in Computer Science*, pp. 35–41, Springer, Berlin, Germany, 2010.

[7] J.-X. Liu, Y. Xu, C.-H. Zheng, H. Kong, and Z.-H. Lai, "RPCA-based tumor classification using gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 4, pp. 964–970, 2015.

[8] C.-H. Zheng, D.-S. Huang, L. Zhang, and X.-Z. Kong, "Tumor clustering using nonnegative matrix factorization with gene selection," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp. 599–607, 2009.

[9] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.

[10] D. Wang, J. X. Liu, Y. L. Gao, C. H. Zheng, and Y. Xu, "Characteristic gene selection based on robust graph regularized non-negative matrix factorization," *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, vol. 108, p. 1, 2015.

[11] J.-X. Liu, J. Liu, Y.-L. Gao, J.-X. Mi, C.-X. Ma, and D. Wang, "A class-information-based penalized matrix decomposition for identifying plants core genes responding to abiotic stresses," *PLoS ONE*, vol. 9, no. 9, Article ID e106097, 2014.

[12] R. Giancarlo and F. Utro, "Speeding up the Consensus Clustering methodology for microarray data analysis," *Algorithms for Molecular Biology*, vol. 6, no. 1, article 1, 2011.

[13] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[14] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[15] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 2030–2048, 2011.

[16] X. Long, H. Lu, Y. Peng, and W. Li, "Graph regularized discriminative non-negative matrix factorization for face recognition," *Multimedia Tools and Applications*, vol. 723, pp. 2679–2699, 2014.

[17] 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.

[18] C. Ding, D. Zhou, X. He, and H. Zha, "R 1-PCA: rotational invariant L 1-norm principal component analysis for robust subspace factorization," in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pp. 281–288, Pittsburgh, Pa, USA, June 2006.

[19] D. D. Lee, "Algorithms for nonnegative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[20] F. Facchinei, C. Kanzow, and S. Sagratella, "Solving quasi-variational inequalities via their KKT conditions," *Mathematical Programming*, vol. 144, no. 1-2, pp. 369–412, 2014.

[21] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[22] S. Yang, C. Hou, C. Zhang, Y. Wu, and S. Weng, "Robust non-negative matrix factorization via joint sparse and graph regularization," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '13)*, pp. 1–5, August 2013.

[23] B. Yang, "Graph regularized non-negative matrix factorization with sparseness constraints," *Computer Science*, vol. 1, no. 40, pp. 218–256, 2013.

[24] M. D. Young, M. J. Wakefield, G. K. Smyth, and A. Oshlack, "Gene ontology analysis for RNA-seq: accounting for selection bias," *Genome Biology*, vol. 11, no. 2, article R14, 2010.

[25] T. Orban, J. M. Sosenko, D. Cuthbertson et al., "Pancreatic islet autoantibodies as predictors of type 1 diabetes in the diabetes prevention trial-type 1," *Diabetes Care*, vol. 32, no. 12, pp. 2269–2274, 2009.

[26] P. Baril, R. Gangeswaran, P. C. Mahon et al., "Periostin promotes invasiveness and resistance of pancreatic cancer cells to hypoxia-induced cell death: role of the β_4 integrin and the PI3k pathway," *Oncogene*, vol. 26, no. 14, pp. 2082–2094, 2007.

[27] B. L. Tang, J. Kausalya, D. Y. H. Low, M. L. Lock, and W. Hong, "A family of mammalian proteins homologous to yeast Sec24p," *Biochemical and Biophysical Research Communications*, vol. 258, no. 3, pp. 679–684, 1999.

[28] M. Erkan, J. Kleeff, A. Gorbachevski et al., "Periostin creates a tumor-supportive microenvironment in the pancreas by sustaining fibrogenic stellate cell activity," *Gastroenterology*, vol. 132, no. 4, pp. 1447–1464, 2007.

[29] Z.-Q. Ye, S. Niu, Y. Yu et al., "Analyses of copy number variation of GK rat reveal new putative type 2 diabetes susceptibility loci," *PLoS ONE*, vol. 5, no. 11, article e14077, 2010.

[30] T. Yamamoto, Y. Kato, M. Karita, M. Kawaguchi, N. Shibata, and M. Kobayashi, "Expression of genes related to muscular dystrophy with lissencephaly," *Pediatric Neurology*, vol. 31, no. 3, pp. 183–190, 2004.

[31] F. Miralles, T. Battelino, P. Czernichow, and R. Scharfmann, "TGF-β plays a key role in morphogenesis of the pancreatic islets of langerhans by controlling the activity of the matrix metalloproteinase MMP-2," *Journal of Cell Biology*, vol. 143, no. 3, pp. 827–836, 1998.

[32] S. R. Bramhall, J. P. Neoptolemos, G. W. H. Stamp, and N. R. Lemoine, "Imbalance of expression of matrix metalloproteinases (MMPs) and tissue inhibitors of the matrix metalloproteinases (TIMPs) in human pancreatic carcinoma," *Journal of Pathology*, vol. 182, no. 3, pp. 347–355, 1997.

[33] J. B. Douglas, D. T. Silverman, M. N. Pollak, Y. Tao, A. S. Soliman, and R. Z. Stolzenberg-Solomon, "Serum IGF-I, IGF-II, IGFBP-3, and IGF-I/IGFBP-3 molar ratio and risk of pancreatic cancer in the prostate, lung, colorectal, and ovarian cancer screening trial," *Cancer Epidemiology Biomarkers and Prevention*, vol. 19, no. 9, pp. 2298–2306, 2010.

[34] E. Karaskov, C. Scott, L. Zhang, T. Teodoro, M. Ravazzola, and A. Volchuk, "Chronic palmitate but not oleate exposure induces endoplasmic reticulum stress, which may contribute to INS-1 pancreatic β-cell apoptosis," *Endocrinology*, vol. 147, no. 7, pp. 3398–3407, 2006.

[35] V. Ellenrieder, B. Alber, U. Lacher et al., "Role of MT-MMPs and MMP-2 in pancreatic cancer progression," *International Journal of Cancer*, vol. 85, no. 1, pp. 14–20, 2000.

[36] X. Y. Huang, Z. L. Huang, J. H. Yang et al., "Erratum to: Pancreatic cancer cell-derived IGFBP-3 contributes to muscle wasting," *Journal of Experimental & Clinical Cancer Research*, vol. 35, pp. 1–13, 2016.

[37] R. J. C. Slebos, J. A. Hoppin, P. E. Tolbert et al., "K-ras and p53 in pancreatic cancer: association with medical history, histopathology, and environmental exposures in a population-based study," *Cancer Epidemiology Biomarkers and Prevention*, vol. 9, no. 11, pp. 1223–1232, 2000.

[38] L.-J. Gu, J. Chen, Z.-H. Lu, L. Li, W.-X. Zhou, and Y.-F. Luo, "Expression of DPC4/Smad4, p21wafI, and p16 in human pancreatic cancer," *Chinese Journal of Cancer*, vol. 21, no. 2, pp. 132–137, 2002.

[39] E.-H. Kim, J.-S. Bae, K. B. Hahm, and J.-Y. Cha, "Endogenously synthesized n-3 polyunsaturated fatty acids in fat-1 mice ameliorate high-fat diet-induced non-alcoholic fatty liver disease," *Biochemical Pharmacology*, vol. 84, no. 10, pp. 1359–1365, 2012.

[40] A. Wee and B. Nilsson, "Highly well differentiated hepatocellular carcinoma and benign hepatocellular lesions: can they be distinguished on fine needle aspiration biopsy?" *Acta Cytologica*, vol. 47, no. 1, pp. 16–26, 2003.

[41] I. Subrungruang, C. Thawornkuno, C.-P. Porntip, C. Pairojkul, S. Wongkham, and S. Petmitr, "Gene expression profiling of intrahepatic cholangiocarcinoma," *Asian Pacific Journal of Cancer Prevention*, vol. 14, no. 1, pp. 557–563, 2013.

[42] S. Tanaka, M. Iwai, Y. Harada et al., "Targeted killing of carcinoembryonic antigen (CEA)-producing cholangiocarcinoma cells by polyamidoamine dendrimer-mediated transfer of an Epstein-Barr virus (EBV)-based plasmid vector carrying the CEA promoter," *Cancer Gene Therapy*, vol. 7, no. 9, pp. 1241–1249, 2000.

[43] M. Tawfik El-Mansi, K. S. Cuschieri, R. G. Morris, and A. R. W. Williams, "Prevalence of human papillomavirus types 16 and 18 in cervical adenocarcinoma and its precursors in Scottish patients," *International Journal of Gynecological Cancer*, vol. 16, no. 3, pp. 1025–1031, 2006.

[44] A. Farina, M. Delhaye, P. Lescuyer, and J.-M. Dumonceau, "Bile proteome in health and disease," *Comprehensive Physiology*, vol. 4, no. 1, pp. 91–108, 2014.

[45] A. Waghray, A. Sobotka, C. R. Marrero, B. Estfan, F. Aucejo, and K. N. Menon, "Serum albumin predicts survival in patients with hilar cholangiocarcinoma," *Gastroenterology Report*, vol. 5, no. 1, pp. 62–66, 2017.

[46] U. Navaneethan, V. Lourdusamy, P. G. Venkatesh, B. Willard, M. R. Sanaka, and M. A. Parsi, "Bile proteomics for differentiation of malignant from benign biliary strictures: a pilot study," *Gastroenterology Report*, vol. 3, pp. 136–143, 2015.

[47] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B. Methodological*, vol. 39, no. 1, pp. 1–38, 1977.