frontiers in Genetics

METHODS
published: 23 February 2021
doi: 10.3389/fgene.2021.621317

Joint Lp-Norm and $L_{2,1}$-Norm Constrained Graph Laplacian PCA for Robust Tumor Sample Clustering and Gene Network Module Discovery

Xiang-Zhen Kong, Yu Song, Jin-Xing Liu, Chun-Hou Zheng*, Sha-Sha Yuan, Juan Wang and Ling-Yun Dai*

School of Computer Science, Qufu Normal University, Rizhao, China

The dimensionality reduction method accompanied by different norm constraints plays an important role in mining useful information from large-scale gene expression data. In this article, a novel method named Lp-norm and $L_{2,1}$-norm constrained graph Laplacian principal component analysis (PL21GPCA) based on traditional principal component analysis (PCA) is proposed for robust tumor sample clustering and gene network module discovery. Three aspects are highlighted in the PL21GPCA method. First, to degrade the high sensitivity to outliers and noise, the non-convex proximal Lp-norm ($0 < p < 1$)constraint is applied on the loss function. Second, to enhance the sparsity of gene expression in cancer samples, the $L_{2,1}$-norm constraint is used on one of the regularization terms. Third, to retain the geometric structure of the data, we introduce the graph Laplacian regularization item to the PL21GPCA optimization model. Extensive experiments on five gene expression datasets, including one benchmark dataset, two single-cancer datasets from The Cancer Genome Atlas (TCGA), and two integrated datasets of multiple cancers from TCGA, are performed to validate the effectiveness of our method. The experimental results demonstrate that the PL21GPCA method performs better than many other methods in terms of tumor sample clustering. Additionally, this method is used to discover the gene network modules for the purpose of finding key genes that may be associated with some cancers.

Keywords: Lp-norm, graph regularization, sparse constraint, principal component analysis, tumor clustering, gene network modules, $L_{2,1}$-norm

OPEN ACCESS

Edited by:
Xianwen Ren,
Peking University, China

Reviewed by:
Xiaojian Shao,
National Research Council Canada (NRC-CNRC), Canada
Yongcui Wang,
Northwest Institute of Plateau Biology (CAS), China

****Correspondence:***
Jin-Xing Liu
sdcavell@126.com
Chun-Hou Zheng
zhengch99@126.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *26 October 2020*
Accepted: *29 January 2021*
Published: *23 February 2021*

Citation:
Kong X-Z, Song Y, Liu J-X, Zheng C-H, Yuan S-S, Wang J and Dai L-Y (2021) Joint Lp-Norm and $L_{2,1}$-Norm Constrained Graph Laplacian PCA for Robust Tumor Sample Clustering and Gene Network Module Discovery. Front. Genet. 12:621317. doi: 10.3389/fgene.2021.621317

INTRODUCTION

High-throughput sequencing technologies, including genome-wide measurements, have enabled large-scale gene expression profiles to accumulate faster (Goodwin et al., 2016). It is of great significance to obtain useful information from these data. Reliable and precise identification of cancer types and obtaining key pathogenic genes are very important for cancer diagnosis and treatment (Koboldt et al., 2012). Generally, gene expression data have a typical characteristic of "high dimension, low sample" size (West, 2003), which is a challenge for most traditional statistical methods. Too many variables and some uncorrelated noise variables in the gene expression data may all have a negative effect on the tumor clustering performance regardless of whether

supervised or unsupervised clustering methods are used. Despite these problems, many researchers have demonstrated the effectiveness of tumor-type identification and feature selection by leveraging many machine learning algorithms (Hochreiter et al., 2010; Lee et al., 2010; Liu J. X. et al., 2015; Bunte et al., 2016; Kong et al., 2017; Wang et al., 2017; Chen et al., 2019). Among them, algorithms based on principal component analysis (PCA) (Collins, 2002; Jolliffe, 2002) have been widely used to process gene expression data successfully (Liu et al., 2013; Liu J. X. et al., 2015; Wang et al., 2017; Feng et al., 2019) for dimension reduction and denoising. However, PCA-based algorithms, including sparse principal component analysis (SPCA) (Zou et al., 2006; Shen and Huang, 2008; Journee et al., 2010; Liu et al., 2016; Feng et al., 2019) and robust principal component analysis (RPCA) (Candès et al., 2009; Liu et al., 2013; Liu J. X. et al., 2015; Wang et al., 2017), mainly deal with data that lie in a linear data manifold (Jiang et al., 2013). Many methods that can handle data lying in a non-linear manifold have been proposed, such as locality preserving projections (LPP) (He et al., 2005), locally linear embedding (LLE) (Roweis and Saul, 2000), local tangent space alignment (Zhang and Zha, 2002), Laplacian eigenmap (LE) (Belkin and Niyogi, 2002, 2003; Spielman, 2007) and latent variable model (LELVM) (Keyhanian and Nasersharif, 2015). The purpose of these non-linear dimensionality reduction techniques is to find a representation of points (samples) in a low-dimensional space, in which all points (samples) still maintain the similarity in the original high-dimensional space.

In recent years, optimization models that combine linear and non-linear dimensionality reduction methods, especially graph Laplacian embedding, have demonstrated their effectiveness. Liu et al. (2017) constructed a graph Laplacian matrix for semisupervised feature extraction. Cai et al. (2011) proposed a method named graph regularized non-negative matrix factorization (GNMF), which combined graph structure and non-negative matrix factorization for an improved compact representation of the original data. Jiang et al. (2013) developed graph-Laplacian PCA (gLPCA), which sought a low-dimensional representation of image data with significant improvement in clustering and image reconstruction by incorporating graph structures and PCA. Feng et al. (2017) employed pgLPCA based on graph Laplacian regularization and Lp-norm for feature selection and tumor clustering. Wang et al. (2019a) used Laplacian regularized low-rank representation (LLRR), which considers the intrinsic geometric structure of gene expression data to cluster the tumor samples. In addition, many methods benefit from norm constraints. For example, Journee et al. (2010) employed the L_0-norm constraint based on PCA to stress the sparse expression of genes in samples. The L_1-norm (Tibshirani, 1996) was introduced as the regularization function in sparse singular value decomposition (SSVD) (Lee et al., 2010; Kong et al., 2017) and the mix-norm optimization model proposed by Wang et al. (2019b). Feng et al. (2016) employed the $L_{1/2}$-norm constraint in their model to select characteristic genes. However, there remain some facets to be improved: for example, the robustness of the algorithm should be enhanced further, and the sparse representation of the data should be highlighted. For these purposes, the Lp-norm (Chartrand, 2012; Nie et al., 2013; Feng et al., 2017; Kong et al., 2017) constraint was used in the optimization model to degrade the sensitivity of outliers of the data. The $L_{2,1}$-norm (Xiang et al., 2012; Yang et al., 2012) constraint was used by Liu et al. (2017) and Wang et al. (2019b) to generate the row sparsity.

Motivated by the literature mentioned above, especially (Tibshirani, 1996; Chartrand, 2012; Xiang et al., 2012; Nie et al., 2013; Feng et al., 2017; Kong et al., 2017), we propose a new method named PL21GPCA incorporating traditional PCA, graph Laplacian embedding and different norm constraints on the loss function and the regularization function for robust tumor sample clustering and gene network module discovery. Five gene expression datasets, including one benchmark dataset, two single-cancer datasets from The Cancer Genome Atlas (TCGA), and two integrated datasets of multiple cancers from TCGA, are used to evaluate the effectiveness of our method. The experimental results demonstrate that the PL21GPCA method outperforms many existing methods in terms of tumor sample clustering. Additionally, this method is employed to discover gene network modules to identify the key genes with close relationships to some cancers.

We organize the rest of this paper as follows. Section "Related Works" provides the related works containing the non-convex proximal Lp-norm, $L_{2,1}$-norm and graph regularized PCA. The optimization model of PL21GPCA is presented, and the solution procedure is detailed in section "Methodology." Section "Experiments and Discussion" presents the parameter selections, experimental results and some discussions. The tumor sample clustering and gene network analysis are also described in this section. In Section "Conclusion and Suggestions," we present the conclusion for this article and propose some suggestions for future research.

RELATED WORKS

Definitions of the Proximal Lp-Norm and $L_{2,1}$-Norm

Let $X \in R^{p \times n}$ be a data matrix, the proximal Lp-norm of X is defined as follows:

$$\| X \|_p = \left(\sum_{i}^{p} \sum_{j}^{n} |x_{ij}|^p\right)^{\frac{1}{p}} \ (0 < p < 1) \tag{1}$$

The Lp-norm with $0 < p < 1$ is a function with three typical characteristics: globally non-differentiable, non-convex, and non-smooth (Chartrand, 2012; Zhang et al., 2015). Many researchers have made suggestions to deal with Lp-norm $(0 < p < 1)$ minimization (Chartrand, 2012; Guo et al., 2013; Qin et al., 2013). Since Lp-norm minimization can result in a sparser solution than the L_1-norm and perform better in terms of robustness to outliers than the L_2-norm in a sense, we use it to constrain the loss function of the PL21GPCA optimization model. The generalized shrinkage operation proposed by Chartrand (2012) is adopted to solve the function effectively in our method.

The $L_{2,1}$-norm of matrix X is as follows:

$$\| X \|_{2,1} = \sum_{i=1}^{p} \sqrt{\sum_{j=1}^{n} x_{ij}^2} = \sum_{i=1}^{p} \| x_i \|_2 \quad (2)$$

where x_i (corresponding to feature i) is the $_i$th row of matrix X. Yang et al. (2012) provided an intuitive explanation of the $L_{2,1}$-norm in the literature. To solve the $L_{2,1}$-norm, we can compute the L_2-norm of each row of X first, record it as a vector $b(X) = \left(\| x_1 \|_2, \| x_2 \|_2, ..., \| x_p \|_2\right)$, and then compute the L_1-norm of vector $b(X)$. The components of vector b indicate the importance of each feature. The $L_{2,1}$-norm favors obtaining a small number of non-zero rows in matrix X, and then feature selection will be achieved.

PCA and Graph Laplacian Embedding

Principal Component Analysis (PCA)

Let $X = (x_1, \cdots, x_n) \in R^{p\times n}$ $(p \gg n)$ be a matrix whose rows represent genes and columns represent samples. PCA is usually used to find the optimal principal directions $V^T = (v_1, \cdots, v_n) \in R^{k\times n}$ $(V^TV = I)$ that define the low-dimensional (k-dim) subspace. And the projected data points in the low subspace Vcan be denoted as the elements of the matrix $U_{p\times k} = (u_1, \cdots, u_k) \in R^{p\times k}$. The traditional PCA finds $U$and V with the squared Frobenius norm:

$$\arg\min_{U,V} \| \text{X-UV}^T \|_F^2 \ \ s.t.\ V^TV = I \quad (3)$$

In our optimization model, the proximal Lp-norm $\|g\|_p$ $(0 < p < 1)$ (Chartrand, 2012; Nie et al., 2013; Feng et al., 2017) is used instead of the traditional quadratic loss function $\|g\|_F$ to reduce the influence of outliers and noise. PCA naturally relates closely to the classic clustering means known as K-means (Ding and He, 2004). The optimal principal components contained in matrix V provide the solution of the K-means clustering method. It inspired us to combine PCA with Laplacian embedding, whose principal purpose is also clustering.

Graph Laplacian Embedding

Principal component analysis can find an approximate set of basis vectors in the case where data usually lie in a linear manifold (Jiang et al., 2013). In consideration of the local invariance of the intrinsic geometric structure of the data distribution, graph Laplacian embedding is a popular method among recent studies in non-linear manifold learning theory (Belkin and Niyogi, 2002, 2003; Spielman, 2007). The assumption of local invariance is that if two points (samples) are close in the intrinsic geometry of the original data distribution, the representations of these two points (samples) in the new coordinate are also close to each other. The local geometric structure can be modeled through a nearest neighbor graph on a scatter of data points. Given the data matrix $X = (x_1, \cdots, x_n) \in R^{p\times n}$, $x_i (i = 1, \cdots, n)$can be regarded as one data point (one vertex in the graph). For each data point x_i, we find its k' nearest neighbors and put edges between x_i and its neighbors. Then, a graph with $_n$ vertices can be constructed, on which the weight matrix $W \in R^{n\times n}$ is defined. w_{ij} is the weight between vertices x_i and x_j, it is used to measure the closeness of two points x_i and x_j, and it is a symmetric similarity matrix. There are three popular choices defining the weight matrix on the graph: heat kernel weighting, 0–1 weighting, and dot-product weighting. If nodes $_i$ and $_j$ are connected, using heat kernel weighting, $w_{ij} = e^{\frac{\|x_i - x_j\|^2}{\sigma}}$, $w_{ij} = 1$ using 0–1 weighting and $w_{ij} = w_i^T w_j$ using dot-product weighting. The different similarity measures are suitable for different situations. Detailed information about the different weighting schemes can be found in the literature (Cai et al., 2011).

Let $Z^T = (z_1, z_2, \cdots, z_n) \in R^{k\times n}$ represent the $_n$data points in the k-dim embedding coordinates $V^T = (v_1, \cdots, v_n) \in R^{k\times n}$ $(V^TV = I)$, i.e., the representation of $_{x_i}$ in the new low-dimensional basis is $z_i = [v_{i1}, \cdots, v_{ik}]$. The "dissimilarity" of the two data points in the low basis can be measured by the Euclidean distance or the divergence distance. The Euclidean distance is adopted in our method. Define the "dissimilarity" of the two points in the low basis as $d(z_i, z_j) = \| z_i - z_j \|^2$, combined with the weight matrix $_W$, and the smoothness of the low-dimensional representation can be measured by minimizing:

$$\begin{aligned} S &= \frac{1}{2}\sum_{i,j=1}^{n} \|z_i - z_j\|^2 w_{ij} \\ &= \sum_{i=1}^{n} z_i^T z_i D_{ii} - \sum_{i,j=1}^{n} z_i^T z_j w_{ij} \\ &= Tr(V^TDV) - Tr(V^TWV) = Tr(V^TLV) \end{aligned} \quad (4)$$

where $_{Tr(\bullet)}$is the trace of a matrix, $_{D=diag(d_1,\cdots,d_n)}$is a diagonal matrix, and $_{d_i=\sum_{j=1}^{n} w_{ij}}$. We call the $_{L=D-W}$ the Laplacian matrix (Spielman, 2007).

METHODOLOGY

The PL21GPCA procedure is presented in this section. **Figure 1** illustrates our general research framework. In brief, our work includes two steps. The first is obtaining the optimal projected matrix $U_{p\times k}$ and the principal directions matrix $V_{k\times n}$ via PL21GPCA. The second is to evaluate the validity of PL21GPCA. In this step, based on the principal directions matrix $V_{k\times n}$ obtained by PL21GPCA, the classic clustering method K-means is employed for tumor sample clustering. According to the projected matrix $U_{p\times k}$, the differentially expressed genes are selected to carry out gene network analysis to find the key genes with close relationships to some cancers.

To summarize, three aspects are highlighted in our method:

(1) To reduce the influence of outliers and noise, the non-convex proximal Lp-norm $\|g\|_p$ $(0 < p < 1)$ is used on the loss function, which could improve the robustness of the optimization model effectively compared with the other constraints.

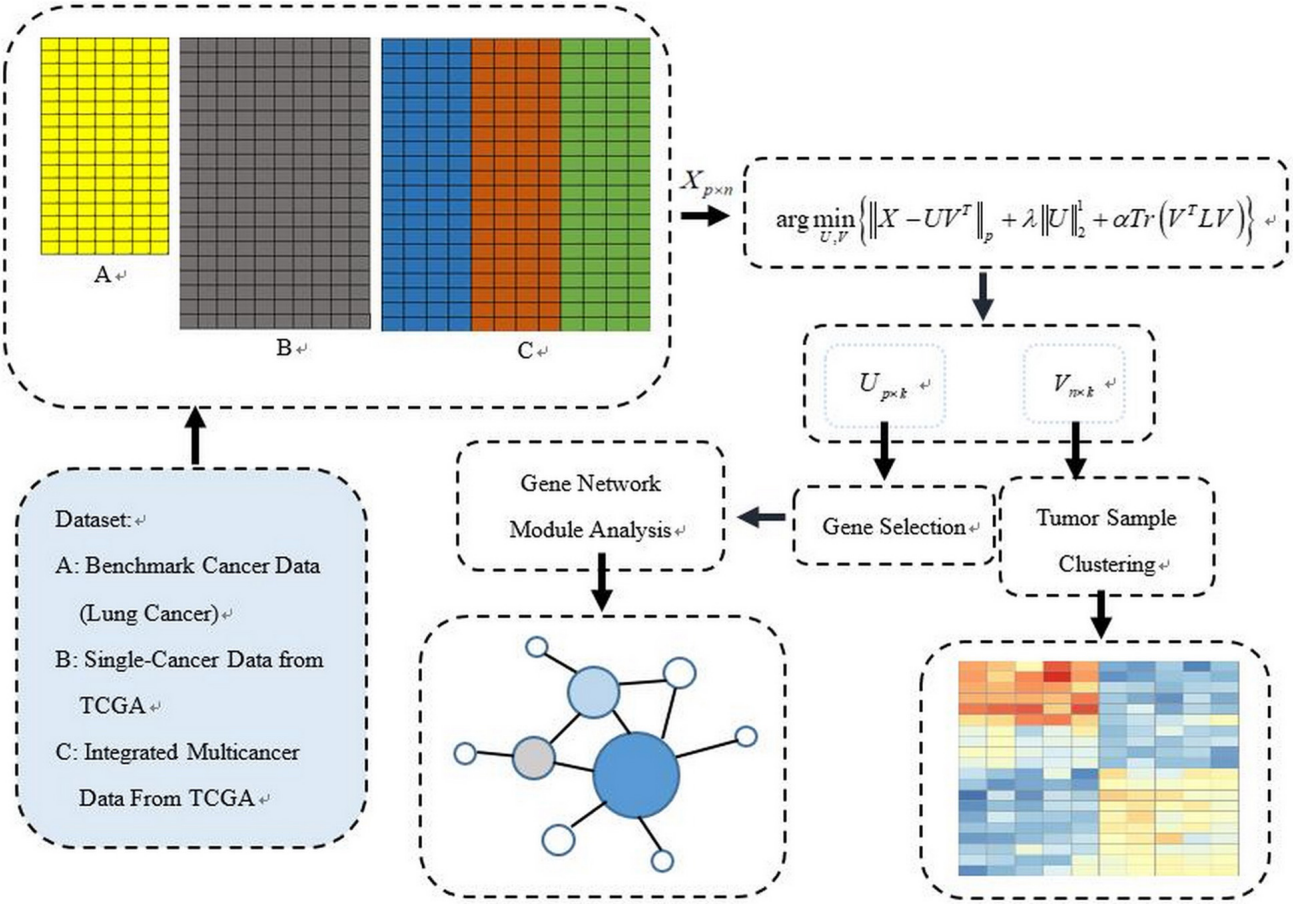

FIGURE 1 | The general schematic framework of the PL21GPCA methodology.

(2) To enhance the sparsity of gene expression in cancer samples, the $L_{2,1}$-norm is used on the projected matrix $U_{p\times k}$.
(3) To retain the intrinsic geometric structure of the data points (samples), the graph regularization item is recommended in the optimization model.

Assume the input matrix $X = (x_1, \cdots, x_n) \in R^{p\times n}$ $(p \gg n)$, which denotes $_p$ genes' expression in $_n$ samples. Our goal is to find the optimal low-dimensional ($_k$-dim) subspace denoted as $V^T = (v_1, \cdots, v_n) \in R^{k\times n}$ $(V^TV = I)$ and the projected matrix $U_{p\times k} = (u_1, \cdots, u_k) \in R^{p\times k}$ in the low subspace. The traditional PCA finds $U$and V with the squared Frobenius norm in the solution. In our optimization model, the proximal Lp-norm $\|g\|_p$ $(0 < p < 1)$ (Chartrand, 2012; Nie et al., 2013; Feng et al., 2017) replaces the traditional quadratic loss function $\|g\|_F$ to reduce the influence of outliers and noise. The L2,1-norm is used on one of the regularization terms to enhance the sparse gene expression in cancer samples. The graph Laplacian regularization item emphasizing the local invariance of the intrinsic geometric structure is recommended in the optimization model.

The objective function of this method is designed as follows:

$$\begin{aligned}&\arg\min_{U,V}\left\{\| \text{X-UV}^T \|_p +\lambda \| U \|_2^1 +\alpha Tr\left(V^TLV\right)\right\}\\ &s.t.\ V^TV = I,\ 0 < p < 1,\ \lambda > 0,\ \alpha > 0\end{aligned} \tag{5}$$

Clearly, the objective function is somewhat intractable because it is non-convex and non-smooth. We adopt the augmented Lagrangian multiplier (ALM) (Hestenes, 1969; Bertsekas, 1982; Spielman, 2007; Lin et al., 2010) to address this optimization problem. Researchers have proven that the ALM algorithm possesses Q-linear convergence properties under some conditions (Bertsekas, 1982).

When using the ALM method to obtain the optimal solution of (5), we replace $X - UV^T$ with $_E$. Eq. (5) can be equivalently written as:

$$\begin{aligned}&\arg\min_{E,U,V}\left\{\| E \|_p +\lambda \| U \|_2^1 +\alpha Tr\left(V^TLV\right)\right\}\\ &s.t.\ \text{E-X} + UV^T = 0,\ V^TV = I\end{aligned} \tag{6}$$

According to the ALM method, eq. (6) is equivalent to minimizing:

$$L_{\mu,Y}(E,U,V) = \| E \|_p + \frac{\mu}{2} \| \text{E-X} + UV^T + \frac{Y}{\mu} \|_F^2 + \lambda \| U \|_2^1 + \alpha Tr\left(V^T L V\right), \quad (7)$$

where $_Y$ is the Lagrangian multiplier, and $_\mu$ is the step size of the update rule. In (7), there are three variables to be solved. The alternating direction method (ADM) (Gabay and Mercier, 1976) is adopted to tackle this thorny problem because the equation with only one variable is easily solved when the others are fixed. By this means, (7) naturally results in three subproblems.

Problem 1: When $_U$ and $_V$ are fixed, (7) is written as follows:

$$L_{\mu,Y}(E,U,V) = \| E \|_p + \frac{\mu}{2} \| \text{E-X} + UV^T + \frac{Y}{\mu} \|_F^2 \quad (8)$$

where $0 < p < 1$. Eq. (8) can be solved by the proximal shrink operator denoted as follows:

$$shrink_p(t,\delta) := \max\{0, |t| - \delta|t|^{p-1}\}\frac{t}{|t|} \quad (9)$$

Let $t = X - UV^T - \frac{Y}{\mu}$, $\delta = \frac{1}{\mu}$. Then, according to the shrinkage operation (soft thresholding) proposed by Chartrand (2012), $_E$ is updated as:

$$E^{r+1} = shrink_p\left\{X - U^r\left(V^t\right)^T - \frac{Y^r}{\mu^r}, \frac{1}{\mu^r}\right\} \quad (10)$$

Problem 2: When $_E$ and $_V$ are fixed, (7) is simplified as follows:

$$L_{\mu,Y}(E,U,V) = \frac{\mu}{2} \| \text{E-X} + UV^T + \frac{Y}{\mu} \|_F^2 + \lambda \| U \|_2^1 \quad (11)$$

To simplify (11), let $H = X - E - \frac{Y}{\mu}$. Then, (11) is written as:

$$L_{\mu,Y}(E,U,V) = \frac{\mu}{2} \| UV^T - H \|_F^2 + \lambda \| U \|_2^1 \quad (12)$$

The partial derivatives of $_L$ with respect to $_U$are:

$$\frac{\partial L}{\partial U} = \mu(UV^T - H)V + 2\lambda QU \quad (13)$$

where $_{Q \in R^{p \times p}}$ is a diagonal matrix with $q_{i,i} = \frac{1}{\|U_{(i,:)}\|_2} (i = 1, \cdots, p)$ (Xiang et al., 2012). Letting (13) be equal to 0, the following update rule for $_U$ is then obtained:

$$U^{r+1} = \left(I + \frac{2\lambda}{\mu^r}Q^r\right)^{-1} H^r V^r \quad (14)$$

To simplify (14), let $A^r = \left(I + \frac{2\lambda}{\mu^r}Q^r\right)^{-1}$, and then (14) is written as:

$$U^{r+1} = A^r H^r V^r \quad (15)$$

Problem 3: When $_E$ and $W$are fixed, (7) is simplified as follows:

$$L_{\mu,Y}(E,U,V) = \frac{\mu}{2} \| \text{E-X} + UV^T + \frac{Y}{\mu} \|_F^2 + \alpha Tr(V^T L V) \quad (16)$$

With respect to the settings $H = X - E - \frac{Y}{\mu}$, (16) can be written equivalently as:

$$\begin{aligned} L_\mu(E,U,V) &= \frac{\mu}{2} \| UV^T - H \|_F^2 + \alpha Tr(V^T L V) \\ &= \frac{\mu}{2} Tr((UV^T - H)(UV^T - H)^T) + \alpha Tr(V^T L V) \end{aligned} \quad (17)$$

Based on (17), $_V$is found by minimizing:

$$V = \arg\min_V Tr(V^T(\frac{\alpha}{\mu}\text{L-H}^T AH)V) \quad (18)$$

Therefore, $_{V^{r+1}}$ can be obtained as follows:

$$V^{r+1} = (v_1, ..., v_k) \quad (19)$$

where $(v_1, ..., v_k)$ are the $_k$eigenvectors corresponding to the smallest $_k$ eigenvalues of the matrix $\frac{\alpha}{\mu}$L-H$^T AH$. Thus, based on the ALM, ADM and the shrinkage operation, the solution to solve the optimization model described in (5) is shown in **Algorithm 1**. In the optimization model, there are six parameters $_k$, $_p$, $_\lambda$, $_\alpha$, $_\rho$, $_\mu$ to be pre-determined, among them. As the parameters used to control the step size in the update rule of AML, we set $\mu = 10^{-2}$ and $\rho = 1.2$ for all gene expression datasets experiments (Feng et al., 2016). The parameter $_k$ is determined refering to the number of prior categories of each dataset. For the three essential parameters $_p$, $_\lambda$, $_\alpha$, to be determined in (5), we choose them corresponding to different situations for the best clustering performance through extensive experiments. Different parameters are chosen for different datasets. Detailed parameter selections and discussions are described in section "Experiments and Discussion."

ALGORITHM 1 | The solution to optimized (5).

Input:

Gene expression data matrix: $X_{p \times n}$,

Parameters: $_k$, $_p$, $_\lambda$, $_\alpha$, $_\rho$, $_\mu$

Output:

$U_{p \times k}$, $V_{n \times k}$

Initialize:

E, Y, U, V

Do

Update $_U$ by (14)

Update $_V$ by (19)

Update $_E$ by (10)

Update $_\mu$ by $\mu = \rho\mu$

Update $_Y$ by $Y^{r+1} = Y^r + \mu^r(E^r - X + U^r(V^r)^T)$

Update $_\mu$ by $\mu^{r+1} = \rho\mu^r$

Until convergence

EXPERIMENTS AND DISCUSSION

Gene Expression Datasets

Five gene expression datasets, which include one benchmark dataset, two single-cancer datasets from TCGA, and two integrated multicancer datasets from TCGA, are used to evaluate

the performance of PL21GPCA. The verified experiments consist of two aspects: "tumor sample clustering" and "gene network module discovery." Based on the optimal low-dimensional ($_k$-dim) subspace denoted as $_{V^T=(v_1,\cdots,v_n)\in R^{k\times n}\ (V^TV=I)}$, the classical clustering method K-means is then used for tumor clustering. For comparison, extensive experiments are also performed using existing dimensionality reduction methods, including SPCA (Journee et al., 2010), RPCA (Candès et al., 2009), gLPCA (Jiang et al., 2013), pgLPCA (Feng et al., 2017) and GNMF (Cai et al., 2011). Among the compared methods, some are based on PCA, and some introduce the graph Laplacian regularization item. Based on the optimal projected matrix $U_{p\times k}$, the differentially expressed genes are selected for gene network analysis to find key genes with close relationships to some cancers.

The details of the five data sets are as follows. The benchmark gene expression dataset is lung cancer data (Bhattacharjee et al., 2001) that have often been employed by researchers to evaluate their algorithms (Lee et al., 2010; Kong et al., 2017), consisting of 12,625 genes of 56 samples. There are four types of lung cancer in the 56 samples: pulmonary carcinoid (20), colon metastases (13), small cell lung carcinoma samples (6) and normal lung samples (17). The two single-cancer datasets and the two integrated multicancer datasets are all from The Cancer Genome Atlas (TCGA) which is known as the largest tumor specimens database. The genomic data provided by TCGA include DNA methylation, microRNA expression, gene expression, protein expression, and DNA copy number, etc. We downloaded gene expression datasets (at level 3) of five different cancers from TCGA: colorectal cancer (CRC), cholangiocarcinoma (CHOL), squamous cell carcinoma of head and neck (HNSC), pancreatic cancer (PAAD), and esophageal cancer (ESCA). Each dataset consists of 20,502 genes expressed in different numbers of samples. In our experiments, CRC and CHOL are used as single-cancer datasets to evaluate the performance of the PL21GPCA method. There are 281 samples for CRC and 45 for CHOL. Each of these two datasets contains two types of cancer samples, "negative" and "positive." "Negative" or "NT" represents normal samples. "Positive" or "TP" represents diseased samples. There are 262 "TP" samples in the CRC data and 36 in the CHOL data, and the rest are "NT" samples. Two integrated datasets are used to further verify the performance of the PL21GPCA method. Each integrated dataset consists of 3 types of cancers. One of the integrated datasets, H_C_P, contains 836 "TP" samples, among which the sample numbers of the three cancers are 398 (HNSC), 262 (CRC), and 176 (PAAD). The other integrated dataset, E_C_C, contains 481 "TP" samples, in which the sample numbers of the three cancers are 183 (ESCA), 36 (CHOL), and 262 (CRC). The statistics of these datasets are summarized in **Table 1**.

Tumor Sample Clustering

Evaluation Metric

Based on the optimal principal directions $V^T = (v_1, \cdots, v_n) \in R^{k\times n}\ (V^TV = I)$, the K-means algorithm is then employed for tumor sample clustering. The accuracy (ACC) and the normalized mutual information (NMI) are the two most commonly used metrics to evaluate the clustering results (Cai et al., 2005). For the $_i$th sample, we use p_i to denote the prior label and r_i to denote the obtained clustering label. The metric ACC is defined as follows:

$$ACC = \frac{\sum_{i=1}^{n} \theta(p_i, map(r_i))}{n}, \tag{20}$$

where $_n$ denotes the total number of samples in every dataset. The function $\theta(x, y)$ equals 1 if $x = y$ and 0 otherwise. The function $map(r_i)$ maps each obtained cluster label r_i to the equivalent prior label. Let $_C$ be the prior set of clusters and C' be the obtained set from our algorithm. Define their mutual information metric $MI(C, C')$ as:

$$MI(C, C') = \sum p(c_i, c_j^{'}) \log_2 \frac{p(c_i, c_j^{'})}{p(c_i) \cdots p(c_j^{'})} \tag{21}$$

where $p(c_i)$ and $p(c_j^{'})$are the probabilities that a sample arbitrarily selected from the dataset belongs to clusters c_i and $c_j^{'}$, respectively, and $p(c_i, c_j^{'})$ is the joint probability. In the experiments, the metric NMI is defined as follows:

$$NMI(C, C^{'}) = \frac{MI(C, C^{'})}{\max(H(C), H(C^{'})} \tag{22}$$

where $H(C)$ and $H(C^{'})$are the entropies of $_C$ and $_{C'}$, respectively. Therefore, the metric $NMI(C, C^{'})$ ranges from 0 to 1. $NMI = 1$ if the two sets of clusters are identical, and if the two sets are independent, $NMI = 0$.

However, a problem that needs to be resolved is that the K-means algorithm may or may not converge to the same solution in each run with random initial conditions. Therefore, the evaluated metrics ACC and NMI obtained by only once-running of k-means is not enough to explain the result. To solve this problem, for the given cluster number $_k$, K-means was run 50 times on each dataset, and the average performance was computed. As a reference, we also recorded the maximum values of ACC and NMI of the 50 runs. Thus, four metrics, ACC_max, ACC_mean, NMI_max and NIM_mean, are used to evaluate our experiments. Generally, the larger the mean value is, the better is the clustering performance, and the better are the stability and robustness of the clustering. This also indicates that the corresponding dimension reduction method has good robustness and sparse effect.

TABLE 1 | Statistical information on the experimental data.

Dataset		# of genes ($_p$)	# of samples ($_n$)	# of classes ($_k$)
Benchmark Data	Lung Cancer	12625	56	4
Single-Cancer Data from TCGA	CRC	20502	281	2
	CHOL	20502	45	2
Integrated Cancer Data from TCGA	H_C_P	20502	836	3
	E_C_C	20502	481	3

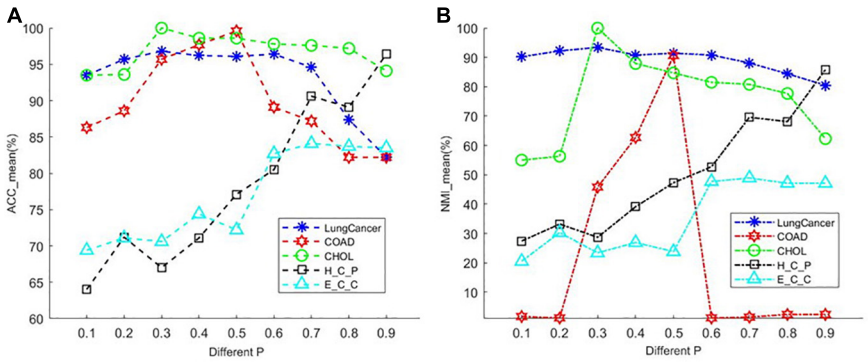

FIGURE 2 | The average performance taking the essential parameter at nine different values from 0.1 to 0.9. **(A)** The mean value of ACC for different cancer datasets. **(B)** The mean value of NMI for different cancer datasets.

Parameter Selection

The PL21GPCA model has three essential parameters, p, λ, and α, which need to be determined in (5). The range of each parameter is $0 < p < 1$, $\lambda > 0$, $\alpha > 0$. When determining the optimal value of one parameter, the other two parameters are fixed. We focus on the influence of the value of p on the performance. PL21GPCA achieves consistently good performance when the two regularization parameters λ and α are varied from 10 to 1,000 on all three datasets. **Figure 2** shows how the average performance varies when taking the essential parameter p at nine different values from 0.1 to 0.9. For every dataset, extensive experiments are carried out to seek the appropriate parameters to achieve the best performance for tumor sample clustering. Thus, different parameters are chosen for different datasets (see **Table 2**).

There is another parameter that is not appear in the objective function of PL21GPCA. However, it is also an important parameter affecting the performance of our method. It is parameter k', the number of nearest neighbors of every point when constructing the graph in the step of graph Laplacian embedding. Setting this parameter too small may cause overfitting, and too large may increase the error. By extensive experiments, we find that the appropriate value for this parameter is near the square root of the sample number for different datasets.

Clustering Results

Tables 3–5 show the clustering results on the lung cancer data, single-cancer data from TCGA (CRC and CHOL datasets), and integrated cancer data (H_C_P and E_C_C datasets), comparing the PL21GPCA-based method with the competitors. For each dataset with a given cluster number k, the K-means algorithm was run 50 times to randomize the experiments. The maximum and the mean value metrics are all presented in the tables. The performance of the PL21GPCA-based method is highlighted in bold in the tables. Regardless of the datasets, the PL21GPCA-based method always results in the best performance on the mean value metrics ACC_mean and NMI_mean. As mentioned above, the mean value is more meaningful than the maximum value, which is for reference only. By leveraging the power of three

TABLE 2 | Values of the three parameters p, λ, and α for different datasets.

Dataset	Lung Cancer	CRC	CHOL	H_C_P	E_C_C
Parameter selections	$p = 0.3$ $\lambda = 10$ $\alpha = 100$	$p = 0.5$ $\lambda = 100$ $\alpha = 100$	$p = 0.3$ $\lambda = 10$ $\alpha = 100$	$p = 0.9$ $\lambda = 100$ $\alpha = 100$	$p = 0.7$ $\lambda = 10$ $\alpha = 100$

TABLE 3 | Clustering performance on lung cancer.

Methods	ACC (%)		NMI (%)	
	ACC_Max	ACC_mean	NMI_Max	NMI_mean
SPCA	100	84.39	100	83.07
RPCA	*100*	*86.25*	*100*	*84.77*
GNMF	85.71	79.71	75.57	69.62
gLPCA	89.29	78.5	80.82	69.86
pgLPCA	100	82	100	80.05
PL21GPCA	**100**	**96.82**	**100**	**93.44**

TABLE 4 | Clustering performance on CRC and CHOL.

Data	Method	ACC (%)		NMI (%)	
		ACC_Max	ACC_mean	NMI_Max	NMI_mean
CRC	SPCA	92.17	87.57	35.3	22.57
	RPCA	*98.22*	*67.95*	*69.82*	*24.33*
	GNMF	88.61	60.5	30.79	18.93
	gLPCA	90.75	87.01	22.7	15
	pgLPCA	94.31	78.65	42.67	20.1
	PL21GPCA	**99.64**	***99.64***	**90.55**	***90.55***
CHOL	SPCA	100	93.38	100	60.65
	RPCA	*100*	*100*	*100*	*100*
	GNMF	100	100	100	100
	gLPCA	100	78.04	100	54.87
	pgLPCA	100	81.87	100	59.83
	PL21GPCA	**100**	**100**	**100**	**100**

TABLE 5 | Clustering performance on H_C_P and E_C_C.

Data	Method	ACC (%)		NMI (%)	
		ACC_Max	ACC_mean	NMI_Max	NMI_mean
H_C_P	SPCA	55.26	51.82	17.85	14.98
	RPCA	*91.87*	*77.3*	*71.43*	*68*
	GNMF	57.3	54.02	29.59	22.59
	gLPCA	55.62	52.96	29.43	16.89
	pgLPCA	86.96	70.26	58.42	45.4
	PL21GPCA	**96.41**	**96.41**	**85.77**	**85.75**
E_C_C	SPCA	71.52	67.9	19.28	15.06
	RPCA	*81.08*	*76.17*	*55.72*	*32.47*
	GNMF	68.4	62.05	19.03	9.29
	gLPCA	70.69	69.58	23.14	19.7
	pgLPCA	79.63	72.72	41.33	31.35
	PL21GPCA	**85.65**	**84.09**	**60.31**	**47.15**

measures, including taking the proximal Lp-norm $\|g\|_p$ $(0 < p < 1)$ on the loss function, employing the $L_{2,1}$-norm regularization item to insure feature selection, and introducing the Laplacian regularization item to emphasize the geometrical structure of the data, the PL21GPCA-based method can always get a better clustering performance.

For the different types of data used in the experiments, a number of meaningful points need to be emphasized further.

The benchmark data

For the lung cancer dataset, **Table 3** shows that the PL21GPCA-based method achieves the same performance as SPCA, RPCA and pgLPCA considering the maximum value metrics (the ACC_max and the NIM_max are also 100%) but is obviously superior to the other methods in terms of the mean value metric (ACC_mean reaches 96.82% and the NIM_mean reaches 93.44%).

Single-cancer data from TCGA

Table 4 shows the clustering performance of the two single-cancer datasets from TCGA. For the CRC dataset, our method presents very superior performance compared with other methods, with the ACC_mean reaching 99.64% as well as the ACC_max. The good average performance shows the robustness of the PL21GPCA method. In addition, the two NMI metrics (all reaching 90.55%) also go far beyond the performance of other methods. For the CHOL dataset, all the methods achieve the same results (100%) when considering the maximum value metrics. Our method achieves the same performance (100%) as GNMF and RPCA in terms of the mean value metrics. A surmise is reported that there may be distinct discriminations for the two kinds of samples in the original CHOL data (Kong et al., 2017).

Integrated multicancer data from TCGA

Table 5 reports the estimation results on the two integrated datasets. It shows that the PL21GPCA method performs much better than the competitors. As highlighted in bold in **Table 5**, for H_C_P data, the ACC_max and the ACC_mean all reach 96.41%, and the NMI_max and the NMI_mean are also superior to the corresponding values for other methods. For E_C_C data, our method is still outstanding; taking the ACC metric as an example, the ACC_max reaches 85.65%, and the ACC_mean reaches 84.09%. Based on the excellent performance on these two integrated datasets, should we speculate that the PL21GPCA method is more suitable for learning the compact representation of higher-dimensional and more complex data than its competitors, which needs further verification.

Finally, as we can see from **Tables 3–5**, among the compared methods, the RPCA method performs second to our method and better than the other competitors, such as SPCA, GNMF, gLPCA, and pgLPCA. The performance of RPCA is in italics in the tables. If the intrinsic geometric structure is introduced to RPCA, will the performance be improved further? This question is also worth further verification.

Embedding Evaluation

To further show the performance of the novel dimensionality reduction method compared others, a visualized data distribution of the low-dimensional embedding corresponding to the first two components of the PCA-based method are demonstrated. Besides the proposed method PL21GPCA, the results of three other methods including SPCA, gLPCA, pgLPCA are compared because these methods are also the direct extensions of PCA. **Figure 3** presents the sample clustering results in a two-dimensional space. We choose two representative datasets CRC data and H-C-P data to show the results. **Figures 3A–D** are the results of the compared methods SPCA, glPCA, gpLPCA and PL21GPCA, respectively, on the CRC dataset. **Figures 3E–H** are the compared results of the four methods on the H-C-P dataset. No matter for the CRC data which contains two types of cancer samples, or for the H-C-P data which contains three types of cancer samples, SPCA and gLPCA make the samples from different categories being mixed together, and the pgLPCA can only separate the samples into categories roughly, so they have unideal clustering results. However, PL21GPCA make the embeddings of samples in clearer distribution. Therefore, the clustering results is better than the compared methods. The visualized results verified the robustness and the flexibility of the proposed model.

Experiments on Simulated data

Experiments on simulation data are also carried out to evaluate the effectiveness of PL21GPCA. The simulation data used in the experiment is a matrix $X_{3000\times80}$ generated by *rand* function in Matlab. In order to simulate the representation of features in different types of samples, based on the generated matrix $X_{3000\times80}$, some changes have also been made. Firstly, we add 1 to the values of columns 1 to 20 in rows i^*30-29 $(i=1,\cdots100)$ of matrix $X_{3000\times80}$, add 2 to the values of columns 21 to 40 in rows i^*30-19 $(i=1,\cdots100)$, add 3 to the values of columns 41 to 60 in rows i^*30-9 $(i=1,\cdots100)$, add 4 to the values of columns 61 to 80 in rows i^*30-5 $(i=1,\cdots100)$, add 2 to the values of columns 21 to 40 in rows i^*30-25 $(i=1,\cdots100)$, add 1 to the values of columns 1 to 20 in rows i^*30-15 $(i=1,\cdots100)$, which means that the 80 samples in the simulation data contain four categories. Secondly, we use the function *imnoise* in matlab to add different sizes of Gaussian white noise

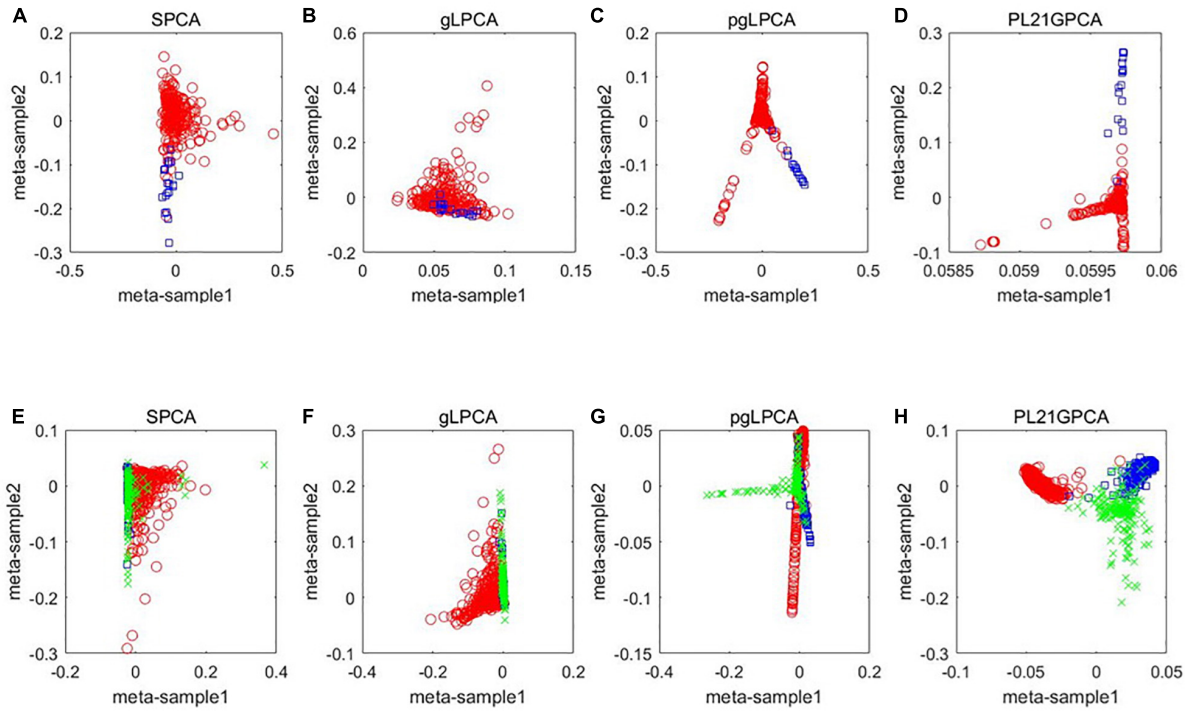

FIGURE 3 | A visualized comparison of low-dimensional embeddings by SPCA, gLPCA, pgLPCA, and PL21GPCA on COAD and H-C-P datasets. **(A–D)** Are the results of the compared methods SPCA, glPCA, gpLPCA, and PL21GPCA respectively on the CRC dataset. **(E–H)** Are the compared results of the four methods on the H-C-P dataset.

to X. The mean value of the added Gaussian white noise is 0 and the variance $_{\sigma^2}$ is chosen in the range of [0.4∼1.2]. Next, we use the proposed method PL21GPCA and the compared methods to reduce the dimension and denoise the simulated data, and then use the K-means method to cluster the denoised data, the evaluation metric ACC_mean mentioned above is used to test the effectiveness of the method. the K-means algorithm is run 50 times to randomize the experiments.

Table 6 shows the experiments results on simulated data. It can be seen evidently that the performances of all methods change with the increase of noise. The best performance of different methods when adding different noises are marked with black bold. Although the performance of pl21GPCA is second only to RPCA when the noise is low ($\sigma^2 = 0.4$), with the increase of Gaussian white noise, the effect of our proposed method is mostly ahead of other methods especially when $\sigma^2 =$ 0.6, 0.8, 1.2, which shows that the new method has better denoising ability and robustness.

TABLE 6 | Clustering performance on simulated data with different Gaussian white noise.

Simulated data	SPCA	RPCA	GNMF	gLPCA	gpLPCA	PL21GPCA
σ^2=0.4	96.6	**99.75**	95.35	87.37	89.47	99.45
σ^2=0.6	94.35	91.33	94.35	84.68	86.45	**97.45**
σ^2=0.8	85.87	91.1	93.85	83.2	85.57	**94.35**
σ^2=1.0	80.12	90.85	**93.4**	86.48	85.67	93.33
σ^2=1.2	70.25	76.43	73.58	85.83	82.12	**87.15**

Gene Network Module Discovery

Due to the outstanding performance of our method on the CRC dataset and the integrated H_C_P dataset, the construction and analysis of the gene network are based on these two datasets. The strategy of gene network module discovery involves two steps. First, the genes for constructing the co-expression gene networks are selected. Second, based on the filtered genes, co-expression networks are established, and then the key genes that may be closely related to some cancers are analyzed.

Gene Selection

In this step, there are two problems to be solved: one is how to select genes, and the other is how many to select. It is known that among thousands of genes, only a handful of them regulate a specific biological process (Delbert et al., 2005; Liu et al., 2013). These minority of genes are called differentially expressed genes (Liu J. et al., 2015). In this article, the differentially expressed genes are selected to carry out gene network analysis according to the projected matrix $U_{p\times k}$. Now, we mark the optimal projected matrix $U_{p\times k}$ as $\tilde{U}$; therefore, these differentially expressed genes can be identified according to $\tilde{U}$ (Liu J. et al., 2015; Feng et al., 2016). We denote $\tilde{U}$as follows:

$$\tilde{U} = \begin{bmatrix} \tilde{u}_{11} & \tilde{u}_{12} & \cdots & \tilde{u}_{1k} \\ \tilde{u}_{21} & \tilde{u}_{22} & \cdots & \tilde{u}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{u}_{p1} & \tilde{u}_{p2} & \cdots & \tilde{u}_{pk} \end{bmatrix} \tag{23}$$

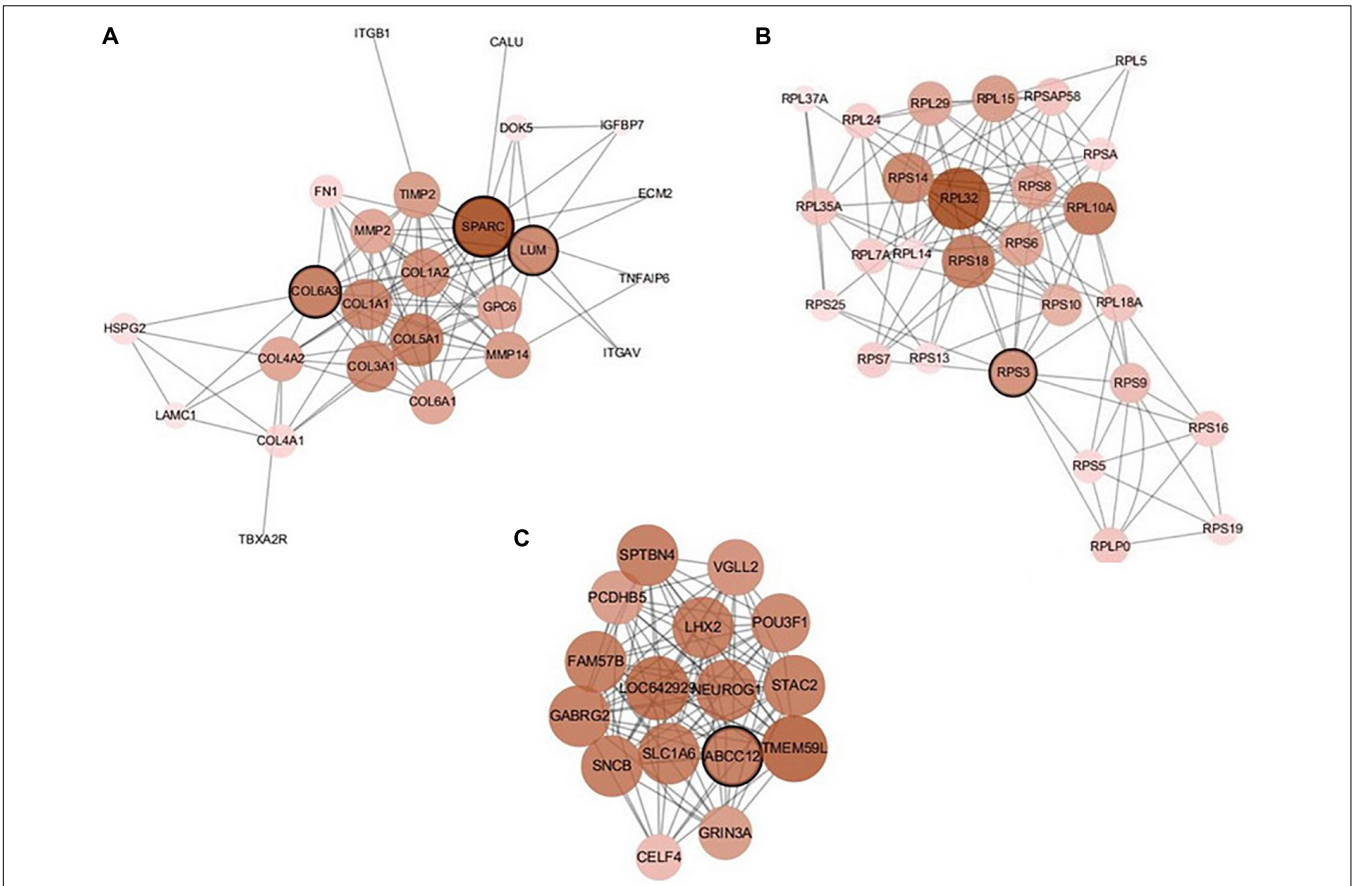

FIGURE 4 | The first three modules of the constructed network based on the CRC data. The five marked genes SPARC, ABCC12, COL6A3, LUM, and RPS3 have been confirmed to be associated with CRC and other cancers. **(A)** Module 1; **(B)** Module 2; **(C)** Module 3.

The upregulated genes are reflected by the positive value in the matrix $\tilde{U}$, and the downregulated genes are reflected by the positive value (Liu et al., 2013). Therefore, the absolute value of the items in $\tilde{U}$ is used to identify the differentially expressed genes. The items of each row in $\tilde{U}$ are summed, and then the evaluating vector denoted as $\hat{U}$ is obtained:

$$\hat{U} = \left[\sum_{j=1}^{k} \tilde{u}_{1j} \ \sum_{j=1}^{k} \tilde{u}_{2j} \ \cdots \ \sum_{j=1}^{k} \tilde{u}_{pj} \right]^T \tag{24}$$

The larger item in $\hat{U}$ indicates the more strongly differentially expressed gene. Therefore, we sort the elements in $\hat{U}$ in descending order and take the top $l(l \ll p)$ elements. In many studies, it has been unclear how many genes should be selected for gene network analysis. Since only a small number of genes can regulate a specific biological process, these genes may play a decisive role in the clustering results of tumor samples. In this paper, the number of genes used for constructing the gene network is determined according to the clustering performance based on the selected genes. Through experimentally investigating the clustering performance with the number of selected genes varied from 500 to 2000, it is found that the clustering results corresponding to **1600** genes are best for the CRC data and **700** for the H_C_P data.

Construction of Gene Networks

Suppose l differentially expressed genes are used to construct the gene network. Let matrix $R_{l\times n}$ denote the l gene expression in $_n$ samples. We use the Pearson correlation coefficient (PCC) (Hou et al., 2019) to measure the correlation of any two genes in $R_{l\times n}$. The values in the PCC matrix vary in the range of [0, 1]. The larger the PCC value is, the higher the correlation is. Based on matrix $R_{l\times n}$, an adjacency matrix $A_{l\times l}$ can be calculated. According to the adjacency matrix, an intuitive visualized graph of the gene interaction network composed of several modules is obtained.

Analysis of Gene Network Modules

There are 39 modules, including 218 nodes and 504 edges, in the constructed network based on the CRC data. We analyzed the top 10 nodes (genes) with higher degrees in the first three modules that retained more relevant interactions. The degree of the node (gene) shows its role in the network modules. The larger the degree of the node (gene) is, the more important the node (gene) is, and such nodes (genes) may retain the

tight connectivity of the network. **Figure 4** shows the main part of the first three gene network modules in which a small number of nodes whose degree is very low have been removed. The roles of the top ten genes in the first three modules are illustrated in **Figure 4**. The degree value of a node in **Figure 4** is represented by its size and color. The larger the node is, the darker its color is, which corresponds to a larger degree of the node. Referring to GeneCard with its website http://www.genecards.org/, we list the annotations of the top ten genes in **Table 7**. Five of the top ten genes have been validated as associated with multiple cancers: SPARC, ABCC12, COL6A3, LUM, and RPS3. The corresponding nodes of these genes are marked with a black outline in **Figure 4** and are also shown in bold in **Table 7**. In the literature (Liu Q. Z. et al., 2015), the gene SPARC has been recommended as a predictor of colorectal cancer. The gene ABCC12 is a human ATP binding cassette (ABC) transporter and is a multidrug resistance protein (MRP9). However, MRP9 has been recognized as an important target for the immunotherapy of breast cancer (Bera et al., 2002). Studies have shown that colorectal cancer can be predicted by the gene COL6A3 because it is overexpressed in samples of colorectal cancer. Therefore, COL6A3 is considered a potential diagnostic and prognostic marker gene for colorectal cancer (Qiao et al., 2015). As one of the members of the leucine-rich proteoglycan family, the gene Lumican (LUM) is overexpressed in many kinds of cancers, including colorectal, neuroendocrine, cervical, carcinoid, breast, and pancreatic cancer. LUM also causes the growth and invasion of pancreatic cancer (Ishiwata et al., 2007). The ribosomal protein gene S3 (RPS3) is also overexpressed in colorectal cancer. Researchers found an increase in ribosome synthesis in patients with colorectal cancer (Pogue-Geile et al., 1991). Although the other five genes RPL32, TMEM59L, LOC642929, LHX2, and TLCD3B have not been identified in clinical studies indicating their effect on cancers, they may be considered candidate oncogenes because of their high ranking in our constructed gene network modules. By constructing co-expression gene network modules based on the CRC dataset, we found some disease-causing genes for colorectal cancer and other related cancers. It shows that constructing gene network modules via the genes filtered based on PL21GPCA can help us discover the key oncogenes.

TABLE 7 | Annotations of the top ten genes in the first three network modules based on CRC data.

Gene	Summary
RPL32	A protein coding gene. Diseases associated with RPL32 include frontal convexity meningioma and retinitis pigmentosa 49
SPARC	Diseases associated with SPARC include osteogenesis imperfecta, type xvii and osteogenesis imperfecta, type iv
TMEM59L	TMEM59L (Transmembrane Protein 59 Like) is a protein coding gene. An important paralog of this gene is TMEM59
LOC642929	LOC642929 (General Transcription Factor II, I Pseudogene) is a pseudogene
ABCC12	Diseases associated with ABCC12 include familial cold autoinflammatory syndrome 1 and episodic kinesigenic dyskinesia 1. An important paralog of this gene is ABCC11
COL6A3	A protein coding gene. An important paralog of this gene is COL6A6
LUM	Among its related pathways are defective ST3GAL3, which causes MCT12 and EIEE15, and keratin sulfate/keratin metabolism
LHX2	LHX2 (LIM Homeobox 2) is a protein coding gene. Diseases associated with LHX2 include schizencephaly and retinitis pigmentosa
TLCD3B	TLCD3B (TLC Domain Containing 3B) is a protein coding gene. An important paralog of this gene is TLCD3A
RPS3	Diseases associated with RPS3 include eumycotic mycetoma and Waardenburg syndrome, type 3

The constructed network based on the integrated data H_C_P includes 157 nodes and 644 edges. We analyzed the five important nodes (genes) with higher degrees in the first three modules that retained more relevant interactions. **Figure 5** illustrates the main part of the first three gene network modules in which the nodes of very low degree have also been removed. Referring to GeneCards, their annotations are listed in **Table 8**. The five genes RPL32, EEF1G, SPRR1B, COL1A2, and MMP2 have been recognized to be related to multiple cancers. The corresponding nodes of these genes are marked with a black outline in **Figure 5**. Wan et al. (2004) conducted large-scale experiments on human liver cancer cells. Research has shown that RPL32 is one of the potential genes that affect human cell growth and cancer formation and provides an important tool for diagnostic markers and drug targets (Wan et al., 2004). EEF1G has been thought to be a characteristic gene for colorectal cancer; it is highly expressed in most colorectal cancers and could be considered a marker gene for colorectal cancer detection (Matassa et al., 2013). In addition, the expression level of EEF1G in pancreatic tumor cells was higher than that in normal cells (Lew et al., 1992). SPRR1B is overexpressed in human oral squamous cells. It has been experimentally proven that SPRR1B overexpression in cells will signal MAP kinases but inhibit MAP kinase signals, so SPRR1B can affect cell growth and maintenance (Michifuri et al., 2013). Kiyoshi Misawa and other researchers mainly studied the expression of COL1A2 in head and neck squamous cell carcinoma (HNSC) and found that hypermethylation of CpG may cause inactivation of the gene COL1A2. Therefore, the COL1A2 gene may affect the formation and development of HNSC and could become a major biomarker (Misawa et al., 2011). As a member of the matrix metalloproteinase (MMP) gene family. MMP2 is relevant to the generation of malignant tumors, including colorectal cancer, lung cancer, and breast cancer (Yu et al., 2002; Arajo et al., 2015; Ren et al., 2015). Analysis through the gene network constructed based on integrated multicancer data is helpful for mining the interrelationships between different cancers and genes. It may provide an important reference for the diagnosis and treatment of various diseases.

CONCLUSION AND SUGGESTIONS

In this article, we propose a new dimensionality reduction method named PL21GPCA based on PCA for robust tumor sample clustering and gene network module discovery. Based

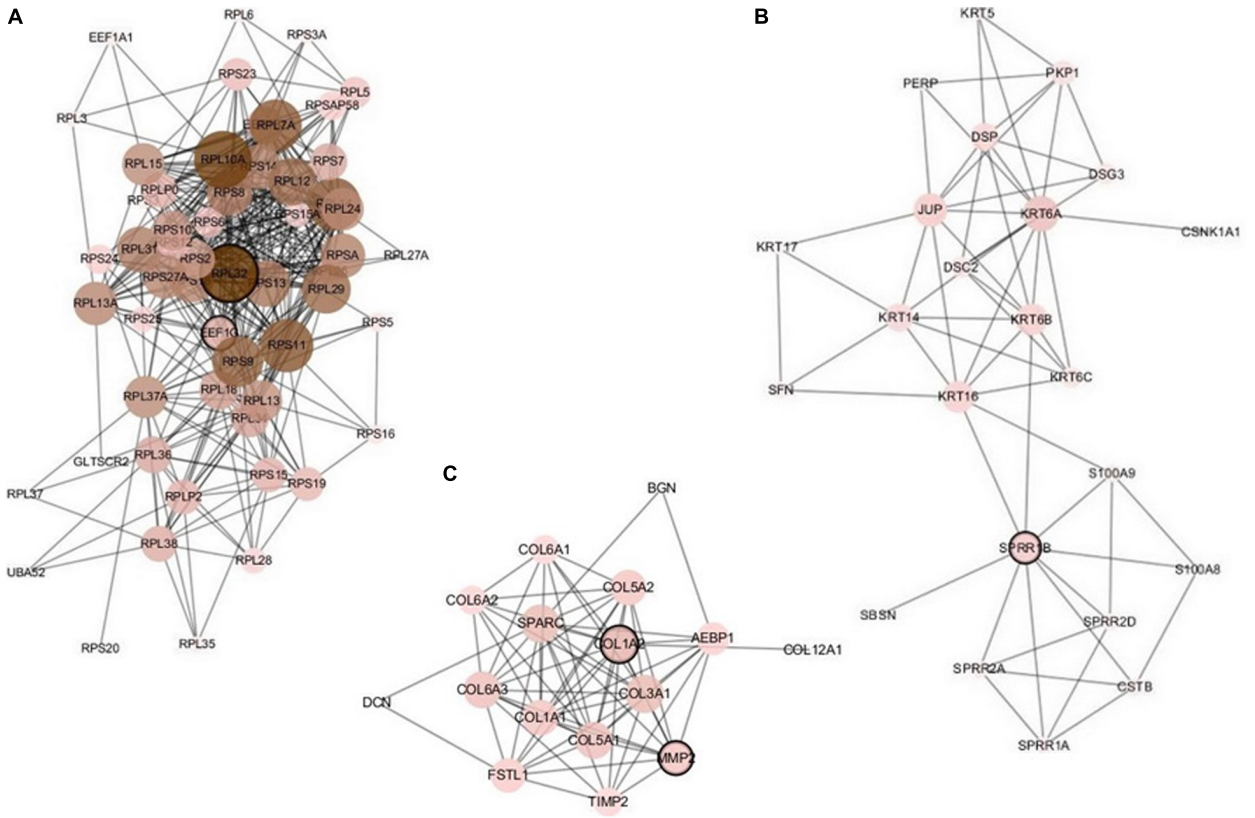

FIGURE 5 | The first three modules of the constructed network based on the H_C_P data. The five marked genes RPL32, EEF1G, SPRR1B, COL1A2, and MMP2 have been confirmed to be associated with multiple cancers. **(A)** Module 1; **(B)** Module 2; **(C)** Module 3.

on the traditional PCA, the non-convex proximal Lp-norm $\|g\|_p$ $(0 < p < 1)$is applied on the loss function to decrease the sensitivity to outliers and noise. The L_2,$_1$-norm is used on the projected matrix to enhance the sparse gene expression in cancer samples. The graph regularization item is introduced to the optimization model to retain the geometric structure of the data. Five gene expression datasets, including one benchmark dataset, two higher-dimensional single-cancer datasets from TCGA, and two integrated multicancer datasets from TCGA, are used to evaluate the performance of our method. The compared experiments demonstrate that the PL21GPCA method outperforms many existing methods in terms of tumor sample clustering. Moreover, this method is employed to discover gene network modules to find the key genes with close relationships to cancers. The results of our study may be a useful reference for clinical diagnosis.

There are some suggestions for future research. First, in the optimization model of PL21GPCA, the constraint used on the loss function is the non-convex proximal Lp-norm $\|g\|_p$ $(0 < p < 1)$, since Lp-norm minimization can result in a sparser solution than the L_1-norm and perform better in terms of robustness to outliers than the L_2-norm. However, in addition to the generalized shrinkage operation proposed by Chartrand (2012), there are some other suggestions to address the Lp-norm $(0 < p < 1)$ minimization (Guo et al., 2013; Qin et al., 2013) problems. Therefore, we will continue to explore other solutions to the optimization model with the Lp-norm $\|g\|_p$ $(0 < p < 1)$. Second, we will evaluate the performance of PL21GPCA as a compact representation method combined with other methods, including supervised and unsupervised clustering methods such as spectral clustering, support vector machine (SVM) or their improved versions. Third, as mentioned above, the PL21GPCA method gets especially outstanding performance

TABLE 8 | Annotations of the most important five genes in the first three network modules based on H_C_P data.

Gene	Summary
RPL32	A protein coding gene. Diseases associated with RPL32 include frontal convexity meningioma and retinitis pigmentosa 49
EEF1G	Diseases associated with EEF1G include gastrointestinal carcinoma. Among its related pathways are viral mRNA translation and gene expression
SPRR1B	A protein coding gene. An important paralog of this gene is SPRR1A
COL1A2	Among its related pathways are ERK signaling and IL4-mediated signaling events
MMP2	Among its related pathways are direct p53 effectors and development endothelin-1/EDNRA signaling

for processing the integrated data, so we will use the PL21GPCA method to process many other integrated data to verify its performance further.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The lung cancer data http://www.unc.edu/∼haipeng. The TCGA data http://www.tcga.org/.

AUTHOR CONTRIBUTIONS

X-ZK conceived and designed the experiments. YS and X-ZK performed the experiments and contributed to the writing of the manuscript. S-SY, JW, and L-YD analyzed the data. J-XL and C-HZ contributed to reagents, materials, and analysis tools. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by grants from the National Natural Science Foundation of China (No. 61702299) and jointly in part by the National Natural Science Foundation of China, Nos. 61872220, 61701279, and 61902215.

ACKNOWLEDGMENTS

Thanks a lot for my co-tutor Yong Xu who is now a professor in Harbin Institute of Technology, Shenzhen, China.

REFERENCES

Arajo, R. F., Lira, G. A., Vilaa, J. A., Guedes, H. G., Leito, M. C. A., Lucena, H. F., et al. (2015). Prognostic and diagnostic implications of MMP-2, MMP-9, and VEGF-a expressions in colorectal cancer. *Pathol. Res. Practice* 211, 71–77. doi: 10.1016/j.prp.2014.09.007

Belkin, M., and Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering. *Adv. Neural Inf. Process. Syst.* 14, 585–591.

Belkin, M., and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* 15, 1373–1396. doi: 10.1162/089976603321780317

Bera, T. K., Iavarone, C., Kumar, V., Lee, S., Lee, B., and Pastan, I. (2002). MRP9, an unusual truncated member of the ABC transporter superfamily, is highly expressed in breast cancer. *Proc. Natl. Acad. Sci. U.S.A.* 99, 6997–7002. doi: 10.1073/pnas.102187299

Bertsekas, D. P. (1982). *Constrained Optimization and Lagrange Multiplier Methods*. Cambridge, MA: Academic Press.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., et al. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. U.S.A.* 98, 13790–13795. doi: 10.1073/pnas.191502998

Bunte, K., Leppaaho, E., Saarinen, I., and Kaski, S. (2016). Sparse group factor analysis for biclustering of multiple data sources. *Bioinformatics* 32, 2457–2463. doi: 10.1093/bioinformatics/btw207

Cai, D., He, X. F., and Han, J. W. (2005). Document clustering using locality preserving indexing. *IEEE Trans. Knowl. Data Eng.* 17, 1624–1637. doi: 10.1109/tkde.2005.198

Cai, D., He, X. F., Han, J. W., and Huang, T. S. (2011). Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 1548–1560. doi: 10.1109/tpami.2010.231

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2009). Robust principal component analysis? *J. ACM* 58, 1–37.

Chartrand, R. (2012). Nonconvex splitting for regularized low-rank + sparse decomposition. *IEEE Trans. Signal Process.* 60, 5810–5819. doi: 10.1109/tsp.2012.2208955

Chen, X. J., Huang, J. Z., Wu, Q. Y., and Yang, M. (2019). Subspace weighting co-clustering of gene expression data. *IEEE-ACM Trans. Comput. Biol. Bioinform.* 16, 352–364. doi: 10.1109/tcbb.2017.2705686

Collins, M. (2002). "A generalization of principal component analysis to the exponential family," in *Proceedings of the 14th International Conference on Advances in Neural Information Processing Systems*, Cambridge, MA.

Delbert, D., Morris, Q. D., and Frey, B. J. (2005). Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics* 21(Suppl. 1), i144–i151.

Ding, C., and He, X. F. (2004). "K-Means Clustering Via Principal Component Analysis,"in *Proceedings of the 21st International Conference on Machine Learning (ICML)*, New York, NY 1, 29.

Feng, C., Gao, Y. L., Liu, J. X., Zheng, C. H., and Yu, J. (2017). PCA based on graph laplacian regularization and P-norm for gene selection and clustering. *IEEE Trans. Nanobiosci.* 16, 257–265. doi: 10.1109/tnb.2017.2690365

Feng, C. M., Liu, J. X., Gao, Y. L., Wang, J., Wang, D. Q., and Du, Y. (2016). "A graph-laplacian pca based on L1/2-norm constraint for characteristic gene selection," in *Proceedings of the 2016th IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM 2016)*, Shenzhen, 1258–1263.

Feng, C. M., Xu, Y., Liu, J. X., Gao, Y. L., and Zheng, C. H. (2019). Supervised discriminative sparse PCA for corn-characteristic gene selection and tumor classification on multiview biological data. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 2926–2937. doi: 10.1109/tnnls.2019.2893190

Gabay, D., and Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.* 2, 17–40. doi: 10.1016/0898-1221(76)90003-1

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi: 10.1038/nrg.2016.49

Guo, S., Wang, Z., and Ruan, Q. (2013). Enhancing sparsity via lp ($0<p<1$) minimization for robust face recognition. *Neurocomputing* 99, 592–602. doi: 10.1016/j.neucom.2012.05.028

He, X., Yan, S., Hu, Y., Niyogi, P., and Zhang, H. J. (2005). Face recognition using laplacian faces. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 328–340.

Hestenes, M. R. (1969). Multiplier and gradient methods. *J. Optim. Theory Appl.* 4, 303–320. doi: 10.1007/bf00927673

Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., et al. (2010). FABIA: factor analysis for bicluster acquisition. *Bioinformatics* 26, 1520–1527. doi: 10.1093/bioinformatics/btq227

Hou, M. X., Gao, Y. L., Liu, J. X., Dai, L. Y., Kong, X. Z., and Shang, J. H. (2019). Network analysis based on low-rank method for mining information on integrated data of multi-cancers. *Comput. Biol. Chem.* 78, 468–473. doi: 10.1016/j.compbiolchem.2018.11.027

Ishiwata, T., Cho, K., Kawahara, K., Yamamoto, T., Fujiwara, Y., Uchida, E., et al. (2007). Role of lumican in cancer cells and adjacent stromal tissues in human pancreatic cancer. *Oncol. Rep.* 18, 537–543.

Jiang, B., Ding, C., Luo, B., and Tang, J. (2013). "Graph-laplacian PCA: closed-form solution and robustness," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland.

Jolliffe, I. (2002). *Principal Component Analysis*. New York, NY: Springer.

Journee, M., Nesterov, Y., Richtarik, P., and Sepulchre, R. (2010). Generalized Power method for sparse principal component analysis. *J. Mach. Learn. Res.* 11, 517–553.

Keyhanian, S., and Nasersharif, B. (2015). "Laplacian eigenmaps latent variable model modification for pattern recognition," in *Procceedings of the 23rd Iranian Conference on Electrical Engineering (ICEE)*, Tehran.

Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi: 10.1038/nature11412

Kong, X. Z., Liu, J. X., Zheng, C. H., Hou, M. X., and Wang, J. (2017). Robust and efficient biomolecular clustering of tumor based on ${p}$ -norm singular value decomposition. *IEEE Trans. Nanobiosci.* 16, 341–348. doi: 10.1109/tnb.2017.2705983

Lee, M., Shen, H., Huang, J. Z., and Marron, J. S. (2010). Biclustering via sparse singular value decomposition. *Biometrics* 66, 1087–1095. doi: 10.1111/j.1541-0420.2010.01392.x

Lew, Y., Jones, D. V., Mars, W. M., Evans, D., Byrd, D., and Frazier, M. L. (1992). Expression of elongation factor-1 gamma-related sequence in human pancreatic cancer. *Pancreas* 7, 144–152. doi: 10.1097/00006676-199203000-00003

Lin, Z. C., Chen, M. M., and Ma, Y. (2010). The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv* [Preprint] 9. doi: 10.1016/j.jsb.2012.10.010

Liu, J., Liu, J. X., Gao, Y. L., Kong, X. Z., Wang, X. S., and Wang, D. (2015). A P-Norm robust feature extraction method for identifying differentially expressed genes. *PLoS One* 10:e0133124. doi: 10.1371/journal.pone.0133124

Liu, J. X., Wang, D., Gao, Y. L., Zheng, C. H., Shang, J. L., Liu, F., et al. (2017). A joint-L2,1-norm-constraint-based semi-supervised feature extraction for RNA-Seq data analysis. *Neurocomputing* 228, 263–269. doi: 10.1016/j.neucom.2016.09.083

Liu, J. X., Wang, Y. T., Zheng, C. H., Sha, W., Mi, J. X., and Xu, Y. (2013). Robust PCA based method for discovering differentially expressed genes. *BMC Bioinformatics* 14 Suppl. 8:S3.

Liu, J. X., Xu, Y., Gao, Y. L., Zheng, C. H., Wang, D., and Zhu, Q. (2016). A class-information-based sparse component analysis method to identify differentially expressed genes on RNA-seq data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 13, 392–398. doi: 10.1109/tcbb.2015.2440265

Liu, J. X., Xu, Y., Zheng, C. H., Kong, H., and Lai, Z. H. (2015). RPCA-based tumor classification using gene expression data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 12, 964–970. doi: 10.1109/tcbb.2014.2383375

Liu, Q. Z., Gao, X. H., Chang, W. J., Wang, H. T., Wang, H., Cao, G. W., et al. (2015). Secreted protein acidic and rich in cysteine expression in human colorectal cancer predicts postoperative prognosis. *Eur. Rev. Med. Pharmacol. Sci.* 19, 1803–1811.

Matassa, D. S., Amoroso, M. R., Agliarulo, I., Maddalena, F., Sisinni, L., Paladino, S., et al. (2013). Translational control in the stress adaptive response of cancer cells: a novel role for the heat shock protein TRAP1. *Cell Death Dis.* 4:e851. doi: 10.1038/cddis.2013.379

Michifuri, Y., Hirohashi, Y., Torigoe, T., Miyazaki, A., Fujino, J., Tamura, Y., et al. (2013). Small proline-rich protein-1B is overexpressed in human oral squamous cell cancer stem-like cells and is related to their growth through activation of MAP kinase signal. *Biochem. Biophys. Res. Commun.* 439, 96–102. doi: 10.1016/j.bbrc.2013.08.021

Misawa, K., Kanazawa, T., Misawa, Y., Imai, A., and Mineta, H. (2011). Hypermethylation of collagen α2 (I) gene (COL1A2) is an independent predictor of survival in head and neck cancer. *Cancer Biomark.* 10, 135–144. doi: 10.3233/cbm-2012-0242

Nie, F. P., Wang, H., Huang, H., and Ding, C. (2013). Joint Schatten -norm and -norm robust matrix completion for missing value recovery. *Knowl. Inf. Syst.* 42, 525–544. doi: 10.1007/s10115-013-0713-z

Pogue-Geile, K., Geiser, J. R., Shu, M., Miller, C., and Pipas, J. M. (1991). Ribosomal protein genes are overexpressed in colorectal cancer: Isolation of a cDNA clone encoding the human S3 ribosomal protein. *Mol. Cell. Biol.* 11, 3842–3849. doi: 10.1128/mcb.11.8.3842

Qiao, J., Fang, C. Y., Chen, S. X., Wang, X. Q., and Liu, F. (2015). Stroma derived COL6A3 is a potential prognosis marker of colorectal carcinoma revealed by quantitative proteomics. *Oncotarget* 6, 29929–29946. doi: 10.18632/oncotarget.4966

Qin, L., Lin, Z., She, Y., and Chao, Z. (2013). A comparison of typical l p minimization algorithms. *Neurocomputing* 119, 413–424. doi: 10.1016/j.neucom.2013.03.017

Ren, F., Tang, R., Xin, Z., Mihiranganee, M. W., Luo, D., Dang, Y., et al. (2015). Overexpression of MMP family members functions as prognostic biomarker for breast cancer patients: a systematic review and meta-analysis. *Plos One* 10:e0135544. doi: 10.1371/journal.pone.0135544

Roweis, S. T., and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* 290, 2323–2326. doi: 10.1126/science.290.5500.2323

Shen, H. P., and Huang, J. H. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multiv. Anal.* 99, 1015–1034. doi: 10.1016/j.jmva.2007.06.007

Spielman, D. A. (2007). "Spectral graph theory and its applications. foundations of computer science, 2007," in *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science FOCS '07*, Providence, RI.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Wan, D., Gong, Y., Qin, W., Zhang, P., Li, J., Wei, L., et al. (2004). Large-scale cDNA transfection screening for genes related to cancer development and progression. *Proc.Natl.Acad.Sci.U.S.A.* 101, 15724–15729. doi: 10.1073/pnas.0404089101

Wang, J., Liu, J. X., Kong, X. Z., Yuan, S. S., and Dai, L. Y. (2019a). Laplacian regularized low-rank representation for cancer samples clustering. *Comput. Biol. Chem.* 78, 504–509. doi: 10.1016/j.compbiolchem.2018.11.003

Wang, J., Liu, J. X., Zheng, C. H., Wang, Y. X., Kong, X. Z., and Wen, C. G. (2019b). A mixed-norm laplacian regularized low-rank representation method for tumor samples clustering. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 172–182. doi: 10.1109/tcbb.2017.2769647

Wang, Y. X., Gao, Y. L., Liu, J. X., Kong, X. Z., and Li, H. J. (2017). Robust principal component analysis regularized by truncated nuclear norm for identifying differentially expressed genes. *IEEE Trans. Nanobiosci.* 16, 447–454. doi: 10.1109/tnb.2017.2723439

West, M. (2003). *Bayesian Factor Regression Models in the "Large p, Small n" Paradigm*, Vol. 7. Oxford: Oxford University Press, 723–732.

Xiang, S. M., Nie, F. P., Meng, G. F., Pan, C. H., and Zhang, C. S. (2012). Discriminative least squares regression for multiclass classification and feature selection. *IEEE Trans. Neural Netw. Learn. Syst.* 23, 1738–1754. doi: 10.1109/tnnls.2012.2212721

Yang, S. Z., Hou, C. P., Nie, F. P., and Wu, Y. (2012). Unsupervised maximum margin feature selection via L_(2,1)-norm minimization. *Neural Comput. Appl.* 21, 1791–1799. doi: 10.1007/s00521-012-0827-3

Yu, C., Pan, K., Xing, D., Liang, G., and Lin, D. (2002). Correlation between a single nucleotide polymorphism in the matrix metalloproteinase-2 promoter and risk of lung cancer. *Cancer Res.* 62, 6430–6433.

Zhang, Z., Xu, Y., Yang, J., Li, X., and Zhang, D. (2015). A survey of sparse representation: algorithms and applications. *IEEE Access* 3, 490–530. doi: 10.1109/access.2015.2430359

Zhang, Z., and Zha, H. (2002). Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *J. Shang. Univ.* 8, 406–424. doi: 10.1007/s11741-004-0051-1

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Stat.* 15, 265–286.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.