Feng et al. BMC Bioinformatics 2019, 20(Suppl 22):716
https://doi.org/10.1186/s12859-019-3229-z

BMC Bioinformatics

RESEARCH Open Access

PCA via joint graph Laplacian and sparse constraint: Identification of differentially expressed genes and sample clustering on gene expression data

Chun-Mei Feng[1,2], Yong Xu[1,3*], Mi-Xiao Hou[1], Ling-Yun Dai[2*] and Jun-Liang Shang[2*]

From International Conference on Data Science, Medicine and Bioinformatics
Nanning, China. 22-24 June 2019

Abstract

Background: In recent years, identification of differentially expressed genes and sample clustering have become hot topics in bioinformatics. Principal Component Analysis (PCA) is a widely used method in gene expression data. However, it has two limitations: first, the geometric structure hidden in data, e.g., pair-wise distance between data points, have not been explored. This information can facilitate sample clustering; second, the Principal Components (PCs) determined by PCA are dense, leading to hard interpretation. However, only a few of genes are related to the cancer. It is of great significance for the early diagnosis and treatment of cancer to identify a handful of the differentially expressed genes and find new cancer biomarkers.

Results: In this study, a new method gLSPCA is proposed to integrate both graph Laplacian and sparse constraint into PCA. gLSPCA on the one hand improves the clustering accuracy by exploring the internal geometric structure of the data, on the other hand identifies differentially expressed genes by imposing a sparsity constraint on the PCs.

Conclusions: Experiments of gLSPCA and its comparison with existing methods, including Z-SPCA, GPower, PathSPCA, SPCArt, gLPCA, are performed on real datasets of both pancreatic cancer (PAAD) and head & neck squamous carcinoma (HNSC). The results demonstrate that gLSPCA is effective in identifying differentially expressed genes and sample clustering. In addition, the applications of gLSPCA on these datasets provide several new clues for the exploration of causative factors of PAAD and HNSC.

Keywords: Differentially expressed genes, Gene expression data, Graph Laplacian, Principal component analysis, Sparse constraint

* Correspondence: yongxu@ymail.com; dailingyun_1@163.com; shangjunliang110@163.com
[1]Bio-Computing Research Center, Harbin Institute of Technology, Shenzhen 518055, Guangdong, People's Republic of China
[2]School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, People's Republic of China
Full list of author information is available at the end of the article

Background

In the field of bioinformatics, research on the difference of expressed genes between cells of different status helps us understand the functions of genes, and what is more, isolate disease related genes. This direction of research is known as identification of differentially expressed genes [1]. It lays a foundation for further research on the relationship between cancer and genes in the molecular level and will improve the efficiency of cancer diagnosis. Sample clustering of gene expression data is another application in bioinformatics [2, 3]. It will facilitate the searching of new cancer subtype, and consequently helps the targeted therapy of tumor.

Matrix decomposition is one of the major techniques to deal with gene expression data [4–8]. At present, many researchers are interested in data modeling, and try to select the differentially expressed genes from a large number of gene expression data [9–12]. It lays a foundation for further research on the relationship between cancer and genes in the molecular level and improves the efficiency of cancer diagnosis. Among them, Principal Component Analysis (PCA) is a basic tool that has been widely used [6, 7]. The traditional linear PCA method considers the global Euclidean structure of the original data, and when the data points are in a manifold structure, the global Euclidean structure cannot exactly describe the real distance between data points.

In recent years, manifold learning has made a lot of progress in theory, algorithm and application [13–16]. The main idea of manifold learning is to establish a nonlinear mathematical model by means of differential calculus and other mathematical tools. The inherent nonlinear geometric structure hidden in the high dimensional data can be revealed by manifold learning. Thus, we can consider introducing the manifold learning in the linear PCA method. Motivated by manifold learning theory, Jiang et al. proposed graph Laplacian PCA (gLPCA) [17]. This method joins a graph Laplacian to the data representation of original data $\mathbf{X}$. The derived low dimensional data can be learned with the cluster information encoded in graph structure $\mathbf{W}$.

Despite its advantage, the PCA joint with graph Laplacian suffers from the fact that the PCs are typically dense [18, 19]. In bioinformatics, the gene expression data involved in PCs have much irrelevant or redundant information. For the study of the pathogenesis of a disease, only a small number of genes are significant. This information plays an important role in early diagnosis of cancer. Hence, the interpretation of the PCs will be facilitated if the derived PCs are sparse, involving a few of nonzero elements. Actually, many sparse PCA methods have been developed. For example, rotation and thresholding are first derived on running and facial spots data to find the sparse PCs [20, 21]. Z-SPCA is designed based on iterative elastic net regression [22]. Good results on biological and regular multivariate data have been achieved by this method. D'Aspremont et al. designed two methods, called DSPCA [23] and PathSPCA [24]. DSPCA finds sparse PCs via semi-definite program (SDP) while PathSPCA directly identifies the non-zero elements one by one. Shen and Huang designed a method called sPCA-rSVD which solves the problem based on low-rank matrix factorization [25]. Sigg and Buhmann considered expectation-maximization (EM) to solve a sparse probabilistic generative model, we call this method as EMSPCA [26]. Journée et al. designed a series of algorithms based on L0 and L1-norm to extract single unit or block unit PCs (GPowerL0, GPowerL1, GPowerL0,m, GPowerL1,m,) on random data and gene expression data to compute the sparse PCs [27]. Lai et al. rewrote the traditional PCA into multilinear regression and sparse regression forms (MSPCA) to deal with tensor data [28]. Motivated by rotation and truncation of PCA basis, Hu et al. proposed an efficient method called SPCArt [29]. Zhao et al. divided the sparse PCA problem into several sub-problems and gave a series of closed-form solutions to compute it. This method is named as block coordinate descent sparse PCA (BCD-SPCA) [30].

Recently, Nie et al. demonstrated that L2,1-norm applying on a matrix can induce sparsity in row [18, 31]. L2,1-norm is defined as $\|\mathbf{X}\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} \mathbf{x}_{ij}^2} = \sum_{i=1}^{n} \|\mathbf{x}^i\|_2$, where $\mathbf{x}^i$ is the i-th row of $\mathbf{X}$. Actually L2,1-norm first computes L2-norm of the row vector $\mathbf{x}^i$ and then calculates L1-norm of the resulting L2-norms $b(\mathbf{X}) = (\|\mathbf{x}^1\|_2, \|\mathbf{x}^2\|_2, ...\|\mathbf{x}^m\|_2)$. Zero rows in $\mathbf{X}$ can be achieved through the effect of L2,1-norm. Thus, considering manifold learning has little effect on identification of differentially expressed genes, the introduction of L2,1-norm to PCA is feasible and effective. Furthermore, as we will show, when the problem is solved iteratively, L2,1-norm can be formulated into a trace form, consequently we can optimize it compatibly with the graph Laplacian.

In this paper, we consider introducing sparsity constraint and graph Laplacian to PCA. A novel method called PCA via joint graph Laplacian and sparse constraint (gLSPCA) is proposed. It not only encodes with the internal geometric structure for clustering purpose, but also imposes sparse constraint on traditional PCA to improve interpretability. As a result, on one hand our method can be applied for sample clustering; on the other hand, it can identify a few of differentially expressed genes. The contributions of this paper can be enumerated as follows:

(i) We proposed a novel method called gLSPCA which simultaneously learns the internal geometric structure and improves the interpretability of PCs. gLSPCA on the one hand can identify differentially expressed genes, on the other hand can be applied for sample clustering.
(ii) The optimization and convergence analysis of gLSPCA are provided.

(iii) The proposed gLSPCA is effective in identifying differentially expressed genes and sample clusters, as demonstrated by experimental results on PAAD and HNSC datasets. gLSPCA provides a tool that is helpful for the study of the pathogenesis of cancer, and the clustering application on samples provides a basis for early diagnosis of cancer.

In what follows, the proposed method and the algorithm for this method are introduced in Methodology section. The properties and convergence analysis of this method are included. Extensive experiments for differentially expressed genes identification and sample clustering are conducted in the section of Results and Discussion, where related sparse PCA methods are compared with our method. The paper is concluded in the section of Conclusions.

Methodology

Mathematical definition

Above all, we define some notations which will be frequently used in following sections. (1) The input data matrix is denoted by $\mathbf{X} = (x_1, ..., x_n) \in \mathbb{R}^{m \times n}$, where n is the number of samples and m is the number of variables, i.e., genes in the gene expression data. (2) The new subspace of projected data points is denoted by $\mathbf{H} = \mathbb{R}^{n \times k}$ and the principal direction is denoted by $\mathbf{U} = (\mathbf{u}_1, ..., \mathbf{u}_k) \in \mathbb{R}^{m \times k}$. (3) The Frobenius norm is denoted as $\|\mathbf{X}\|_F$. (4) The $L_{2,1}$-norm is denoted as $\|\mathbf{X}\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} \mathbf{x}_{ij}^2} = \sum_{i=1}^{n} \|\mathbf{x}^i\|_2$. (4) The trace of matrix $\mathbf{Z}$ is denoted as $\mathrm{Tr}(\mathbf{Z})$.

The classical PCA and Graph-Laplacian PCA

In this subsection, we briefly review the classical PCA and gLPCA. PCA finds the new subspace of projected data points $\mathbf{H}$ and principal direction $\mathbf{U}$ by solving the following optimization problem [7]:

$$\min_{\mathbf{U},\mathbf{H}} \left\|\mathbf{X}-\mathbf{U}\mathbf{H}^T\right\|_F^2 \quad s.t.\ \mathbf{H}^T\mathbf{H} = \mathbf{I}. \tag{1}$$

In gene expression data, each column x_i is a linearized vector of sample. The basic PCA model cannot recover non-linear structure of data. gLPCA incorporates the geometric manifold information to find the non-linear structure of data [7]. Considering $\mathbf{H}$ is the embedding matrix, the gLPCA is formulated as follows:

$$\min_{\mathbf{U},\mathbf{H}} \left\|\mathbf{X}-\mathbf{U}\mathbf{H}^T\right\|_F^2 + \alpha \mathrm{Tr}\left(\mathbf{H}^T\mathbf{L}\mathbf{H}\right) \quad s.t.\ \mathbf{H}^T\mathbf{H} = \mathbf{I}, \tag{2}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the graph Laplacian matrix. $\mathbf{D} = diag(d_1, ..., d_n)$ is a diagonal matrix whose elements are column or row sums of $\mathbf{W}$ ($\mathbf{W}$ is a symmetric nonnegative weight matrix). It can be expressed as $d_i = \sum_j \mathbf{W}_{ij}$. The definition of $\mathbf{W}_{ij}$ is listed as follows:

$$\mathbf{W}_{ij} = \begin{cases} 1 & if\ \mathbf{x}_i \in \mathbf{N}_k(\mathbf{x}_j)\ \ or\ \ \mathbf{x}_j \in \mathbf{N}_k(\mathbf{x}_i), \\ 0 & otherwise, \end{cases} \tag{3}$$

where $\mathbf{N}_k(\mathbf{x}_i)$ is the k nearest neighbours of $\mathbf{x}_i$ [24]. The authors also presented a robust version to improve the robustness of this method. Since our paper focuses on the sparsity of the gLPCA method, we will not elaborate this robust version further.

The proposed method: PCA via joint graph Laplacian and sparse regularization (gLSPCA)

Recently, sparse representation has been widely applied in the field of bioinformatics. It decomposes a set of high-dimensional data into a series of linear combinations of low dimensional codes, and hopes the combination coefficients to be zero as much as possible. The PCA suffers from the fact that the PCs are typically dense. The interpretation of the PCs might be facilitated if the idea of sparse constraint has been utilized. We consider introducing $L_{2,1}$-norm constraint on the PCs $\mathbf{H}$ to improve the interpretability of PCA based method. Since the $L_{2,1}$-norm can induce sparsity in rows, the PCs can be sparser and more easily explained [25]. Then, the quality of the decomposition is improved. The proposed method (gLSPCA) solves the following minimization problem:

$$\min_{\mathbf{U},\mathbf{H}} \left\|\mathbf{X}-\mathbf{U}\mathbf{H}^T\right\|_F^2 + \alpha \mathrm{Tr}\left(\mathbf{H}^T\mathbf{L}\mathbf{H}\right) + \gamma \|\mathbf{H}\|_{2,1} \quad s.t.\ \mathbf{H}^T\mathbf{H} = \mathbf{I}, \tag{4}$$

where α and γ are scalar parameters to balance the weights of graph Laplacian and sparse constraint respectively.

Optimization

It is hard to obtain a closed solution from Eq. (4). Thus, we solve the problem via iterative optimization. The solution of $\mathbf{U}$ is obtained by calculating partial derivatives at first. Then, the solution of $\mathbf{H}$ can be obtained by performing eigen-decomposition, after these two variables $\mathbf{U}$ and $\mathbf{H}$ are integrated into one variable $\mathbf{H}$ to substitute the original objective function. Obtaining convergence after a number of iterations, we finally get the PCs with internal geometry and sparsity which were ignored in previous studies. Firstly, following an optimization technique of L2,1-norm [25, 26], the optimization of original problem can be approximated by the following problem:

$$\min_{\mathbf{U},\mathbf{H}} \left\|\mathbf{X}-\mathbf{U}\mathbf{H}^T\right\|_F^2 + \alpha \mathrm{Tr}\left(\mathbf{H}^T\mathbf{L}\mathbf{H}\right) + \gamma \mathrm{Tr}\left(\mathbf{H}^T\mathbf{D}\mathbf{H}\right) \quad s.t.\ \mathbf{H}^T\mathbf{H} = \mathbf{I}, \tag{5}$$

where $\mathbf{D}$ is a diagonal matrix with elements:

$$\mathbf{D}_{ii} = \frac{1}{2\|\mathbf{h}_i\|_2}. \tag{6}$$

Then, to get the solution of $\mathbf{U}$, we fix $\mathbf{H}$ and the derivative of $\mathscr{L}(\mathbf{U},\mathbf{H},\mathbf{D})$ respect to $\mathbf{U}$ is

$$\frac{\partial \mathscr{L}(\mathbf{U},\mathbf{H},\mathbf{D})}{\partial \mathbf{U}} = -2\mathbf{X}\mathbf{H} + 2\mathbf{U}, \tag{7}$$

By setting the derivative of $\mathbf{U}$ to zero, we have

$$\mathbf{U} = \mathbf{X}\mathbf{H}. \tag{8}$$

Substituting the solutions of $\mathbf{U}$ into Eq. (5), we have

$$\begin{aligned} &\mathrm{Tr}\left(\mathbf{X}-\mathbf{X}\mathbf{H}\mathbf{H}^T\right)\left(\mathbf{X}-\mathbf{X}\mathbf{H}\mathbf{H}^T\right)^T + \alpha \mathrm{Tr}\left(\mathbf{H}^T\mathbf{L}\mathbf{H}\right) + \gamma \mathrm{Tr}\left(\mathbf{H}^T\mathbf{D}\mathbf{H}\right) \\ &= -\mathrm{Tr}\left(\mathbf{H}^T\mathbf{X}^T\mathbf{X}\mathbf{H}\right) + \|\mathbf{X}\|_F^2 + \alpha \mathrm{Tr}\left(\mathbf{H}^T\mathbf{L}\mathbf{H}\right) + \gamma \mathrm{Tr}\left(\mathbf{H}^T\mathbf{D}\mathbf{H}\right) \\ &= \mathrm{Tr}\left(\mathbf{H}^T\left(-\mathbf{X}^T\mathbf{X} + \alpha\mathbf{L} + \gamma\mathbf{D}\right)\mathbf{H}\right) + \|\mathbf{X}\|_F^2. \end{aligned} \tag{9}$$

Therefore, Eq. (8) is equivalent to the following problem:

$$\ell(\mathbf{H}) = \min_{\mathbf{H}^T\mathbf{H}=\mathbf{I}} \mathrm{Tr}\left(\mathbf{H}^T\mathbf{A}\mathbf{H}\right), \tag{10}$$

where $\mathbf{A} = -\mathbf{X}^T\mathbf{X} + \alpha\mathbf{L} + \gamma\mathbf{D}$. Thus, the optimal $\mathbf{H}$ is the eigenvectors corresponding to the first k smallest eigenvalues of the matrix $\mathbf{A}$.

In the following, for convenience of parameter setting, we transform $\mathbf{A}$ to another equivalent form. We use η_k to denote the largest eigenvalue of matrix $\mathbf{X}^T\mathbf{X} - \gamma\mathbf{D}$. For Laplacian matrix $\mathbf{L}$, we use η_s to represent the largest eigenvalue of $\mathbf{L}$. We then set

$$\alpha = \frac{\beta}{1-\beta}\frac{\eta_k}{\eta_s}, \tag{11}$$

so that the tuning of α becomes the tuning of β. Thus, (4) can be rewritten as follows:

$$\min_{\mathbf{H}} \mathrm{Tr}\ \mathbf{H}^T\left[(1-\beta)\left(\mathbf{I} - \frac{\mathbf{X}^T\mathbf{X} + \gamma\mathbf{D}}{\eta_k}\right) + \beta\frac{\mathbf{L}}{\eta_s}\right]\mathbf{H} \quad s.t.\ \mathbf{H}^T\mathbf{H} = \mathbf{I}. \tag{12}$$

In this way, the solution of $\mathbf{H}$ can be obtained by computing the first k smallest eigenvalues of matrix $\mathbf{A}_1$:

$$\mathbf{A}_1 = (1-\beta)\left(\mathbf{I} - \frac{\mathbf{X}^T\mathbf{X} + \gamma\mathbf{D}}{\eta_k}\right) + \beta\frac{\mathbf{L}}{\eta_s}. \tag{12}$$

The range of β is $0 \le \beta \le 1$. In particular, when $\beta = 0$ and $\gamma = 0$, gLSPCA degrades to classical PCA. When $\beta = 1$ and $\gamma = 0$, it equals to Laplacian Embedding (LE). We summarize the algorithm of the proposed gLSPCA approach in Algorithm 1.

Input: Data matrix $\mathbf{X} = (x_1, ..., x_n) \in R^{m\times n}$, parameters γ and β.

Output: Matrix $\mathbf{U}$ and $\mathbf{H}$.

1: Initialize $\mathbf{D} = \mathbf{I}_{n\times n}$;

2: **repeat**

Construct weight matrix $\mathbf{W}$;
Compute the diagonal matrix $\mathbf{D}$, graph Laplacian $\mathbf{L}$;
Compute $\mathbf{H}$ by the eigenvectors corresponding to the first k smallest eigenvalues of matrix $\mathbf{A}_1$;
Compute the optimal $\mathbf{U}$ according to Eq. (8);
Compute diagonal matrix $\mathbf{D}$ according to Eq. (6);

Until converges

Convergence analysis

We would like to show the objective value does not increase in each iteration of the proposed gLSPCA algorithm. Firstly, a simple lemma is provided [32].

Lemma 1. For any non-zero vectors $\mathbf{a},\mathbf{b} \in \mathbb{R}^m$:

$$\|\mathbf{a}\|_2 - \frac{\|\mathbf{a}\|_2^2}{2\|\mathbf{b}\|_2} \le \|\mathbf{b}\|_2 - \frac{\|\mathbf{b}\|_2^2}{2\|\mathbf{b}\|_2}. \tag{14}$$

The convergence analysis of gLSPCA is summarized as Theorem 1.

Theorem 1: The optimization procedure of the proposed gLSPCA algorithm will monotonically decrease the objective function in each iteration.

Proof. Following the algorithm of gLSPCA, when we fix $\mathbf{D}^t$ in the t-th iteration and optimize $\mathbf{U}^{t+1}$, $\mathbf{H}^{t+1}$, we have:

$$\begin{aligned} &\left\|\mathbf{X}-\mathbf{U}^{t+1}\left(\mathbf{H}^{t+1}\right)^T\right\|_F^2 + \alpha \mathrm{Tr}\left(\left(\mathbf{H}^{t+1}\right)^T\mathbf{L}\mathbf{H}^{t+1}\right) + \gamma \mathrm{Tr}\left(\mathbf{H}^{t+1}\mathbf{D}^t\mathbf{H}^{t+1}\right) \\ &\le \left\|\mathbf{X}-\mathbf{U}^{t}\left(\mathbf{H}^{t}\right)^T\right\|_F^2 + \alpha \mathrm{Tr}\left(\left(\mathbf{H}^{t}\right)^T\mathbf{L}\mathbf{H}^{t}\right) + \gamma \mathrm{Tr}\left(\mathbf{H}^{t}\mathbf{D}^t\mathbf{H}^{t}\right). \end{aligned} \tag{15}$$

Since $\|\mathbf{H}\|_{2,1} = \sum_{i=1}^n \|\mathbf{h}_i\|_2$, this inequality indicates

$$\begin{aligned} &\left\|\mathbf{X}-\mathbf{U}^{t+1}\left(\mathbf{H}^{t+1}\right)^T\right\|_F^2 + \alpha \mathrm{Tr}\left(\left(\mathbf{H}^{t+1}\right)^T\mathbf{L}\mathbf{H}^{t+1}\right) + \gamma \sum_{i=1}^n \left(\frac{\|\mathbf{h}_i^{t+1}\|_2^2}{2\|\mathbf{h}_i^t\|_2} - \|\mathbf{h}_i^{t+1}\|_2\right) \\ &\le \left\|\mathbf{X}-\mathbf{U}^{t}\left(\mathbf{H}^{t}\right)^T\right\|_F^2 + \alpha \mathrm{Tr}\left(\left(\mathbf{H}^{t}\right)^T\mathbf{L}\mathbf{H}^{t}\right) + \gamma \sum_{i=1}^n \left(\frac{\|\mathbf{h}_i^{t}\|_2^2}{2\|\mathbf{h}_i^t\|_2} - \|\mathbf{h}_i^{t+1}\|_2\right). \end{aligned} \tag{16}$$

According to Lemma 1, we know that

$$\frac{\|\mathbf{h}_i^{t+1}\|_2^2}{2\|\mathbf{h}_i^t\|_2} - \|\mathbf{h}_i^{t+1}\|_2 \ge \frac{\|\mathbf{h}_i^{t}\|_2^2}{2\|\mathbf{h}_i^t\|_2} - \|\mathbf{h}_i^{t}\|_2. \tag{17}$$

Thus, we have the following result.

$$\left\|\mathbf{X}-\mathbf{U}^{t+1}\left(\mathbf{H}^{t+1}\right)^{T}\right\|_{F}^{2}+\alpha\mathrm{Tr}\left(\left(\mathbf{H}^{t+1}\right)^{T}\mathbf{L}\mathbf{H}^{t+1}\right)+\gamma\left\|\mathbf{H}^{t+1}\right\|_{2,1}$$
$$\leq\left\|\mathbf{X}-\mathbf{U}^{t}\left(\mathbf{H}^{t}\right)^{T}\right\|_{F}^{2}+\alpha\mathrm{Tr}\left(\left(\mathbf{H}^{t}\right)^{T}\mathbf{L}\mathbf{H}^{t}\right)+\gamma\left\|\mathbf{H}^{t}\right\|_{2,1}. \tag{18}$$

This inequality proves that the objective function of (4) will monotonically decrease in each iteration.

Results and discussion

The primary goal of our method is to improve the sparsity of gLPCA because the PCs of this method are dense. We evaluate the performance of the proposed method with the other five related methods, including four sparse PCA methods: Z-SPCA [22], GPower [27], PathSPCA [24], SPCArt [29], and a graph Laplacian PCA method: gLPCA [17]. There are two deflation algorithms and two block algorithms for GPower. In practice, the results of the four algorithms are not different significantly. We choose one of these algorithms as comparison method in our experiments. The experiments are mainly divided into two aspects:

(i) Identifying differentially expressed genes. The joint effect of sparse constraint and graph Laplacian in our method can be evaluated by the identification of differentially expressed genes. Firstly, the new oncogenes can be found in these discovered differentially expressed genes. Then, the function and interacting proteins network analysis of these new oncogenes are given. Finally, pathway analysis explains the combined biological processes of the identified differentially expressed genes.
(ii) Tumour sample clustering. Since the sparse PCs are encoded with the internal geometric structure for clustering purpose, tumour sample clustering can be used to evaluate how well it works. Clustering the data according to the similarity of each data point provides a basis for accurate subtype of cancer.

Experimental settings

We set $r = 2$ to be the number of reduced dimensions. The similarity matrix is constructed by k-nearest neighbour graph with Gaussian kernel, where we set $k = 5$ and the σ of Gaussian kernel to be 1. We set λ to infinity as the parameter value of Z-SPCA method, thus soft thresholding can be conducted to compute the sparse PCs for the gene expression data with high dimension and small sample. For GPower and PathSPCA method, we use the default parameter values suggested by the authors. For SPCArt method, we set $\lambda^{*} = 1/\sqrt{m}$ to guarantee the sparsity and avoid truncating to zero vectors. For our method, the best parameters are selected in the range of $\beta = (0.1, ..., 0.9)$ and $\gamma = (10^{-30}, ..., 10^{30})$. We report the best results with the optimal parameters for all compared methods.

Datasets

The details of the two datasets used in our experiments are described in Table 1. The dataset of pancreatic cancer (PAAD) and head and neck squamous carcinoma (HNSC) are downloaded from The Cancer Genome Atlas (TCGA). This database is an open comprehensive multi-dimensional map of the key genomic changes in 33 types of cancer dataset. These two datasets have thousands of genes but only a small number of samples. Much irrelevant or redundant information is contained in such gene expression data. The following experiments on identification of differentially expressed genes and tumour sample clustering are particularly important in the cancer study.

Identifying differentially expressed genes

In bioinformatics, the PCs involve a large number of genes. In cancer study, only a small number of genes are significant for early diagnosis of cancer and accurate subtype of cancer. These genes can be defined as differentially expressed genes. We can analyze the identified differentially expressed genes to evaluate the effectiveness of the sparsity constraint in our method.

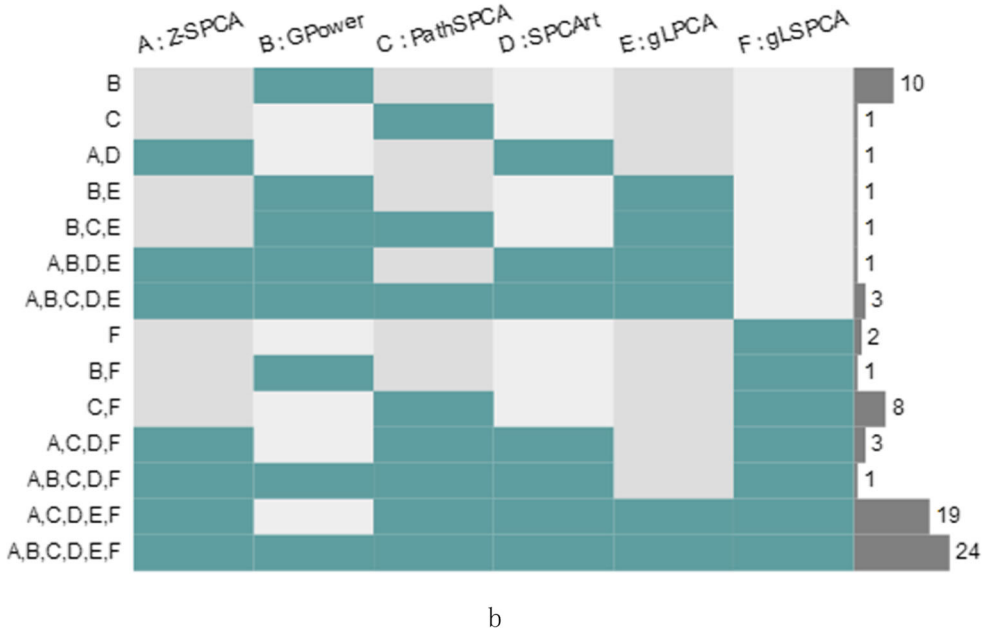
Firstly, we compute the scores for all genes in descending order. Then, the index set of differentially expressed genes is formed by the corresponding indices. To be fair, all methods extract the largest 100 values. The extracted genes with high scores in data representation can be deemed as differentially expressed genes. We match the selected differentially expressed genes to the pathogenic genes of PAAD and HNSC published on GeneCards. The public available website of GeneCards is http://www.genecards.org/, which is an open, integrative database that provides comprehensive, useful information on all predicted and annotated human genes [33].

Matching results of each method on PAAD and HNSC datasets are listed in an additional file (see Additional file 1). Additional file 1 shows the differentially expressed genes identified by all compared methods, as well as the relative scores of each gene associated with the disease. The unique genes of each method are also marked in bold in this file. These unique genes are the differentially expressed genes

Table 1 Summary of the two datasets

Data sets	Number of		class distribution	
	Samples	Genes	Normal	disease
PAAD	180	20,502	4	176
HNSC	418	20,502	20	398

a



b

Fig. 1 Overlap among the differentially expressed genes identified by the compared methods

that one method can identify while the other methods cannot.

To visualize the overlap among the differentially expressed genes identified by the methods, we send the results to OmicsBean to generate a Venn diagram. OmicsBean is a multi-group data analysis system, and its public address is http://www.omicsbean.com:88/. The overlap result of the differentially expressed genes identified by the methods is visualized by Venn diagram in Fig. 1. In this figure, (a) is the overlap result on PAAD dataset and (b) is the overlap result on HNSC dataset. The left coordinate represents the different permutations and combinations of various methods. The right coordinate represents the number of unique genes that are excavated by one method (only one method on the left coordinate) or the same genes by several methods (multiple methods on the left coordinate).

From Fig. 1, it can be concluded that the number of unique gene mined by GPower on the two datasets is the most. But it also loses a large number of important genes, which can be seen from the results of the Additional file 1. These missing genes lead to a poor overall effect in identifying differentially expressed genes. There are also a few unique genes mined by PathSPCA. But the scores of these genes are not high enough, they are not important pathogenic genes. Thus, there is no further research of these genes. gLSPCA finds two unique genes in each dataset, and these genes are highly related to disease. These genes can be defined as the new oncogenes identified by our method. It is necessary to carry out further studies on these genes and their detailed analysis is discussed in the next subsection.

Here, we first detect the efficiency of the identified differentially expressed genes. The identification accuracy (IA) and the total relevance scores (TRS) of these genes are listed in Table 2. The best results are in bold. The IA is defined as follows:

$$IA = \frac{\Sigma_{i=1}^{n} \psi(g_s, g_n)}{n} \times 100\%, \quad (19)$$

Table 2 Results on identification accuracy (IA) and total relevance score (TRS) of six methods on PAAD and HNSC dataset

Methods	PAAD		HNSC	
	IA	TRS	IA	TRS
Z-SPCA	77.00	901.67	53.00	540.91
GPower	77.00	922.27	43.00	378.70
PathSPCA	61.00	682.56	**60.00**	579.06
SPCArt	77.00	901.67	53.00	540.91
gLPCA	75.00	878.39	50.00	513.94
gLSPCA	**80.00**	**927.70**	**60.00**	**591.31**

Table 3 The function of differentially expressed genes on PAAD dataset identified by gLSPCA but not the other methods

Gene name	Function	Relevance score
PPY	This gene encodes a member of the neuropeptide Y (NPY) family of peptides.	21.13
CD24	This gene encodes a sialoglycoprotein that is expressed on mature granulocytes and B cells and modulates growth and differentiation signals to these cells.	8.27

where $\psi(x, y) = 1$ if $x = y$ and $\psi(x, y) = 0$ otherwise. g_s is the selected genes from our method and g_n represents the pathogenic genes of disease from GeneCards. Larger IA indicates the identification performance of differentially expressed genes is better. TRS denotes the total relevance scores of all identified differentially expressed genes, which is computed by GeneCards.

From this table, it can be concluded that gLSPCA has higher IA and TRS results than the other methods over the two datasets. The results of Z-SPCA, SPCArt and gLPCA methods are relatively stable, while those of GPower and PathSPCA methods are unstable. The IA result of PathSPCA on HNSC equals to that of gLSPCA, but the result on PAAD dataset is the worst. Since the differentially expressed genes identified by our method have higher correlation with disease, the TRS of PathSPCA is lower than gLSPCA. For GPower method, the IA and TRS performances have much difference on the two datasets. It shows that the adaptability of GPower to different datasets is not satisfactory. From the above discussion, we can conclude that the proposed method gLSPCA performs better than the other methods on identifying differentially expressed genes.

Function and interacting proteins network analysis

To detect the correlation between identified oncogenes with disease, we summarize the functions of these genes in Tables 3 and 4. Table 3 lists the function of differentially expressed genes on PAAD dataset identified by gLSPCA but not the other methods. The relevance score of PPY on PAAD indicates that it is an important virulence gene of PAAD. Published article has proved that PPY responses to a mixed meal in PAAD [34]. Thus, the medical study of PAAD is based on the biological changes

Table 4 The function of differentially expressed genes on HNSC dataset identified by gLSPCA but not the other methods

Gene name	Function	Relevance score
HSPA1A	This intronless gene encodes a 70 kDa heat shock protein which is a member of the heat shock protein 70 family.	7.67
COL6A1	The collagens are a superfamily of proteins that play a role in maintaining the integrity of various tissues.	4.05

of PPY. CD24 as a potential oncogene interferes the RNA treatment of PAAD cancer cells has been studied [35]. And its relevance score with PAAD is 8.27. Table 4 lists the functions and relevance scores of the differentially expressed genes on HNSC dataset identified by gLSPCA but not the other methods. The high relevance scores reflect the close relationship between HSPA1A, COL6A1 and HNSC. But there are few biological researches on this issue, which provides a great space to study it.

Since these genes belong to protein coding genes, it is useful to send them to GeneCards for finding the interacting proteins network. We find three interacting proteins networks. The results can be found in Fig. 2, where (a) is the interacting proteins network of PPY, (b) is HSPA1A and (c) is COL6A1. Each graph shows the most significant five interacting genes. In this figure, each network node represents proteins result by a single and protein-coding gene locus. The edges in this figure are the protein-protein associations. These associative proteins jointly promote a shared function, which does not necessarily mean they are binding each other in physics. The specific known interactions are explained in this figure. The graph of protein networks are helpful to carry out a deeper biological study of these differentially expressed genes.

Pathway analysis

The combined biological processes of the identified differentially expressed genes can be explained by pathways. Pathways help us understand the advanced functions of organisms and biological systems at the molecular level. We send these genes to KEGG: http://www.kegg.jp/. The pathways of highest overlap on these two datasets are presented in additional file [36]. The pathway of focal adhesion plays an essential role in biological processes. This pathway is discovered from

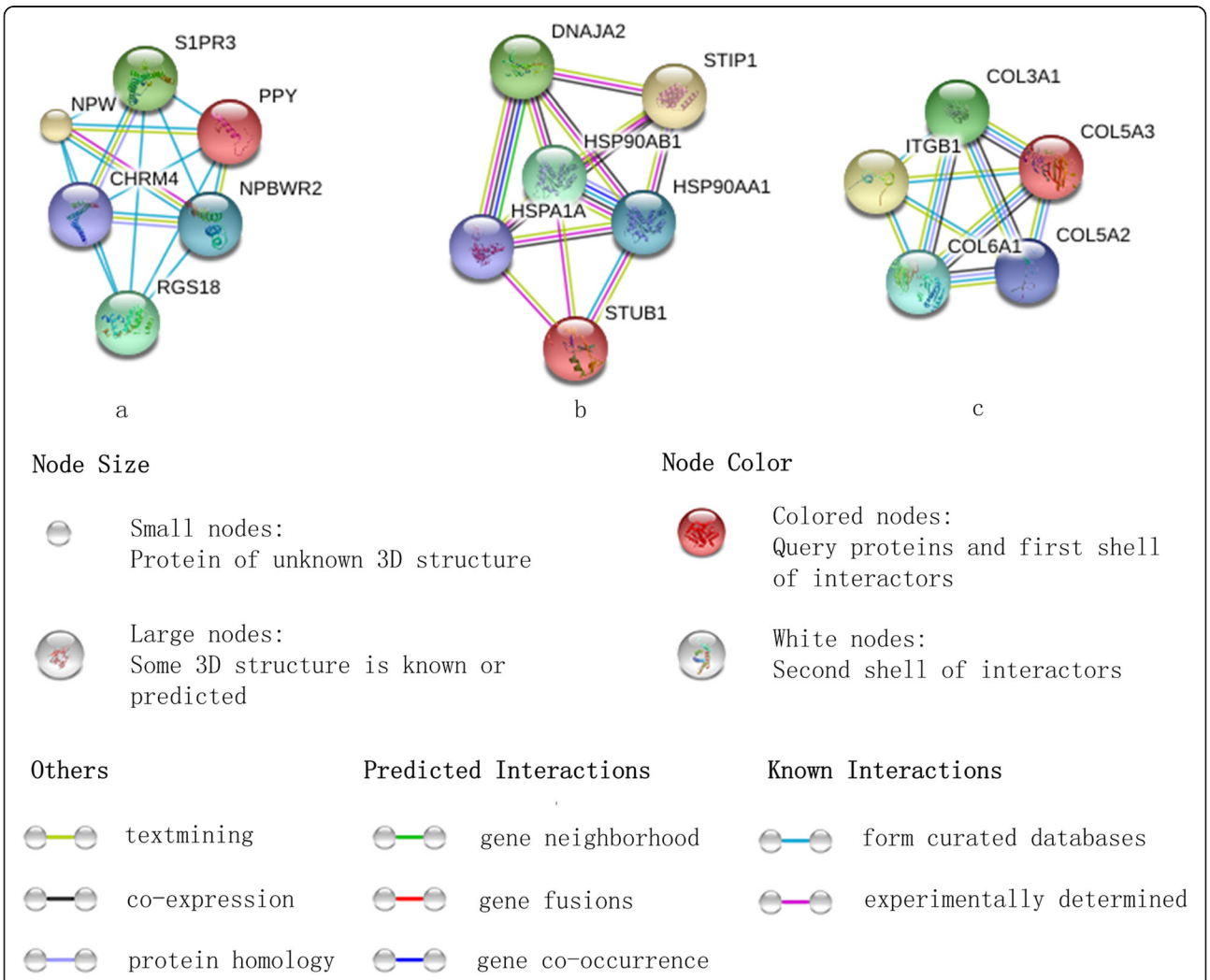


Fig. 2 The interacting proteins network of the identified differentially expressed genesGraphical presentation of the interacting proteins network of the differentially expressed genes identified by gLSPCA but not the other compared methods.

Table 5 ACC performance of all methods

Datasets	All-Ge	Z-SPCA	GPower	PathSPCA	SPCArt	gLPCA	gLSPCA
PAAD	83.09	95.00	95.00	95.00	96.35	95.00	**97.22**
HNSC	78.23	75.84	77.51	72.73	75.84	79.43	**92.88**

Notes: "All-Ge" denotes all features cluster without any dimension reduction processing

the identified differentially expressed genes of gLSPCA on PAAD dataset. The published article concludes that the modification of focal adhesion and integration might be a novel therapeutic approach for the treatment of pancreatic cancer [37]. The pathway of ECM-receptor interaction is identified by the result of gLSPCA on HNSC dataset. The extracellular matrix (ECM) is made up of various structural and functional macromolecules, and it also plays a vital role in tissue and organ morphogenesis, as well as the holding of cell and tissue structure. Shim et al. hypothesize the over expression of cortical protein leads to the degradation of ECM in HNSC [38]. Researches show the biological system in these pathways is an important part for the study of PAAD and HNSC.

Tumour sample clustering results

Sample clustering based on gene expression data is helpful for the detection of tumour subtypes. Since the graph Laplacian is introduced to the proposed method, the geometric structure of data is explored. However, the encoded internal geometric structure is for clustering purpose. It is useful to evaluate whether the explored geometric structure benefits sample clustering. And the discovery new subtype of tumour is helpful for the targeted therapy of tumour. In this experiment, filtering out redundant information by the proposed method, the corresponding results are obtained by K-means clustering. Following the related clustering work, we adopt clustering accuracy (ACC) as evaluation criteria in our experiments [39]. The criteria of ACC can be calculated by

$$ACC = \frac{\Sigma_{i=1}^{n}\delta(p_i, map(q_i))}{n} \times 100\%, \quad (20)$$

where q_i is the clustering label obtained by the algorithm and p_i is the truth label. $\delta(p_i, map(q_i))$ is given by

$$\delta(x, y) = \begin{cases} 1, & x = y, \\ 0, & otherwise, \end{cases} \quad (21)$$

where $map(q_i)$ is the best mapping function. Table 5 summarizes the ACC results of all compared methods, in which "All-Ge" denotes all gene clusters without any dimension reduction processing. The best results are highlighted in bold. From the results, the observations can be summarized as follows:

(i) Since the graph Laplacian is introduced to gLSPCA for clustering purpose, the gLSPCA method performs better than the other methods.
(ii) All-Ge has the lowest ACC result on PAAD dataset, and has intermediate result on HNSC. On PAAD dataset, the clustering results might be interfered by too much irrelevant and redundant information if dimensionality reduction is not employed. However, seldom irrelevant and redundant information contained in HNSC dataset, as well as much information loss in some sparse PCA methods, might lead to the intermediate result of All-Ge on this dataset.

Conclusions

In this paper, we have proposed a new PCA method called gLSPCA by joint graph Laplacian and sparse constraint. The most distinguished characteristics of the new method are that gLSPCA not only considers the internal geometric structure in the data representation, but also adds sparse constraint to PCA. Specifically, we obtain PCs to represent the data meanwhile transform the PCs to approximate the cluster membership indicators in K-means method. The algorithm as well as the convergence analysis of this method has also been developed. The effectiveness of our method has been demonstrated on differentially expressed genes identification and tumour sample clustering comparing with currently available sparse and graph based PCA methods. Finally, we have evaluated the identified differentially expressed genes in the way of co-expression (pathways) and interacting proteins network.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-019-3229-z.

Additional file 1. The pathways of highest overlap on PAAD and HNSC datasets, the pathway of focal adhesion and ECM-receptor interaction. Matching results of each method on PAAD and HNSC datasets.

Abbreviations

ACC: clustering accuracy; All-Ge: denotes all genes cluster without any dimension reduction processing; BCD_SPCA: block coordinate descent sparse PCA; DSPCA and PathSPCA: two sparse PCA method designed by D'Aspremont; EMSPCA: expectation-maximization sparse PCA; gLSPCA: PCA via joint graph Laplacian and Sparse regularization; GPower: a serious of sparse PCA method based on L_0- and L_1-norm with single unit or block units ($GPower_{L0}$, $GPower_{L1}$, $GPower_{L0,m}$, $GPower_{L1,m}$); HNSC: head and neck squamous carcinoma; IA: differentially expressed genes identification performance; MSPCA: Multilinear regression and Sparse regression PCA;

PAAD: pancreatic cancer; PCA: Principal component analysis; PCs: Principal Components; SDP: semidefinite program; sPCA-rSVD: low-rank matrix factorization sparse PCA; SPCArt: sparse PCA Motivated based on rotation matrix and sparse basis; TCGA: The Cancer Genome Atlas; TRS: total relevance scores; Z-SPCA: a sparse PCA method designed by Zou et al.

Acknowledgements
We would like to thank the following individual for their comments and contributions to this article.
Zhenfang Hu, College of Computer Science and Technology, Zhejiang University, China.

About this supplement
This article has been published as part of BMC Bioinformatics Volume 20 Supplement 23, 2019: Proceedings of the Joint International GIW & ABACBS-2019 Conference: bioinformatics. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-23.

Availability of data and material
The datasets that support the findings of this study are available in https://cancergenome.nih.gov/.

Funding declarations
Publication costs are funded by the NSFC under grant Nos. 61572284, 61872220, and 61702299.

Authors' contributions
CMF designed the method. MXH implemented and performed the experiments. LYD and JLS analyzed the obtained results and wrote the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Bio-Computing Research Center, Harbin Institute of Technology, Shenzhen 518055, Guangdong, People's Republic of China. [2]School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, People's Republic of China. [3]Key Laboratory of Network Oriented Intelligent Computation, Shenzhen 518055, People's Republic of China.

Published: 30 December 2019

References

1. Yuan Q, Song C, Gao L, Zhang H, Yang C, Sheng J, Ren J, Chen D, Wang Y. Transcriptome de novo assembly and analysis of differentially expressed genes related to cytoplasmic male sterility in onion. Plant Physiol Biochem. 2018;125:35.
2. Zaslavsky L, Ciufo S, Fedorov B, Tatusova T. Clustering analysis of proteins from microbial genomes at multiple levels of resolution. BMC Bioinform. 2016;17(S8):276.
3. Sharma A, Shigemizu D, Boroevich KA, López Y, Kamatani Y, Kubo M, Tsunoda T. Stepwise iterative maximum likelihood clustering approach. BMC Bioinform. 2016;17(1):319.
4. Guo K, Liu L, Xu X, Xu D, Tao D. GoDec+: fast and robust low-rank matrix decomposition based on maximum correntropy. IEEE Trans Neural Netw Learn Syst. 2018;29(6):2323–36.
5. Wang J, Liu JX, Zheng CH, Wang YX, Kong XZ, Weng CG. A mixed-norm Laplacian regularized low-rank representation method for tumor samples clustering. IEEE/ACM Trans Comput Biol Bioinform. 2017;PP(99):1–1.
6. Feng CM, Gao YL, Liu JX, Zheng CH, Yu J. PCA based on graph Laplacian regularization and P-norm for gene selection and clustering. IEEE Trans Nanobiosci. 2017;16(4):257–65.
7. Feng CM, Gao YL, Liu JX, Wang J, Wang DQ, Wen CG. Joint L1/2-norm constraint and graph-Laplacian PCA method for feature extraction. BioMed Res Int. 2017;2017(2, part 2):1–14.
8. Feng C-M, Xu Y, Liu J-X, Gao Y-L, Zheng C-H. Supervised discriminative sparse PCA for com-characteristic gene selection and tumor classification on multiview biological data. IEEE Trans Neural Netw Learn Syst. 2019;30:1–12.
9. Trigeorgis G, Bousmalis K, Zafeiriou S, Schuller BW. A deep matrix factorization method for learning attribute representations. IEEE Trans Pattern Anal Mach Intell. 2017;39(3):417–29.
10. Liu JX, Kong XZ, Zheng CH, Shang JL, Zhang W. Sparse singular value decomposition-based feature extraction for identifying differentially expressed genes. In: IEEE international conference on bioinformatics & biomedicine; 2017. p. 1822–1827.
11. Češka M, Dannenberg F, Paoletti N, Kwiatkowska M, Brim L. Precise parameter synthesis for stochastic biochemical systems. Acta Informatica. 2017;54(6):589–623.
12. Feng C-M, Xu Y, Li Z, Yang J. Robust classification with sparse representation fusion on diverse data subsets. arXiv preprint arXiv:190611885; 2019.
13. Zhao Y, You X, Yu S, Xu C, Yuan W, Jing XY, Zhang T, Tao D. Multi-view manifold learning with locality alignment. Pattern Recogn. 2018:154–66.
14. Moon KR, Iii JSS, Burkhardt D, Dijk DV, Wolf G, Krishnaswamy S. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. Curr Opin Syst Biol. 2018;7:36–46.
15. Zhu B, Liu JZ, Cauley SF, Rosen BR, Rosen MS. Image reconstruction by domain-transform manifold learning. Nature. 2017;555(7697):487.
16. Feng C-M, Wang K, Lu S, Xu Y, Kong H, Shao L. Coupled-projection residual network for MRI super-resolution. arXiv preprint arXiv:190705598; 2019.
17. Bo J, Ding C, Luo B, Jin T. Graph-Laplacian PCA: closed-form solution and robustness. In: Computer vision & pattern recognition; 2013. p. 3492–3498.
18. Zhao Z, He X, Cai D, Zhang L, Ng W, Zhuang Y. Graph regularized feature selection with data reconstruction. IEEE Trans Knowl Data Eng. 2016;28(3):689–700.
19. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017;14(10):979–82.
20. Benidis K, Sun Y, Babu P, Palomar DP. Orthogonal sparse PCA and covariance estimation via procrustes reformulation. IEEE Trans Signal Process. 2016;64(23):6211–26.
21. Duforetfrebourg N, Luu K, Laval G, Bazin E, Blum MGB. Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. Mol Biol Evol. 2016;33(4):1082–93.
22. Merola GM. SPCA: sparse principal component analysis. Pattern Recogn Lett. 2014;34(9):1037–45.
23. D'Aspremont A, Ghaoui LE, Jordan MI, Lanckriet GRG. A direct formulation for sparse PCA using semidefinite programming. SIAM Rev. 2007;49(3):434–48.
24. D'Aspremont A, Bach F, Ghaoui LE. Full regularization path for sparse principal component analysis. 2008;99(6):1015–1034.
25. Shen H, Huang JZ. Sparse principal component analysis via regularized low rank matrix approximation. J Multivar Anal. 2008;99(6):1015–34.
26. Sigg CD, Buhmann JM. Expectation-maximization for sparse and non-negative PCA. In: International conference on machine learning; 2008. p. 960–967.
27. Journée M, Nesterov Y, Richtárik P, Sepulchre R. Generalized power method for sparse principal component analysis. Core Discuss Pap. 2010;11(2008070):517–53.
28. Lai Z, Xu Y, Chen Q, Yang J, Zhang D. Multilinear sparse principal component analysis. IEEE Trans Neural Netw Learn Syst. 2014;25(10):1942–50.
29. Hu Z, Gang P, Wang Y, Wu Z. Sparse principal component analysis via rotation and truncation. IEEE Trans Neural Netw Learn Syst. 2016;27(4):875.
30. Qian Z, Meng D, Xu Z, Gao C. A block coordinate descent approach for sparse principal component analysis. Neurocomputing. 2015;153:180–90.
31. Gui J, Sun Z, Ji S, Tao D, Tan T. Feature selection based on structured sparsity: a comprehensive study. IEEE Trans Neural Netw Learn Syst. 2016;28(7):1490–507.
32. Hou C, Nie F, Li X, Yi D, Wu Y. Joint embedding learning and sparse regression: a framework for unsupervised feature selection. IEEE Trans Cybern. 2014;44(6):793–804.
33. Safran M, Dalah I, Alexander J, Rosen N, Stein TI, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H. GeneCards version 3: the human gene integrator. Database. 2010;2010(1):baq020.

34. Hart PA, Baichoo E, Bi Y, Hinton A, Kudva YC, Chari ST. Pancreatic polypeptide response to a mixed meal is blunted in pancreatic head cancer associated with diabetes mellitus. Pancreatology. 2015;15(2):162–6.
35. Eyal S, Alex S, Uri R, Rami K, Peter A, Timothy W, Nadir A. Targeting CD24 for treatment of colorectal and pancreatic cancer by monoclonal antibodies or small interfering RNA. Cancer Res. 2013;68(8):2803–12.
36. Zhang H-j, Tao J, Sheng L, Hu X, Rong R-m, Xu M, Zhu T-y. RETRACTED: Twist2 promotes kidney cancer cell proliferation and invasion via regulating ITGA6 and CD44 expression in the ECM-Receptor-Interaction pathway. Biomed Pharmacother. 2016;81(Issue 1):453–9.
37. Kleinschmidt EG, Schlaepfer DD. Focal adhesion kinase signaling in unexpected places. Curr Opin Cell Biol. 2017;45:24–30.
38. Passer D, Vandevrugt A, Atmanli A, Domian I. Atypical protein kinase C-dependent polarized cell division is required for myocardial trabeculation. Cell Rep. 2016;14(7):1662–72.
39. Pehlivanlı AÇ. A novel feature selection scheme for high-dimensional data sets: four-staged feature selection. J Appl Stat. 2016;43(6):1140–54.

Publisher's Note