RPCA-Based Tumor Classification Using Gene Expression Data

Jin-Xing Liu, Yong Xu, Chun-Hou Zheng, Heng Kong, and Zhi-Hui Lai

Abstract—Microarray techniques have been used to delineate cancer groups or to identify candidate genes for cancer prognosis. As such problems can be viewed as classification ones, various classification methods have been applied to analyze or interpret gene expression data. In this paper, we propose a novel method based on robust principal component analysis (RPCA) to classify tumor samples of gene expression data. Firstly, RPCA is utilized to highlight the characteristic genes associated with a special biological process. Then, RPCA and RPCA+LDA (robust principal component analysis and linear discriminant analysis) are used to identify the features. Finally, support vector machine (SVM) is applied to classify the tumor samples of gene expression data based on the identified features. Experiments on seven data sets demonstrate that our methods are effective and feasible for tumor classification.

Index Terms—Classification, data mining, feature selection, principal component analysis, sparse method

1 Introduction

Monitoring gene expression levels on a genomic scale becomes possible with the rapid development of gene microarray technologies. These techniques have been used to deep delineate cancer groups or to identify candidate genes for cancer prognosis and therapeutic targeting. As such problems can be viewed as classification ones, various classification methods have been applied to analyze or interpret gene expression data resulting from DNA microarrays [1], [2], [3], [4], [5], [6], [7], [8].

Gene expression data are characterized by thousands of variables (genes) and a small number of samples (the patients), which often implies a high degree of multi-collinearity. We usually name such a problem as high-dimension-small-sample-size one, which makes it difficult to directly use classical classification methods. In a classification framework, one solution is to reduce the dimensionality of the data either by feature selection, or by feature extraction that summarizes most of the information. Feature selection of gene expression data has been extensively studied in the last few years. In the most commonly used methods of feature selection, a score for each gene is independently calculated at first, then the genes with high scores are selected [9]. Such methods are often denoted as univariate feature selection (UFS). The virtues of UFS methods are intuitive and computationally simple [10]. But a common disadvantage of UFS methods is that they separately considered each feature, thereby ignored the dependencies among features. Multivariate feature selection (MFS) methods, i.e., feature extraction, have been proposed to overcome the disadvantage. Until now, many MFS methods, such as principal component analysis (PCA) and partial least squares (PLS), have been used to analyze gene expression data. Ma and Kosorok used PCA to identify differential gene pathways [11]. PLS was used to analyze the high-dimensional genomic data by Boulesteix and Strimmer [12].

However, such traditional methods still have some drawbacks, e.g., the principal components (PCs) of PCA or latent vectors (LVs) of PLS are usually dense, which makes it difficult to interpret PCs or LVs without subjective judgment. Researchers proposed many sparse versions of MFS methods to overcome these drawbacks. Such methods have significant advantages, while losing little statistical efficiencies [13], [14], [15], [16]. Journée et al. proposed a generalized power method for sparse PCA [13]. Allen and Maleti-Savati used sparse PCA for metabolomics [17]. However, due to the holistic nature of PCA, the resulting components are global interpretations and lack intuitive meanings. Cao et al. used sparse PLS for biologically relevant feature selection [14]. However, because PLS regression is very sensitive to the presence of outliers in the data, it is susceptible to noise.

Recently, a new method of matrix recovery, namely robust principal component analysis (RPCA), has been introduced into the signal processing field [18]. To overcome the drawbacks of above MFS, we propose a novel RPCA-based method for classifying tumor samples. Assuming that all the data points are stacked as column vectors of a matrix $\mathbf{D}$, and the matrix (approximately) has low rank, the RPCA proposed by Candes et al. can recover a low-rank matrix $\mathbf{A}$ from highly corrupted measurements $\mathbf{D}$ [18]. Although the method has been successfully applied to remove shadows from face images and to model background from surveillance video [18], its validity for analyzing gene expression data still needs to be studied. For a special biological process, the expression profiles of most of the genes are flat. All these genes are considered as non-differential expression. It is natural to treat these data of non-differentially expressed genes as approximately low rank. Only a small number of genes are relevant to a special biological process [19], so the data of these differentially expressed genes can be treated as sparse perturbation signals.

In this paper, RPCA is applied to extract a subset of genes associated with a special biological process. Then, support vector machine (SVM) is used to classify tumor samples based on the extracted features. A modified method is also proposed to enhance the classification performance based on RPCA and linear discriminant analysis (RPCA+LDA).

The main contributions of our work are as follows: first, it proposes, for the first time, the idea and method based on RPCA for classifying tumor samples; second, it provides a large number of tumor classification experiments on gene expression data.

The rest of the paper is organized as follows. The proposed method based on RPCA is given in Section 2. The results are shown in Section 3. Section 4 gives the conclusions.

2 Methodology

The proposed method for classifying tumor samples can be divided into two steps. First, RPCA is used to extract a subset of genes from original gene expression data. Then, support vector machine [20] is used to classify the tumor samples based on the identified features.

2.1 RPCA Model for Gene Expression Data

RPCA was originally proposed by Candes et al. [18]. Our goal of using RPCA to model gene expression data is to classify tumor samples based on the characteristic genes that are identified by

- *J.-X. Liu is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, Guangdong 518055, China, and the School of Information Science and Engineering, Qufu Normal University, Rizhao, Shandong 276826, China. E-mail: sdcavell@126.com.*
- *Y. Xu is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, and the Key Laboratory of Network Oriented Intelligent Computation, Shenzhen, Guangdong 518055, China. E-mail: yongxu@ymail.com.*
- *C.-H. Zheng is with the College of Electrical Engineering and Automation, Anhui University, Hefei, Anhui 230039, China. E-mail: zhengch99@126.com.*
- *H. Kong is with the Department of General Surgery, Nan Shan District People's Hospital, Shenzhen, Guangdong 518055, China. E-mail: generaldoc@126.com.*
- *Z.-H. Lai is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, Guangdong 518055, China. E-mail: lai_zhi_hui@163.com.*

Manuscript received 9 Jan. 2014; revised 18 Nov. 2014; accepted 11 Dec. 2014. Date of publication 24 Dec. 2014; date of current version 4 Aug. 2015.
Recommended for acceptance by S. Draghici.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2014.2383375

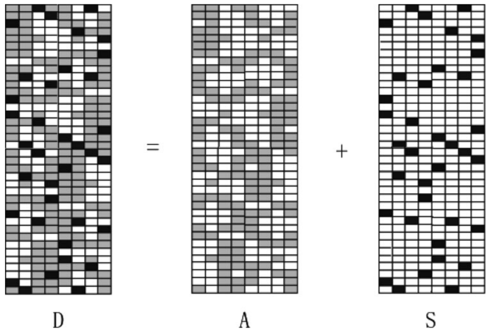

Fig. 1. The schematic model to decompose gene expression data by the RPCA. In this figure, D, A and S denote observation matrix, low-rank matrix and sparse perturbation signals, respectively.

our method. Considering the matrix $\mathbf{D}$ of gene expression data with size $m \times n$, each row of $\mathbf{D}$ represents the transcriptional responses of a gene in all the n samples, and each column of $\mathbf{D}$ represents the expression levels of all the m genes in one sample. Supposing that $\mathbf{D}$ is given by $\mathbf{D} = \mathbf{A} + \mathbf{S}$, RPCA solves the following optimization problem:

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{A}\|_* + \lambda\|\mathbf{S}\|_1 \\ &\text{subject to} \quad \mathbf{D} = \mathbf{A} + \mathbf{S}, \end{aligned} \tag{1}$$

where λ is a positive regulation parameter, $\|\mathbf{A}\|_* := \sum_i \sigma_i(\mathbf{A})$ denotes the nuclear norm of the matrix $\mathbf{A}$, that is, the sum of its singular values, and $\|\mathbf{S}\|_1 := \sum_{ij} |S_{ij}|$ denotes the l_1-norm of $\mathbf{S}$.

A Lagrange multiplier Φ is introduced to remove the equality constraint of the RPCA problem in Eq. (1). According to [21], the augmented Lagrange multiplier method can be applied on the Lagrangian function:

$$\begin{aligned} L(\mathbf{A}, \mathbf{S}, \Phi, \mu) = &\|\mathbf{A}\|_* + \lambda\|\mathbf{S}\|_1 \\ &+ \langle \Phi, \mathbf{D} - \mathbf{A} - \mathbf{S} \rangle + \frac{\rho}{2}\|\mathbf{D} - \mathbf{A} - \mathbf{S}\|_F^2, \end{aligned} \tag{2}$$

where ρ is a positive scalar and $\|.\|_F^2$ denotes the Frobenius norm. Lin et al. gave a method to solve the RPCA problem, which can be referred to [21].

As mentioned in Introduction, it is reasonable to regard the differentially expressed genes as sparse signals, so the differential and non-differential ones can be regarded as the sparse perturbation signals $\mathbf{S}$ and the low-rank matrix $\mathbf{A}$, respectively. When the observation matrix $\mathbf{D}$ has been decomposed into the sparse perturbation signals $\mathbf{S}$ and the low-rank matrix $\mathbf{A}$, the genes of differential expression can be identified by exploiting the perturbation signals $\mathbf{S}$. Fig. 1 shows the schematic model to decompose gene expression data by the RPCA.

In Fig. 1, each row represents the transcriptional responses of a gene in all the samples and each column represents a sample. The white and the gray blocks denote zero and nearly zero, respectively. The black blocks denote the perturbation signals. As shown in Fig. 1, the matrix $\mathbf{S}$ of differentially expressed genes (black blocks) can be recovered from the matrix $\mathbf{D}$ of gene expression data.

Given an appropriate parameter λ, RPCA can decompose the observation matrix $\mathbf{D}$ and give the sparse perturbation matrix $\mathbf{S}$. Most of entries in $\mathbf{S}$ are zero (as white blocks shown in Fig. 1). The genes corresponding to non-zero entries in $\mathbf{S}$ can be considered as ones of differential expression.

2.2 Gene Selection by RPCA

As already mentioned above, the goal of this paper is to propose a RPCA-based method to classify tumor samples based on the subset of genes that are selected by our method. Corresponding to data matrix $\mathbf{D}$, each row of $\mathbf{S}$ represents the transcriptional responses of a gene in all the n samples, and each column of $\mathbf{S}$ represents the expression levels of all the m genes in one sample. After RPCA decomposed the observation matrix $\mathbf{D}$, the sparse perturbation matrix $\mathbf{S}$ can be obtained. Therefore the characteristic genes can be determined by analyzing the sparse matrix $\mathbf{S}$.

The sparse matrix $\mathbf{S}$ can be denoted as

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mn} \end{bmatrix}. \tag{3}$$

The positive and the negative entries in the sparse matrix $\mathbf{S}$ may reflect up- and down-regulated genes of differential expression, respectively. Here, we just consider the amplitude of signals, i.e., the absolute value of entries in $\mathbf{S}$, to identify the differentially expressed genes. So the following two steps are performed: firstly, the absolute values of entries in the sparse matrix $\mathbf{S}$ are calculated; secondly, the matrix $\mathbf{S}$ is summed by rows to obtain the evaluating vector $\tilde{S}$. Mathematically, it can be formulated as follows:

$$\tilde{S} = \left[\sum_{i=1}^{n} |s_{1i}| \quad \cdots \quad \sum_{i=1}^{n} |s_{mi}| \right]^T. \tag{4}$$

Consequently, the entries in $\tilde{S}$ are sorted in descending order to obtain the new evaluating vector $\hat{S}$. Without loss of generality, we suppose that the first c_1 entries in $\hat{S}$ are nonzero, that is,

$$\hat{S} = \left[\hat{s}_1, \quad \ldots, \quad \hat{s}_{c_1}, \underbrace{0, \quad \ldots, \quad 0}_{m-c_1} \right]^T. \tag{5}$$

The principle of the proposed feature extraction method is listed as follows: if some elements of the evaluating vector are zero, deleting the associated input variables does not change the optimal hyperplane of the remaining variables. Even if the element of the evaluating vector is not zero, when its value is small enough, the deletion of the associated input variable does not affect the optimal hyperplane very much.

Generally, the larger the element in $\hat{S}$ is, the more differential the gene expression is, and the more important the gene is. So, the genes associated with only the first num_1 ($num_1 \leq c_1$) entries in $\hat{S}$ are picked out as characteristic ones.

2.3 Model of RPCA+LDA for Gene Data

The naive RPCA method cannot utilize label information of samples, that is, it is an unsupervised method. Therefore it does not always achieve a good classification performance. A two-stage selection method has been proposed to obtain the sparse representation of the test sample [22]. In this research, similar to Xu et al. [22], our scheme can be divided into three stages to improve the classification performance. In the first stage, RPCA is used to extract the characteristic genes from gene expression data. In the second stage, in order to utilize the label information of samples, we use linear discriminant analysis [23] to refine the subset of characteristic genes. Finally, SVM [20] is applied to classify the samples based on the identified features.

Next, we show how to extract the subset of characteristic genes by using LDA [23].

TABLE 1
Summary of the Five Two-Class Data Sets

Data sets	Training set			Testing set			Genes
	Class 1	Class 2	All	Class 1	Class 2	All	
Acute leukemia	7	18	25	4	9	13	5,000
Colon	26	14	40	14	8	22	2,000
Gliomas	18	14	32	10	8	18	12,625
Medulloblastoma	6	16	22	3	9	12	5,893
Prostate	39	51	90	20	26	46	12,600

According to [23], the objective function of LDA is given as follows:

$$\begin{aligned} \mathbf{w} &= \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}} \\ S_b &= \sum_{k=1}^{c} m_k (\mu^{(k)} - \mu)(\mu^{(k)} - \mu)^T \\ S_w &= \sum_{k=1}^{c} \left[\sum_{i=1}^{m_k} \left(x_i^{(k)} - \mu^{(k)}\right)\left(x_i^{(k)} - \mu^{(k)}\right)^T \right], \end{aligned} \tag{6}$$

where μ is the mean vector of total samples, m_k is the number of samples in the kth class, $\mu^{(k)}$ is the average vector of the kth class, and $x_i^{(k)}$ is the ith sample in the kth class. After the characteristic genes have been identified by using RPCA, LDA is utilized to bring label information of samples into classification model. Then, a subset of refined genes are identified according to discriminant vectors of LDA.

When obtaining $\mathbf{w}$ by LDA, the characteristic genes can be extracted by analyzing the projection vectors $\mathbf{w}$.

The projection matrix $\mathbf{W}$ can be denoted as

$$\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r] = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1r} \\ w_{21} & w_{22} & \cdots & w_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mr} \end{bmatrix}. \tag{7}$$

Like the method of Section 2.2, we calculate the absolute values of the projection matrix $\mathbf{W}$ and sum it by rows. Mathematically, it can be formulated as follows:

$$\tilde{W} = \left[\sum_{i=1}^{r} |w_{1i}| \quad \cdots \quad \sum_{i=1}^{r} |w_{mi}| \right]^T. \tag{8}$$

Consequently, to obtain the new evaluating vector $\hat{W}$, the entries in $\tilde{W}$ are sorted in descending order. Without loss of generality, we suppose that the first c_2 entries in $\hat{W}$ are non-zero, that is,

$$\hat{W} = \left[\hat{w}_1, \quad \ldots, \quad \hat{w}_{c_2}, \underbrace{0, \quad \ldots, \quad 0}_{m-c_2} \right]^T. \tag{9}$$

Generally, the larger the element in $\hat{W}$ is, the more differential the gene expression is. So, the genes associated with only the first num_2 ($num_2 \leq c_2$) entries in $\hat{W}$ are selected as characteristic ones.

3 Results

In this section, the effectiveness of our methods is demonstrated by classifying human tumor samples in seven data sets. Three kinds of measures are used to evaluate the performances of these methods. The first one is the leave-one-out cross-validation (LOO-CV) performance which is estimated only by making use of the training data sets for tuning the parameters. The second one is the accuracy which measures the classification performance by using the percentage of correctly classified samples. The third one is the AUC (Area Under the Curve of ROC) which is suitable for evaluating the performance of binary classification [24].

TABLE 2
Summary of the Two Multi-Class Data Sets

Data sets	Number of			Description
	Classes	Samples	Genes	
11_Tumor	11	174	12,533	11 various tumor types
Brain_Tumor	5	90	5,920	5 brain tumor types

3.1 Data Sets

We use seven publically available data sets of gene expression to evaluate the performances of our methods: acute leukemia data [25], colon cancer data [26], gliomas data [27], medulloblastoma data [28], prostate cancer data [29], 11_Tumors data [30] and Brain_Tumor data [31]. All the tumor samples in the seven data sets are randomly divided into training and testing sets. In all the data sets, two-thirds of the samples in each class are assigned to the training set and the rest to the testing set. Tables 1 and 2 list overviews of the characteristics of the five two-class data sets and the two multi-class data sets, respectively.

The acute leukemia data set contains gene expression levels for $N = 5{,}000$ genes in 38 samples and consists of 27 cases of acute lymphoblastic leukemia and 11 cases of acute myeloid leukemia. The gene expression data are summarized by a $5{,}000 \times 38$ matrix.

The colon cancer data set contains gene expression levels for $N = 2{,}000$ genes in 62 colon tissues, of which 40 and 22 samples are tumor and normal, respectively. Colon cancer is the fourth most common cancer for males and females and the second most frequent cause of death.

The gliomas data set contains gene expression levels for $N = 12{,}625$ genes in 50 samples, including 28 glioblastomas and 22 anaplastic oligodendrogliomas.

The medulloblastoma data set is derived from childhood brain tumors. The pathogens of these tumors are not well understood, but it is generally accepted that there are two known histological subclasses: classic and desmoplastic, whose differences can be clearly seen under the microscope. The data set contains gene expression levels for $N = 5{,}893$ genes in 34 samples, including nine classic and 25 desmoplastic.

The prostate data set consists of 136 prostate tissues, including 59 normal and 75 tumor samples. The number of genes is 12,600.

The 11_Tumor data set consists of 174 samples in 11 various human tumors, including ovary, bladder, breast, colorectal, gastroesophagus, kidney, liver, prostate, pancreas, lung adeno and lung squamous.

The Brain_Tumor data set contains gene expression levels for $N = 5{,}920$ genes in 90 samples, including medulloblastoma, malignant glioma, AT/RT, normal cerebellum and PNET.

3.2 Experimental Process

The classification models are constructed by using the training set and their classification performances are tested by using the testing set.

The RPCA is firstly performed on training set, denoted as $\mathbf{D}_{tr}$, to produce the non-differentially expressed matrix $\mathbf{A}_{tr}$ and differentially expressed matrix $\mathbf{S}_{tr}$, which can be shown as follows:

$$\mathbf{D}_{tr} = \mathbf{A}_{tr} + \mathbf{S}_{tr}. \tag{10}$$

Since each row of $\mathbf{S}_{tr}$ represents the transcriptional responses of a gene in all the n samples, the differentially expressed genes can be identified based on matrix $\mathbf{S}_{tr}$. In this research, the sparsity-controlling parameter λ is set to $[0.3 * \max(m,n)]^{-0.5}$ when using RPCA [32]. Then characteristic genes are identified according to $\mathbf{S}_{tr}$ by using the method described in Section 2.2. The extracted data set are denoted by $\mathbf{D}'_{tr}$ (training data) and $\mathbf{D}'_{tt}$ (testing data).

When using the model of RPCA+LDA, we refine the characteristic genes by LDA, which is given in Section 2.3. The extracted data subsets are denoted by $\mathbf{D}''_{tr}$ (training data) and $\mathbf{D}''_{tt}$ (testing data).

After processing the gene expression data by using RPCA or RPCA+LDA, the final step is to classify the tumor samples. In this research, $\mathbf{D}''_{tr}$ (or $\mathbf{D}'_{tr}$) is used to train the SVM, and $\mathbf{D}''_{tt}$ (or $\mathbf{D}'_{tt}$) is used to test the classification performance.

Consequently, all numerical experiments are performed with thirty random partitions on the seven original data sets. These randomizations are the same for all numerical experiments on one dataset. Each randomized training and testing sets contains the same amount of samples of each class compared to the original training and testing sets.

3.3 Results of Tumor Classification

In this subsection, characteristic genes are identified by using RPCA or RPCA+LDA, and then these genes are utilized to classify the tumor samples. For comparison, SVM and LDA are directly used to classify the original data, respectively. And sparse PCA (SPCA) and PLS are used for feature extraction with SVM for classification (SPCA+SVM, PLS+SVM). For fair comparison with RPCA+LDA, another competitive method, SPCA+LDA+SVM, is involved in our experiments. In this scheme, first, SPCA is used to identify a subset of genes. Then, similar with RPCA+LDA, LDA is used to refine a subset of genes. For comparison, we also use Sparse Representation-based Classification (SRC) [33], k-Nearest Neighbor Classification (kNNC) [34] and Sparse Logistic Regression with Elastic Net penalty (SLREN) [35] methods to classify the tumor samples.

SPCA that maximizes the sample variance seeks to project the data onto the linear combination of variables. Suppose $\mathbf{D}$ has zero mean, the sparse principal component z_k of $\mathbf{D}$ can be calculated by using the following objective function:

$$\Psi_{l_1}(\gamma) \overset{def}{=} \max_{z \in B^p} \sqrt{z^T \Sigma z} - \gamma \|z\|_1 \tag{11}$$

with sample covariance matrix $\Sigma = \mathbf{D}^T \mathbf{D}$ and sparsity-controlling parameter $\gamma \geq 0$, $\|\cdot\|_1$ denotes the l_1-norm. Then a subset of features can be selected according to the sparse principal components $\mathbf{Z}$. Detailed description of gene selection by SPCA can be found in [36].

Let $\mathbf{X}$ denote an $n \times m$ matrix, which consists of the n observations of the m features. Let $\mathbf{Y}$ be an $n \times 1$ vector to denote the classification labels. PLS method performs a simultaneous decomposition of $\mathbf{X}$ and $\mathbf{Y}$ with the constraint to search for a set of components (called latent vectors) that explain as much as possible of the covariance between $\mathbf{X}$ and $\mathbf{Y}$. PLS method decomposes $\mathbf{X}$ and $\mathbf{Y}$ into the form:

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \\ \mathbf{Y} &= \mathbf{U}\mathbf{Q}^{\mathbf{T}} + \mathbf{F}, \end{aligned} \tag{12}$$

where the $\mathbf{T}$, $\mathbf{U}$ are $n \times k$ matrices of the k extracted score vectors (components). The $m \times k$ matrix $\mathbf{P}$ and the $1 \times k$ matrix $\mathbf{Q}$ represent the matrices of loadings and the $n \times m$ matrix $\mathbf{E}$ and the $n \times 1$ matrix $\mathbf{F}$ are the matrices of residuals. Detailed description of PLS can be found in [12].

SRC has been successfully used to recognize human faces [33]. Let $\mathbf{D}$ be an $m \times n$ matrix of m genes in n samples of all k object classes

$$\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_k] = \left[d_{1,1}, d_{1,2}, \ldots, d_{k,n_k}\right] \tag{13}$$

any test sample $\mathbf{y} \in \mathrm{R}^m$ can be approximately represented in terms of all training samples as $\mathbf{y} = \mathbf{D}\alpha_0$, where $\alpha_0 = \left[0, \ldots, 0, \alpha_{i,1}, \ldots, \alpha_{i,n_i}, 0, \ldots, 0\right]^T \in \mathrm{R}^n$ is a coefficient vector whose entries are zero except those associated with the ith class.

According to [33], SRC solves the following l_1-minimization problem to seek the sparse solution:

$$\hat{\alpha} = \arg\min \|\alpha\|_1 \quad \text{subject to} \quad \mathbf{D}\alpha = \mathbf{y}. \tag{14}$$

The details of the SRC algorithm can be referred to [33].

SLREN is a sparse version of logistic regression which solves the following problem [35]:

$$\min_{(\beta_0,\beta) \in R^{m+1}} \left[-\ell_Q(\beta_0, \beta) + \eta P_\theta(\beta)\right] \tag{15}$$

where

$$\ell_Q(\beta_0, \beta) = -\frac{1}{2n} \sum_{i=1}^{n} w_i \left(z_i - \beta_0 - x_i^T \beta\right)^2 + C\left(\tilde{\beta}_0, \tilde{\beta}\right)^2 \tag{16}$$

$$P_\theta(\beta) = \frac{1}{2}(1 - \theta)\|\beta\|_{l_2}^2 + \theta \|\beta\|_{l_1} \tag{17}$$

$$z_i = \tilde{\beta}_0 + x_i^T \tilde{\beta} + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)(1 - \tilde{p}(x_i))}, \tag{18}$$

$$w_i = \tilde{p}(x_i)(1 - \tilde{p}(x_i)). \tag{19}$$

$P_\theta(\beta)$ is the elastic-net penalty [35], which is a tradeoff between the ridge-regression penalty ($\theta = 0$) and the lasso penalty ($\theta = 1$). Detailed description of the SLREN algorithm can be referred to [37].

The classification results are listed in Table 3. In this table, ACC denotes accuracy performance, and the second and third columns give the percentage values of accuracies, and the last column gives the AUC values of these methods. For each experiment, we run the program thirty times with random splits of the data set and calculate the mean and standard deviation of the classification accuracies. The numerical values in this table represent the means with standard deviations.

On acute leukaemia data set, the testing accuracy of RPCA + LDA + SVM is better than others. On colon cancer and gliomas data sets, the testing accuracies of RPCA-based methods are better than others. On medulloblastoma data set, the LOO-CV and testing performances of RPCA+LDA+SVM are better than others. On prostate cancer data set, the LOO-CV performances of methods 4, 6 and 9 are better than others, but the testing performances of methods 6, 7 and 9 are better than others.

Then, on acute leukaemia and 11_Tumor data sets, all the methods (Except method 1, 2 and 9) give much higher testing accuracies (Especially RPCA+LDA+SVM shows above 99 percent performance), while on gliomas data set, all the methods give much lower testing performances. It may reflect that different data sets have different separabilities. Here, the data sets of acute leukaemia and 11_Tumor may be of high separability, and the gliomas data set may be of low separability.

Furthermore, on some data sets, such as acute leukaemia data, gliomas data, and medulloblastoma data, method 1 (SVM) gives the very lower testing performance. These results may reflect the over-training of classifier, but other methods, especially our methods, can avoid the over-training by using feature selection. For example, our methods can obtain the highest testing performances on colon and gliomas data.

In addition, from the last column of this table, we can see that on all the data sets, RPCA+LDA+SVM can give the greatest measure and RPCA+SVM can give the competitive results.

TABLE 3
Summary of the Classification Results

Experimental Method	LOO-CV performance	ACC on testing set	AUC
Acute leukaemia data			
1 SVM	79.20 ± 7.00	80.77 ± 4.05	0.82 ± 0.12
2 SPCA+SVM	89.60 ± 5.72	85.38 ± 6.74	0.85 ± 0.11
3 PLS+SVM	98.80 ± 1.93	96.92 ± 3.97	0.99 ± 0.01
4 SPCA+LDA+SVM	99.60 ± 1.26	95.38 ± 6.49	**1.00 ± 0.00**
5 RPCA+SVM	94.40 ± 2.07	93.08 ± 6.74	0.99 ± 0.01
6 RPCA+LDA+SVM	100.00 ± 0.00	**99.23 ± 2.43**	**1.00 ± 0.00**
7 SRC	96.40 ± 2.95	96.15 ± 4.05	0.94 ± 0.07
8 kNNC	93.16 ± 5.14	93.08 ± 6.74	0.93 ± 0.09
9 SLREN	88.22 ± 2.23	89.23 ± 10.38	0.84 ± 0.17
10 LDA	96.00 ± 2.67	96.15 ± 4.05	0.94 ± 0.07
Colon cancer data			
1 SVM	83.25 ± 4.42	84.55 ± 2.35	0.88 ± 0.06
2 SPCA+SVM	83.25 ± 3.34	85.91 ± 4.52	0.92 ± 0.04
3 PLS+SVM	89.25 ± 1.69	**90.91 ± 3.03**	0.92 ± 0.02
4 SPCA+LDA+SVM	89.25 ± 3.13	89.55 ± 4.31	0.92 ± 0.03
5 RPCA+SVM	88.75 ± 2.43	**90.91 ± 2.14**	**0.93 ± 0.03**
6 RPCA+LDA+SVM	91.25 ± 1.77	**90.45 ± 3.35**	**0.93 ± 0.03**
7 SRC	87.75 ± 2.19	85.00 ± 3.74	0.84 ± 0.04
8 kNNC	83.39 ± 4.50	82.73 ± 5.59	0.92 ± 0.02
9 SLREN	80.06 ± 4.08	88.18 ± 3.83	0.88 ± 0.05
10 LDA	87.50 ± 2.36	89.55 ± 3.74	0.89 ± 0.04
Gliomas data			
1 SVM	69.38 ± 7.34	72.22 ± 7.41	0.68 ± 0.05
2 SPCA+SVM	68.75 ± 6.75	70.00 ± 7.03	0.76 ± 0.08
3 PLS+SVM	76.56 ± 6.46	76.11 ± 6.95	0.79 ± 0.08
4 SPCA+LDA+SVM	71.88 ± 8.96	72.78 ± 9.24	0.79 ± 0.05
5 RPCA+SVM	76.25 ± 6.11	**76.67 ± 4.38**	0.87 ± 0.06
6 RPCA+LDA +SVM	82.19 ± 4.67	**77.78 ± 5.24**	**0.89 ± 0.05**
7 SRC	76.88 ± 7.25	72.78 ± 7.61	0.72 ± 0.07
8 kNNC	80.20 ± 5.20	**76.67 ± 8.73**	0.88 ± 0.10
9 SLREN	74.12 ± 7.77	76.11 ± 9.82	0.76 ± 0.10
10 LDA	74.69 ± 8.13	73.33 ± 5.74	0.73 ± 0.06
Medulloblastoma data			
1 SVM	72.73 ± 0.00	75.00 ± 0.00	0.83 ± 0.10
2 SPCA+SVM	73.18 ± 1.44	80.00 ± 4.30	0.86 ± 0.12
3 PLS+SVM	95.45 ± 3.71	**88.33 ± 5.83**	0.91 ± 0.12
4 SPCA+LDA+SVM	83.18 ± 9.10	85.83 ± 4.03	0.93 ± 0.09
5 RPCA+SVM	73.18 ± 1.44	86.67 ± 5.83	0.91 ± 0.11
6 RPCA+LDA+SVM	98.64 ± 3.07	**88.33 ± 5.83**	**0.94 ± 0.11**
7 SRC	86.36 ± 6.78	83.33 ± 6.80	0.73 ± 0.15
8 kNNC	80.00 ± 5.15	83.33 ± 8.78	0.75 ± 0.12
9 SLREN	76.44 ± 5.33	81.67 ± 8.61	0.67 ± 0.18
10 LDA	80.45 ± 7.44	83.33 ± 3.93	0.72 ± 0.06
Prostate cancer data			
1 SVM	79.67 ± 4.72	78.91 ± 4.59	0.74 ± 0.09
2 SPCA+SVM	80.00 ± 4.35	82.17 ± 7.59	0.89 ± 0.05
3 PLS+SVM	77.89 ± 3.90	77.83 ± 8.31	0.84 ± 0.06
4 SPCA+LDA+SVM	91.22 ± 1.69	87.61 ± 4.47	**0.94 ± 0.04**
5 RPCA+SVM	87.22 ± 2.73	87.39 ± 4.32	**0.94 ± 0.04**
6 RPCA+LDA+SVM	89.44 ± 1.68	**88.26 ± 3.43**	**0.94 ± 0.04**
7 SRC	88.78 ± 2.69	**88.02 ± 3.90**	0.91 ± 0.04
8 kNNC	86.69 ± 2.36	83.26 ± 3.70	0.91 ± 0.04
9 SLREN	90.86 ± 1.78	**88.91 ± 7.35**	0.88 ± 0.08
10 LDA	88.11 ± 2.10	85.87 ± 5.50	0.86 ± 0.06
11_Tumor data			
1 SVM	49.56 ± 6.35	87.38 ± 2.76	-
2 SPCA+SVM	68.67 ± 2.84	88.02 ± 2.15	-
3 PLS+SVM	75.84 ± 3.23	98.02 ± 1.15	-
4 SPCA+LDA+SVM	73.45 ± 2.09	98.52 ± 0.93	-
5 RPCA+SVM	79.91 ± 3.26	**99.34 ± 1.15**	-
6 RPCA+LDA+SVM	78.50 ± 3.91	**99.34 ± 1.15**	-
7 SRC	93.27 ± 1.73	**100.00 ± 0.00**	-
8 kNNC	82.70 ± 4.89	85.41 ± 3.13	-
9 SLREN	64.46 ± 1.86	88.66 ± 1.12	-
10 LDA	92.21 ± 1.61	98.84 ± 0.52	-
Brain_Tumor data			
1 SVM	69.66 ± 1.67	82.50 ± 7.82	-
2 SPCA+SVM	75.00 ± 4.08	86.18 ± 2.97	-
3 PLS+SVM	82.07 ± 3.65	86.88 ± 2.87	-
4 SPCA+LDA+SVM	80.17 ± 5.70	88.13 ± 4.11	-
5 RPCA+SVM	81.03 ± 3.81	**92.19 ± 3.97**	-
6 RPCA+LDA+SVM	84.83 ± 2.41	**92.50 ± 2.64**	-
7 SRC	88.45 ± 2.94	**92.01 ± 4.01**	-
8 kNNC	86.33 ± 6.02	89.69 ± 3.31	-
9 SLREN	70.98 ± 1.13	84.15 ± 5.62	-
10 LDA	86.90 ± 2.47	91.50 ± 3.33	-

Finally, on multi-class tumor data sets (11_Tumor and Brain_Tumor), SRC and our methods can give the highest testing accuracies.

From the results on the seven data sets, it can be drawn that the RPCA-based methods are effective and feasible for tumor classification.

3.4 Results of Feature Selection

In this section, RPCA+LDA is used to highlight characteristic genes. The method is set to the same initial parameters as above. We retrieve them from the web of National Center for Biotechnology Information (NCBI) [38] to further understand the identified genes. Here, due to the space limitation, results of only two data sets are given in the following.

On the acute leukaemia data set, Table 4 lists the top twenty genes identified by our method. From Table 4, it can be seen that

TABLE 4
The Top 20 Genes of the Acute Leukemia Data Identified by RPCA+LDA

No.	Official Symbol	Summary of function
254	CEBPD	It is important genes involved in immune responses.
1,331	STAB1	It may function in lymphocyte homing.
1,555	CFD	This gene has a role in immune system biology.
2,261	NR4A2	It mutations have been associated with disorders related to dopaminergic dysfunction.
3,208	TCL1	It was implicated in the T cell leukemia.
3,543	MPO	It is a protein synthesized during myeloid differentiation.
4,814	GLUL	Antigen binding; immune response.
43	CD2	It is a surface antigen of the T-lymphocyte lineage.
202	IL8	It plays a role in the pathogenesis of bronchiolitis.
1,124	BLK	It has a role in B-cell receptor signaling.
1,865	LTB	An inducer of the inflammatory response system.
2,654	S100A13	It is widely expressed in various types of tissues with a high expression level in thyroid gland.
3,437	SELL	Its defects causes the leukocyte adhesion deficiency.
3,774	GATA3	Its defects cause hypoparathyroidism with sensorineural deafness and renal dysplasia.
4,454	ELANE	It is expressed during neutrophil differentiation.
4,828	CST3	Its mutation was associated with amyloid angiopathy.
168	ZNF787	Zinc ion binding; regulation of transcription.
182	RHOG	It promotes reorganization of the actin cytoskeleton and regulate cell shape, attachment, and motility.
646	SNRPN	Its deletion is responsible for Angelman syndrome or Prader-Willi syndrome.
1,374	CD79A	It is necessary for the B-cell antigen receptor.

TABLE 5
Enrichment Analysis of the Top 100 Genes in the Acute Leukemia Data Set

Rank	Item Name	P-value	Term in Query	Term in Genome
1	Human Leukemia_Ben-Dor00_143genes	1.95E-21	20	129
2	Genes down-regulated in hematopoietic progenitor cells (HPC) of T lymphocyte	9.29E-15	13	63
3	Genes in module_33	2.39E-10	23	372
4	Hematopoietic organ development	4.91E-10	23	625
5	Human Leukemia_Chiaretti10_405genes	5.83E-10	17	316
6	Defense response	1.02E-09	30	1,190
7	Regulation of immune system process	1.07E-09	27	943
8	Hemopoiesis	1.34E-09	22	589
9	Immune system development	1.39E-09	23	657
10	Genes in module_45	4.65E-09	26	563

TABLE 6
The Top 20 Genes of the Medulloblastoma Data Identified by RPCA+LDA

No.	Official Symbol	Summary of function
2,952	NEUROG1	Neuron differentiation; DNA binding.
3,392	UCHL1	It may be associated with Parkinson disease.
3,928	SHROOM2	It is in amiloride-sensitive sodium channel activity.
4,477	GAD1	It may also play a role in the stiff man syndrome.
5,037	CA4	It participates in a variety of biological processes, e.g. respiration, saliva, and gastric acid.
5,093	POU3F2	It enhances the activation of corticotropin-releasing hormone regulated genes.
5,373	TYRO3	It is involved in controlling cell survival and immunoregulation and phagocytosis.
20	AFFX-	AFFX-HUMRGE/M10098_5_at (Control)
1,538	THRA	It is a nuclear hormone receptor.
2,285	LHX2	It may function as a transcriptional regulator of a unique cysteine-rich zinc-binding domain.
3,416	PRKCB	Its functions include B cell activation, apoptosis induction, endothelial cell proliferation, etc.
4,167	IGBP1	It is in proliferation and differentiation of B cells.
4,457	MBP	It is a constituent of the myelin sheath of Schwann cells and oligodendrocytes in the nervous system.
4,611	RAD23A	It plays a role in nucleotide excision repair.
4,940	IGH	It responds foreign antigens and initiate immune.
5,442	SLC6A8	Its defects can result in X-linked creatine deficiency syndrome.
1,152	MIF	It encodes a lymphokine involved in cell-mediated immunity, immunoregulation, and inflammation.
1,279	POLR2J	It is for synthesizing messenger RNA.
1,447	TTR	Its mutation caused cardiomyopathy, etc.
2,113	RND3	It encodes a protein which is a member of the small GTPase protein superfamily.

TABLE 7
Enrichment Analysis of the Top 100 Genes in the Medulloblastoma Data Set

Rank	Item Name	P-value	Term in Query	Term in Genome
1	Genes in module_100 (C4 - CM: Cancer Modules)	9.47E-13	29	529
2	generation of neurons	3.28E-10	31	1,200
3	neurogenesis	1.48E-09	31	1,270
4	behavior	2.18E-08	21	597
5	regulation of multicellular organismal development	3.49E-08	29	1,249
6	positive regulation of RNA metabolic process	3.55E-08	28	1,161
7	Genes up-regulated in circulating endothelial cells (CEC) from cancer patients compared to those from healthy donors.	9.89E-07	11	158
8	neuron part	2.74E-05	20	978
9	protein dimerization activity	6.85E-05	21	1,047
10	axon	9.03E-05	12	362

most of genes have affinity with acute leukaemia. For example, GATA3, TCL1, CD2, BLK and CD79A are associated with T-cell and B-cell which play an important role in endothelial cell biology. CEBPD and LTB may play roles in degenerative and inflammatory diseases by its proteolysis of collagen-IV.

On the acute leukaemia data set, to further study the biological function of the identified genes, we also perform the functional enrichment analysis of the top 100 genes identified by our method on the web site http://toppgene.cchmc.org/enrichment.jsp. The result is listed in Table 5. From this table, we can see that there are 129 genes in the item of "Human Leukemia_Ben-Dor00_143genes", in which twenty genes are included in these top 100 genes. This item has the lowest p-value, so it is considered as the most probable enrichment item. Some other items with the most significance are also listed in the table.

On medulloblastoma data set, the top twenty genes identified by RPCA+LDA are listed in Table 6. From Table 6, we can see that most of genes are related to medulloblastoma. For instance, UCHL1, GAD1, TYRO3, POU3F2, THRA, SLC6A8, TTR and TAD23A are involved in a variety of neurophysiologic processes. IGBP1, IGH, MBP, MIF, TYRO3 and PRKCB are involved in immune responses progresses.

On medulloblastoma data set, the top 100 genes are also analyzed by the tool of functional enrichment analysis on the same web site. The result is listed in Table 7.

4 Conclusion

In this paper, we propose the novel methods based on RPCA and RPCA+LDA for tumor classification using gene expression data. The main objective of our paper is to study tumor classification by feature selection. First, RPCA is used to highlight the characteristic genes associated with a special biological process. Then, the feature selection is following by using RPCA and RPCA+LDA. Finally, SVM is used for tumor classification. The results demonstrate that the proposed methods are effective and feasible for tumor classification.

In the future, we will focus on the biological meanings of gene selection.

Acknowledgments

This work was supported in part by the NSFC under grant Nos. 61370163, 61233011, 61332011, 61375012, 61203376 and 61272339; Shandong Provincial Natural Science Foundation, under grant No. ZR2013FL016; China Postdoctoral Science Foundation funded project, No. 2014M560264; Shenzhen Municipal Science and Technology Innovation Council (Nos. JCYJ20140417172417174 and JCYJ20130329151843309); Key Project of Anhui Educational Committee, under Grant No. KJ2012A005. Yong Xu is the corresponding author.

References

[1] L. H. Sobin and I. D. Fleming, "TNM classification of malignant tumors," *Cancer*, vol. 80, no. 9, pp. 1803–1804, 1997.

[2] R. Hewett and P. Kijsanayothin, "Tumor classification ranking from microarray data," *BMC Genomics*, vol. 9, no. Suppl 2, p. S21, 2008.

[3] C.-H. Zheng, L. Zhang, T. Ng, and C. Shiu, "Metasample based sparse representation for tumor classification," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 8, no. 5, pp. 1273–1282, Sep./Oct. 2011.

[4] C. H. Zheng, V. To-Yee Ng, L. Zhang, C. K. Shiu, and H. Q. Wang, "Tumor classification based on non-negative matrix factorization using gene expression data," *IEEE Trans. NanoBiosci.*, vol. 10, no. 2, pp. 86–93, Jun. 2011.

[5] X. Hang and F.-X. Wu, "Sparse representation for classification of tumors using gene expression data," *J. Biomed. Biotechnol.*, vol. 2009, no. 403689, pp. 1–6, 2009.

[6] S.-L. Wang, Y.-H. Zhu, W. Jia, and D.-S. Huang, "Robust classification method of tumor subtype by using correlation filters," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 9, no. 2, pp. 580–591, Mar./Apr. 2012.

[7] J.-X. Liu, Y.-L. Gao, Y. Xu, C.-H. Zheng, and J. You, "Differential expression analysis on RNA-seq count data based on penalized matrix decomposition," *IEEE Trans. NanoBiosci.*, vol. 13, no. 1, pp. 12–18, Mar. 2014.

[8] S.-L. Wang, X.-L. Li, and J. Fang, "Finding minimum gene subsets with heuristic breadth-first search algorithm for robust tumor classification," *BMC Bioinformat.*, vol. 13, no. 1, p. 178, 2012.

[9] S. Dudoit, J. P. Shaffer, and J. C. Boldrick, "Multiple hypothesis testing in microarray experiments," *Stat. Sci.*, vol. 18, no. 1, pp. 71–103, 2003.

[10] Y. Saeys, I. Inza, and P. Larra Aga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[11] S. Ma and M. R. Kosorok, "Identification of differential gene pathways with principal component analysis," *Bioinformatics*, vol. 25, no. 7, pp. 882–889, 2009.

[12] A. L. Boulesteix and K. Strimmer, "Partial least squares: A versatile tool for the analysis of high-dimensional genomic data," *Briefings Bioinformat.*, vol. 8, no. 1, pp. 32–44, Jan. 2007.

[13] M. Journée, Y. Nesterov, P. Richtarik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *J. Mach. Learn. Res.*, vol. 11, pp. 517–553, 2010.

[14] K. A. Le Cao, S. Boitard, and P. Besse, "Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems," *BMC Bioinformat.*, vol. 12, no. 1, p. 253, 2011.

[15] J.-X. Liu, C.-H. Zheng, and Y. Xu, "Extracting plants core genes responding to abiotic stresses by penalized matrix decomposition," *Comput. Biol. Med.*, vol. 42, no. 5, pp. 582–589, 2012.

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

[17] G. I. Allen and M. Maleti Savati, "Sparse non-negative generalized PCA with applications to metabolomics," *Bioinformatics*, vol. 27, no. 21, pp. 3029–3035, 2011.

[18] E. J. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," Arxiv Preprint ArXiv:0912.3599, 2009.

[19] W. R. Scheible, R. Morcuende, T. Czechowski, C. Fritz, D. Osuna, N. Palacios-Rojas, D. Schindelasch, O. Thimm, M. K. Udvardi, and M. Stitt, "Genome-wide reprogramming of primary and secondary metabolism, protein synthesis, cellular growth processes, and the regulatory infrastructure of Arabidopsis in response to nitrogen," *Plant Physiology*, vol. 136, no. 1, pp. 2483–2499, 2004.

[20] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[21] Z. Lin, M. Chen, L. Wu, and Y. Ma. (2010). The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," [Online]. Available: http://Arxiv.org/abs/100055v2

[22] Y. Xu, D. Zhang, J. Yang, and J.-Y. Yang, "A two-phase test sample sparse representation method for use with face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 9, pp. 1255–1262, Sep. 2011.

[23] D. Cai, X. He, and J. Han, "SRDA: An efficient algorithm for large-scale discriminant analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 1, pp. 1–12, Jan. 2008.

[24] A. L. Tarca, M. Lauria, M. Unger, E. Bilal, S. Boue, K. K. Dey, J. Hoeng, H. Koeppl, F. Martin, and P. Meyer, "Strengths and limitations of microarray-based phenotype prediction: Lessons learned from the improver diagnostic signature challenge," *Bioinformatics*, vol. 29, no. 22, pp. 2892–2899, 2013.

[25] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 15, 1999.

[26] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Nat. Acad. Sci. USA*, vol. 96, no. 12, pp. 6745–6750, Jun. 8, 1999.

[27] C. L. Nutt, D. R. Mani, R. A. Betensky, P. Tamayo, J. G. Cairncross, C. Ladd, U. Pohl, C. Hartmann, M. E. McLaughlin, T. T. Batchelor, P. M. Black, A. Von Deimling, S. L. Pomeroy, T. R. Golub, and D. N. Louis, "Gene expression-based classification of malignant gliomas correlates better with survival than histological classification," *Cancer Res.*, vol. 63, no. 7, pp. 1602–1607, Apr. 1, 2003.

[28] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 12, pp. 4164–4169, Mar. 23, 2004.

[29] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, and J. P. Richie, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.

[30] A. I. Su, J. B. Welsh, L. M. Sapinoso, S. G. Kern, P. Dimitrov, H. Lapp, P. G. Schultz, S. M. Powell, C. A. Moskaluk, and H. F. Frierson, "Molecular classification of human carcinomas by use of gene expression signatures," *Cancer Res.*, vol. 61, no. 20, pp. 7388–7393, 2001.

[31] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, and C. Lau, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.

[32] J.-X. Liu, Y.-T. Wang, C.-H. Zheng, W. Sha, J.-X. Mi, and Y. Xu, "Robust PCA based method for discovering differentially expressed genes," *BMC Bioinformat.*, vol. 14, no. 8, pp. 1–10, 2013.

[33] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[34] A. P. James and S. Dimitrijev, "Nearest neighbor classifier based on nearest feature decisions," *Comput. J.*, vol. 55, no. 9, pp. 1072–1087, 2012.

[35] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Royal Stat. Soc.: Ser. B (Stat. Methodol.)*, vol. 67, no. 2, pp. 301–320, 2005.

[36] J.-X. Liu, Y. Xu, C.-H. Zheng, Y. Wang, and J.-Y. Yang, "Characteristic gene selection via weighting principal components by singular values," *Plos One*, vol. 7, no. 7, p. e38873, 2012.

[37] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Softw.*, vol. 33, no. 1, pp. 1–22, 2010.

[38] NCBI (2012, Oct. 10) National center for biotechnology information [Online]. Available: http://www.ncbi.nlm.nih.gov/