Iraqi Journal of Science, 2021, Vol. 62, No. 4, pp: 1321-1333
DOI: 10.24996/ijs.2021.62.4.28

ISSN: 0067-2904

Enhanced Supervised Principal Component Analysis for Cancer Classification

Ghadeer JM Mahdi [1,*], Bayda A. Kalaf [1], Mundher A. Khaleel [2]
[1]Department of Mathematics, College of education for Pure Sciences- ibn Al-Haitham, University of Baghdad, Iraq
[2]Department of Mathematics, Faculty of Computer Science and Mathematics, University of Tikrit, Iraq

Received: 11/10/2020 Accepted: 20/11/2020

Abstract

In this paper, a new hybridization of supervised principal component analysis (SPCA) and stochastic gradient descent techniques is proposed, and called as SGD-SPCA, for real large datasets that have a small number of samples in high dimensional space. SGD-SPCA is proposed to become an important tool that can be used to diagnose and treat cancer accurately. When we have large datasets that require many parameters, SGD-SPCA is an excellent method, and it can easily update the parameters when a new observation shows up. Two cancer datasets are used, the first is for Leukemia and the second is for small round blue cell tumors. Also, simulation datasets are used to compare principal component analysis (PCA), SPCA, and SGD-SPCA. The results show that SGD-SPCA is more efficient than other existing methods.

Keywords: classification, cancer diagnostic, Hilbert-Schmid, stochastic gradient descent, principal component analysis.

تحسين طريقة تحليل العنصر الرئيسي لتصنيف السرطان

غدير جاسم محمد مهدي[1]*، بيداء عطية خلف[1]، منذر عبد الله خليل[2]
[1]قسم الرياضيات ، كلية التربية للعلوم الصرفة ، ابن الهيثم، جامعة بغداد، العراق
[2]قسم الرياضيات ، جامعة تكريت ، جامعة تكريت، العراق

الخلاصة:

في هذا البحث، تم اقتراح طريقة جديد لتحليل المكون الرئيسي الخاضع للإشراف (SPCA) وتقنيات الانحدار العشوائي يسمى SGD−SPCA لمجموعات البيانات الكبيرة الحقيقية التي تحتوي على عدد صغير من العينات في مساحة عالية الأبعاد. تصبح SGD−SPCA أداة مهمة يمكن استخدامها لتشخيص وعلاج السرطان بدقة. عندما يكون لدينا مجموعات بيانات كبيرة تتطلب العديد من المعلمات، فإن SGD−SPCA هي طريقة رائعة، ويمكنها بسهولة تحديث المعلمات عند اضافة عينات جديدة. تم استخدام مجموعتين من بيانات للسرطان: النوع الأول لوكيميا الدم والنوع الثاني أورام الخلايا الزرقاء الصغيرة المستديرة. بالإضافة إلى ذلك، تُستخدم مجموعات بيانات المحاكاة لمقارنة تحليل المكونات الرئيسية (PCA) و SPCA و(SGD−SPCA) . تظهر النتائج أن−SGD SPCAأكثر كفاءة من الطرق الأخرى الموجودة.

*Email: mahdighadeer@gmail.com

1. Introduction

Cancer classification is one of the main research areas in the biostatistics field. Usually, cancer datasets have a small number of samples and high dimensions. Most variables (genes) are not related to cancer classification, so dimension reduction and variable selection are required. The fundamental problems of cancer diagnosis and treatment can be explained using gene expression data. Accurate prediction of cancer type has a great appreciation in providing a better treatment on patients [1]. Morphological and clinical-based methods were always used for classification, but the diagnostic ability for these methods was reported to have many limitations/constraints [2].

Many statistical classification methods have been applied to cancer classification, but not all of them are efficient methods. Some of them have limited diagnostic abilities [3]. Due to some points, cancer classification methods are non-trivial tasks. Firstly, the gene expression data has a very high dimensionality, and it usually contains thousands of genes. The second point is that the data size is small, where some sets have less than one hundred genes. In the third point, most genes are insignificant to characterizing cancer. The statistical methods that have been used previously are not designed to deal efficiently with this kind of datasets [4]. In the area of statistics and machine learning, classification problems have been broadly studied. In the past, many classification algorithms had been proposed, such as Naïve Bayes [5], linear decrement analysis [6], neural network [7], Decision Tree [8], support vector machine [9], k-nearest neighbor [10], etc. For most of these algorithms, the authors did not pay more attention to the time, and they were only concerned with classification accuracy. In reality, due to high dimensionality, many classification methods are computationally expensive and not accurate [11]. Depending on the dataset structure, an appropriate classification method can be implemented. For example, the number of individuals, variables, and the type of data; i.e., whether variables are quantitative or qualitative [12]. This work deals with quantitative variables where the number of variables is much larger than the number of individuals; hence the PCA is a suitable analysis method for dimensionality reduction.

In this work, a modified SPCA is presented by using stochastic gradient descent techniques for cancer classification. It is applied to two different types of cancer datasets which are small round blue cell tumors [13] and leukemia [14] datasets.

The paper is organized as follows. In sections 2 and 3, PCA and SPCA are explained, respectively. The process of using different kernels SPCA is discussed in section 4. Section 5 introduces the enhanced SPCA using Stochastic Gradient Descent. Simulation studies and experimental results are given in sections 6 and 7, respectively. In section 8, the discriminatory selection feature is discussed. Section 9 presents the discussion of the results.

All computations conducted for this article are run in R on a single processor, without any distributed computing.

2. Principal Component Analysis

Let $D = \{(x_i, y_i)\}_{i=1}^{n}$ be a set of data points with dimensionality $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, and small size n; $d \gg n$. We define the input data points as $X = [x_1, \dots, x_n] \in \mathbb{R}^{d \times n}$ and the observations (labels) as $Y = [y_1, \dots, y_n] \in \mathbb{R}^{1 \times n}$. PCA is a method for transforming the data from high dimensional space, d, to low dimensional space, k, where $k \ll d$. In other words, PCA searches out to replace the set of d input variables, $x_1, x_2, \dots, x_d$, which are unordered and correlated variables (original data space), by a group of k linear projections, $u_1, u_2, \dots, u_k$, which are ordered and uncorrelated (component space).

The new variables, $p_1, p_2, \dots, p_k$, that form a new coordinate system are called principal components (PCs). Usually, there are at most d PCs because they are orthogonal linear transformations of the original variables. Still, not all the d PCs have used a subset of k PCs. To approximate the space spanned by the actual data points $X = [x_1, x_2, \dots, x_n]^t$, a set of k PCs, $\{u_1, u_2, \dots, u_k\}$, are chosen, based on the percentage of the variance of the actual data. u_1 is called the first PCs and has the highest variance in the data, and hence it is the most significant PCs. u_2 is called the second PCs and has the second-highest variance in the data, and so on until u_k that has the minimum variance in the data [11].

A matrix $U \in \mathbb{R}^{d\times n}$ can be constructed using a set of principals $u_1, u_2, \dots, u_k$, so an observation x can be projected onto the column space of $U = [u_1, u_2, \dots, u_k]$. The projection of x onto U can be seen as a linear system of equations, i.e.,

$$\hat{x} = U\beta \tag{1}$$

where $\hat{x} \in \mathbb{R}^d$ and $\beta \in \mathbb{R}^k$ are unknown parameters. Eq.1 is a linear system, and it has an exact solution, $x = \hat{x}$, if x lies in the column space of U ($col(U)$), or $span\{u_1, u_2, \dots, u_k\}$. Otherwise, there is no solution for Eq.1, then it should be solved for projection onto $col(U)$ or $span\{u_1, u_2, \dots, u_k\}$, and then its reconstruction. Let us define the difference between x and $\hat{x}$ to be the residual ($r = x - \hat{x}$). The value of r needs to be small, which can be gained when r is orthogonal to $col(U)$. Hence,

$$r \perp U \rightarrow x - \hat{x} \perp U \rightarrow x - U\beta \perp U \rightarrow U^T(x - U\beta) = 0$$

Therefore,

$$\beta = (U^T U)^{-1} U^T x \tag{2}$$

Since U is a set of orthonormal vectors, then $U^T = U^{-1}$ and $U^T U = I$, so form Eq.1 and 2:

$$\hat{x} = U(U^T U)^{-1} U^T x = U^T U x \tag{3}$$

For n projected data points, $\{\hat{x}\}_{i=1}^n$:

$$\sum_{i=1}^{n} \|\hat{x}_i\|_2^2 = \sum_{i=1}^{n} \hat{x}_i \hat{x}_i^T = \sum_{i=1}^{n} u^T x_i x_i^T u = u^T \sum_{i=1}^{n} x_i x_i^T \, u = u^T S u \tag{4}$$

where $S = X^T X$ is a covariance matrix and $u^T S u$ is the variance of the projected data points onto the PCA subspace. Our goal is to find a projection direction u that maximizes the variance of projection (squared length of reconstruction), i.e.,

$$maximize \; u^T S \, u, \text{subject to } u^T u = 1 \tag{5}$$

Using the Lagrange multiplier conversion, it follows that:

$$L(u, \lambda) = u^t S u - \lambda(u^t S u - 1)$$

where λ is constant. By taking the derivative and setting it to be equal to zero, we get $2Su - 2\lambda u = 0$; Consequently, $Su = \lambda u$, where λ is the eigenvalue of the sample covariance matrix S and u is the corresponding eigenvector. i.e., $u^t S u = u^t \lambda u = \lambda u^t u = \lambda$, where $u^t u = 1$. As a result, the total data variance can be composed by $\sum_{i=1}^{n} Var(u_i) = \sum_{i=1}^{n} \lambda_i = Trace(S)$.

$Var(u_i)$ is maximized if λ_i is the maximum eigenvalue of S, and the first principal component is the corresponding eigenvector. In general, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ such that $Var(u_1) \geq Var(u_2) \geq \cdots \geq Var(u_n)$. The ratio $R = \frac{\lambda_{(k+1)} + \cdots + \lambda_n}{\lambda_1 + \cdots + \lambda_n}$ is a goodness-of-fit that measures how $u_1, u_2, \dots, u_k$ represent the n original variables lower-dimensional space. R should be small, and a large proportion of the total variation in X is explained by the first principal component [7].

3. Supervised Principal Component Analysis

As stated in the discussion in the previous section, PCA finds the direction of maximum variation of n-dimensional space; this can be used as a reduction and pre-processing operation for classification. PCA is an unsupervised classifier, unlike Fisher Discernment Analysis (FDA) [7]; however, SPCA is a generalized method of PCA. SPCA has some advantages over FDA, and it can use label information for classification tasks. The sequence of principle components that have the maximum dependency on the response variable can be estimated using the SPCA [3].

Suppose that $\{x_i, y_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^k, y_i \in \mathbb{R}^l$. y_i is not restricted to binary classes, therefore it is not required that y_i has only discrete values, and hence the model can be used for regression as well. In regular PCA, a lower-dimensional subspace $S = U^T X$ has been looked at, where matrix X is the covariate matrix and U is an orthogonal projection matrix. However, in SPCA, the projection matrix U should be determined where $P(Y|X) = P(Y|U^T X)$, which means a subspace that contains approximately the same information as the original. Between the original covariate X and Y, the predictive information must exist. If X and Y are entirely independent, the regression or classification could not be processed.

Using $S_j = X_j^T\, Y / \sqrt{X_j^T\, X}$, the steps of SPCA can be achieved as follows; first, the standard regression coefficients for each j can be computed. Then, corresponding to all the columns, the data matrix X_0 is reduced where $|S_j| > \theta$, and θ can be found by cross-validation [3]. Now, for the reduced data matrix, X_0, compute the first principal component which can be used in a regression model or a classification algorithm to produce the outcome. SPCA is consistent, unlike standard PCA; PCA takes different directions for the component as the number of data points increases [15]. The modified SPCA can be derived using the Hilbert-Schmidt independence criterion that is discussed below.

3.1. Kernel Supervised PCA

The linear projection for PCA might not be completely effective when the data points exist in nonlinear space. Two options are applied to handle this problem. First, PCA should be changed to be a nonlinear method. Second, the data points should be changed to fall on a linear or close to linear subspace. The second solution can be achieved by mapping the data points to space with higher dimensionality, hoping that it falls on a linear manifold.

To find a linear transformation U such that $U^T X$ has maximum dependence of Y, a linear kernel on $U^T X$ and a kernel over Y (call it B) can be made. We attempt to find U that maximizes $\frac{1}{(n-1)^2} Tr(KHBH) = \frac{1}{(n-1)^2} Tr(X^T UU^T XHBH)$ which is objective to $max_U Tr(X^T UU^T XHBH)$. It implies that $max_U Tr(U^T XHBHX^T U)$ adds a constraint $U^T U = I$, where U will be the top k eigenvectors of $XHBHX^T$. Notice that, if $B = I$ is chosen, then $XHBHX^T = XHHX^T = (X - \bar{X})(X - \bar{X})^T$, which is the covariance matrix of X, and hence it can be concluded that PCA is a special case from SPCA [16]. The possible kernel functions that can be chosen are the linear kernel, polynomial kernel, Gaussian kernel, and delta kernel. The following algorithm shows the necessary steps for the SPCA.

Algorithm 1: SPCA

Input: X, n, K, K_{test}, L, the testing data, data size, kernel matrices of training, testing datasets, and target variable, respectively.

Output: Reduced dimensional datasets.

1. Compute $H = I - \frac{1}{n} ee^T$
2. $Q = KHLHK$
3. Compute basis: $B \leftarrow$ generalized eigenvectors of (Q, K) corresponding to the top d eigenvalues.
4. Evaluate training data: $Z \leftarrow B^T[\Phi(X)^T \Phi(X)] = B^T K$
5. Evaluate test example: $z \leftarrow B^T[\Phi(X)^T \Phi(X) = B^T K_{test}$

3.2. Hilbert-Schmidt Independence Criterion (HSIC)

HSIC was introduced by Gretton [17]. It is an independent criterion that measures the independence of variables X and Y. Barshon used it for a supervised PCA technique [3]. If $z = \{(x_1, y_1), \dots, (x_n, y_n)\} \in X \times Y$ is a series of n independent observations drawn from $P_{(X,Y)}(X, Y)$, then we calculate $(n-1)^{-2} Tr(KHBH)$ as an estimator of HSIC, where $H, K, B \in \mathbb{R}^{n \times n}$, $K_{ij} = k(x_i, x_j), B = b(y_i, y_j)$, and $H = I - \mathrm{n}^{-1} Xee^T$, such that k and b are positive semidefinite function and $e = [1 \dots 1]^T$. By subtracting the mean of each row, then XH centralized version of X, i.e. $XH = X(I - n^{-1} ee^T) = X - n^{-1} Xee^T$, where each entry in row i of $n^{-1} Xee^T$ is the mean of i^{th} row of X. The idea is based on the useful features that show the maximized independence between two distributions [12]. Measuring the independence between two distributions can be performed using different techniques. In general, two distributions are different if their means are different, but if the two means are the same, then the second moment of these distributions needs to be checked. Now, by calculating the difference between $E(\Phi(X))$ and $E(\Phi(Y))$, we can find out whether the two random variables X and Y have the same distribution or not, i.e. X and Y have the same distribution if $\|E(\Phi(X)) - E(\Phi(Y))\|^2$ is equal to 0.

5. Stochastic Gradient Descent SPCA (SGD-SPCA)

The goal is to find the directions in which the variance, $\mathbb{E}(XX^T)$, is maximum. Let u be denoted as the unit vector direction along which the variance is maximum. The variance along this direction is given by:

$$\sigma_{u^2}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i.u)^2 = \frac{1}{n}(xu)^T(xu) = \frac{1}{n}u^T x^T xu = u^T vu, where\ v = \frac{1}{n}x^T x$$

Evaluate the gradients $w.r.t.$ to u, then it follows that:

$$L(u,\lambda) = \sigma_u^2 - \lambda(u^T u - 1),$$

and set it to zero to find the optimum values.

$$\frac{\partial L}{\partial \lambda} = u^T u - 1 \ \& \ \frac{\partial L}{\partial u} = 2vu - 2\lambda u$$

We solve the optimization problem whose objective function is given in the following equation:

$$minimize_{U,\beta}\sum_{i=1}^{n} l(y_i, Ux_i, \beta) + \lambda\|X - XU^T U\|_F^2\ s.t.\ UU^T = I_k \qquad (10)$$

where $l(.)$ is a loss function, $X_{n\times p}$ is data matrix, $Y_{n\times q}$ are dependent variables, U is the basis for the learned subspace, β is learned coefficient for prediction, and $\lambda > 0$ is a trade-off parameter. If we consider the case where $l(y_i, Ux_i, \beta) = \|y_i - \beta^T Ux_i\|_F^2$ is the Frobenius error loss, then Eq.10 becomes

$$minimize_{U,\beta}\ \|Y - XU^T\beta\|_F^2 + \lambda\|X - XU^T U\|_F^2\ s.t.\ UU^T = I_k \qquad (11)$$

The derivative of the objective function is

$$\frac{\partial f}{\partial U} = -2(XU^T)^{+}YY^T P_{XU^T}^{\perp} X - 2\lambda UX^T X \qquad (12)$$

where $P_{XU^T}^{\perp}$ is the projection matrix onto the orthogonal complement of the span of XU^T. The retraction is the key notation for applying the SGD-based optimization method to the manifold optimization [18]. A step in the direction of the negative gradient is taken for a retraction. For moving between two points on the manifold, if a closed-form expression is available, the updated step can be taken directly, which is called the geodesic step [19]. Edelman gave an expression for the geodesic step, as follows:

$$L_{t+1} = L_t^T\ V_t \cos(\mu_t \Sigma_t)\ V_t^T\ + V_t \sin(\mu_t \Sigma_t) V_t^T \qquad (13)$$

where Armijo backtracking line search chooses the step size μ_t [20]. The processes of SGD-SPCA at each iteration can be summarized as follows:

i. At the current iteration, calculate the Euclidean gradient.

ii. Obtain the Riemannian gradient by projecting the negative Euclidean onto the tangent space.

iii. For the resulting matrix, compute the singular value decomposition.

vi. With Armijo line search, update *L* by taking a geodesic step.

The SGD-SPCA can be summarized in the following algorithm:

Algorithm 2: SGD-SPCA

Input: $X, Y, k, \lambda,$ and S_0 are initializations that are generated by PCA

Output: R is a radiused dataset

1. $i = 0$
2. While $\|grad(S_i)\|_F < \epsilon$
3. Calculate the gradient: $\nabla_i = \left.\frac{\partial f}{\partial S}\right]_{S=S_i}$ using Eq.12
4. Calculate: $grad(S_i)^T = \left(I_p - S_i^T S_i\right)\nabla_i^T$
5. Compute the SVD: $U_i, \Sigma_i, V_i = SVD(-grad(S_i)^T)$
6. Update S_i^T using Eq.13
7. $i = i + 1$
8. End $While$
9. Generate the reduced data: $R = XS_i^T$

6. Simulation Studies

The performance for the methods (PCA, SPCA, and SGD-SPCA) was checked using three simulation datasets, generated using the same developed model used in Bair [1]. They considered 500 variables (genes) and three different values of individuals (50, 100, 200). The response was designed to have two classes. A comparison between PCA, SPCA, and SGD-PCA is presented in Table 1. As can be shown, SGD-SPCA performs better than SPCA and PCA. For instance, in the 200 samples dataset, the first principal component (PC1) captures 62.56% of the variation using SPCA, whereas it captures only 48.51% of the variation using PCA.

Table 1- Comparison of PCA, SPCA, and SGD-SPCA based on the percentage of variation for each principal component.

Dataset size	Principal components	PCA	SPCA	SGD-SPCA
50	PC1	38.21	40.38	**50.56**
	PC2	8.34	10.93	**12.41**
	PC3	7.56	8.21	**8.38**
100	PC1	48.51	55.38	**58.56**
	PC2	9.74	11.93	**15.41**
	PC3	5.55	7.21	**9.38**
200	PC1	48.51	54.38	**62.56**
	PC2	9.73	9.93	**11.41**
	PC3	7.65	6.21	**9.38**

The sensitivity (also called the true positive rate) and the specificity (also called the true negative rate) were checked. The best scenario of the ROC curve exists when the area under the curve (AUC) is equal to 1. PCA, SPCA, and SGD-PCA were applied to the simulation datasets with 200 samples. Figure-1 shows that SGD-SPCA satisfies the best scenario with 82.9% AUC. It indicates that SGD-SPCA provides the best classification.

7. Experimental results

SPCA and SGD-SPCA were applied on two real datasets, which are Leukemia and SRBCT datasets, downloaded from the UCI Machine Learning repository (*https://archive.ics.uci.edu/ml/index.php*). In the following two sections, a brief description for each dataset is outlined.

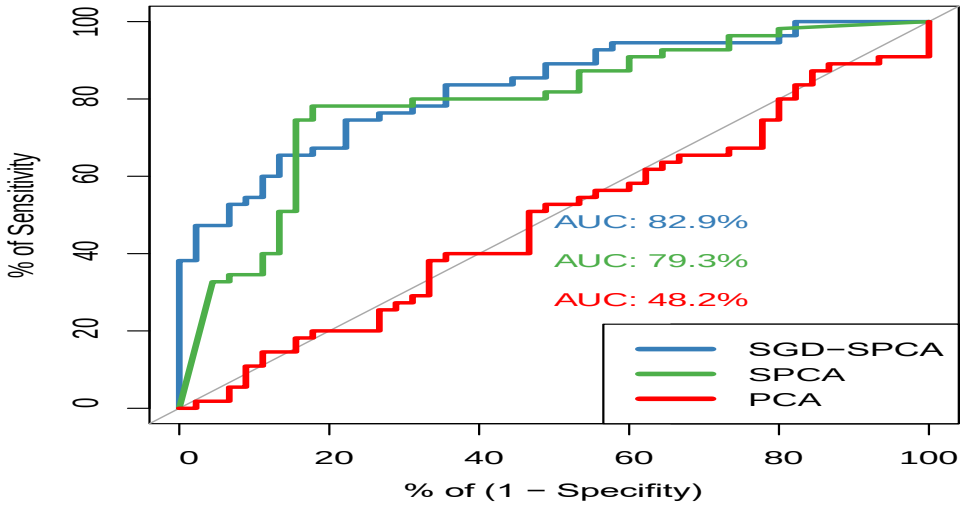


Figure 1-ROC curve for PCA, SPCA, and SGD-SPCA.

7.1 Leukemia datasets

Blood and bone marrow cells can be affected by blood cancers, which can change their efficiency and

behavior. The three types of blood cancer, namely leukemia, lymphoma, and myeloma, seriously damage the circulatory and lymphatic systems. They are classified into different types which affect different types of white blood cells. Part of this work focuses on leukemia, where patients generate high numbers of abnormal white blood cells that are not functional and hence cannot fight infections. Based on the impacts and growth of white blood cells, leukemia is divided into four types, which are Acute Lymphocytic Leukemia (ALL), Acute Myeloid Leukemia (AML), Chronic Lymphocytic Leukemia (CLL), and Chronic Myeloid Leukemia (CML) [21]. The gene expression dataset that was analyzed in this paper includes data from AML and ALL patients, published by Golub [14]. The data are derived from a proof-of-concept study, and it shows how the gene expression monitoring (via a DNA microarray) can classify the new cases of cancer, providing a common approach for assigning tumors to known classes and identifying new cancer classes. Using this type of dataset, patients were classified into AML and ALL categories. The complete leukemia dataset has 3051 genes and 72 observations. The low number of observations does not allow much flexibility for supervised methods, given the need to split the dataset into training and testing parts. The raw dataset was processed by following the following steps, based on the original paper [22]; (a) thresholding, with floor of 100 and ceiling of 16000; (b) filtering, with exclusion of genes with $max/min \leq 5$ or $\max - min \leq 500$, where max and min refer to the maximum and minimum intensities for a particular gene across the 72 samples; (c) base 10 logarithmic transformation; and (d) scaling the data observation wise. The dataset was split into 50 training samples (17 AML, 33 ALL) and 22 testing samples (8 AML, 14 ALL).

7.2 SRBCT dataset

SRBCT dataset contains the gene expression of 83 observations (patients) with 2308 variables (genes). The correct clinical diagnosis is extremely challenging for the four different childhood tumors because of the similar appearance on routine histology. The tumor types include the Ewing family (EWS), rhabdomyosarcoma (RMS), neuroblastoma (NB), and Burkitt lymphomas (BL). In this paper, the distinction between these four tumors is achieved based on gene expression values. The dataset was split into 63 training samples (23 EWS, 20 RMS, 12 NB, and 8 BL) and 20 testing samples (6 EWS, 5 RMS, 6 NB, 3 BL).

8. Selection of Discriminatory Features

Working with large datasets faces many difficulties, such as time consumption and inefficient results. To analyze leukemia and SRBCT datasets, we selected the most significant genes for cancer type; in other words, the genes that are differentially expressed across classes. The HSIC process (section 3.2) was used for leukemia datasets, which demonstrated only 329 genes as significant. As could be observed from Figure-2, only few genes are apparently interesting. The active genes are colored in black and supplementary genes are colored in gray.

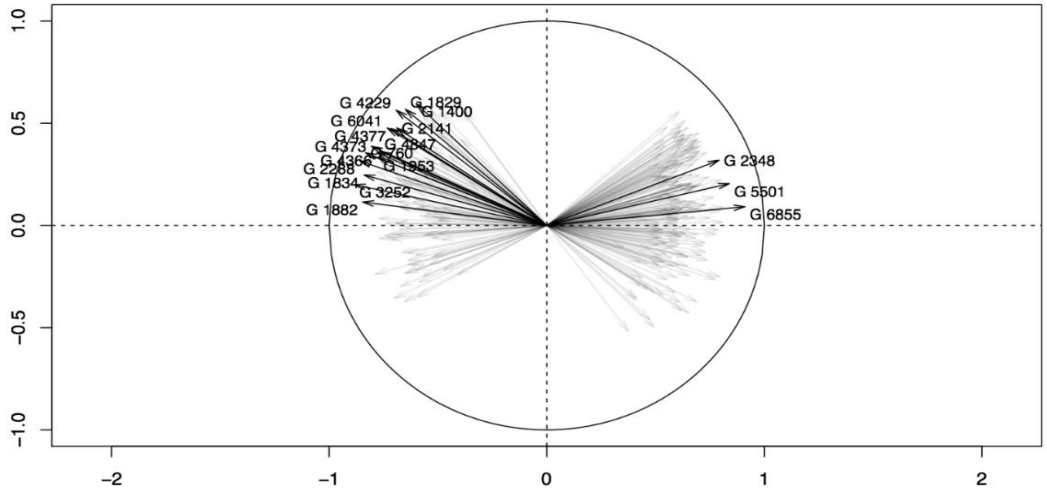


Figure 2-The most significant genes in leukemia dataset; the significant genes are marked with black and the nonsignificant genes are marked with gray.

8.1 Modified t-test

In this section, a modified t-test is used to select the most significant genes in the SRBCT data set. The common t-test was proposed by Welch. It is used to measure the difference between two groups of samples. Based on Eq.14, t-test calculates a score, t_i, that represents gene i.

$$t_i = \sqrt{\frac{(\bar{x}_{i1} - \bar{x}_{i2}) n_1 n_2}{\left[s_{i1}^2(n_1 - 1) + s_{i2}^2(n_2 - 1)\right](n_1 + n_2)}} \tag{14}$$

where $S_{ik}^2 = \frac{1}{n_k}\sum_{i=1}^{n_k}(x_{ik} - \bar{x}_{ik})^2,\ k = 1, 2.$ (15)

Here, $\bar{x}_{i1}$ and $\bar{x}_{i2}$ are the mean expression values for a gene in two different classes. n_1 and n_2 denote the number of samples. There are two limitations to the usage of t-test. First, t-test solves problems with only two classes. Second, from Eq.15, if the mean of the two classes is equal, the value of $t_i = 0$, and then the gene i will be removed as an irrelevant gene, whereas it might be able to provide classification information for samples.

t-test was modified to overcome the abovementioned two problems, as follows:

$$t_i = \max\{\frac{1}{2}\left|\frac{\bar{x}_{ik} - \bar{x}_i}{\sqrt{\frac{1}{n_k} + \frac{1}{n}} \cdot s_i}\right| + \frac{1}{2}\ln\left(\frac{s_{ik}^2 + s_i^2}{2 s_{ik} s_i}\right), k = 1, 2, \dots, K\} \tag{16}$$

where $\bar{x}_{ik} = \sum_{j \in c_k} \frac{\bar{x}_{ij}}{n_k}$, $\bar{x}_i = \sum_{j=1}^{n} \frac{x_{ij}}{n}$, and $s_i = \sqrt{\frac{1}{n-K}\sum_k \sum_{j \in C_k}\left(x_{ij} - \bar{x}_{ik}\right)^2}$.

K and n refer to the number of classes and samples, respectively. Class k, that includes n_k samples, is denoted by c_k. s_i is the pooled within-class standard deviation for gene i. $\bar{x}_i$ is the mean expression value for gene i, $\bar{x}_{ik}$ is the mean expression value for gene i in class k, and $\bar{x}_{ij}$ is the mean expression value for gene i in sample j. Eq.16 is used to calculate t_i, which is the score for each gene. The genes with high scores were selected for further processing because they are more relevant to the classification. 208 genes were determined as essential genes in the SRBCT dataset. The most significant 7 genes in the SRBTC dataset are listed with their descriptions in Table-2. Figure-3 illustrates the correlation between the selected genes for training data in Leukemia and SRBCT datasets. As can be seen, some genes are highly correlated.

Table 2-Description of some selected genes for SRBCT dataset

Gene ID	Gene Expression Product
G 770394	Fc fragment of IgG, receptor, transporter, alpha
G 377461	caveolin 1, caveolae protein, 22kD
G 796258	sarcoglycan, alpha
G 784224	fibroblast growth factor receptor 4
G 325182	cadherin 2, N-cadherin
G 812105	transmembrane protein
G 241412	E74-like factor 1

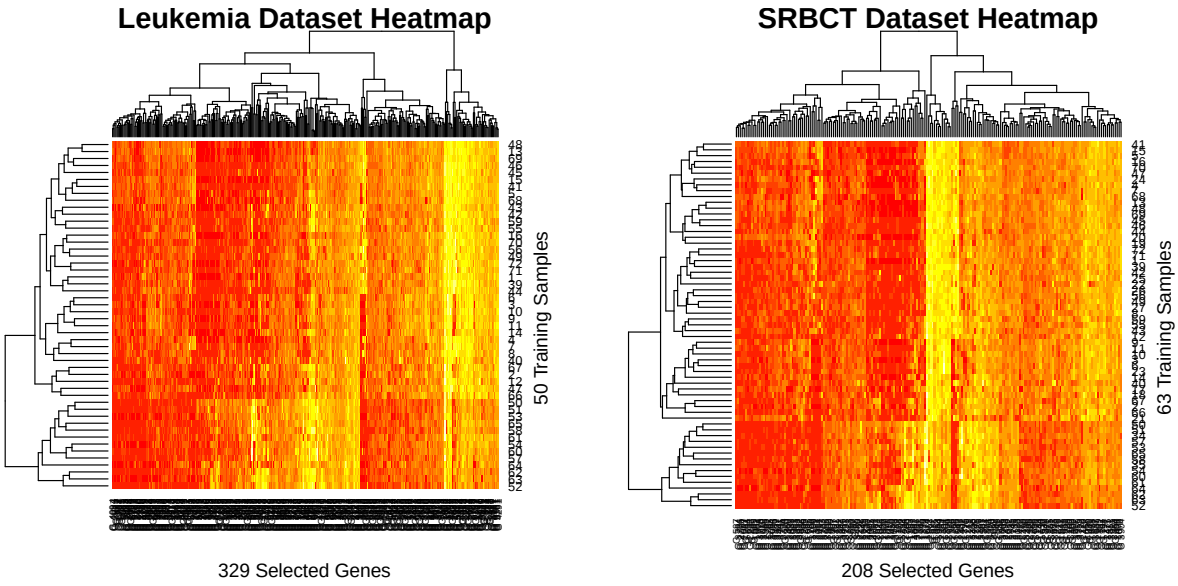
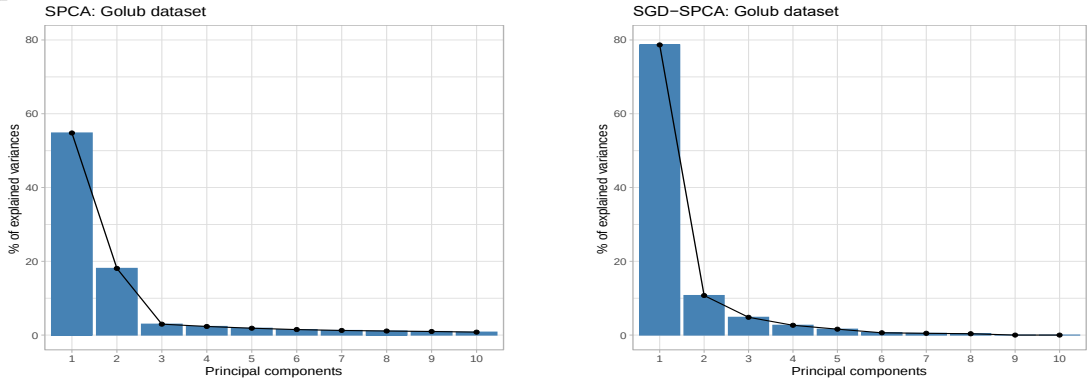


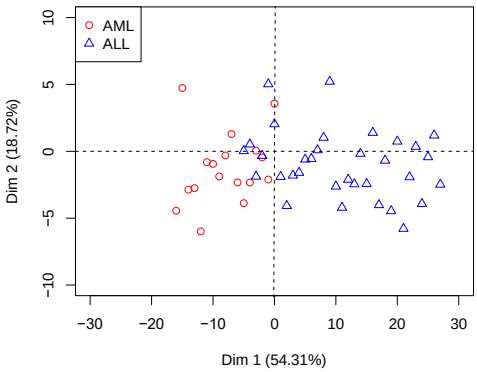
Figure 3-Heatmaps of selected genes for training data in leukemia and SRBCT datasets.

9. Results and Discussion

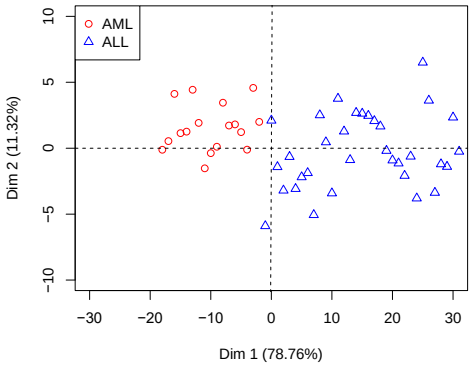
The SPCA and SGD-SPCA were applied to two reduced datasets. The top two plots in Figures- 4 and 5 show the first 10 principal components with their percentage of variation. SGD-SPCA performs better than SPCA for leukemia and SRBCT datasets. Now, how well do the principal components separate the classes? Graphically, from Figures-4 and 5, it can be agreed the SPCA and SGD-SPCA both performed excellently. The individuals in a 2-dimensional plane can be visualized. By looking at the individual plot, we can predict the class of the observation. Support Vector Machine (SVM) with Gaussian kernel is used to classify classes in leukemia and SRBCT datasets. By testing our two test datasets using Figure-4.d and Figure-5.d, the results are shown in Figure-6. It can be seen that almost all the testing individuals lie in the right class. Table-3 summarizes the classification accuracy for training and testing in both datasets. With only two principal components, we obtained above 90% accuracy, while a value of around 95% could be achieved with four principal components. Comparisons among Naive Bayes, KNN, Decision Tree, and SGD-PCA for both Leukemia and SRBCT datasets are given in Table-4. $93-94\%$ accuracy is gained using only 4 principal components, and the total real-time shows that the SGD-SPCA is the fastest compared with the other methods.



a. Principal components in SPCA b. Principal components in SGD-PCA

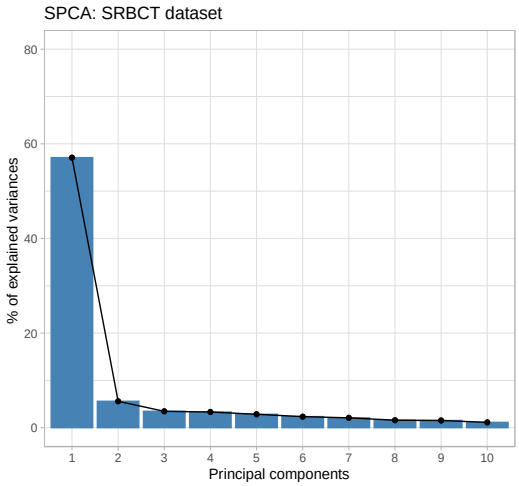


c. SPCA

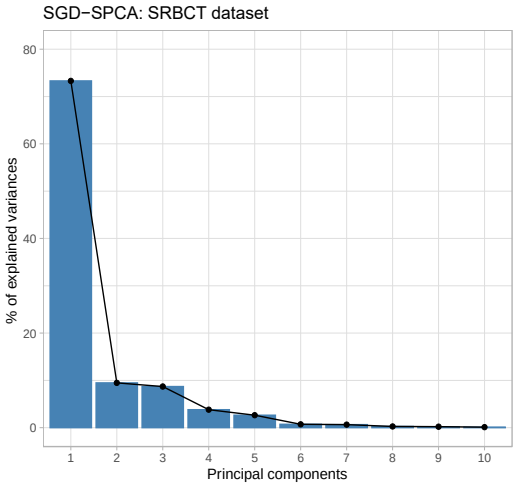


d. SGD-SPCA

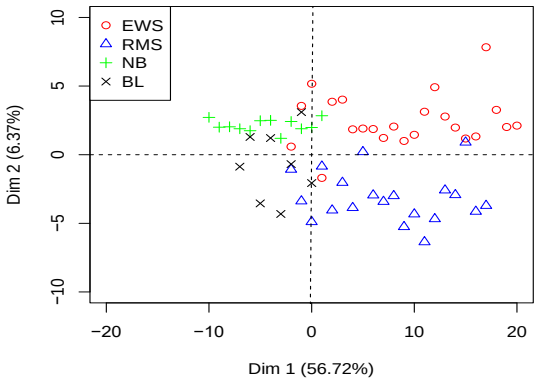
Figure 4-SPCA and SGD-PCA for leukemia dataset.



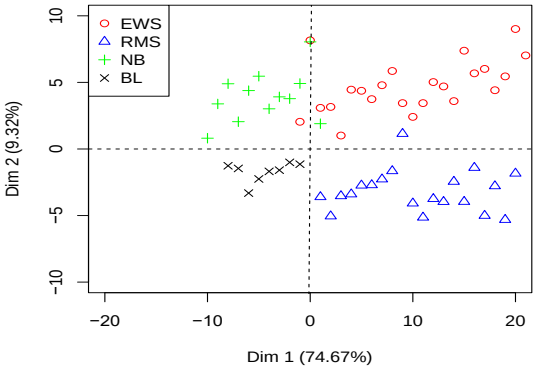
a. Principal components in SPCA



b. Principal components in SGD-PCA

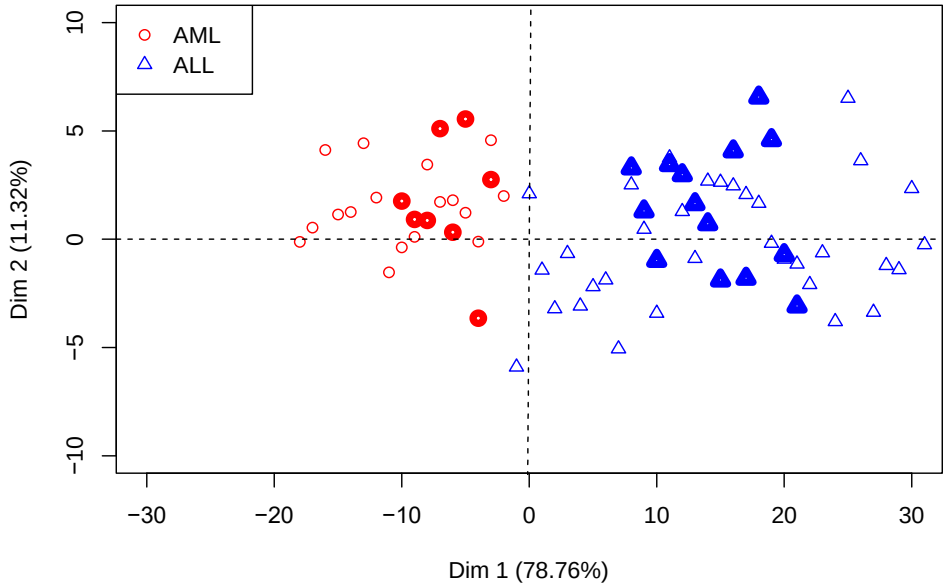


c. SPCA

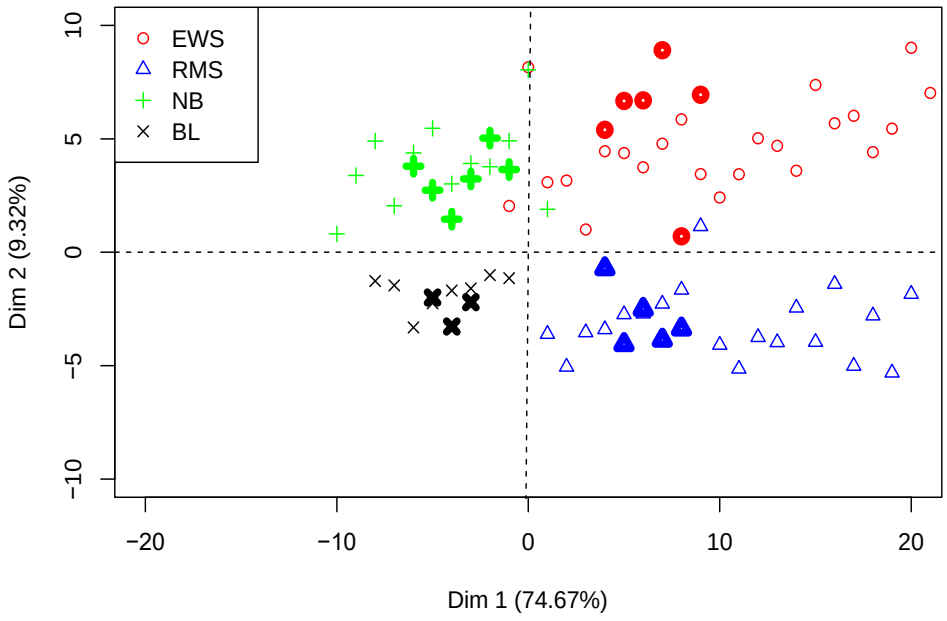


d. SGD-SPCA

Figure 5-SPCA and SGD-PCA for SRBCT dataset.



a. Leukemia dataset



b. SRBCT dataset

Figure 6-Training and testing individuals for leukemia and SRBCT datasets.

Table 3-Classification accuracy for training and testing in both leukemia and SRBCT datasets

Method	Number of PCs	**Leukemia dataset**		**SRBCT dataset**	
		Training	Testing	Training	Testing
SGD-PCA	2	96.60%	97.34%	95.71%	94.82%
	3	98.34%	96.73%	97.69%	96.53%
	4	**99.3%**	**99.1%**	**99.2%**	**99.7%**

Table 4- Comparisons among Naive Bayes, KNN, Decision Tree, and SGD-PCA for both Leukemia and SRBCT datasets

Method	**Leukemia dataset**			**SRBCT dataset**		
	Training	Testing	Total Time	Training	Testing	Total Time
Naive Bayes	84.60%	83.42%	3m 42s	83.94%	82.62%	4m 2s
KNN	86.80%	91.55%	2m 21s	85.33%	84.55%	3m 10s
Decision Tree	87.10%	90.39%	4m 33s	87.10%	89.39%	5m 43s
SGD-PCA	**95.5%**	**94.4%**	**1m 19s**	**95.2%**	**94.7%**	**1m 7s**

Conclusions

The present work proposed a new SGD-SPCA method for reducing the dimensionality of large real cancer datasets. Stochastic gradient descent was used to modify the SPCA techniques. The experimental result show accuracy values between 93 and 94 percent using four principal components for both leukemia and SRBCT datasets. A comparison between the modified and some other existing methods proves that SGD-PCA satisfies the criteria of best accuracy and less time.

References

1. Bair, E., Hastie, T., Paul, D., & Tibshirani, R. **2006** . Prediction by supervised principal components. *Journal of the American Statistical Association*, **101**(473): 119–137.
2. Reis-Filho, J. S., & Pusztai, L. **2011**. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet*, **378**(9805): 1812–1823.
3. Barshan, E., Ghodsi, A., Azimifar, Z., & Jahromi, M. Z. **2011**. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, **44**(7): 1357–1371.
4. Arenas-Garcia, J., Petersen, K. B., Camps-Valls, G., & Hansen, L. K. **2013**. Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods. *IEEE Signal Processing Magazine*, **30**(4): 16–29
5. Xu, S., **2018**. Bayesian Naïve Bayes classifiers to text classification. Journal of Information Science, **44**(1): 48-59.
6. Devi, R. G., & Sumanjani, P. **2015.** Improved classification techniques by combining KNN and Random Forest with Naive Bayesian classifier. 2015 IEEE International Conference on Engineering and Technology (ICETECH), 1-4.
7. Al-RawiK. R., & AL-RawiS. K. **2020**. Smart Doctor: Performance of Supervised ART-I Artificial Neural Network for Breast Cancer Diagnoses. *Iraqi Journal of Science*, **61**(9): 2385-2394. https://doi.org/10.24996/ijs.2020.61.9.25
8. Zahedi, J., & Rounaghi, M. M. **2015**. Application of artificial neural network models and principal component analysis method in predicting stock prices on Tehran Stock Exchange. *Physica A: Statistical Mechanics and Its Applications*, **438**: 178-187.
9. Mahdi G.J. **2020.** A Modified Support Vector Machine Classifiers Using Stochastic Gradient Descent with Application to Leukemia Cancer Type Dataset. *BSJ*. **17**(4): 1255-1266. DOI: 10.21123/ bsj. 2020.17.4.1255
10. Battineni, G., Chintalapudi, N. and Amenta, F., **2019**. Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM). Informatics in Medicine

Unlocked, 16, p.100200.

11. Chen, Y., & Hao, Y. **2017**. A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. *Expert Systems with Applications*, **80**:340–355.

12. Piironen, J., & Vehtari, A. **2017**. Iterative supervised principal components. *ArXiv Preprint ArXiv:1710.06229*.

13. Mahdi, G. J. M., Chakraborty, A., Arnold, M. E., & Rebelo, A. G. **2019**. Efficient Bayesian modeling of large lattice data using spectral properties of Laplacian matrix. *Spatial Statistics*, **29**: 329–350.

14. Gregorio, A., Corrias, M. V, Castriconi, R., Dondero, A., Mosconi, M., Gambini, C., Moretta, A., Moretta, L., & Bottino, C. **2008**. Small round blue cell tumours: diagnostic and prognostic usefulness of the expression of B7-H3 surface molecule. *Histopathology*, **53**(1): 73–80.

15. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., & Caligiuri, M. A. **1999**. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439): 531–537.

16. Vandenrijt, J.-F., Lièvre, N., & Georges, M. P. **2014** . Improvement of defect detection in shearography by using principal component analysis. *Interferometry XVII: Techniques and Analysis*, *9203*, 92030L.

17. Arenas-Garcia, J., Petersen, K. B., Camps-Valls, G., & Hansen, L. K. **2013**. Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods. *IEEE Signal Processing Magazine*, **30**(4): 16–29.

18. Wang, T. and Li, W., **2018**. Kernel learning and optimization with Hilbert–Schmidt independence criterion. *International Journal of Machine Learning and Cybernetics*, **9**(10): 1707-1717.

19. Ali Y. H., & MedhatR. A. **2018**. Enhancement of Principal Component Analysis using Gaussian Blur Filter. *Iraqi Journal of Science*, **59**(3B): 1509-1517. Retrieved from http://scbaghdad.edu.iq /eijs/ index.php/eijs/article/view/402

20. Ritchie, A., Scott, C., Balzano, L., Kessler, D., & Sripada, C. S. **2019**. Supervised principal component analysis via manifold optimization. *Proceedings of 2019 IEEE Data Science Workshop (DSW)*.

21. Edelman, A., Arias, T. A., & Smith, S. T. **1998**. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, **20**(2): 303–353.

22. Zhang Q, Mahdi G, Tinker J, Chen H **2020**. A graph-based multi-sample test for identifying pathways associated with cancer progression. *Computational Biology and Chemistry*. **26**: 107285. DOI: 10.1016/j.compbiolchem.2020.107285.

23. Mei, Q., Zhang, H. and Liang, C., **2016**. A discriminative feature extraction approach for tumor classification using gene expression data. *Current Bioinformatics*, **11**(5): 561-570.