Two Novel Criteria for Feature Selection based on Sparse Principal Components Analysis

Shaohong Luo, Yingxin Lin, Tingwei Li

Abstract

To remove the redundant features in datasets and identify the most important features in breast cancer diagnosis, we propose two feature selection methods called Importance Criterion and PCs Criterion based on Sparse Principal Components Analysis for analyzing the important attributes in breast data diagnosis. Compared to standard PCA and other feature selection methods applied in the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, our method can have the highest classification accuracy on three classical classification algorithms with the fewest selected feature. Besides, rather than working as the black box, our methods can list the important features, which can be used to select more significant features for the breast cancer diagnosis.

1. Introduction

Breast cancer is one of the leading causes of death in women. It is also one of the types of cancer that are particularly important to diagnose and detect early. Traditionally, breast tumors are diagnosed through a full biopsy. A biopsy is done when mammograms, other imaging tests, or a physical exam shows a breast change that may be cancer. Fine needle aspirations (FNAs) provide a way to examine a small amount of tissue from the tumor[1]. By carefully examining the characteristics of individual cells and important background features, such as the size of cell clumps, physicians have been able to successfully diagnose using FNAs.

However, many different features are thought to be associated with malignancy, and the process remains highly subjective, depending on the skill and experience of the physician. To improve the speed, correctness, and objectivity of the diagnostic process, researchers have begun to machine learning techniques to discover breast cancer.

Feature selection is used to reduce irrelevant data and find the most relevant features that would increase classification accuracy. Many features selection like Correlation-based Feature Selection[2], Consistency-based Subset Evaluation[3], Information Gain[4], etc. methods has been proposed. Specifically, in medical domains, people have used feature selection to solve medical problems, such as ovarian cyst[5], liver disorder[6], breast cancer[7], etc. We review some novel feature selection methods aimed at cancer which are proposed in recent years. Aalaei et al.[8] proposed procedure that uses a GA-based feature selection with an artificial neural network (ANN). As the result, PS-classifier and GA-classifier select the best subset of features that produce the highest classification accuracy. Reddy et al.[9] used SVD-Entropy in Ovarian, lung, and breast cancer datasets to remove the redundant attributes in genes and predict the probability of cancer. Rao et al.[10] combined bee colony and gradient boosting decision tree to select feature. To verify their method, they applied it to six different cancer datasets and gain promising results.

Sparse PCA[11]–[15] is a good way for feature selection. Compared to standard PCA, one of the advantages of Sparse PCA is easier to be interpreted. Different from each principal component in standard PCA is the linear combination of all the original variables, many of the loadings of Sparse PCA are zero. The sparsity gives chance to filter out some important features from the dataset by studying the principal components of the Sparse PCA of the dataset. To extract the key features in the dataset, researchers have proposed several algorithms to control the non-zero loadings[16], [17]. Besides, the important features selection criteria have been discussed in some papers. For example, Gravuer et al.[18] used Sparse PCA to study a range of human, biogeographic, and biological influences on the invasion of Trifolium species into New Zealand. Specifically, Gravuer et al. [18] defined the absolute value of loading which is greater than 0.2 as the important feature. Then an Aggregated Boosted Trees model is applied to analyze the relationship between selected features and the raw features.

In this paper, based on the interpretability of Sparse PCA, we proposed two new criteria to analyze the principal components of Sparse PCA, which help to extract the important features. Due to various algorithms to solve Sparse PCA and each way can get different principal components, without loss of generality, this paper analyzed the result of four different Sparse PCA algorithms(i.e. SCotLASS[15], GPower[13], SPCA[11], and sPCA_rSVD[12]). After using our two criteria extracting the important features separately, in order to evaluate the importance of these features and the feasibility of our criteria, we fed the selected features to three classical classifiers (i.e. Logistics Regression, SVM, and Naïve Bayes) as the experimental group and got corresponding testing accuracy and f1 score. In the meanwhile, we fed the dataset after standard PCA and the original dataset into the same classifiers as the control group. The results showed that the accuracy and f1 of the experimental group are close to the control group, which can indicate that our criteria can filter out about 30% of features from the original dataset as important features and do not hurt the accuracy.

In the following, we briefly review standard PCA, LASSO, four different algorithms for Sparse PCA and introduce our two novel criteria for features selection in section 2. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset is introduced and experimental results with this dataset are given in section 3. Section 4 presents the findings and highlights the feasibility of our two novel criteria. The last section provides a discussion for future application. Some supplement tables and graphs are provided in the appendix.

2. Methodology

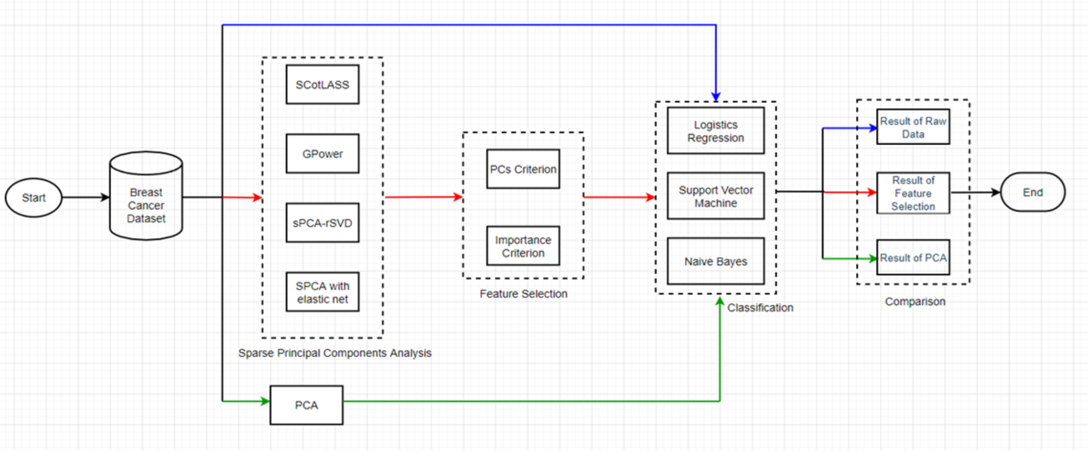

Figure 1 Process of the Research

The process of our research is shown in the figure. Firstly, we used four sparse PCA algorithms to process the dataset. Then our two feature selection criteria were applied to filter out some features. After that, As the experimental group, we put these new features into the three different classifiers to get the accuracy. In the meantime, the raw data and the data after standard PCA were fed into the same classifier as the control group. And the final step is a comparison and the result can indicate the feasibility of our two novel feature selection methods.

2.1 Dataset Description

In this project, the Wisconsin Diagnostic Breast Cancer (WDBC) dataset from UCI Machine Learning Repository is used. They have been collected at the University of Wisconsin–Madison Hospitals. There are 569 instances (62.74% benign, 37.26% malignant) that each record has thirty attributes in addition to patient id number and the binary diagnosis result (benign and malignant). Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nucleus present in the image. “The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image” [1], so there will be 30 features per image. The extracted features are as follows.

Table 1. Summary of Wisconsin Diagnostic Breast Cancer dataset[19]

Attribute Number	Attribute Description	Range		
		Mean	Standard error	Largest value
1	Radius	6.98 - 28.11	0.11- 2.87	7.93 - 36.04
2	Texture	9.71 - 39.28	0.36 - 4.89	12.02 - 49.54
3	Perimeter	43.79 - 188.50	0.76 - 21.98	50.41 - 251.20
4	Area	143.50 - 2501.00	6.80 - 542.20	185.20 - 4254.00
5	Smoothness	0.05 - 0.16	0.00 - 0.03	0.07 - 0.22
6	Compactness	0.02 - 0.35	0.00 - 0.14	0.03 - 1.06
7	Concavity	0.00 - 0.43	0.00 - 0.40	0.00 - 1.25
8	Concave points	0.00 - 0.20	0.00 - 0.05	0.00 - 0.29
9	Symmetry	0.11 - 0.30	0.01 - 0.08	0.16 - 0.66
10	Fractal dimension	0.05 - 0.10	0.00 - 0.03	0.06 - 0.21

2.2 The Review of PCA and LASSO

2.2.1 PCA

PCA is a common method for dimensional reduction in data science. By using PCA, the high dimensional data will project onto the first few components but preserve majority variance. Consider X is a $n \times p$ matrix, n is the number of observations, p is the number of variables.

Assume X is normalized, the first principal component can be written by $Z_1 = \sum_{j=1}^{p} \alpha_{1j} X_j$,

$\alpha_1 = (\alpha_{11}, ..., \alpha_{1p})^T$ where,

$$\alpha_1 = \arg\max_{\alpha} \alpha^T \Sigma \alpha \text{, subject to } |\alpha_1| = 1$$

Σ is the covariance matrix, $\Sigma = \frac{1}{n}(X^T X)$. The rest principal components can be written as:

$$\alpha_{k+1} = \arg\max_{\alpha} \alpha^T \Sigma \alpha \text{ , subject to } |\alpha| = 1 \text{ and } \alpha^T \alpha_l = 0, \forall 1 \le l \le k$$

There is another form to represent the principal components. Review the connection between PCA and SVD(Singular Value Decomposition), let

$$X = UDV^T$$

Where D is the diagonal matrix and the entries $\sigma_i = D_{ii}$, σ is in descending order, U and V are orthonormal, the columns of V are the eigenvectors of Σ. V is the loading matrix of the principal components. By $XV = UD$, we know that $Z_k = U_k \mathrm{d}_k$ where U_k is the kth column of U . Let x_i be the i th row of X. V_k is the first k components joint, where $V_k = [V_1, V_2, ..., V_k]$, $V_k \in R^{n \times k}$.

The approximation $\hat{X}$ can be written as:

$$\hat{X} = V_k V_k^T x_i$$

To minimize the L2 regularization is:

$$\min \sum_{i=1}^{n} ||X - \hat{X}||^2 = \min_{V_k} \sum_{i=1}^{n} ||x_i - V_k V_k^T x_i||^2$$

2.2.2 LASSO and its Variants

Consider the linear regression. Let's set up a true value vector $Y = (y_1, y_2, ..., y_n)^T$ and predictor variables $X = [X_1, X_2, ..., X_p]$ $(X \in R^{n \times p})$, $X_j = [x_{1j}, x_{2j}, ..., x_{nj}]^T$, $j = 1, 2, ..., p$. In Linear Regression, we try to minimize the following expression:

$$\arg\min_{\beta} || Y - \sum_{j=1}^{p} X_j \beta_j ||^2$$

Ridge Regression is similar to the least square method, but implement the L2 penalty,

$$\arg\min_{\beta} || Y - \sum_{j=1}^{p} X_j \beta_j ||^2 + \lambda \sum_{j=1}^{p} || \beta_j ||^2, \lambda \geq 0$$

In this case, LASSO implement the L1 penalty and shown as follows,

$$\arg\min_{\beta} || Y - \sum_{j=1}^{p} X_j \beta_j ||^2 + \lambda \sum_{j=1}^{p} || \beta_j ||_1, \lambda \geq 0$$

In the LASSO criterion, when solving the optimization problem above, due to the nature of the L1 penalty, some coefficients will become zero and generate sparsity if λ is large enough. Thus, LASSO is also a way to perform variables selection.

Elastic Net regularization is the method that linearly combines L1 and L2 penalty. In linear regression,

$$\arg\min_{\beta} || Y - \sum_{j=1}^{p} X_j \beta_j ||^2 + \lambda_2 \sum_{j=1}^{p} || \beta_j ||^2 + \lambda_1 \sum_{j=1}^{p} || \beta_j || \quad \lambda_1, \lambda_2 \geq 0$$

Zou et al.[20]improve the Elastic Net by multiplying the estimated coefficients, which can lower bias and improve the accuracy of prediction.

$$(1 + \lambda_2) \arg\min_{\beta} || Y - \sum_{j=1}^{p} X_j \beta_j ||^2 + \lambda_2 \sum_{j=1}^{p} || \beta_j ||^2 + \lambda_1 \sum_{j=1}^{p} || \beta_j ||$$

2.3 Sparse PCA

2.3.1 SCoTLASS (Simplified Component Technique-LASSO)

SCoTLASS[15], the Simplified Component Technique-LASSO, extends the standard PCA by taking the variance maximization perspective of the PCA.The proposed method performs the maximization

$$a_k^T \hat{\mathbf{\Sigma}} a_k$$

subject to

$$a_k^T a_k = 1 \text{ and (for } k \geq 2 \text{) } a_h^T a_k = 0, h < k$$

under the extra constraints

$$\sum_{j=1}^{p} |a_{kj}| \leq t$$

for some tuning parameter t, where a_{kj} is the j-th element of the k-th vector $a_k(k = 1,2,....., p)$. SCoTLASS differs from PCA in the inclusion of the constraints, so a decision must be made on the value of the tuning parameter, t. It is easy to see that for t greater than or equal to $\sqrt{p}$, we get PCA.

"As t decreases from $\sqrt{p}$, we move progressively away from PCA and eventually to a solution where only one variable has a nonzero loading on each component." The geometry of SCoTLASS in the case when p = 2 is shown in Figure 1 with vector a's elements a_1 and a_2. For SCoTLASS algorithm, we are restricted to the part of the circle inside the dotted square.

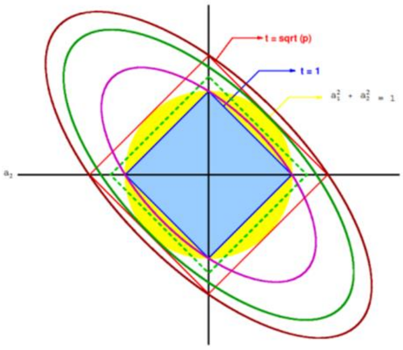

Figure 2 The Two-Dimensional SCoTLASS[15]

2.3.2 GPower (A Generalized Power Method)

Journée at el. [13] propose a generalized power method (GPower), which uses two single-unit and two block optimization formulations of the sparse PCA problem. Its goal is "extracting a single sparse dominant principal component of a data matrix, or more components at once, respectively" . [13] GPower consider the first principal component. By the variance maximization definition, a direct formulation of $l1$ constrained sparse principal component is

$$\begin{array}{ll} \arg \max\limits_{\|\alpha\|=1} & \alpha^T \boldsymbol{X}^T \boldsymbol{X} \alpha \\ \text{subject to} & \| \alpha \|_1 \leq t. \end{array}$$

Equivalently, we can solve

$$ar\,g \max_{\|\alpha\|=1} \sqrt{\alpha^T \boldsymbol{X}^T \boldsymbol{X} \alpha} \tag{1}$$

Journée et al. [13] considered the Lagrangian form of (1)

$$ar\,g \max_{\|\alpha\|=1} \sqrt{\alpha^T \boldsymbol{X}^T \boldsymbol{X} \alpha} - \lambda \| \alpha \|_1 \tag{2}$$

An equivalent formulation of is

$$\begin{array}{c} (U^*, \alpha^*) = \arg \max\limits_{U,\alpha} U^T \boldsymbol{X} \alpha - \lambda \| \alpha \|_1 \\ \text{subject to } \| U \| = 1, \| \alpha \| = 1 \end{array}$$

For any U, the optimal α and $X^T U$ must share the same sign for each component. Let $z_j = |\alpha_j|$, and

Z = |α|. Then, the optimal Z* must satisfy

$$Z^{*} = \arg \max_{Z} \sum_{j=1}^{p} \left(|\boldsymbol{X}^{T}U|_{j} - \lambda\right) z_{j}$$
$$\text{subject to } z_{j} \geq 0, \sum_{j=1}^{p} z_{j}^{2} = 1.$$

After inserting the Z* back to the function (y), a new optimationz criterion of U can be obtained and so does α. In Generalized Power Method, the reformulated problems are of the form of maximization of a convex function on a compact set. This algorithm seems to be faster if the objective function or feasible set is strongly convex. The dimensionality of the feasible set does not depend on n, but on p and on the number m of components to be extracted. If p ≪ n, this is a very desirable property. "On random and real-life biological data, our methods systematically outperform the existing algorithms both in speed and trade-off performance". In the case of biological data, the components obtained by this block algorithm provide the richest biological interpretation.

2.3.3 SPCA

Based on the minimum reconstruction error expression of Standard PCA, Zou reconstructed the product coefficient matrix $V'V$ into two matrices (A, B). After applying Lagrange multipliers λ_0 to drop the constraint of $\mathrm{A} = \mathrm{B}$, for $\lambda_0 > 0$, the first k principal components can be written as

$$\min_{A,B} \sum_{i=1}^{n} \|x_i - AB'x_i\|^2 + \lambda_0 \sum_{j=1}^{k} \|\beta_j\|^2$$
$$\text{subject to } A'A = I_{k\times k}$$

where $A_{p\times k} = [\alpha_1, \dots, \alpha_k]$, $B_{p\times k} = [\beta_1, \dots, \beta_k]$

In order to obtain sparsity, $L1$ penalty is added on β, and the formula becomes

$$\min_{A,B} \sum_{i=1}^{n} \|x_i - AB'x_i\|^2 + \lambda_0 \sum_{j=1}^{k} \|B_j\|^2 + \sum_{j=1}^{k} \lambda_j \|B_j\|_1 \qquad (3)$$
$$\text{subject to } A'A = I_k$$

where λ_0 controls all principal components and different λ_j are allowed for penalizing the loadings of different principal components. The structure of $L1 + L2$ penalty is also called as elastics net.

An alternating minimization algorithm can be used to solve (3), the detail of the calculation can be found in[11], here we briefly overview the key point. We fix a variable one time to transform the problem into the convex optimization problem. Specifically, to get B, the B can be treated as constant value. Each β_j in B is obtained via

$$\hat{\beta}_j = \arg\min_{\beta_j} \|Y_j^{*} - \boldsymbol{X}\beta_j\|^2 + \lambda_0 \|\beta_j\|^2 + \lambda_{1,j} \|\beta_j\|_1$$

where $Y_j^{*} = \boldsymbol{X}\alpha_j$. To obtain A, the problem can be expressed as after fixing B

$$\arg \underset{A}{min} \sum_{i=1}^{n} \|x_i - AB^T x_i\|^2 = \|X - XBA^T\|^2 \text{ , subject to } A'A = I_k$$

This is a reduced rank form of the Procrustes Rotation problem, suppose the SVD decomposition of $X^T XB$ is UDV^T, then the $\hat{A} = UV^T$.

2.3.4 sPCA-rSVD

Different from SPCA, to reconstruct the product coefficient matrix $V'V$, sPCA-rSVD uses best rank-1 approximation. Specifically, defined the SVD decomposition of the data matrix is $X = U\Sigma V^T$, according to the Eckart-Young theorem, let $\widetilde{U} = U_1$, $\widetilde{V} = V_1$, where U_1 and V_1 is the first row and first column of U and V respectively. The best rank-1 approximation for X is

$$\underset{\widetilde{U},\widetilde{V}}{min} \|\boldsymbol{X} - \widetilde{U}\widetilde{V}^T\|_1^2 \text{ , subject to } \|\widetilde{U}\| = 1$$

Then to generate the sparsity, Shen and Huang added three different types of penalty respectively (i.e. soft thresholding, hard thresholding, and smoothly clipped absolute deviation (SCAD) penalty). It can be expressed as the following optimization problem

$$(\widehat{U}, \widehat{V}) = \arg \underset{U,V}{min} \|X - UV^T\|_1^2 + \lambda p(V), \text{ subject to } \|U\| = 1$$

where the notation $p(\,)$ means the penalty. Similar to SPCA, to solve this problem, we also can use an alternating minimization algorithm. There are two variables in this optimization problem, we firstly fix V and try to get U. When the V is given, the optimization problem can be simplified as

$$\widehat{U} = ar\,g \underset{U}{min} \|X - UV^T\|_1^2 \tag{4}$$

The progress to solve (4) is highly similar as (3). the optimal U is $U = \frac{XV}{\|XV\|_1}$. On the other hand, while U is given, the optimal V can be expressed by soft-thresholding operator $V = S\left(\boldsymbol{X}^T U, \frac{\lambda}{2}\right)$. This means the solution of optimal V depends on a different type of penalty.

2.4 Classifiers

2.4.1 Logistics Regression

It is a supervised machine learning model widely used for binary classification. The main aim of this algorithm is to come up with a decision boundary that separates the data points or instances into either of the two classes. The Logistic Response Function is the hypothesis used in this algorithm that converts the decision boundary into the probability of belonging to a particular class. The hypothesis or the logistic response function is as follows:

$$h_\Phi(x) = \frac{1}{1 + e^{-\Phi^T X}}$$

where $\Phi^T X = \phi_0 + \phi_1 x_1 + \phi_2 x_2 + \cdots + \phi_n x_n$

2.4.2 SVM

The intuition of SVM is to find the hyperplane that maximizes the margin of separation between positive and negative examples. For the binary classification task, a linear hyperplane is feasible. However, for the multi-classification problem, kernel function should be used to map feature space into higher dimensional space in order to transform nonlinear separable space into linear separable space. Given the training set $T = \{(x_i, y_i), i = 1, \dots, N, x_i \in R^M, y_i \in \{1, -1\}\}$, the hyperplane based on kernel function can be defined as

$$f(X) = \sum_{i=1}^{M} Y_i a_i k(X_i, X_i) + b$$

where $k(X_i, X_i)$ is the kernel function.

2.4.3 Naïve Bayes

Naive Bayes is a supervised machine learning algorithm based on the concept of the Bayes theorem.

The intuition of Naive Bayes is to maximize the posterior probability, the feature is classified to the type which has maximum posterior probability. Assume the dataset has n data, a features, and y types, Let $X = [x_1, \dots, x_n]$, $A = [a_1, \dots, a_k]$, $Y = [y_1, \dots, y_i]$. The posterior probability of each type can be written as

$$P(Y = y_i \mid X) = \frac{P(X \mid Y = y_i)P(Y = y_i)}{P(X)} = \frac{P(a_1, \dots, a_k \mid Y = y_i)P(Y = y_i)}{P(X)}$$

To simplify this problem, Naive Bayes assumes conditional independence of each feature. Then the equation can be transformed to

$$P(Y = y_i \mid X) = \frac{P(Y = y_i) \prod_{j=1}^{k} P(a_j \mid Y = y_i)}{P(X)}$$

calculate the posterior probability of each feature a_j, the type of feature corresponds to the maximum posterior probability $\hat{y}$.

$$\hat{y} = \underset{y_i}{\operatorname{argmax}} P(Y = y_i) \prod_{j=1}^{k} P(a_j \mid Y = y_i)$$

Since the independence of each feature is unrealistic in a real-world application, the accuracy of Naïve Bayes is not ideal. To improve its performance, consider the data conformed to Gaussian Distribution to replace the strong constraint of independence. Gaussian naïve Bayes theorem is given below:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y{}^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y{}^2}\right)$$

2.5 Feature selection criteria

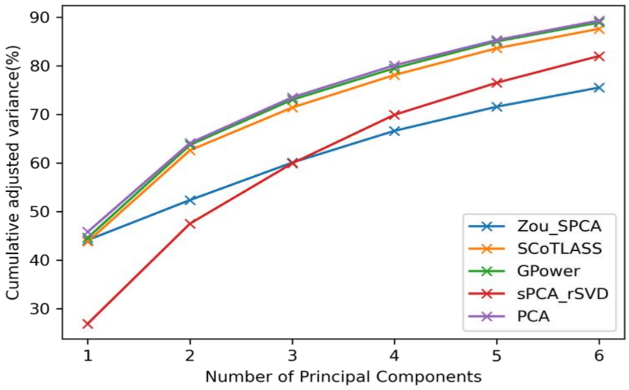

Figure 3 Comparison the Cumulative adjusted variance between different methods for SPCA and Standard PCA

Both in PCA and Sparse PCA, the first few principal components can represent the most cumulative variance, as is shown in the figure. Besides, compared to standard PCA, the principal components are sparse, which makes the PCA more interpretable. Inspired by these properties, to select the important features in the dataset, we propose two criteria based on Sparse PCA. Given α is a hyperparameter:

1. Choose the loadings whose absolute values are greater than α in the first two principal components. (PCs Criterion)

2. Choose the first 10 loadings whose absolute values are greater than α. The priority: Order of Principal components > absolute value of loadings. (Important Criterion)

The motivation of criterion 1 is that the first two principal components can express most of the variance. So, we speculate that the features which have higher weights in the first two principal components would be important. And the definition of higher weights is greater than α, which α could be the average, the third quartile, etc. The selection of α based on the distribution of the dataset. Note that after feature selection by criterion 1, a different number of figures can be generated by different Sparse PCA algorithms. This may bring unbalance in comparison. To solve this problem, criterion 2 chooses the constant number of features, which makes each Sparse PCA generates the same number of features.

3. Result

As we discussed above, firstly we applied SPCA, sPCA_rSVD, SCoTLASS, and GPower respectively to WDBC dataset. The loadings of the first 6 Principal Components of each Sparse PCA algorithm are shown in Table 2-Table 5 sequentially. These tables can be found in the appendix.

The selected features from the WDBC dataset for each Sparse PCA algorithms which are based on Importance Criterion where the hyperparameter α is 0.25 are displayed in Table 6. Similarly, Table 7 shows the extracted features based on PCs criterion where the hyperparameter α is 0.25. Due to the limitation of space, Table 7 is placed in the appendix.

Table 6 Selected Feature for Different Sparse PCA Algorithms Base on Important Criterion(α=0.25)

=SPCA	GPower	SCoTLASS	sPCA-rSVD
perimeter_mean	concave points_mean	concavepoints_mean	perimeter_worst
concavity_mean	concavity_mean	concavity_mean	radius_worst
area_mean	concave points_worst	concavepoints_worst	perimeter_mean
compactness_se	fractal_dimension_mean	perimeter_worst	area_mean
concavity_se	fractal_dimension_se	fractal_dimension_mean	radius_mean
radius_worst	fractal_dimension_worst	fractal_dimension_se	area_worst
radius_se	texture_se	fractal_dimension_worst	compactness_mean
symmetry_se	smoothness_se	compactness_se	concavity_worst
compactness_worst	symmetry_se	smoothness_se	compactness_worst
fractal_dimension_worst	radius_se	radius_se	compactness_se

To evaluate the features in Table 6 can represent the important features in the dataset, define the original dataset which contains 30 features, and the dataset after standard PCA which keep its first six principal components as the control group, the selected features based on PCs Criterion as experimental group 1 and the selected features based on Important Criterion as experimental group 2. Then, the control group and experimental groups are fed into Logistics Regression, SVM, and Naïve Bayes respectively. With 10-fold cross-validation to ensure its credibility, we got the mean test accuracy for 10-fold cross-validation shown in Table 8 - Table 10 for the control group and experimental groups. The accuracy is calculated as

$$\text{Accuracy } = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP: True Positive, TN: True Negative, FP: False Positive, and FN: False Negative.

Table 8. The Testing Accuracy and for Control Group

No.	Model	Testing Accuracy
1	LR	95.61%
2	SVM	93.86%
3	NB	91.22%
4	PCA-LR	96.49%
5	PCA- SVM	94.74%
6	PCA- NB	91.42%

Table 9 The Accuracy and F1 Score for Experimental Group 1

No.	Classification Method	Zou_SPCA (f=8)	GPower (f=6)	SCoTLASS (f=8)	sPCA-rSVD(f=12)	Average Performanceof Classifiers
1	Logistics Regression	94.68%	94.85%	95.90%	94.85%	95.07%
2	SVM	96.96%	97.08%	97.96%	96.43%	97.11%
3	Naïve Bayes	91.47%	88.60%	92.11%	93.86%	91.51%
	Average Performance of Sparse PCA algorithm	94.37%	93.51%	95.32%	95.05%	

Table 10 The Accuracy and F1 Score for Experimental Group 2

No.	Classification Method	Zou_SPCA (f=10)	GPower (f=10)	SCoTLASS (f=10)	sPCA-rSVD(f=10)	Average Performance of Classifiers
1	Logistics Regression	94.68%	96.13%	96.61%	95.50%	95.73%
2	SVM	97.96%	97.13%	97.61%	97.44%	97.54%
3	Naïve Bayes	92.98%	91.22%	93.12%	93.30%	92.66%
	Average Performance of Sparse PCA algorithm	95.21%	94.83%	95.78%	95.41%	

4. Discussion

To make the comparison, based on the average accuracy in Table 6 – Table 8, Figure 1 presents the accuracy for different groups. The accuracy for Importance Criterion and PCs Criterion is the mean of accuracy for four Sparse PCA algorithms in the condition of the same classifier.

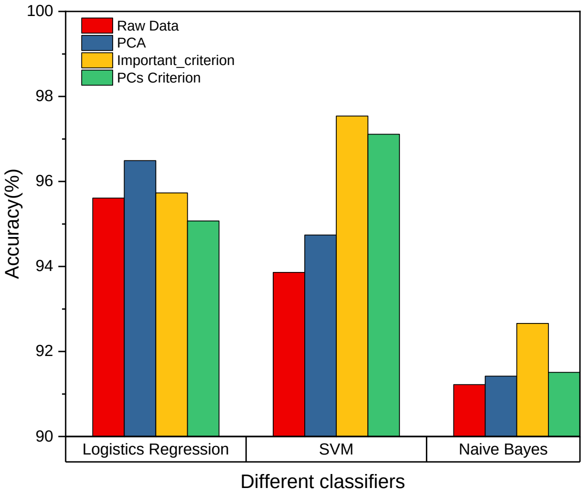

Figure 4 The comparison of accuracy among different classifiers

Firstly, compared the experimental groups with the control group, The accuracy of the SVM and Naïve Bayes classifier for selected features based on both Importance Criterion and PCs Criterion has the higher accuracy than the Raw Data and Data after processing by Standard PCA. Although in the Logistics Regression classifier, the performance of experimental groups is not better than the control group, the difference of accuracy among them is very small, which is less than 2%. Since 1/3 features of the original dataset are used for the classification task and the accuracy for three different classifiers is close to the accuracy after standard PCA and raw dataset, it can conclude that the selected features based on Importance Criterion and PCs criterion are indeed the key features for breast cancer diagnosis. Therefore, Importance Criterion and PCs criterion are both proven useful for feature extraction in Sparse PCA, even for different Sparse PCA algorithms.

Furthermore, making the comparison between Important Criterion and PCs Criterion, it can be found that Important Criterion performs better than PCs Criterion in three classifiers. So, this result suggests that Important Criterion is more effective to find the key features in Sparse PCA.

We also compare our two feature selection criteria with other five feature selection methods including K-means, F-score, ACO, Genetic Algorithm, and Particle Swarm Optimization on the WDBC dataset by 10-fold cross-validation in Table 11[21-23]. Compared to K-means, F-score, ACO, Genetic Algorithm, and Particle Swarm Optimization, in terms of feature selection rate, our method (PCs Criterion) reduces more features than other feature selection algorithms. This indicates that our criterion does a better job in selecting the most important feature from all features. Concerning classification accuracy, both of our two methods are superior to most of the other methods. Besides, the accuracy of our results is the average accuracy of four different Sparse PCA algorithms, which means that our criteria are stable and generalized, which can be feasible in many popular Sparse

PCA algorithms. In the industrial application, we can just pick the optimal Sparse PCA algorithms based on computational time and classification accuracy for dimension reduction, which can get even higher performance than average accuracy. The computation time for WDBC dataset of SPCA, SCoTLASS, GPower, and sPCA-rSVD is shown in Table 10. With the loss of generality, the different number of principal components of Sparse PCA algorithms keep is considered.

Table 11 Classification accuracy and Feature Section Rate of seven different Algorithm with SVM

Feature Selection Methods	Classifier	Accuracy (%)	Feature Selection Rate (%)
This Paper (Importance Criterion)	**SVM**	**97.11**	**33.3**
This Paper (PCs Criterion)	**SVM**	**97.54**	**30.0**
K-means[21]	SVM	97.38	33.3
F-score[22]	SVM	97.56	30.0
ACO[23]	SVM	95.96	50.0
Genetic Algorithm[23]	SVM	97.19	60.0
Particle Swarm Optimization[23]	SVM	97.37	56.7

Table 12 The computational time for four Sparse PCA algorithms

No.	Dimension Reduction Method	PCs=6	PCs=10	PCs=20	PCs=30
1	Zou_SPCA	3.76	4.19	5.95	6.55
2	SCoTLASS	26.67	41.04	65.83	74.25
3	GPower	2.16	0.41	0.69	1.08
4	sPCA-rSVD	0.06	0.09	0.19	0.30

Although researchers have developed some feature selection methods on breast cancer data, few of them gave the name of important features. This makes their feature selection method a black box, which can only work as an intermediary for classifiers. In our method, the name of important features is given (Table 6 - Table 7) benefits from the advantage of the interpretability of Sparse PCA and our two novel feature selection criteria. It provides a new thought on how to define features for breast cancer diagnosis.

5. Conclusion and Future Work

Data from a cancer diagnosis can be high dimensional and features of such data are usually highly redundant. Identifying important features not only circumvent the curse of dimensionality but give guidance to choose the features for cancer diagnosis. To filter out the informative features from the raw dataset, we propose two novel feature selection criteria based on Sparse PCA. Our method reduces removes the features with a low correlation and keeps the high accuracy. Besides, compared to other feature selection algorithms which are used as a medium for the classifiers, our method can tell the feature's name, which makes it more possible for real-world application.

Experiments are conducted with the WDBC breast cancer dataset from the public data repository. By applying our proposed two novel criteria, it is demonstrated that the number of features is successfully reduced without sacrificing the classification accuracy. To prove the feasibility of our criterion, in comparison to the standard PCA and the original dataset, the features selected by our criteria can get the higher accuracy in SVM and Naïve Bayes accuracy and close accuracy in Logistics Regression. To show our priority to other feature selection algorithms, we compare other state-of-the-art methods on the WDBC breast cancer dataset. The result shows that our method can use the fewest features to get the almost highest accuracy.

The usefulness of the Importance criterion and PCs criterion provides a guideline to investigate the interpretability of Sparse PCA algorithms. In our future work, we plan to explore different cancer datasets and use different classifiers such as random forest, neural network, etc. Besides, in terms of the improvement of our feature selection criteria, an automated method to choose hyperparameter faced to the different datasets is necessary.

Contribution

Shaohong Luo:

Propose the idea and design the pipeline. Code to validate the algorithms.

Contribution in Paper writing:

Methodology: SPCA, sPCA-rSVD, classifiers, feature Selection Criteria

Result, Discussion, Conclusion, Appendix

Yingxin Lin:

Contribution in Paper writing:

Introduction

Methodology: Dataset Description, SCoTLASS, GPower

Tingwei Li:

Contribution in Paper writing:

Methodology: The Review of PCA and LASSO

Proofread the paper and fix some format problems

All of us enjoy this research and MATH123. Thank you!

Reference

[1] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, “Nuclear feature extraction for breast tumor diagnosis,” San Jose, CA, Jul. 1993, pp. 861–870. doi: 10.1117/12.148698.

[2] M. A. Hall, “Correlation-based Feature Selection for Machine Learning,” p. 198.

[3] M. Dash and H. Liu, “Consistency-based search in feature selection,” *Artif. Intell.*, vol. 151, no. 1–2, pp. 155–176, Dec. 2003, doi: 10.1016/S0004-3702(03)00079-1.

[4] J. R. Quinlan, “Induction of decision trees,” *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.

[5] A. S. Md. Sohail, Md. M. Rahman, P. Bhattacharya, S. Krishnamurthy, and S. P. Mudur, “Retrieval and classification of ultrasound images of ovarian cysts combining texture features and histogram moments,” in *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, Rotterdam, Netherlands, 2010, pp. 288–291. doi: 10.1109/ISBI.2010.5490352.

[6] J. J. Lin and P.-C. Chang, “A particle swarm optimization based classifier for liver disorders classification,” p. 3.

[7] R. Sheikhpour, M. A. Sarram, and R. Sheikhpour, “Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer,” *Appl. Soft Comput.*, vol. 40, pp. 113–131, Mar. 2016, doi: 10.1016/j.asoc.2015.10.005.

[8] S. Aalaei, H. Shahraki, A. Rowhanimanesh, and S. Eslami, “Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets,” *Iran J Basic Med Sci*, vol. 19, no. 5, p. 7, 2016.

[9] U. Janardhan Reddy, B. Venkata Ramana Reddy, and B. Eswara Reddy, “Unsupervised Feature Selection Approach for Cancer Prediction,” *IETE J. Res.*, pp. 1–6, Feb. 2021, doi: 10.1080/03772063.2021.1878062.

[10] H. Rao *et al.*, “Feature selection based on artificial bee colony and gradient boosting decision tree,” *Appl. Soft Comput.*, vol. 74, pp. 634–642, Jan. 2019, doi: 10.1016/j.asoc.2018.10.036.

[11] H. Zou, T. Hastie, and R. Tibshirani, “Sparse Principal Component Analysis,” *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 265–286, Jun. 2006, doi: 10.1198/106186006X113430.

[12] H. Shen and J. Z. Huang, “Sparse principal component analysis via regularized low rank matrix approximation,” *J. Multivar. Anal.*, vol. 99, no. 6, pp. 1015–1034, Jul. 2008, doi: 10.1016/j.jmva.2007.06.007.

[13] M. Journee and M. Journee, “Generalized Power Method for Sparse Principal Component Analysis,” p. 37.

[14] D. M. Witten, R. Tibshirani, and T. Hastie, “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis,” *Biostatistics*, vol. 10, no. 3, pp. 515–534, Jul. 2009, doi: 10.1093/biostatistics/kxp008.

[15] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin, “A Modified Principal Component Technique Based on the LASSO,” *J. Comput. Graph. Stat.*, vol. 12, no. 3, pp. 531–547, Sep. 2003, doi: 10.1198/1061860032148.

[16] A. Rahoma, S. Imtiaz, and S. Ahmed, “A new criterion for selection of non-zero loadings for sparse principal component analysis (SPCA),” *Can. J. Chem. Eng.*, vol. 99, no. S1, Oct.

2021, doi: 10.1002/cjce.24026.

[17] S. Gajjar, M. Kulahci, and A. Palazoglu, “Selection of non-zero loadings in sparse principal component analysis,” *Chemom. Intell. Lab. Syst.*, vol. 162, pp. 160–171, Mar. 2017, doi: 10.1016/j.chemolab.2017.01.018.

[18] K. Gravuer, J. J. Sullivan, P. A. Williams, and R. P. Duncan, “Strong human association with plant invasion success for Trifolium introductions to New Zealand,” *Proc. Natl. Acad. Sci.*, vol. 105, no. 17, pp. 6344–6349, Apr. 2008, doi: 10.1073/pnas.0712026105.

[19] H. Wang and S. W. Yoon, “Breast Cancer Prediction Using Data Mining Method,” p. 12.

[20] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[21] B. Zheng, S. W. Yoon, and S. S. Lam, “Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms,” *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1476–1482, Mar. 2014, doi: 10.1016/j.eswa.2013.08.044.

[22] M. F. Akay, “Support vector machines combined with feature selection for breast cancer diagnosis,” *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3240–3247, Mar. 2009, doi: 10.1016/j.eswa.2008.01.009.

[23] Y. Prasad, K. K. Biswas, and C. K. Jain, “SVM Classifier Based Feature Selection Using GA, ACO and PSO for siRNA Design,” in *Advances in Swarm Intelligence*, vol. 6146, Y. Tan, Y. Shi, and K. C. Tan, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 307–314. doi: 10.1007/978-3-642-13498-2_40.

Appendix

Table 2 The loadings of first 6 Principal Components of SPCA

No.	Features	PC1	PC2	PC3	PC4	PC5	PC6
1	radius_mean	-0.195					
2	texture_mean	-0.010	0.022		-0.555		
3	perimeter_mean	-0.273					
4	area_mean	-0.242	0.394				
5	smoothness_mean	-0.061			0.036	0.530	
6	compactness_mean	-0.228	-0.007				
7	concavity_mean	-0.268					
8	concavepoints_mean	-0.250					
9	symmetry_mean	-0.055	-0.068	0.077	0.028	0.065	0.186
10	fractal_dimension_mean	-0.039	-0.295			0.262	
11	radius_se	-0.206		0.390			
12	texture_se	0.045	-0.104	0.327	-0.360		
13	perimeter_se	-0.214		0.275			
14	area_se	-0.208	0.155	0.251			
15	smoothness_se	0.013	-0.141	0.317	-0.018	0.128	
16	compactness_se	-0.220	-0.323				
17	concavity_se	-0.227	-0.313			-0.227	
18	concave points_se	-0.238	-0.129		0.031		
19	symmetry_se		-0.161	0.380	0.002		
20	fractal_dimension_se	-0.161	-0.296				
21	radius_worst	-0.236	0.573				
22	texture_worst		0.025	-0.093	-0.741		
23	perimeter_worst	-0.236	0.052				
24	area_worst	-0.216	0.010				
25	smoothness_worst	-0.016		-0.086	-0.016	0.760	
26	compactness_worst	-0.191	-0.054	-0.360	-0.090		
27	concavity_worst	-0.232	-0.033	-0.300	-0.026		
28	concave points_worst	-0.239					
29	symmetry_worst	-0.007					0.983
30	fractal_dimension_worst	-0.109	-0.172	-0.342	-0.009		
	Number of nonzero loadings	28	20	12	12	6	2
	Adjusted variance	41.11%	11.24%	7.71%	6.54%	4.97%	3.93%
	Cumulative adjusted variance	41.11%	52.35	60.06%	66.60%	71.57%	75.5%

Table 3 The loadings of first 6 Principal Components of sPCA-rSVD

No.	Features	PC1	PC2	PC3	PC4	PC5	PC6
1	radius_mean	-0.380			-0.009		
2	texture_mean				0.486	0.060	0.444
3	perimeter_mean	-0.389			-0.014		
4	area_mean	-0.385					
5	smoothness_mean			0.485	-0.001		
6	compactness_mean		-0.438				
7	concavity_mean	-0.059	-0.296		0.043		
8	concavepoints_mean	-0.245		0.159			
9	symmetry_mean			0.455			-0.023
10	fractal_dimension_mean		-0.076	0.278	-0.114	-0.171	0.017
11	radius_se	-0.062		0.109	0.377		-0.350
12	texture_se				0.388	-0.288	0.062
13	perimeter_se	-0.055		0.120	0.383		-0.366
14	area_se	-0.126		0.016	0.292		-0.277
15	smoothness_se			0.158	0.041	-0.655	0.164
16	compactness_se		-0.392	-0.009		-0.105	
17	concavity_se		-0.293			-0.076	
18	concave points_se		-0.230		0.116	-0.047	-0.104
19	symmetry_se			0.205	0.070	-0.319	-0.189
20	fractal_dimension_se		-0.172			-0.421	
21	radius_worst	-0.391			-0.009		
22	texture_worst				0.431	0.205	0.517
23	perimeter_worst	-0.396			-0.010		
24	area_worst	-0.376					
25	smoothness_worst			0.437	-0.044		0.298
26	compactness_worst		-0.393		-0.019	0.164	
27	concavity_worst		-0.407			0.162	
28	concave points_worst	-0.136	-0.036	0.172		0.096	
29	symmetry_worst			0.376	-0.033	0.232	
30	fractal_dimension_worst		-0.268	0.021	-0.120		0.189
	Number of nonzero loadings	12	11	14	20	14	13
	Adjusted variance	26.9%	20.6%	12.4%	10.0%	7.5%	5.5%
	Cumulative adjusted variance	26.9%	47.5%	59.9%	69.0%	76.5%	82.0%

Table 4 The loadings of first 6 Principal Components of SCoTLASS

No.	Features	PC1	PC2	PC3	PC4	PC5	PC6
1	radius_mean	0.236	-0.211				
2	texture_mean	0.078		-0.086	0.570		0.001
3	perimeter_mean	0.245	-0.190				
4	area_mean	0.238	-0.209				
5	smoothness_mean	0.117	0.191		-0.112	0.124	0.524
6	compactness_mean	0.239	0.153	-0.024	-0.034		
7	concavity_mean	0.268	0.031				
8	concavepoints_mean	0.276	0.000		-0.048		0.016
9	symmetry_mean	0.110	0.178			0.432	-0.016
10	fractal_dimension_mean	0.010	0.430				
11	radius_se	0.211	-0.025	0.324		0.057	0.036
12	texture_se	0.000	0.071	0.218	0.577	0.014	0.118
13	perimeter_se	0.217	-0.001	0.314		0.053	
14	area_se	0.210	-0.085	0.250	-0.009	0.043	0.021
15	smoothness_se		0.216	0.357	0.023		0.309
16	compactness_se	0.149	0.263	0.039		-0.205	-0.271
17	concavity_se	0.131	0.214	0.093		-0.276	-0.274
18	concave points_se	0.171	0.129	0.205		-0.103	-0.027
19	symmetry_se	0.000	0.192	0.244		0.554	-0.257
20	fractal_dimension_se	0.062	0.330	0.153		-0.272	-0.136
21	radius_worst	0.246	-0.192				
22	texture_worst	0.078		-0.235	0.553		0.044
23	perimeter_worst	0.255	-0.163				
24	area_worst	0.242	-0.192		-0.001		
25	smoothness_worst	0.100	0.167	-0.193	-0.081	0.002	0.573
26	compactness_worst	0.204	0.135	-0.297	-0.016	-0.041	-0.059
27	concavity_worst	0.230	0.076	-0.208	-0.062	-0.096	-0.072
28	concave points_worst	0.263		-0.136	-0.098		0.000
29	symmetry_worst	0.094	0.120	-0.301		0.506	-0.201
30	fractal_dimension_worst	0.099	0.307	-0.288		-0.113	
	Number of nonzero loadings	29	27	19	13	16	18
	Adjusted variance	43.8%	18.8%	8.8%	6.7%	5.5%	4.0%
	Cumulative adjusted variance	43.8%	62.6%	71.4%	78.1%	83.6%	87.6%

Table 5 The loadings of first 6 Principal Components of GPower

No.	Features	PC1	PC2	PC3	PC4	PC5	PC6
1	radius_mean	0.220	-0.233				
2	texture_mean	0.103	-0.052		0.622		
3	perimeter_mean	0.228	-0.214				
4	area_mean	0.222	-0.231				
5	smoothness_mean	0.142	0.187		-0.132	0.392	-0.302
6	compactness_mean	0.239	0.154	-0.070			
7	concavity_mean	0.258	0.062			-0.081	
8	concavepoints_mean	0.261				0.059	
9	symmetry_mean	0.138	0.191			0.323	0.354
10	fractal_dimension_mean	0.063	0.367				-0.113
11	radius_se	0.206	-0.106	0.282	-0.077	0.132	
12	texture_se	0.000	0.094	0.383	0.389	0.118	
13	perimeter_se	0.211	-0.090	0.278		0.101	
14	area_se	0.203	-0.153	0.227	-0.091	0.111	
15	smoothness_se		0.204	0.327		0.205	-0.348
16	compactness_se	0.170	0.233	0.142		-0.293	0.077
17	concavity_se	0.153	0.197	0.162		-0.364	
18	Concave points_se	0.183	0.130	0.223	-0.087	-0.201	
19	symmetry_se	0.041	0.183	0.300		0.235	
20	fractal_dimension_se	0.102	0.280	0.204		-0.283	
21	radius_worst	0.229	-0.219				
22	texture_worst	0.104			0.650		
23	perimeter_worst	0.237	-0.199				
24	area_worst	0.226	-0.218				
25	smoothness_worst	0.128	0.175	-0.237		0.344	-0.389
26	compactness_worst	0.210	0.146	-0.245		-0.109	
27	concavity_worst	0.229	0.101	-0.184		-0.177	
28	concave points_worst	0.251		-0.170			
29	symmetry_worst	0.123	0.144	-0.265		0.269	0.490
30	fractal_dimension_worst	0.132	0.278	-0.238		-0.089	-0.078
	Number of nonzero loadings	28	27	17	7	19	8
	Adjusted variance	44.6%	19.1%	9.3%	6.5%	5.5%	4.0%
	Cumulative adjusted variance	44.6%	63.7%	73.0%	79.5%	85.0%	88.9%

Table 7 Selected Feature for Different Sparse PCA Algorithms Base on PCs Criterion (α=0.25)

Zou_SPCA		GPower		SCoTLASS		sPCA-rSVD	
PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
perimeter_mean	area_mean	concavity_mean	fractal_dimension_mean	concavity_mean	fractal_dimension_mean	radius_mean	compactness_mean
concavity_mean	fractal_dimension_mean	concave points_mean	fractal_dimension_se	concavepoints_mean	compactness_se	perimeter_mean	compactness_se
	compactness_se	concave points_worst	fractal_dimension_worst	perimeter_worst	fractal_dimension_se	area_mean	concavity_se
	concavity_se			concavepoints_worst	fractal_dimension_worst	radius_worst	compactness_worst
	fractal_dimension_se					perimeter_worst	concavity_worst
	radius_worst					area_worst	fractal_dimension_worst