MicroRNA data reduction of esophageal cancer

Wei Li¹, Zhongyu Su, Wenshuang Li, Tianjia Liu and Pengcheng Guo
School of Science, Dalian Ocean University, Dalian, 116023, China

¹Eamil: wwli@dlou.edu.cn

Abstract. Esophageal cancer is one of the most aggressive and lethal type of malignant digestive tract tumor. MicroRNA plays pivotal roles in cell homeostasis. The dysregulation of microRNA can lead to the development of cancer. Because of the large amount of microRNA data for esophageal cancer patients, it is difficult to analyse data. Hence there is a pressing need to reduce data for early diagnosis of esophageal cancer. In this paper, an efficient method is provided to reduce the microRNA data of esophageal cancer patients. Principal component analysis is firstly implemented to reduce the microRNA data dimension. Based on the results of principal component analysis, clustering analysis is used for further data reduction. With this comprehensive method, the 1046 microRNA data of 198 esophageal cancer patients is comprehensively analyzed. This method has great potential to be a useful tool for the preoperative diagnosis and prognostic prediction of esophageal cancer.

1. Introduction
Large amounts of high-dimensional data is becoming increasingly common in many modern applications such as imaging analysis and data analysis. The massive size of such data creates big challenges in terms of computing speed and computer memory [1-2]. Principal component analysis (PCA) as an effective multivariate statistical method, can reduce the problem in high dimension space to that in low dimension space, making the problem simple and intuitive. These fewer composite indicators are uncorrelated and can provide most of the information of the original indicator [3-4]. In this paper, PCA is used to effectively reduce the microRNA data dimension of esophageal cancer patients. Combined with cluster analysis, microRNA data of esophageal cancer patients is comprehensively analyzed. This method has great potential to be a useful tool for preoperative diagnosis and prognostic prediction of esophageal cancer.

Esophageal cancer is now one of the most aggressive and lethal type of malignant digestive tract tumor [5-6]. China is a country with a high incidence of esophageal cancer, which accounts for about one-fifth of the cancer deaths each year. Early patients should be preferred for surgical treatment. In addition to surgical treatment for middle and advanced cancer, radiotherapy, chemotherapy, traditional Chinese medicine and laser therapy can be used [7]. MicroRNA is an endogenous small RNA with a length of about 20-25 nucleotides, which has many important regulatory functions in the cell. MicroRNA plays pivotal roles in cell homeostasis and dysregulation of microRNA can lead to the development of cancer [8-10]. So it is necessary to analyze microRNA data for the study of esophageal cancer.

This paper consists of four sections. Section 1 is the introduction, including the research purpose and significance; section 2 is the theoretical introduction, including the basic theory and algorithm steps of PCA and cluster analysis; in Section 3, based on the theoretical guidance in section 2, using
PCA, with SPSS software, combined with clustering analysis, microRNA data of esophageal cancer patients is efficiently analyzed; Section 4 is the summary and discussion.

2. Methods
An efficient method combined PCA and clustering analysis is provided in this paper. Firstly, PCA is implemented to reduce the microRNA data dimension and obtain the main components. Secondly, clustering analysis is used to classify the main components in PCA for further data reduction.

2.1. Principal component analysis
Principal component analysis (PCA), is a multivariate statistical method that converts multiple indicators into several comprehensive indexes by using the idea of dimensionality reduction and minimizing the loss of information [11-12]. Usually called main component, the comprehensive indicators of the transformation to generate each principal component are a linear combination of the original variables and unrelated, and have better properties than the original variables. So only using a few principal components without loss of too much information, can solve the complex high-dimensional problems [13-14].

The algorithm steps are as follows:
(1) Standardize the original data to eliminate the influence of variables on the order of magnitudes and dimension;
(2) Calculate the correlation coefficient matrix R according to the normalized data matrix;
(3) Find the characteristic roots and eigenvectors of R matrix;
(4) Determine the principal component and give appropriate interpretation to the information contained in the main components;
(5) Synthesize the main components and obtain the comprehensive evaluation values.

2.2. Clustering analysis
Clustering is the process of dividing an object set into several groups, so that the data objects in the same group have a higher similarity, while the data objects in different groups are not similar. Similar or not similar are defined based on the value of attribute variables, which are generally expressed by the distance between each object. A cluster is a collection of similar groups of objects that are often treated as an object [15-16]. In this paper, the system clustering method is implemented to analyze the main components obtained by PCA. The algorithm steps are as follows:
First, n samples are n classes (category). One class contains one sample. Second, two classes with the closest properties are merged into a new class. So that n-1 classes are obtained. Sequentially, merging two closest classes yields n-2 classes, etc. Finally, all the samples are classified to one category.

3. Results
Firstly, PCA is implemented to reduce the microRNA data dimension and obtain the main components, with SPSS software. Secondly, clustering analysis is used to classify the main components in PCA for further data reduction.

3.1. Principal component analysis

3.1.1. Preliminary work. This part is the verification that the microRNA data of esophageal cancer is suitable for PCA. The original data is from the American UCSC database [17]. It is 1046 microRNA data from 198 esophageal cancer patients as table 1:

Table 2 shows that KMO statistics is 0.994 and P=0.000<0.01, which are suitable for PCA. (Note: KMO is a ratio of correlation coefficient and partial correlation coefficient, a statistical value used to compare correlation coefficient and partial correlation coefficient. The closer this value is to 1, indicates that the more suitable this data is for PCA. Bartlett sphericity test is to determine whether the
correlation coefficient matrix is the statistical value for the unit matrix. The associated probability is less than the significant (0.05 or 0.1) shows that the correlation coefficient matrix is not a unit matrix, suitable for PCA [18].

The results of KMO (Kaiser-Meyer-Olkin) and bartlett test are obtained using SPSS in table 2:

Table 1. MicroRNA data of esophageal cancer patients.

| sample          | TCGA LNA5 U701 | TCGA JYA6 FH01 | TCGA L5A8 NU01 | TCGA L7A5 6G01 | TCGA IGA3 YC01 | TCGA L5A4 OE01 | TCGA L5A8 NM01 |
|-----------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| has-mir-1322    | 0             | 0             | 0             | ...           | 0             | 0             | 0             |
| has-mir-1323    | 0             | 0             | 0             | 0             | 0             | 0             | 0             |
| has-mir-3195    | 0             | 0             | 0             | 0             | 0             | 0             | 0             |
| has-mir-1321    | 0             | 0             | 0             | 0             | 0             | 0             | 0             |
| has-mir-3193    | 0.4691        | 0.7343        | 0.1256        | 0             | 0             | 0.3863        | 1.2238        |
| ...             | ...           | ...           | ...           | ...           | ...           | ...           | ...           |
| has-mir-3909    | 0.4691        | 0.4133        | 0.348         | 0.6624        | 0.7826        | ...           | ...           |
| has-mir-600     | 0             | 0             | 0             | 0             | 0             | 0             | 0.5313        |

Table 2. KMO and Bartlett test.

|                          | KMO sampling adequacy | .994 |
|--------------------------|-----------------------|------|
| Bartlett test of sphericity | The approximate chi-square  | 3.1.2. Analysis results. | Based on the above work, the factor extraction, scree plot and components scores are analyzed. And the score of each variable factor and the score of principal component are calculated.

Factor extraction

Solving factor solution is chosen as the main component analysis method. The basis for determining the number of factors is that the cumulative contribution rate of variance is over 85%, as shown in table 3 and figure 1.

Table 3. Total variance interpretation.

| composition          | Initial eigenvalue | Extract the sum of squares of the load. |
|----------------------|--------------------|--------------------------------------|
| TCGALNA5U701         | 188.355            | 188.355                              |
| TCGAJYA6FH01         | 2.326              | 2.326                                |
| TCGAL5A8NU01         | 1.024              | 1.024                                |
| TCGAL7A56G01         | .593               | .593                                 |
| TCGAL5A4OE01         | .001               | .001                                 |
| TCGAL5A8NM01         | .001               | .001                                 |

Scree plot
According to figure 1, there is a steep inflection point to the third principal component eigenvalue. The ingredients after it are flat. The eigenvalues of first principal component and the rest of the main components present obvious distinction. The basic features after the fourth principal component drive nearly flat. And table 3 shows that the first three principal component eigenvalues account for 96.821% of all information, which occupies most of all information. We can only use these three principal components to express the majority of the information [19].

Calculation of the score of each component
Using SPSS to calculate the score matrix of each component, the output results are shown in table 4.

Table 4. Factor score coefficient matrix.

| composition     | 1     | 2     | 3     |
|-----------------|-------|-------|-------|
| TCGALNA5U701    | 0.005 | -0.049| 0.118 |
| TCGAJYA6FH01    | 0.005 | 0.035 | 0.076 |
| TCGAL5A8NU01    | 0.005 | 0.020 | -0.102|
| TCGAL7A56G01    | 0.005 | -0.019| 0.071 |
| ...             | ...   | ...   | ...   |
| TCGAIGA3YC01    | 0.005 | -0.040| -0.017|
| TCGAL5A4OE01    | 0.005 | 0.043 | 0.026 |
| TCGAL5A8NM01    | 0.005 | 0.066 | 0.048 |

The score of each principal component on each index is the weight. The expression of the principal component is:
First principal component:
PRIN1=0.005*1+0.005*2+0.005*3+0.005*4+0.005*5+...+0.005*196+0.005*197+0.005*198;
Second principal component:
PRIN2=-0.049*1+0.035*2+0.020*3-0.019*4+0.054*5+...-0.040*196+0.043*197+0.066*198;
Third principal component:
PRIN3=0.118*1+0.076*2-0.102*3+0.071*4+0.084*5+...-0.017*196+0.026*197+0.048*198.

Calculation of the score of each variable factor
Table 5. Factor score of microRNA data.

| sample          | factor1 | factor2 | factor3 |
|-----------------|---------|---------|---------|
| has-mir-1322    | -0.56662| 0.01485 | -0.26765|
| has-mir-1323    | -0.51456| 0.15082 | 0.06874 |
| has-mir-3195    | -0.56277| -0.00202| -0.31573|
| ...             | ...     | ...     | ...     |
| has-mir-3908    | -0.56777| 0.01502 | -0.27375|
| has-mir-3909    | -0.43791| -0.01901| 0.10046 |
| has-mir-600     | -0.54546| 0.08406 | -0.20423|

Calculation of the principal component score of each variable
Multiplying the values of factor 1, factor 2 and factor 3 by the arithmetic square root of their respective variance, yields the score of principal component 1, principal component 2 and main component 3 as table 6.

Table 6. Scores of the principal components of microRNA data.

| sample          | factor1 | factor2 | factor3 | principal components1 | principal components2 | principal components3 |
|-----------------|---------|---------|---------|------------------------|------------------------|------------------------|
| has-mir-1322    | -0.56662| 0.01485 | -0.26765| -7.78                  | 0.02                   | -0.27                  |
| has-mir-1323    | -0.51456| 0.15082 | 0.06874 | -7.06                  | 0.23                   | 0.07                   |
| has-mir-3195    | -0.56277| -0.00202| -0.31573| -7.72                  | 0.00                   | -0.32                  |
| ...             | ...     | ...     | ...     | ...                    | ...                    | ...                    |
| has-mir-3908    | -0.56777| 0.01502 | -0.27375| -7.79                  | 0.02                   | -0.28                  |
| has-mir-3909    | -0.43791| -0.01901| 0.10046 | -6.01                  | -0.03                  | 0.10                   |
| has-mir-600     | -0.54546| 0.08406 | -0.20423| -7.49                  | 0.13                   | -0.21                  |

3.2. Clustering analysis of table 5
Based on the principal component results in 3.1, with SPSS software, clustering analysis is implemented to further analyze the data. The number of cluster for the data of three principal components of 198 esophageal cancer patients is 6, as table 7:

Table 7. Cluster analysis based on PCA.

| sample          | Six clustering | sample          | Six clustering |
|-----------------|----------------|-----------------|----------------|
| hsa-mir-1322    | 1              | hsa-mir-203     | 4              |
| hsa-mir-1323    | 1              | hsa-mir-202     | 1              |
| ...             | ...            | hsa-mir-205     | 5              |
| hsa-mir-191     | 2              | ...             | ...            |
| hsa-mir-190     | 1              | hsa-mir-215     | 6              |
| hsa-mir-192     | 3              | ...             | ...            |
| ...             | ...            | hsa-mir-600     | 1              |

4. Conclusions
In this paper, the microRNA data of esophageal cancer patients is efficiently reduced by a comprehensive method combined PCA and cluster analysis. Through PCA, 1046 microRNA data of 198 patients is reduced to three principal components. Combining with clustering analysis, the PCA results are reduced to 6 classes.

The results provide a more efficient and accurate basis for the data analysis of esophageal cancer patients. They are the important preliminary data for the analysis of the potential value and research on the early detection, early prevention and early diagnosis for esophageal cancer. Moreover, our future
research is to explore whether the microRNA in the results of cluster analysis influence the formation of esophageal cancer cells.

References

[1] Wang D, Shen H and Truong Y 2016 Efficient dimension reduction for high-dimensional matrix-valued data Neurocomputing 190 25-34

[2] Pan Q 2017 Research on dimensionality reduction method in high dimensional longitudinal data analysis The financial times 26 19-20

[3] Hou W 2006 An improved method for comprehensive evaluation of principal component analysis [J] Journal of liaoning normal university (natural science edition) 2 211-214

[4] Meng S W 1992 The main component analysis method is used to evaluate the multi-index [J] Statistical study 4 86-87

[5] Hu W, Liang Y, Zhang S, Hu Y, Liu J 2014 The significance of subcarinal dissection in esophageal cancer surgery [J] Asia Pac J Clin Oncol 183-189

[6] Lin J J, Wang W Y, Xu Q Q, et al 2016 MiR-218 increases sensitivity to cisplatin in esophageal cancer cells via targeting survivin expression [J] Open Med 11(1) 31-35

[7] Weng J Y and Young D S 2017 Some dimension reduction strategies for the analysis of survey data [J] Journal of Big Data 12 4-43

[8] Billeter A T, Barnett R E, Druen D, et al 2012 MicroRNA as a new factor in lung and esophageal cancer [J] Journal of huangshan college 5 20-22

[9] Guo Y J 2002 Comprehensive evaluation theory and method [M] (Beijing: science press) 111-117

[10] Du Q, Jia L Y and Yan X F 2014 Spss statistical analysis from introduction to mastery [M] (People's post office) 288-302

[11] Li P Z 2008 The application of cluster analysis and principal component analysis in the evaluation of regional comprehensive consumption level [D] Nankai university

[12] Usoro A E and Moffat I U 2015 Principal Component Analysis of Nigeria Value of Major Imports [J] American Journal of Economics 5 508-512