PCA, K-Means and EFA Example with Breast Cancer Coimbra Data Set

Jiachen Zhang

目录

Abstract: Principle component analysis, K-Means, exploratory factor analysis are all unsupervised algorithms. In this homework, these algorithms are performed upon 9 variables concerning health, and the results are interpreted.

Keywords: Principle Component Analysis; K-Means; Exploratory Factor Analysis

(The main body is about 8 pages)

1 Read the Data

Read the data *dataR2.csv*. The classification variable is the variable to be predicted, so it is transformed to the factor variable. There is no missing values in the data.

1.1 Data Information

The data *dataR2.csv* contains 10 variables, and the sampl size is 116. The information of the variables is as follows:

1. 9 quantitative variabless: Age(年龄) (years), BMI(身体质量指数) (kg/m2), Glucose(葡萄糖) (mg/dL), Insulin(胰岛素) (µU/mL), HOMA(空腹胰岛素), Leptin(瘦蛋白) (ng/mL), Adiponectin(脂肪连接蛋白) (µg/mL), Resistin(抵抗素) (ng/mL), MCP.1(单核细胞趋化蛋白) (pg/dL)
2. 1 qualitative variable: Classification (Labels: 1=Healthy controls; 2=Patients)

2 Data Exploration

2.1 Summary Statistics

The frequency table of 'Classification' is as follows. In the sample, there are 52 healthy persons and 64 patients. The summary statistics of the 9 continuous variables by group and the correlation plot are shown in Appendix.

表 1: Frequency table of the target variable

Classification	Frequency
1	52
2	64

3 PCA - Principle Component Analysis

Use principle component analysis to reduce the number of dimensions by constructing principal components (PCs). In this case, there are 9 variables, which is actually not a big number, thus the principle component analysis might not be necessary. In practice, there is a tradeoff between the reduction of dimension and the loss of variation.

In this section, the principal components is computed and visualized, and then the appropriate number of principal components is chosen, and the value of the principal components for classification by decision tree.

3.1 Compute the Principal Components

There are two general methods to perform PCA in R, which are spectral decomposition and singular value decomposition. The function `princomp()` uses the spectral decomposition approach, and the function `prcomp()` uses the singular value decomposition (SVD). SVD has slightly better numerical accuracy. Therefore, the function `prcomp()` is used here.

```
pca_res <- prcomp(data_all[,-10], center = TRUE, scale = TRUE)
```

表 2: Importance of components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	1.75	1.234	1.081	1.052	0.850	0.811	0.664	0.541	0.179
Proportion of Variance	0.34	0.169	0.130	0.123	0.080	0.073	0.049	0.033	0.004
Cumulative Proportion	0.34	0.509	0.639	0.762	0.842	0.915	0.964	0.996	1.000

Each of the 9 principal components explains a percentage of the total variation in the dataset. That is to say, PC1 explains 34% of the total variance, which means that more than one-third of the information in the 9 variables can be encapsulated by the one principal component. PC2 explains 16.9% of the variance.

3.2 Information of the PCA result

The PCA object denoted as "pca_res" in the code contains the following information:

1. x: the values of each sample in terms of the principal components

The value of the first 5 components are used for classification, and the details can be found in Appendix.

2. rotation: the matrix of variable loadings (columns are eigenvectors)

It is shown in the "The Interpretation of the Principal Components" section.

表 3: The value of the first 4 individuals in terms of the principal components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
1	-1.98	-0.125	-0.361	-0.273	-0.491	0.343	0.277	0.199	0.249
2	-1.32	-0.247	-1.331	1.886	-0.175	-0.086	0.084	0.439	-0.019
3	-1.21	-0.988	-0.361	0.910	1.309	-1.382	0.545	-0.156	-0.044
4	-1.19	0.271	-1.752	0.271	-0.162	-0.499	0.647	0.736	0.048

3.3 PCA visualization

3.3.1 Screeplot and Cumulative Variance Plot

Use the screeplot cumulative variance plot to select the principal components to keep. The screeplot and cumulative variance plot are shown below.

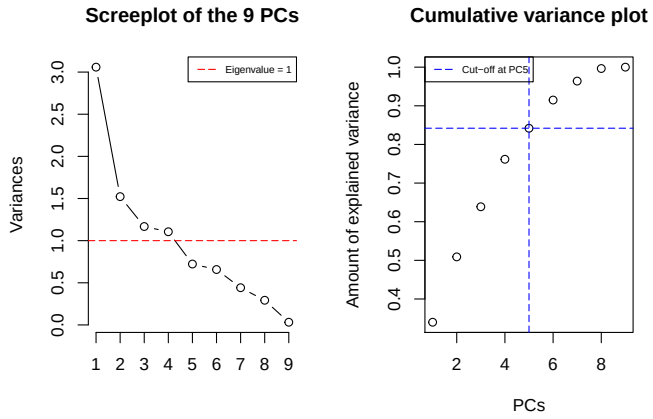

图 1: Screeplot and cumulative variance plot

Use the two following rules to select the principal components:

1. Kaiser rule: pick PCs with eigenvalues of at least 1.

Since the data is standardized, an eigenvalues less than 1 would mean that the principal component actually explains less than a single explanatory variable, and the corresponding components are likely to be discarded.

2. Proportion of variance plot: the selected PCs should be able to describe at least 80% of the variance.

The plots show that the first 4 components has an Eigenvalue more than 1 and explains more than 70% of variance. However, as the first five components explains more than 80%, the first five components are kept.

3.3.2 Biplot

The biplot merge an usual PCA score plot with a plot of loadings. It includes the position of each sample in terms of PC1 and PC2 and how the initial variables map onto the two principal components. The `ggbiplot` package is used to plot biplots and the biplot is as follows.

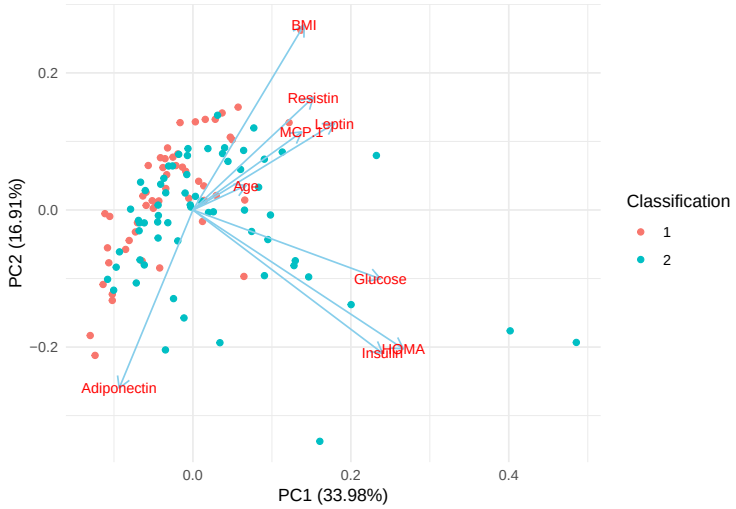

图 2: Biplot

The interpretation of the biplot in detail can be found in Appendix.

3.3.3 The Interpretation of the Principal Components

The first five components are selected. The loading matrix below shows the connection between the components and the variables and interpretation of the principal components is based on finding which variables are most strongly correlated with each component.

表 4: Loading matrix

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Age	0.125	0.066	-0.207	0.821	0.253	-0.295	0.308	-0.127	-0.029
BMI	0.260	0.499	0.426	-0.071	-0.232	-0.277	-0.053	-0.599	0.062
Glucose	0.439	-0.186	-0.131	0.126	0.200	-0.030	-0.807	-0.084	-0.201
Insulin	0.444	-0.386	0.094	-0.060	-0.298	0.122	0.391	-0.094	-0.614
HOMA	0.493	-0.375	-0.012	-0.006	-0.139	0.068	0.131	-0.016	0.758
Leptin	0.331	0.234	0.583	0.058	0.288	-0.022	0.054	0.636	-0.031
Adiponectin	-0.173	-0.481	0.282	-0.277	0.529	-0.488	0.103	-0.232	-0.005
Resistin	0.282	0.304	-0.289	-0.303	0.598	0.421	0.244	-0.243	-0.015
MCP.1	0.255	0.210	-0.497	-0.359	-0.119	-0.633	0.089	0.301	-0.046

The first component increases with Insulin(胰岛素), HOMA(空腹胰岛素) and Glucose(葡萄糖). Because Insulin increases with Glucose in blood increases, the first component can be viewed as a measure of blood sugar.

The second component increases with BMI(身体质量指数) and decreases with Adiponectin(脂肪连接蛋白), so the second component can be viewed as a measure of body shape.

The third component increases with BMI(身体质量指数) and Leptin(瘦蛋白), and decreases with MCP.1(单核细胞趋化蛋白), so the third component can be viewed as a measure of health status concerning body fat.

The fourth component is strongly connected with Age(年龄), so the fourth component can be viewed as a measure of age.

The fifth component increases with Adiponectin(脂肪连接蛋白) and Resistin(抵抗素), which are both secreted by fat cells, so the fifth component can be viewed as a measure of health status concerning dat cells.

4 K-Means

Clustering is a broad set of techniques for finding subgroups of observations within a data set. Clustering allows us to identify which observations are alike, and potentially categorize them. In this section, K-means clustering is used for splitting the dataset into a set of k groups. This is an unsupervised method, which implies that it seeks to find relationships between the n observations without being trained by a response variable. However, in this case, there is a response variable indicating the health of the individual, and the 9 variables are all connect to the response variable, so the result of clustering is compared with the response variable in the end of the section.

4.1 The Distance Matrix

The classification of observations into groups requires some methods for computing the distance between each pair of observations. Here Euclidean distance is applied.

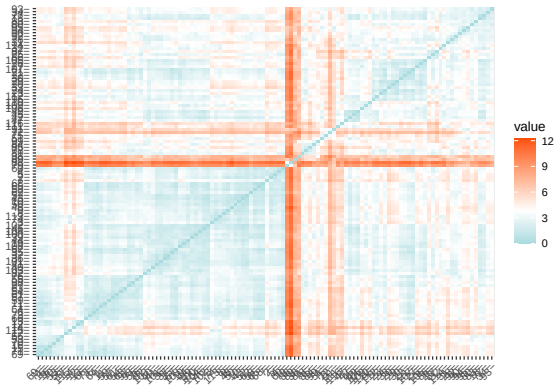

图 3: Distance matrix visualization

The function `fviz_dist` is used for visualizing the distance matrix, and a color closes to red represents a further distance. The plot of distance shows that some individuals are quite similiar, while others have large dissimilarities.

4.2 Compute K-Means Clustering

Here the data is grouped into two clusters, specified by the parameter `centers = 2`. The kmeans function also has an nstart option that attempts multiple initial configurations and reports on the best one. For example, adding nstart = 10 will generate 10 initial configurations.

```
k_res <- kmeans(data_k, centers = 2, nstart = 10)
```

4.3 Information of the K-Means Result

The output of kmeans is a list includingh the following information:

1. cluster: a vector of integers from 1 to the number of clusters, indicating the cluster to which each point is allocated

```
## For example, The first 6 individuals belong to 2 2 2 2 2 2 clusters
```

2. centers: a matrix of cluster centers

表 5: Cluster centers

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
0.058	0.849	0.629	0.704	0.658	0.843	-0.211	0.504	0.388
-0.032	-0.464	-0.344	-0.385	-0.360	-0.461	0.116	-0.276	-0.212

3. size: the number of points in each cluster

There are 41 in the first cluster, 75 in the second cluster. Although the numbers of samples with value 1 and 2 of classification variable are quite close (52 and 64 respectively), there are significantly more samples in cluster 2 than in cluster 1.

4.4 K-Means Visualizetion

View the k-means results by using `fviz_cluster`. The two clusters have a small overlap area.

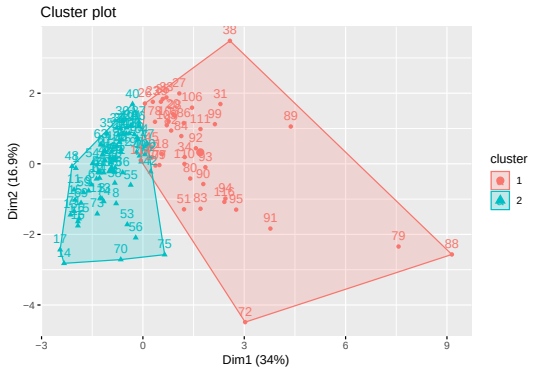

图 4: Clustering by k-means

Using the methods of determining the optimal clusters (shown in Appendix), 2 clusters can be regarded as the optimal number of clusters.

4.5 Comparison of the Clustering Results and the Value of Classification Variable

Compare the results of 2 clustering with the groups divided by the original classification variable. There are some equal values in the clustering results and the value of classification variable. This might indicate that the clustering result partly reflect the health status of the individual.

```
##         Classification
## Clusters  1  2
##        1 14 27
##        2 38 37
```

5 EFA - Exploratory Factor Analysis

Exploratory factor analysis (EFA) is a method used to uncover the underlying structure of a relatively large set of variables. In this case, there are 9 variables and there is not any prior knowledge about the latent variables. Thus, in this section, the optimal number of factors is determined, and then the factors are interpreted.

According to the methods of determining the number of factors to extract (shown in Appendix), 4 factors are computed.

5.1 Compute Factor Analysis with a Varimax Rotation

Use `factanal` function for factor analysis. The function performs maximum-likelihood factor analysis on a covariance matrix. The number of factors to be fitted is specified by the argument `factors`. Here

```
fac_res <- factanal(data_fac, factors = 4, rotation = "varimax")
```

5.2 Information of the Factor Analysis Result

1. uniqueness: the proportion of variability, which can not be explained by a linear combination of the factors

表 6: Uniqueness

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
0.892	0.011	0.256	0.036	0.005	0.547	0.877	0.795	0.005

A high uniqueness for a variable indicates that the factors do not account well for its variance. The uniqueness of age, Leptin, Adiponect and Resistin is relatively high.

2. communality: the fraction of the variable's total variance explained by the factor

表 7: Communality

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
0.108	0.989	0.744	0.964	0.995	0.453	0.123	0.205	0.995

3. loadings: the contribution of each original variable to the factor

```
##
## Loadings:
##             Factor1 Factor2 Factor3 Factor4
## Age                                 0.326
## BMI                  0.979   0.156
## Glucose      0.475   0.138   0.219   0.671
## Insulin      0.974   0.107
## HOMA         0.932           0.160   0.308
## Leptin       0.243   0.595           0.194
## Adiponectin         -0.285  -0.188
## Resistin     0.104   0.157   0.352   0.213
## MCP.1        0.109           0.989
##
##                Factor1 Factor2 Factor3 Factor4
## SS loadings      2.126   1.461   1.242   0.747
## Proportion Var   0.236   0.162   0.138   0.083
## Cumulative Var   0.236   0.399   0.537   0.620
```

Notice there is no entry for certain variables, because R does not print loadings less than 0.1. The table beneath the loadings shows the proportion of variance explained by each factor. The row Cumulative Var gives the cumulative proportion of variance explained. The row Proportion Var gives the proportion of variance explained by each factor, and the row SS loadings gives the sum of squared loadings. According to Kaiser's rule, a factor is worth keeping if the SS loading is greater than 1, thus, the optimal number of factors is 3 with Kaiser's rule.

5.3 Visualization and Interpretation of the factors

The loadings can be visualized as follows.

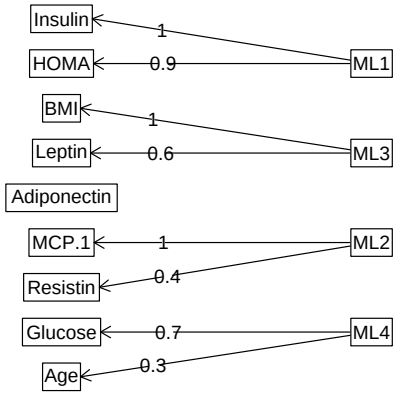

图 5: Factor Analysis

The loadings in the plot is rounded. The plot shows that the first factor has large loadings on Insulin(胰岛素) and HOMA(空腹胰岛素), so the first factor represents the aspect of Insulin. The second factor has large loadings on BMI(身体质量指数)

and Leptin(瘦蛋白), so the second factor can be viewed as the representation of the body shape. The third factor has large loadings on MCP.1(单核细胞趋化蛋白) and Resistin(抵抗素), so the third factor is connected with inflammation(炎症). The fourth factor has large loadings on Glucose(葡萄糖), so the fourth factor can be viewed as the representation of the blood sugar level(血糖).

Note that Adiponectin(脂肪连接蛋白) and Age(年龄) are not well explained by the four factors. This is consistent with the high uniqueness of these two variables, which are 0.876554 and 0.891852 respectively.

6 Conclusions and Shortcomings

In this homework, principal component analysis, k-means and exploratory factor analysis are performed upon the breast cancer Coimbra data set. Overall, the results are not so good: the classification using pca reduces the accuracy, the clustering result and the original classification variable are not a perfect match, and the age and adiponectin variables are not well explained by the factors. This might be because the number of variables and the sample size are not too large. But on the other hand, the results are acceptable.

7 Reference

[1]Team, B. (2018, September 18). How to read PCA biplots and scree plots - BioTuring Team. Medium. https://bioturing.medium.com/how-to-read-pca-biplots-and-scree-plots-186246aae063

[2]K-means Cluster Analysis • UC Business Analytics R Programming Guide. (n.d.). University of Cincinnati. Retrieved May 29, 2021, from https://uc-r.github.io/kmeans_clustering

[3]A Simple Example of Factor Analysis in R. (n.d.). Freie Universitaet Berlin. Retrieved May 29, 2021, from https://www.geo.fu-berlin.de/en/v/soga/Geodata-analysis/factor-analysis/A-simple-example-of-FA/index.html

[4]Neto, J. (n.d.). Factor Analysis. Unknown. Retrieved May 29, 2021, from http://www.di.fc.ul.pt/~jpn/r/factoranalysis/factoranalysis.html

[5]Intro - Basic Exploratory Factor Analysis | QuantDev Methodology. (n.d.). PennState. Retrieved May 29, 2021, from https://quantdev.ssri.psu.edu/tutorials/intro-basic-exploratory-factor-analysis

[6]Gilles Raiche and David Magis (2020). nFactors: Parallel Analysis and Other Non Graphical Solutions to the Cattell Scree Test. R package version 2.4.1. https://CRAN.R-project.org/package=nFactors

[7]11.4 - Interpretation of the Principal Components | STAT 505. (n.d.). PennState: Statistics Online Courses. Retrieved May 29, 2021, from https://online.stat.psu.edu/stat505/lesson/11/11.4

[8] 朱彦萍. (2017). 2 型糖尿病患者血清抵抗素单核细胞趋化蛋白-1 和 c 反应蛋白与颈总动脉内膜-中层厚度关系探讨及干预研究. 中国药物与临床 (7).

8 Appendix

8.1 Some Summary Statistics

8.1.1 Summary Statistics of the 9 Continuous Variables by Group

Caculate the summary statistics of the 9 continuous variables by group. The mean of the mean of age, BMI, leptin, adiponectin for Group 1 and Group 2 are relatively close.

表 8: Summary statistics of Group 1 and 2

	mean	sd	median	min	max	range
Group 1						
Age1	58.1	19.0	65.0	24.00	89.0	65.0
BMI1	28.3	5.4	27.7	18.67	38.6	19.9
Glucose1	88.2	10.2	87.0	60.00	118.0	58.0
Insulin1	6.9	4.9	5.5	2.71	26.2	23.5
HOMA1	1.6	1.2	1.1	0.47	7.1	6.6
Leptin1	26.6	19.3	21.5	4.31	83.5	79.2
Adiponectin1	10.3	7.6	8.1	2.19	38.0	35.9
Resistin1	11.6	11.4	8.9	3.29	82.1	78.8
MCP.11	499.7	292.2	471.3	45.84	1256.1	1210.2
Group 2						
Age2	56.7	13.5	53.0	34.00	86.0	52.0
BMI2	27.0	4.6	27.4	18.37	37.1	18.7
Glucose2	105.6	26.6	98.5	70.00	201.0	131.0
Insulin2	12.5	12.3	7.6	2.43	58.5	56.0
HOMA2	3.6	4.6	2.0	0.51	25.1	24.5
Leptin2	26.6	19.2	18.9	6.33	90.3	84.0
Adiponectin2	10.1	6.2	8.4	1.66	33.8	32.1
Resistin2	17.2	12.6	14.4	3.21	55.2	52.0
MCP.12	563.0	384.0	465.4	90.09	1698.4	1608.3

8.1.2 Correlation Plot

The correlation plot between the 9 continuous variables is as follows. The HOMA variable has a strong correlation with both Insulin and Glucose, and the other correlation coefficients are close to zero.

8.2 The Information of Biplot

The following information can be concluded from the biplot:

1. How the samples relate to one another in our PCA, in other words, which samples are similar and which are different:

The sample points are in different colors according to the target variable. The samples of classfication 1 is likely to distributed in the top left in the plot.

2. How each variable contributes to each principal component:

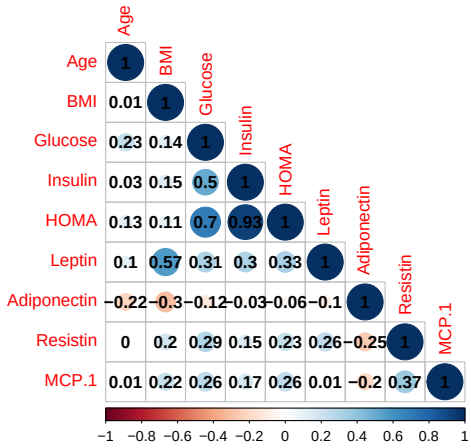

图 6: Correlation plot

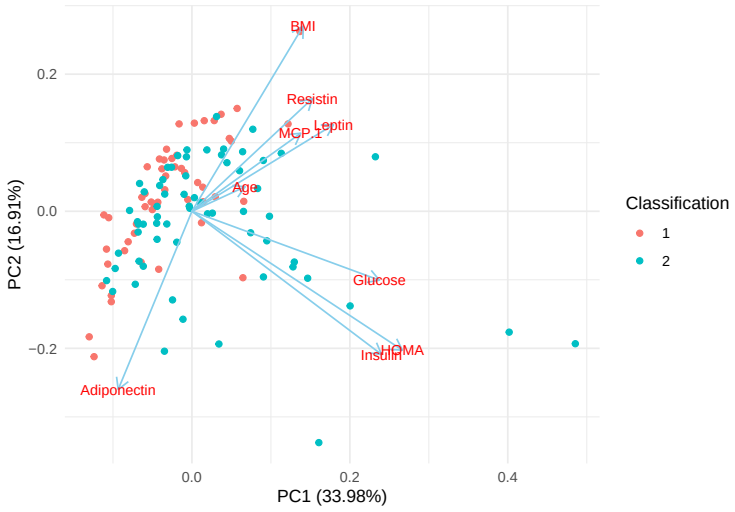

图 7: Biplot

The vectors are pinned at the origin of PCs (PC1 = 0 and PC2 = 0), and their project values on each PC show how much weight they have on that PC. In this case, Glucose, HOMA, and Insulin strongly influence PC1, while Adiponectin, BMI, Insulin and HOMA have more say in PC2.

3. The correlation between the variables:

The angles between the vectors show how variables correlate with each another.

- When two vectors are close, forming a small angle, the two variables they represent are positively correlated. Example: Insulin and HOMA.
- When the angle between two vectors are close to 90°, they are not likely to be correlated. Example: Adiponectin and Glucose.
- When the angle between two vectors are close to 180°, they are negative correlated. Example: Adiponectin and BMI.

Note that PC1 and PC2 can only explain about 50% of the variance, so the correlation between the variables concluded from the biplot might not match the actual correlation well. For example, the correlation between Adiponectin and HOMA or Insulin is closer to zero than the correlation between Adiponectin and Glucose, but the angle of the latter is closer to 90°.

8.3 Use the First 5 Principal Components and the Original 9 variables Respectively for Classification

According to the last homework, methods including logistic regression, support vector machine, random forest, linear discriminant analysis, k-nearest neighbors, naive Bayes and decision tree can all be used for classification. In this homework, the decision tree is applied.

In this section, the first 5 principal components are used for classification, and then the original 9 variables are also used for a comparison. The code is not shown in the report due to the limited space.

表 9: Comparison of classification with pca and original variables

	5 components	9 original variables
Accuracy	0.50000	0.68182
Precision	0.57143	0.69231
Recall	0.33333	0.75000
F1	0.42105	0.72000
AUC	0.51905	0.67949

The evaluation parameters of classification with the 5 principal components are all lower than with the 9 original variables. Remember that in the biplot, some points from different group are overlapped, and they are hard to be classified. Although the first 5 components keep more than 80% of the variance, the accuracy of classification is still reduced a lot.

8.4 Determining Optimal Clusters

8.4.1 Visual Assessment

Here because there are two levels in classification variables, the number of clusters is set to be 2. However, there might be different groups within the healthy individuals (Group1) and the patients (Group2), so in this section, several different numbers of clusters is used and the difference in the results is shown below.

First, try the same K-Means process for 3, 4, and 5 clusters, and the results are shown in the figure.

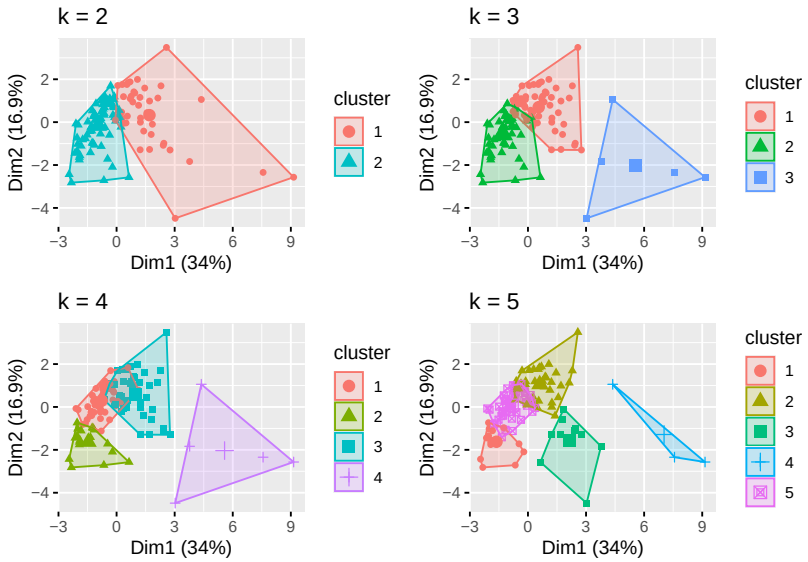

图 8: K-Means with different number of clusters

Although this visual assessment shows where true dilineations occur between clusters, it does not show what the optimal number of clusters is.

8.4.2 The Optimal Number of Clusters

The most popular methods for determining the optimal clusters are Elbow method and average Silhouette method, and they are tried below.

Method 1: Elbow Method

The process of the Elbow method is as follows:

- Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
- For each k, calculate the total within-cluster sum of square.
- Plot the curve of wss according to the number of clusters k.
- The location of a bend in the plot is generally considered as an indicator of the appropriate number of clusters.

However, there is not a ideal bend in the plot, so the optimal number of clusters cannot be determined by the elbow method.

Method 2: Average Silhouette Method

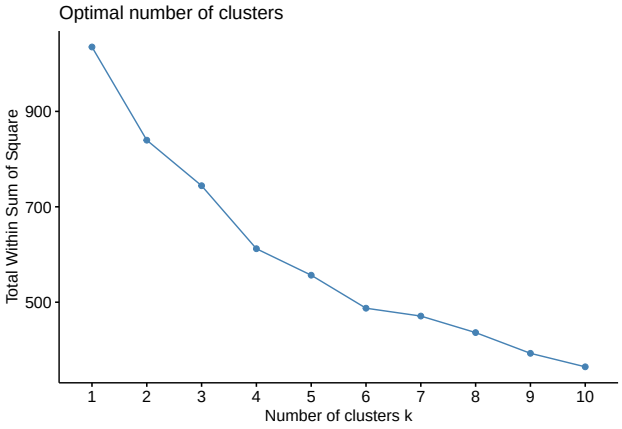

图 9: Plot of elbow method

The average silhouette method computes the average silhouette of observations for different values of number of clusters. The optimal number of clusters k is the one that maximizes the average silhouette over a range of possible values for number of clusters.

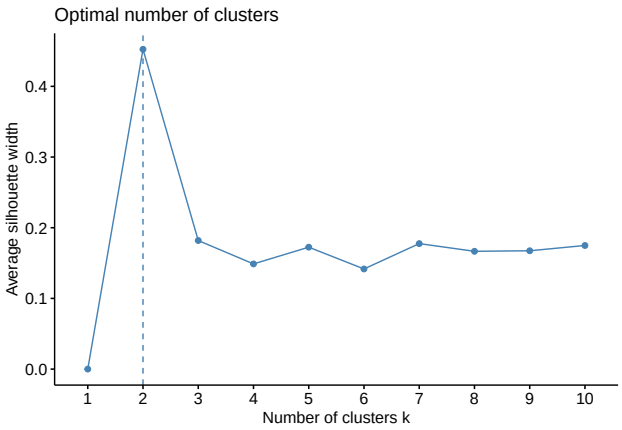

图 10: Plot of average Silhouette method

The results show that 2 clusters maximize the average silhouette values with 3 clusters coming in as second optimal number of clusters.

In conclusion, 2 clusters can be regarded as the optimal number of clusters.

8.5 Determining the Number of Factors to Extract

The `nScree` function returns an analysis of the number of factors to retain in an exploratory principal component or factor analysis. Different solutions are given, including the Kaiser rule, the parallel analysis, optimal coordinates and acceleration factor. Further details can be found in the description of package "nFactors" [6].

```
##   noc naf nparallel nkaiser
## 1   2   1         4       4
```

In this case, the optimal numbers of factors to extract are 2 with optimal coordinates solution, 1 with acceleration factor, 4 with the parallel analysis and 4 with the Kaiser rule. Because the Kaiser rule and parallel analysis are more widely used, in this case 4 factors are computed.

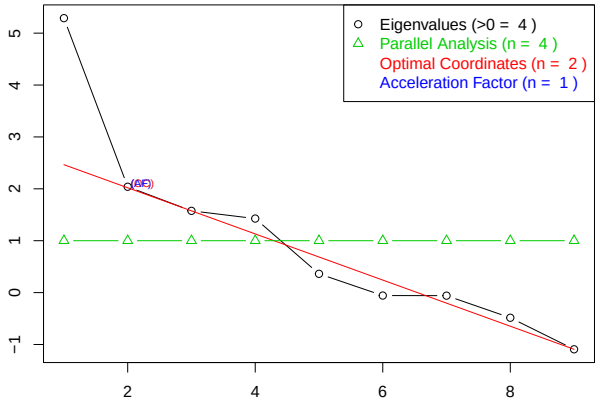

图 11: Solutions to Scree Test