Introduction to Statistical Learning Coursework

JIAQI LI

8^{th} March 2022

Contents

1 Introduction

1.1 Background

The latest global cancer burden data released by the World Health Organization's International Agency for Research on Cancer (IARC) for 2020 shows 19.29 million new cancer cases worldwide in 2020. One of the most notable changes is the rapid increase in the number of new cases of breast cancer to 2.26 million, officially replacing lung cancer (2.2 million) for the first time as the number one cancer worldwide, accounting for 11.7% of all new cancer cases. In addition, globally, 9.23 million new cancers will occur in women in 2020, accounting for 48% of the total. Of these, 2.26 million are new cases of breast cancer, far exceeding other cancer types in women (e.g., colorectal cancer 870,000, lung cancer 770,000, cervical cancer 600,000, etc.) [2].

Although overall breast cancer will only be ranked fifth in the number of global cancer deaths in 2020, it still ranks first in the number of female cancer deaths with 680,000 cases.

1.2 Data

1.2.1 Reason to choose

As a woman, I am very concerned about the breast cancer, and identifying benign or malignant is very important in breast cancer screening. In order to better screen for benign and malignant tumors by disease quickly, I will study in this project whether 30 indexes such as radius mean and texture mean of cysts can better identify benign and malignant tumors.

For this report I have used the Breast Cancer Wisconsin (Diagnostic) Data Set from the UCI Machine Learning Repository website [1]. In the data there are ten real-valued features calculated for each cell nucleus: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. And the mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius [3]. And in this report I will explore whether these 30 FEATURES are a better predictor of whether the patient's breast disease is benign or malignant and if that could be reduce to lower dimension for draft prediction. To test the credibility of my results, I will also use the second column of data from the cited data (indicating whether it is actually benign or malignant) [1].

1.2.2 Prior to deal with

At the very beginning I pre-processed the data first. In the initial data, the diagnosis column has two parameters, "M" and "B", where "M" stands for malignant breast cancer and "B" for benign breast cancer. In order to make it easier to process the data, I changed all the string parameters to number parameters, where "1" represents malignant breast cancer and "2" represents benign breast cancer.

1.2.3 Structure and dealing of data set

We can see the structures of the data set of the first 6 columns in the Figure 1 [1].

In order for the subsequent calculations to run smoothly, I need to process the data before starting, for example by removing the first and second columns that do not belong to index, doing standardizing, naming each row with their corresponding ID, and storing the determination of whether the disease is benign or malignant (i.e. the second column) separately.

2 Data Analysis

2.1 Multidimensional Scaling

Multidimensional scaling takes the distance or dissimilarity matrix, in this project I use the Classical Scaling on Euclidean distances and the ordinal scaling on Euclidean distances.

Here I wanted to explore whether these features would give a better estimate of the nature of the tumour, but I found a total of 30 features here, which was too many, so I chose to downscale here using Multidimensional Scaling to find the few dimensions with the greatest correlation.

2.1.1 Checking Eigenvalue

We have 30 features here, so we chose n' = 31, but really we need to look properly at the eigenvalues. By Figure 2[2], we can find that the dimension of the configuration is chosen to be 2, because the first two eigenvalues are significantly bigger than the rest.

The Log Abs Eigen graph which is shown in Figure 3[3] can show that apart from the first two, there are as many negative eigenvalues that the points in graph of log eig are roughly symmetric.

2.1.2 Projection

1. Classical Scaling
   The projection in classical scaling of 569 patients on the selected 2 dimensions was drawn with R to show the distribution of the 569 patients, where red represents benign and black represents malignant, and from Figure 4[4] we can see that there are two clusters on the graph, which prompted us to use k-means clusters to divide these patients into groups to confirm whether these features could roughly predict the type of breast tumour.
2. Ordinal Scaling
   Similar projection in ordinal scaling of 569 patients on the selected 2 dimensions was drawn with R, and from Figure 5[5] we can see that there are two clusters on the graph too.

Since whether it is benign or not is a kind of a judgment, it is more appropriate to choose ordinal scaling here and it looks more compact in the picture, so I choose ordinal scaling for the subsequent analysis.

2.2 Procrustes Analysis

the isoMDS function prints out the stress value as it changes every five iterations, by this we can find that there is a significant reduction in stress, so we'd expect the scaling solution to be different, and we can check the expectation is true in Figure 6[6]. Here the black shows the classical scaling and the red shows the ordinal scaling.

I identified the two most variate dimensions in the original data and assumed that they were the two selected dimensions, from which we can see a high degree of overlap in Figure 7[7] drawn by Procrustes Analysis. Also the red here shows the classical scaling and the black shows the ordinal scaling.

2.3 k-means clustering

With Figure 8[8] we can find the optimal k. After choosing k=2, these patients can be well divided into two categories. And in the Figure 9[9] I show the 2 clusters that can be obtained with this model.

2.4 superimposition

Now based on the true clustering, which is the second column of the original data as a marker, I superimposed the two clusters, as shown in Figure 10[10], and the difference between theses two is small, which shows that the classification is reasonable and can be accurate in predicting whether this breast tumour is benign or malignant by the FEATURES of these patients.

3 conclusion

In the determination of benign versus malignant breast cancer, these 30 features can successfully reduce the dimension of the data from 30 to 2. Although in fact such lower dimension estimates are still subject to small error, while retaining the salient features that facilitate clustering, they can give a good rough estimate of the nature of the tumour prior to surgery.

References

[1] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[2] IARC. Cancer fact sheets. `https://gco.iarc.fr/today/fact-sheets-cancers`.

[3] UCI Machine Learning. Breast cancer wisconsin (diagnostic) data set. `https://gco.iarc.fr/today/fact-sheets-cancers`.

4 Graph

Figure 1: First 6 columns of the initial data set

id <int>	diagnosis <int>	radius_mean <dbl>	texture_mean <dbl>	perimeter_mean <dbl>	area_mean <dbl>
842302	1	17.990	10.38	122.80	1001.0
842517	1	20.570	17.77	132.90	1326.0
84300903	1	19.690	21.25	130.00	1203.0
84348301	1	11.420	20.38	77.58	386.1
84358402	1	20.290	14.34	135.10	1297.0
843786	1	12.450	15.70	82.57	477.1
844359	1	18.250	19.98	119.60	1040.0
84458202	1	13.710	20.83	90.20	577.9
844981	1	13.000	21.82	87.50	519.8
84501001	1	12.460	24.04	83.97	475.9

1–10 of 569 rows | 1–6 of 32 columns Previous 1 2 3 4 5 6 ... 57 Next

Figure 2: Eigenvalue Graph

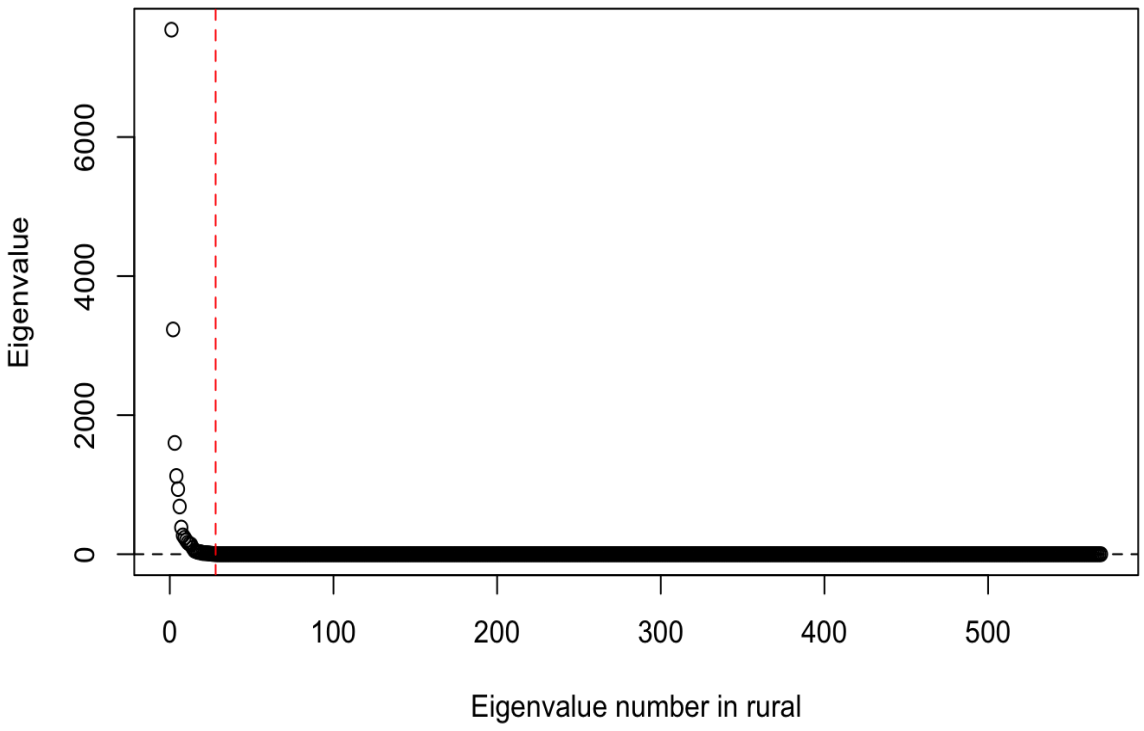

Figure 3: Log eigenvalue graph

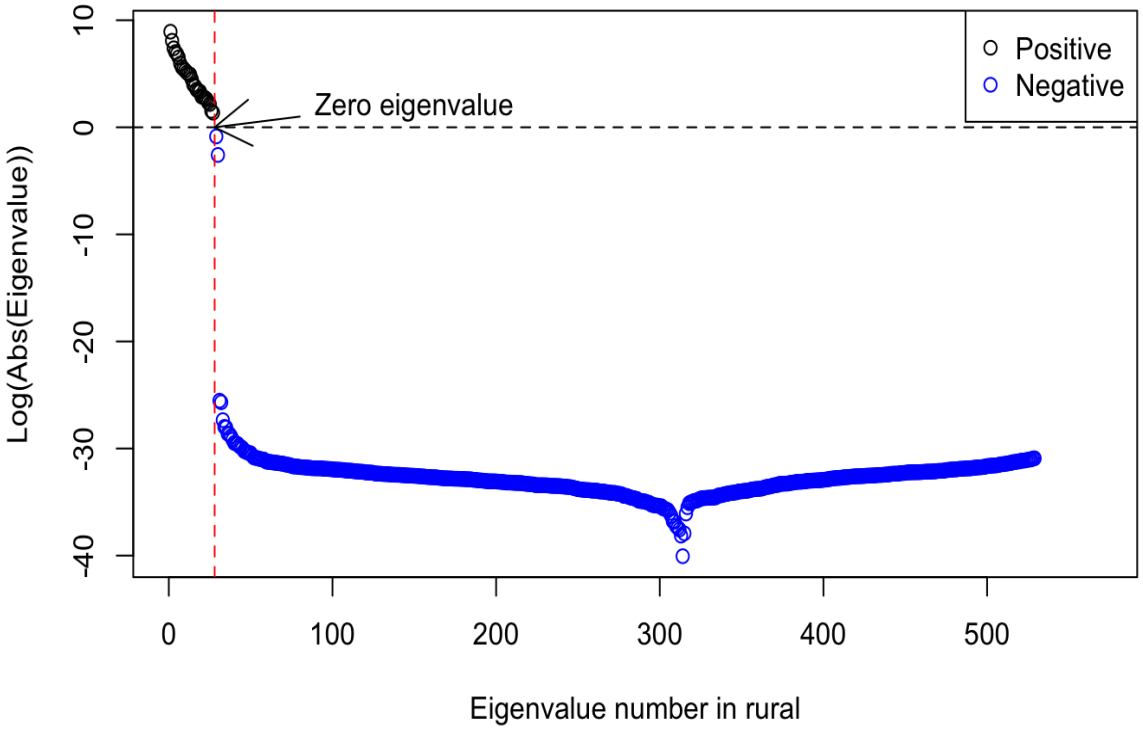

Figure 4: Classical Euclidean 2D Projection

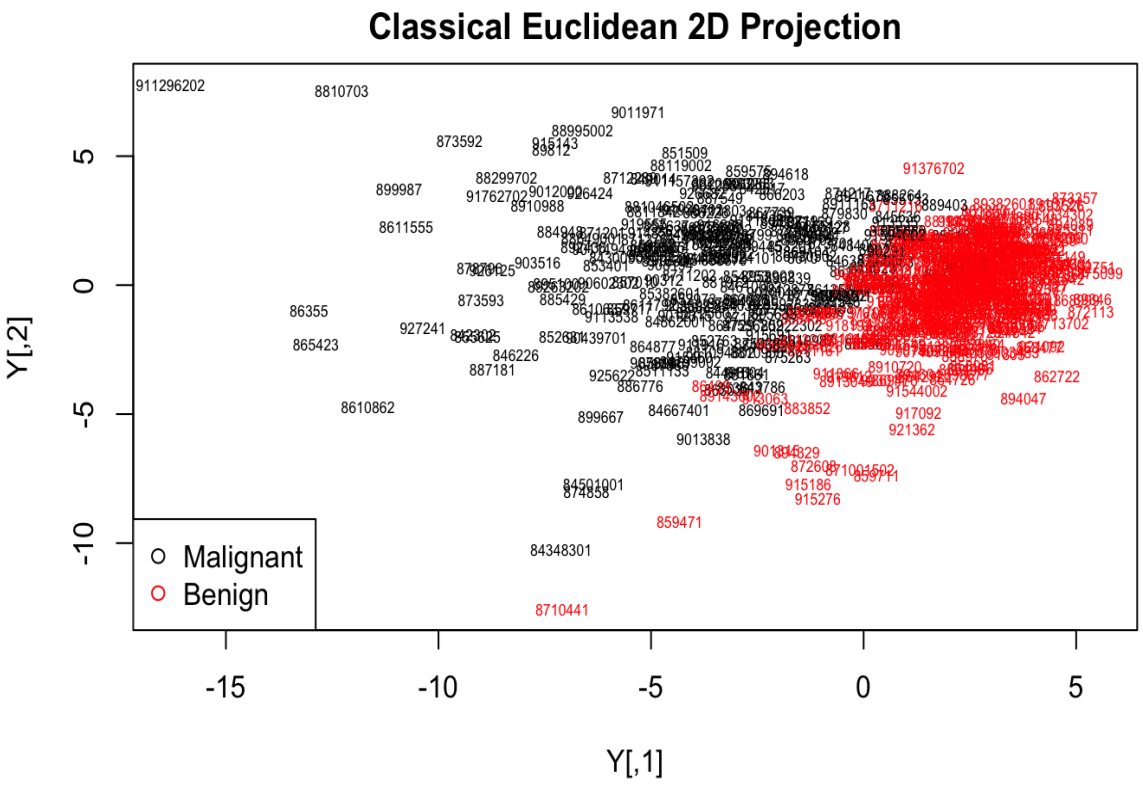

Figure 5: Ordinal Euclidean 2D Projection

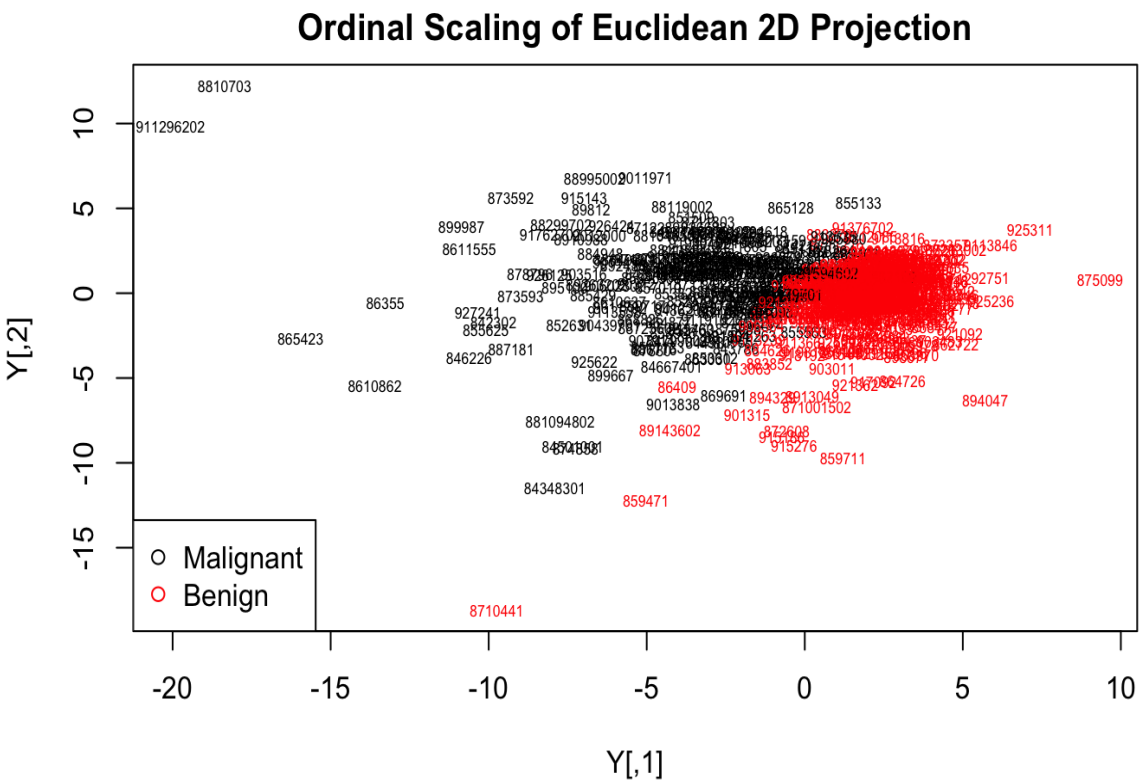

Figure 6: Procustes Analysis with the Ordinal and Classical Scaling

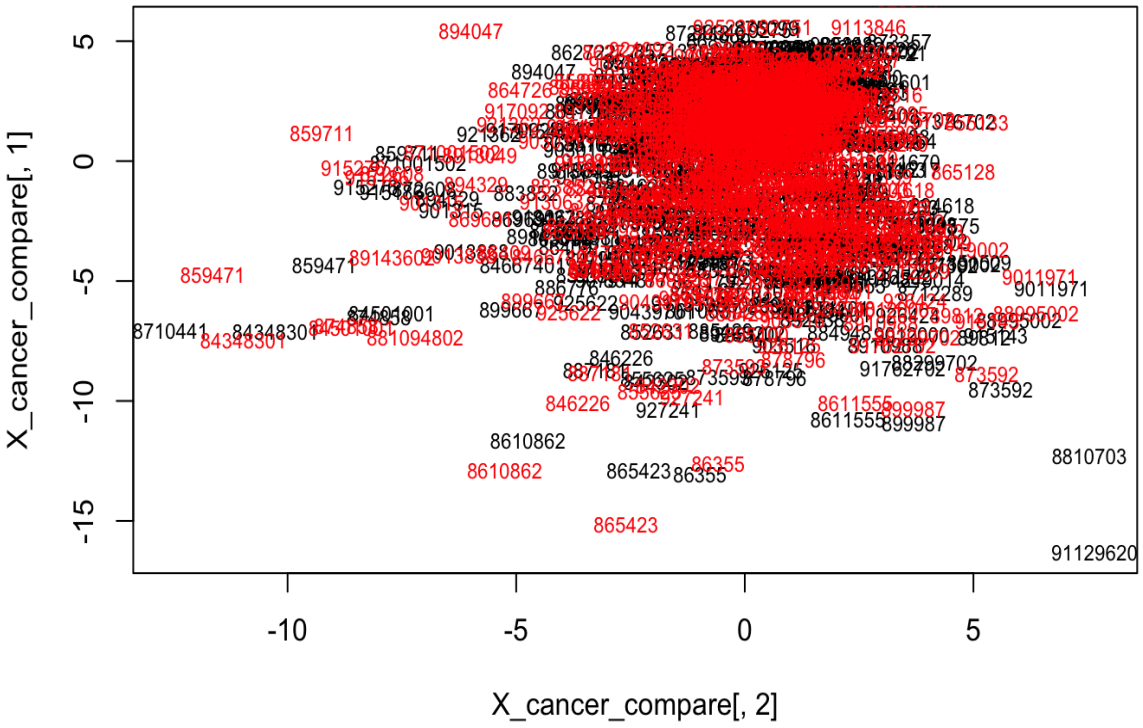

Figure 7: Procustes Analysis with the Ordinal scaling and Original data

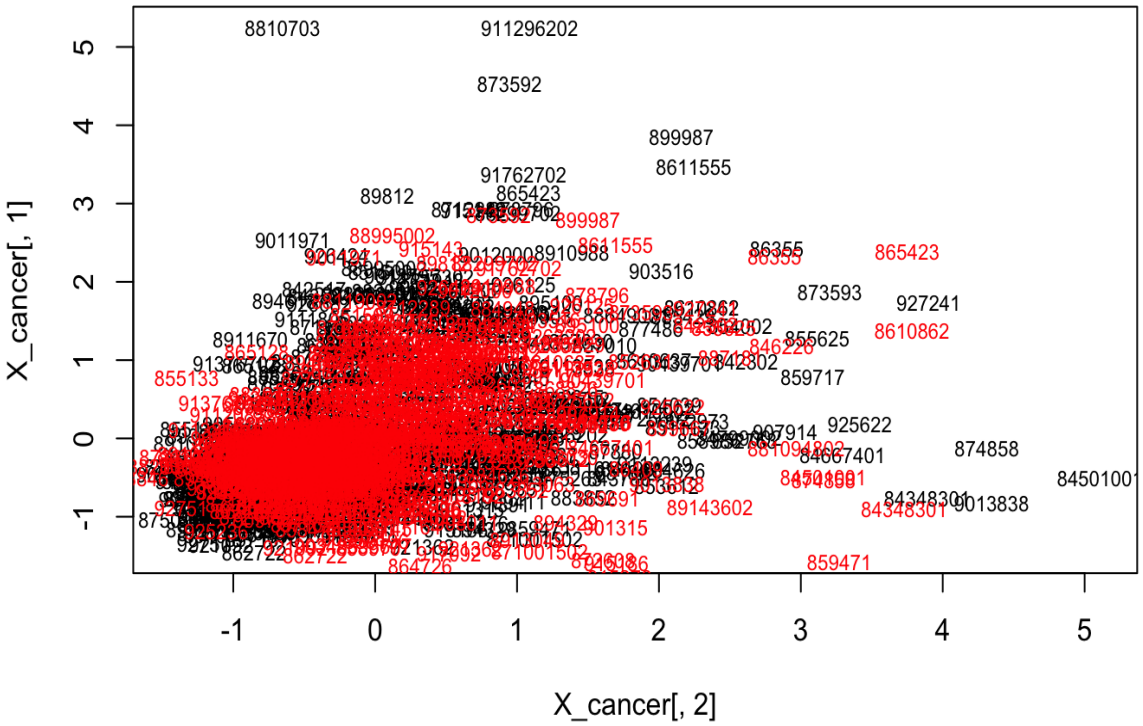

Figure 8: Choose the optimal K

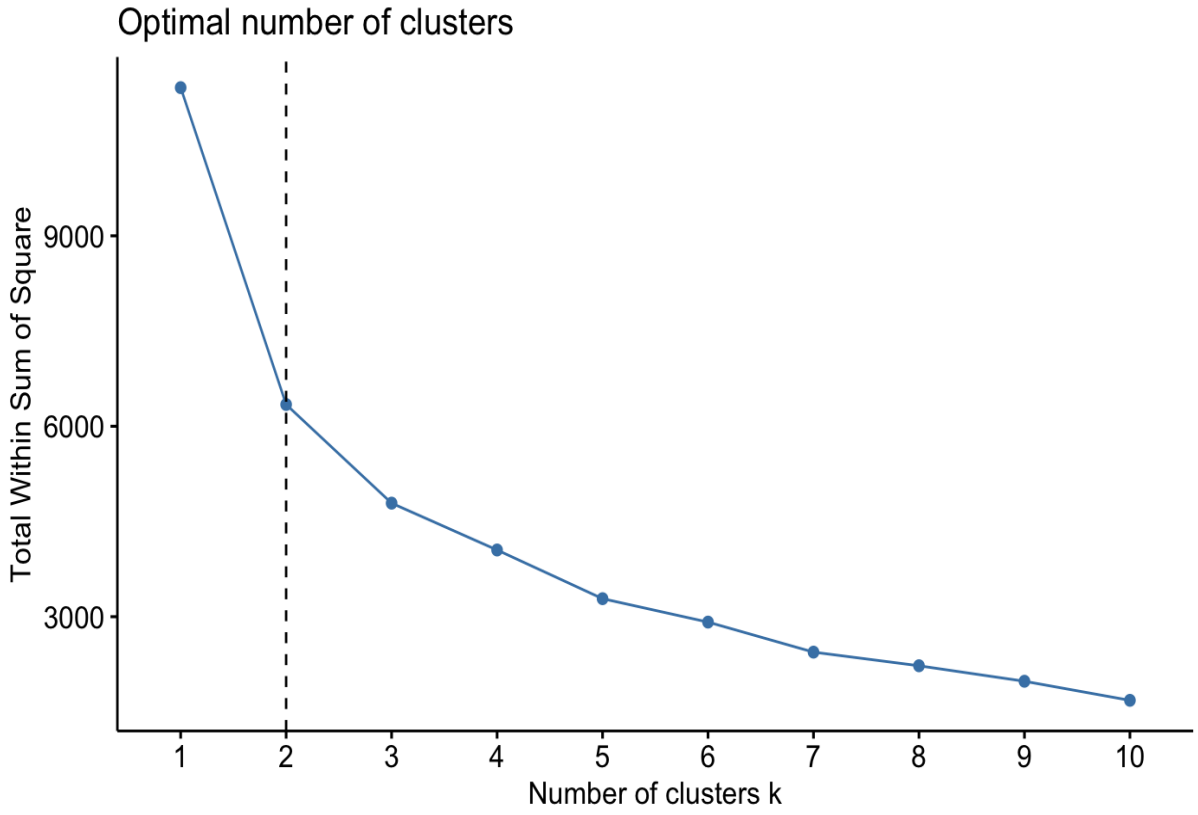

Figure 9: cluster graph

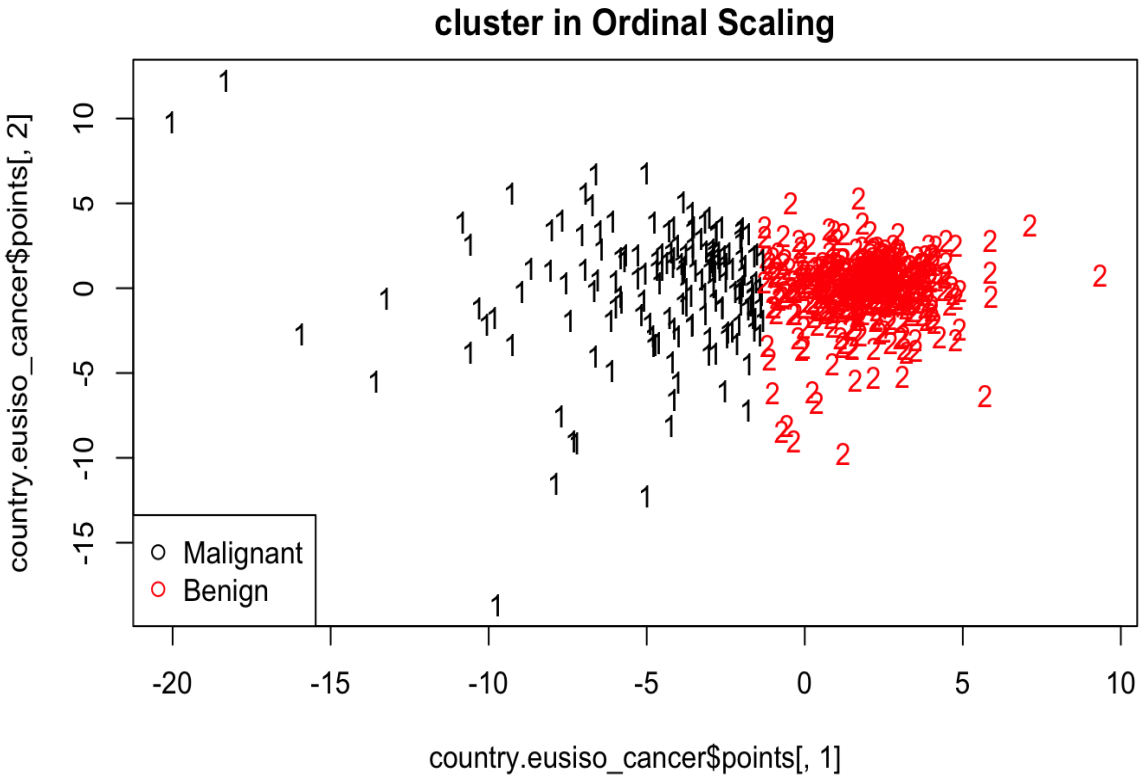

Figure 10: comparison cluster graph