Analysis of Breast Cancer dataset

I. Introduction

I analyze Breast Cancer data obtained from Kaggle[1], which has been previously analyzed by others. This dataset has **569** observations of Breast Cancer cell's measured data, which includes the following 31 variables:
mean radius, mean texture, mean perimeter, mean area, mean smoothness, mean compactness, mean concavity, mean concave points, mean symmetry, mean fractal dimension, radius error, texture error, perimeter error, area error, smoothness error, compactness error, concavity error, concave points error, symmetry error, fractal dimension error, worst radius, worst texture, worst perimeter, worst area, worst smoothness, worst compactness, worst concavity, worst concave points, worst symmetry, worst fractal dimension, **Target (1/0 represent whether the patients need to do targeted therapy)**

I conducted two-cluster clustering using the EM algorithm. Obtained one cluster (cluster 1) with **171** observations and another cluster (cluster 2) with **398** observations. I modified what was needed. Column "**Target**" was not used in the EM algorithm, but it will help us to determine the cluster that requires a targeted therapy later. So, the number of variables that we are going to use in clustering will be 30.

II. Cluster Visualization

Since we have lots of variables, scatterplots views using different pairs of variables is recommended. I have selected a few pairs of variables to visualize them to help us compare and understand the difference between cluster 1 and cluster 2.

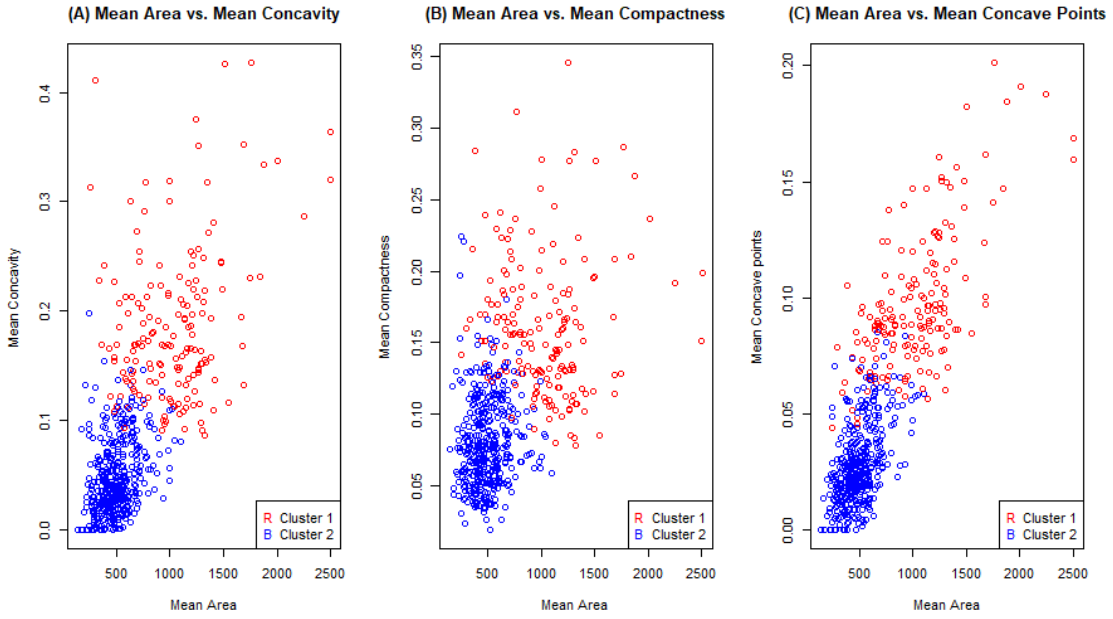

Figure 1: Viewing the clusters through the eyes of the different dimensions of the data

Scatterplot (A) in Figure 1 shows that the Mean Area and the Mean Concavity of the breast Cancer cell partition the data into a cluster of lower Mean Area with lower Mean concavity and another cluster with higher Mean Area with higher Mean Concavity. Scatterplot (B) of Figure 1 shows that the Mean Area and the Mean Compactness of the breast Cancer cell partition the data into a cluster of lower Mean Area with Lower Mean Compactness and another cluster of higher Mean Area with higher Mean Compactness. Finally, Scatterplot (C) of Figure 1 shows that the Mean Area and Mean Concavity Points of the cell partition the data into a cluster of lower Mean Area with lower Mean Concavity Points, but with a little bit overlap between the clusters. The plot indicates again that the Mean Area certainly helps distinguish two groups in the data.

Thus, by the visual inspection, I would conclude that there is one cluster of lower Mean Area (around 0-750) and another cluster of higher Mean Area (around 750+). We will further do some Numerical Analysis below.

III. Numerical Summaries

Firstly, compare the means of the variables above. Since we have too many variables, I only show the variables that we are exploring above.

Mean	mean.area	mean.concavity	mean.compactness	mean.concave.points
Cluster 1	1028.0181	0.18625292	0.16105439	0.09766041
Cluster 2	494.5749	0.04692854	0.07997417	0.02797755

Table 1. Means of a part of clustering variables by cluster

Table 1 confirms what we observed in Figure 1 (A, B, C), that the clusters differ in average Mean area. I would label cluster 1 as the patients who have higher Breast Cancer cells' mean area, and I would label cluster 2 as the patients who have lower Breast cancer cells' mean area. This is a good clustering for the doctors to decide whether and what therapy we should do for the patients. If we further look at the mean concavity, mean compactness and mean concave points, we can also find that the cells with higher mean area in cluster 1 usually have higher mean concavity and higher mean compactness and higher mean concave points. Also, the cells with lower mean area in cluster 2 usually have lower mean concavity, lower compactness, and lower concave points.

Secondly, we check the median of them.

Median	mean.area	mean.concavity	mean.compactness	mean.concave.points
Cluster 1	1027.0000	0.16900000	0.15530000	0.09029000
Cluster 2	476.4000	0.03980500	0.07578500	0.02537500

Table 2. Median of a part of clustering variables by cluster

By looking at the Median table, it has the similar indication as the Mean Table 1 above.

Thirdly, we check the Standard Deviation of them.

Standard Deviation	mean.area	mean.concavity	mean.compactness	mean.concave.points
Cluster 1	389.4628	0.07029639	0.04929546	0.03053978
Cluster 2	162.7916	0.03359003	0.03113455	0.01708406

Table 3. Standard Deviation of a part of clustering variables by cluster

By looking at the above table 3 of standard deviation of the variables, we found that Cluster 1 tends to have more variability in Mean area, mean concavity, mean compactness, and mean concave points. Another cluster has lower variability in Mean area, mean concavity, mean compactness, and mean concave points.

III.1 Labeling by Mean Area?

Based on the scatterplots A, B, C from Figure 1, we see that it is clear that the Mean Area partitions the data into a cluster of significantly higher Mean Area and another cluster of significantly lower Mean Area. Furthermore, by comparing the Mean, Median, and Standard deviation of the Mean Area in two clusters, we can also find that Mean Area in two clusters are significantly different. The mean of the Mean Area in cluster 1 is 1028.02, and the mean of the Mean Area in cluster 2 is 494.575.

Thus, we can conclude that we are able to label these two clusters by the Mean Area.

IV. Principal Components Analysis

Since we have totally 30 variables in our analysis, we are going to try Principal Component Analysis to see if we can reduce the number of variables.

First, we centered and scaled the data because there are variables with different units. Then, we calculate the Variance covariance matrix of the centered and scaled data. The first few variables of the matrix is shown below:

	mean.radius	mean.texture	mean.perimeter	mean.area	mean.smoothness
mean.radius	1	0.32378189	0.9978553	0.9873572	0.17058119
mean.texture	0.3237819	1	0.3295331	0.3210857	-0.02338852
mean.perimeter	0.9978553	0.32953306	1	0.9865068	0.20727816
mean.area	0.9873572	0.32108570	0.9865068	1	0.17702838
mean.smoothness	0.1705812	-0.02338852	0.2072782	0.1770284	1

Table 4. First few variables of the variance covariance matrix of Xcs

From the above table, we see that there are high correlations between variables. Thus, we conclude that this dataset is **worth** doing a Principal Component Analysis.

Secondly, we calculated the Eigenvectors and eigenvalues of variance-covariance. The result of the eigenvalues are given below:

λ1	**λ2**	**λ3**	**λ4**	**λ5**	**λ6**
13.2816	5.6913	2.8179	1.9806	1.6487	1.2073
λ7	**λ8**	**λ9**	**λ10**	**λ11**	**λ12**
0.6752	0.4766	0.4168	0.3506	0.2939	0.2611
λ13	**λ14**	**λ15**	**λ16**	**λ17**	**λ18**
0.2413	0.1570	0.0941	0.0798	0.0593	0.05261
λ19	**λ20**	**λ21**	**λ22**	**λ23**	**λ24**
0.0494	0.0311	0.0299	0.0274	0.0243	0.0180
λ25	**λ26**	**λ27**	**λ28**	**λ29**	**λ30**
0.0154	0.0081	0.0069	0.0015	0.0007	0.00013

Table 5. Eigenvalues of the variance-covariance matrix

According to the ordered eigenvalues, if we want to preserve close to 90% variance, we should choose the first 7 variables. Just like the plot below:

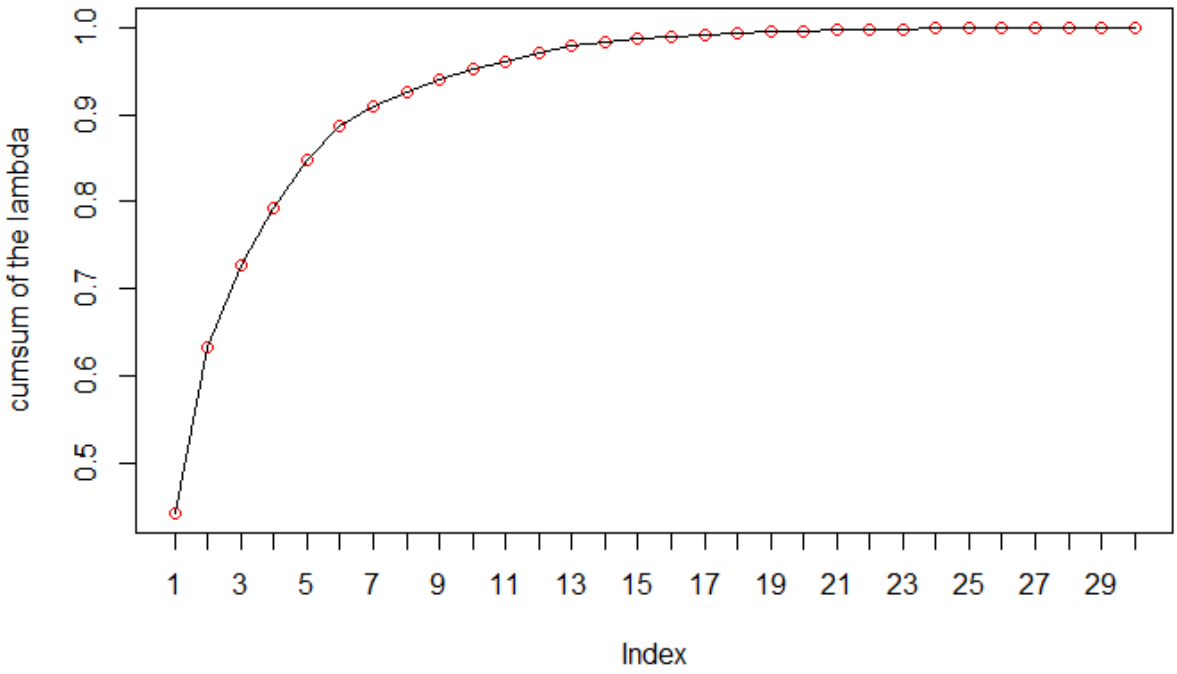

Figure 2. Cumulative sum of the PCs

Thus, if we want to keep 90% of the original variability in the data, we would need 7 PCs.

V. Estimation of Variable Distribution

Now, we are going to select an important variable and determine the probability model for this variable. We are going to use Newton Algorithm to find the Maximum likelihood Estimation of the parameters and asymptotic confidence intervals for the parameters.

Based on the above analysis, I think that the most important variable is “mean.area”. Now, let’s look at the histogram of data of this variable.

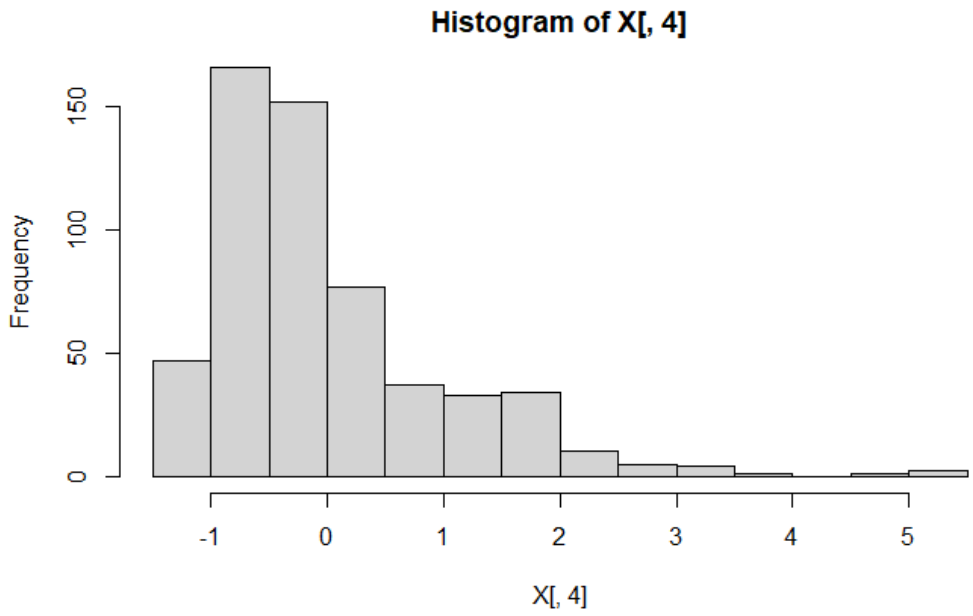

Figure 3. Histogram of variable mean.area of the data

By looking at the rough curve in the histogram, I decided to use Gamma distribution to apply to this variable.

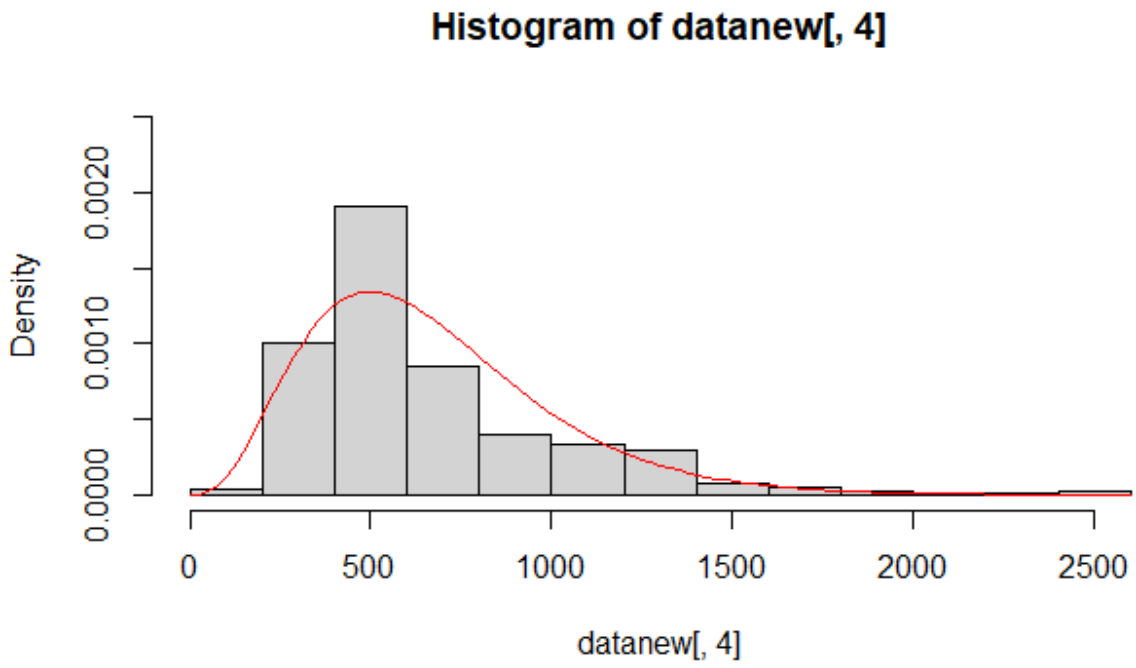

Figure 4. Histogram of variable mean.area with the Gamma distribution

By using the Newton algorithm to find the MLE of the parameters, the estimates of alpha and lambda are given by 4.282259349 and 0.006538908 after 10 iterations.

The final gradient vector is given by (1.682087e-07 -1.146033e-05)

The final hessian is given by:

-149.5832	87017.59
87017.5929	-56986873.43

Table 6. Final Hessian of the Newton Algorithm

The confidence interval for parameter alpha is (3.802783, 4.283036)

The confidence interval for parameter lambda is (-0.47293785, 0.00731573)

Then, we plot the curve into the histogram to double check.

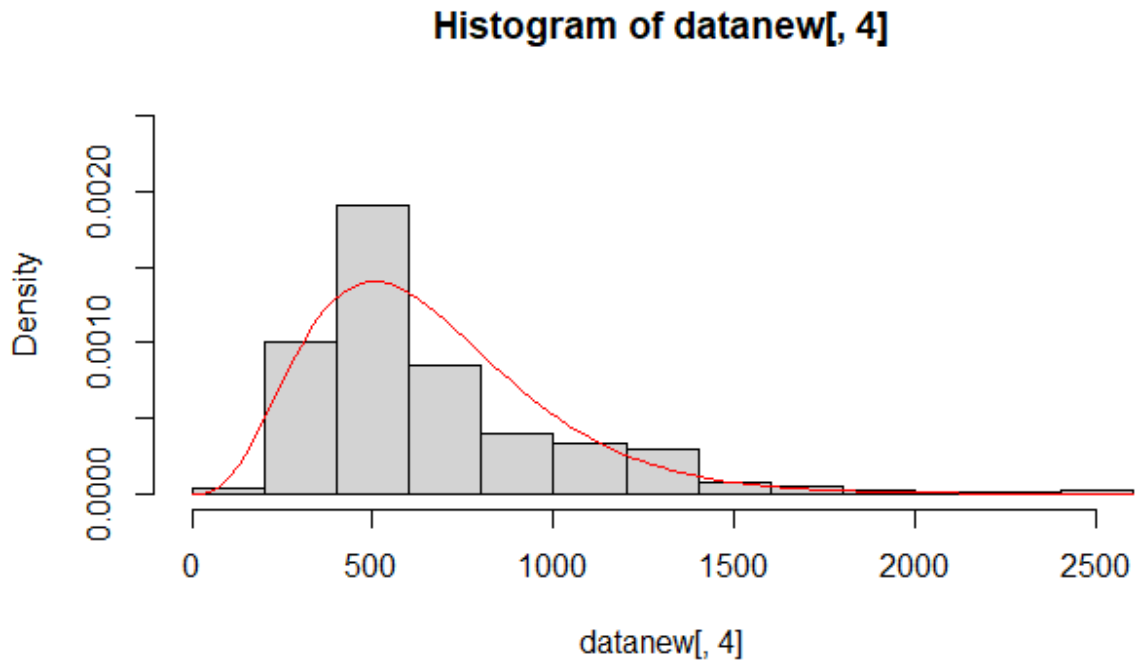

Figure 5. New curve comparing to Figure 4

VI. Conclusion

In this research, we have analyzed 569 observations of Breast Cancer cell's measured data with 31 variables.

1. From the clustering analysis by EM algorithm, we conclude that the data can be divided into two clusters. One cluster with significantly higher Mean Area, while another with lower Mean Area. I also chose to label these two clusters by the variable Mean Area. More importantly, we found that the mean value of variable "**Target**" in **cluster 1** is **0.041**, and the mean value of variable "**Target**" in **cluster 2** is **0.88**. This means that almost all patients in cluster 2 need to do **targeted therapy**, and compared with this, most patients in **cluster 1** do **not** have to do the **targeted therapy**. We can then conclude that this cluster analysis helps doctors to determine if a Breast Cancer patient has to do targeted therapy based on their cancer cells measured data.
2. From the Principal Component Analysis, we found that there are high correlations between variables, and so it is worth doing a **PCA**. After doing the principal components analysis, we conclude that we can reduce **23 PCs** and remain **7 PCs** to keep **90%** of the original variability in the data.
3. The variable that I think is most important is mean.area. I decided to apply a **Gamma** probability model to this variable. By using Newton algorithm, we found the MLE of the parameters alpha and lambda are 4.282259349 and 0.006538908 after 10 iterations. We also found that the confidence interval for parameter alpha is (3.802783, 4.283036), and the confidence interval for parameter lambda is (-0.47293785, 0.00731573) at 5% significance level.

Reference

[1]https://www.kaggle.com/krishnabalanagu/clustering?select=ANN_Quiz_Data_Breast_Cancer.csv

[3] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/