Analysis of Breast Cancer Wisconsin Diagnosis

1 Introduction

In this paper, I will analyze Breast Cancer Wisconsin (Diagnostic) data set from UC Irvine Machine Learning Repository [1], which contains characteristics of the cell nuclei present in the image of a fine needle aspirate (FNA) of a breast mass. The data set has 569 observations of images with 32 variables. Each observation contains the ID number, diagnosis and the mean, standard error and "worst" or largest of each characteristic for each cell nucleus.

The 10 ordered characteristics of cell nuclei computed for each image of breast mass and its description provided are:

	Characteristics	Descriptions
a	radius	mean of distances from center to points on the perimeter
b	texture	standard deviation of gray-scale values
c	perimeter	
d	area	
e	smoothness	local variation in radius lengths
f	compactness	$perimeter^2$/ area - 1.0
g	concavity	severity of concave portions of the contour
h	concave points	number of concave portions of the contour
i	symmetry	
j	fractal dimension	mean for "coastline approximation" - 1

Table 1. Characteristics (Features) of cell nuclei for each image of breast tumor

The variables are:

No	Variable	Descriptions
1	id	ID number
2	diagnosis	The diagnosis of breast tissues (M = malignant, B = benign)
3 - 12	mean of characteristics	
13 - 22	standard error of characteristics	
23 - 32	worst of characteristics	

Table 2. Variables in data set

For example, variable 3 is the radius_mean, variable 13 is the radius_se and variable 23 is radius_worst. Data analysis was conducted without the first 2 variables as they are not interval variables. In the following subsection 1.1, I will discuss about the missing values and outliers in the data set.

1.1 Missing Values and Outliers in Data set

Upon observing the data set, there are no missing values in the data set.

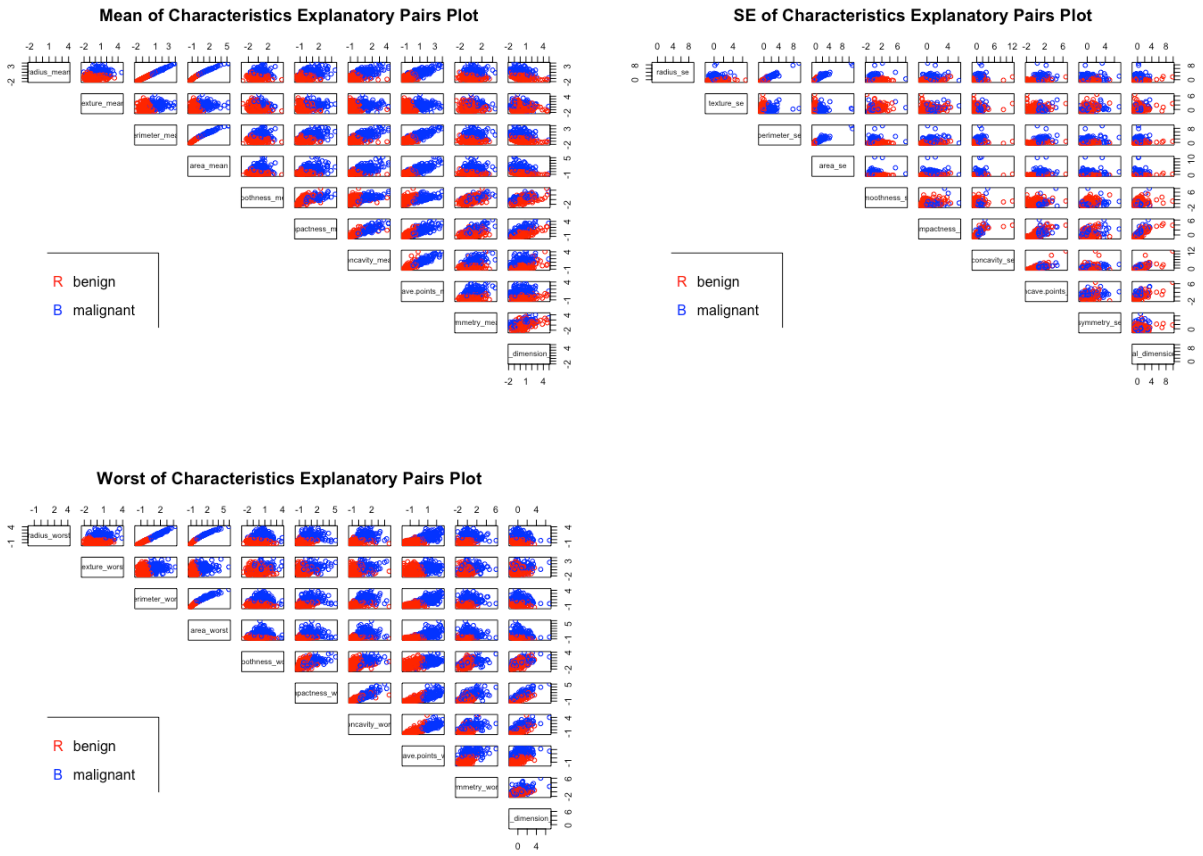

Figure 1. Pair plots of variables before removing outliers

To identify the outliers, I plotted the scaled variables as seen in Figure 1. Although there are a lot of observations, notice that there are many points that are beyond -3 and +3, indicating that those points are outliers, which are removed for the analysis.

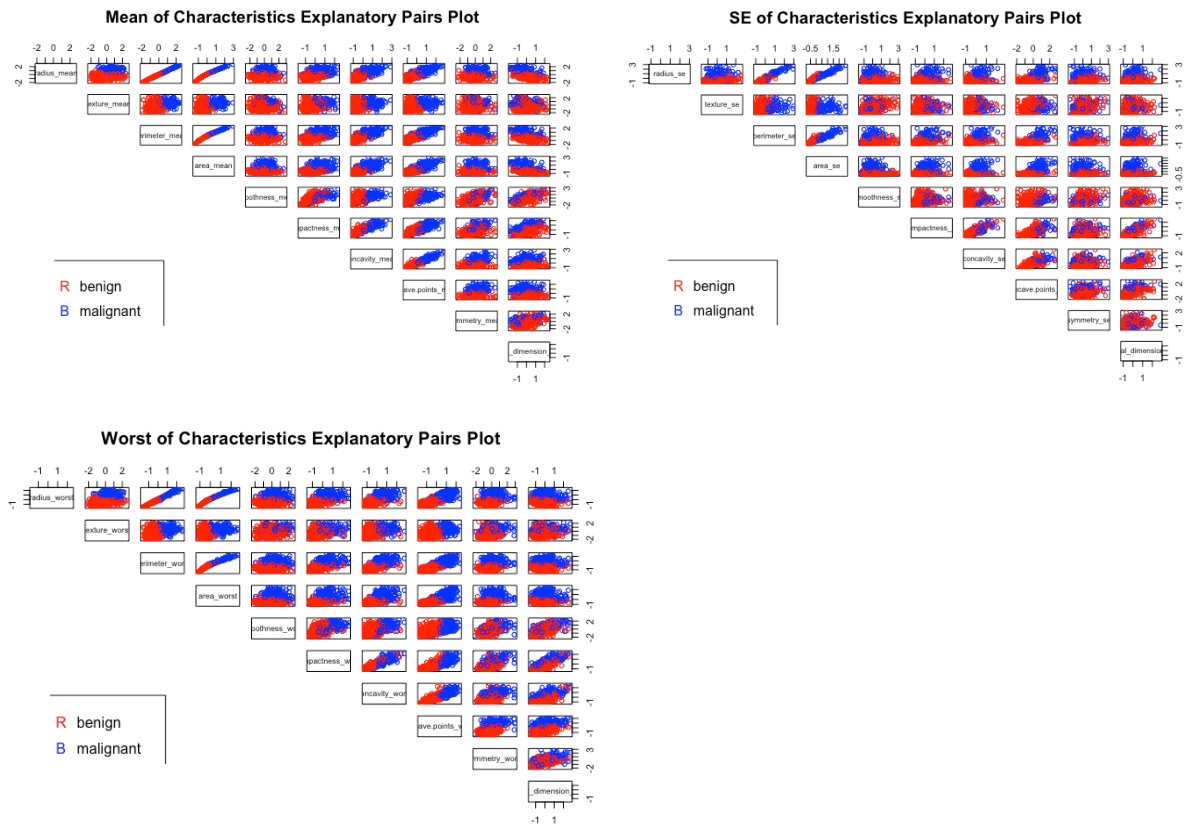

Figure 2. Pair plots of variables after removing outliers

From Figure 2, we can confirm that the outliers were removed from the data set as all points fall within -3 and 3. A total of 74 observations were removed and the data set now contains 495 observations.

I then performed the analysis of the new scaled data set using cluster analysis and principal of components in Section 2 and 3 respectively. Then, I chose the most important variable and determined a model in Section 4. Lastly, I conclude my findings in Section 5. In the following Section 2, I will discuss my findings from doing cluster analysis

2 Cluster analysis

I conducted a cluster analysis using the k-means algorithm as given in the program for lesson I-10 of my Stat 102B class [2]. I obtained the results that cluster 1 and cluster 2 has 355 and 140 breast mass images respectively. I modified what was needed and began my analysis using the program. In Section 2.1, discuss the summary statistics obtained and in Section 2.2, I will be using scatterplots to further support the finding. In Section 2.3, I discuss whether the type of tumor (malignant or benign) plays a role in clustering and conclude in Section 2.4.

2.1 Numerical Summaries

variables		mean		median		standard deviation	
		cluster 1	cluster 2	cluster 1	cluster 2	cluster 1	cluster 2
mean	**radius**	**12.4922**	**17.3598**	**12.4300**	**17.5850**	**1.8695**	**2.6250**
	texture	18.1021	20.8871	17.6400	20.7800	3.8963	3.4122
	perimeter	**80.2467**	**114.4910**	**79.7800**	**116.7000**	**12.3600**	**17.3416**
	area	**490.8580**	**956.7071**	**476.3000**	**972.1000**	**148.7425**	**281.8441**
	smoothness	0.0922	0.1024	0.0906	0.1026	0.0124	0.0107
	compactness	0.0762	0.1417	0.0728	0.1347	0.0265	0.0334
	concavity	**0.0421**	**0.1553**	**0.0374**	**0.1485**	**0.0276**	**0.0484**
	concave points	0.0263	0.0865	0.0240	0.0867	0.0149	0.0241
	symmetry	0.1721	0.1915	0.1707	0.1906	0.0219	0.0208
	fractal dimension	0.0615	0.0624	0.0610	0.0620	0.0053	0.0062
se	**radius**	**0.2819**	**0.5769**	**0.2562**	**0.5539**	**0.1149**	**0.2277**
	texture	1.1569	1.1408	1.0660	1.0685	0.4877	0.3853
	perimeter	**1.9676**	**4.0579**	**1.8170**	**3.8055**	**0.7831**	**1.5601**
	area	**21.8651**	**65.5829**	**19.9100**	**60.0550**	**10.1376**	**32.5459**
	smoothness	0.0068	0.0067	0.0061	0.0062	0.0026	0.0020
	compactness	0.0182	0.0320	0.0152	0.0301	0.0102	0.0130
	concavity	**0.0213**	**0.0415**	**0.0181**	**0.0385**	**0.0148**	**0.0162**
	concave points	0.0093	0.0150	0.0089	0.0143	0.0041	0.0041
	symmetry	0.0195	0.0191	0.0184	0.0179	0.0061	0.0061
	fractal dimension	0.0030	0.0041	0.0027	0.0039	0.0015	0.0016
worst	**radius**	**13.8384**	**20.9396**	**13.6400**	**21.1400**	**2.1860**	**3.4709**
	texture	23.9609	28.1736	23.2100	27.9550	5.4990	4.9257
	perimeter	**89.8736**	**139.6831**	**88.3300**	**140.2000**	**14.6833**	**22.5316**
	area	**600.7580**	**1378.4270**	**567.9000**	**1360.5000**	**193.5279**	**443.4641**
	smoothness	0.1255	0.1444	0.1256	0.1436	0.0192	0.0194
	compactness	0.1769	0.3690	0.1696	0.3583	0.0781	0.1179
	concavity	**0.1601**	**0.4458**	**0.1453**	**0.4172**	**0.1029**	**0.1340**
	concave points	0.0774	0.1808	0.0786	0.1785	0.0352	0.0339
	symmetry	0.2712	0.3156	0.2688	0.3092	0.0417	0.0540
	fractal dimension	0.0776	0.0914	0.0766	0.0897	0.0111	0.0169

Table 3. Means, Median and Standard deviations of clustering variables by cluster

Since there are many variables in the data set, I reported all the summary statistics for all variables by cluster. In particular, I bolded several notable characteristics of the breast mass in Table 3, that are, area, perimeter, radius and concavity. Referring to Table 3, these variables average, median and standard deviations have a strikingly large difference by cluster. For instance, the average area_mean is much higher in cluster 2 compared to cluster 1. The median and standard deviation of area_mean is also much higher in cluster 2 compared to cluster 1. The same pattern follows for area_se and area_worst. As the summary statistics are different for different cluster, the area of the cell nuclei of a breast mass can be used to characterize the clusters. Similarly, the concavity, perimeter and radius of the cell nuclei of a breast mass can be used to characterize the clusters because summary statistics are much higher in cluster 2 compared to cluster 1. This is expected for radius and perimeter as area is dependent on perimeter and radius.

Furthermore, from the numerical summaries in Table 3, in all aspects of the characteristics of the cell nuclei, cluster 1 always has a smaller value compared to cluster 2. Hence, I think cell nuclei of cluster 1 is generally smaller in size and has less abnormal shape/structure compared to that of cluster 2.

2.2 Cluster visualization

As mentioned in Section 2.1, since area is dependent on perimeter and radius, I will be choosing only the area and the concavity of the cell nuclei in the breast mass image to focus on.

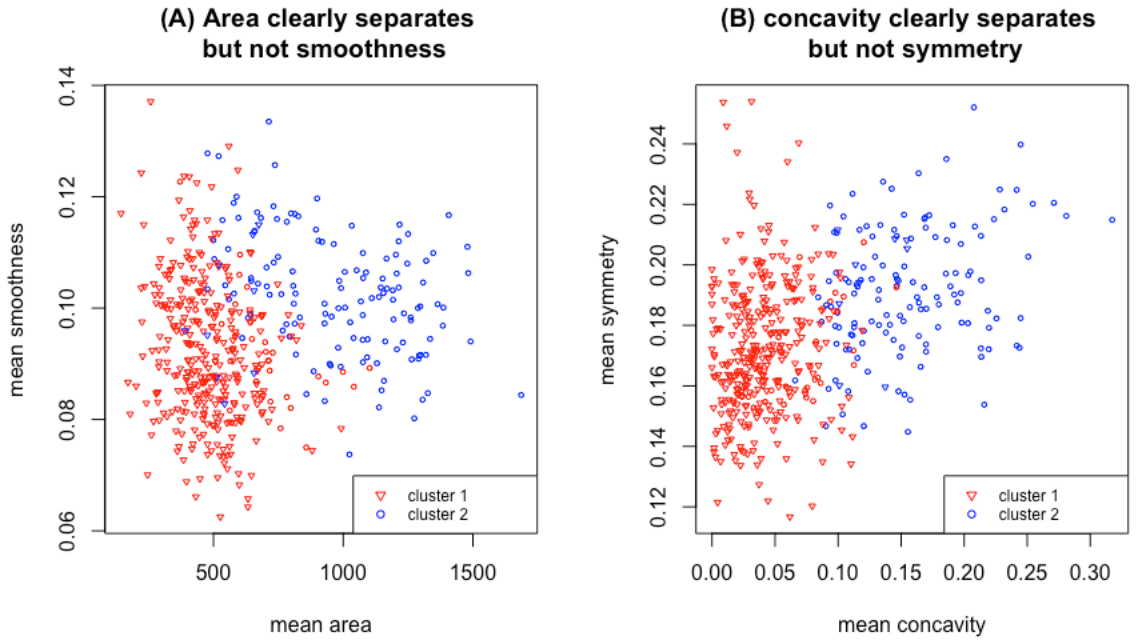

Figure 3. Viewing the clusters through the eyes of different dimensions

From Figure 3, we can see that it supports findings in Section 2.1, that concavity and area are important labels for the clusters. Although there is a slight mixture om clusters, we can see from

(A) and (B) that area and concavity separates the cluster well with cluster 1 having smaller average area and concavity while the symmetry and smoothness of the cell nuclei is all over the place for both clusters.

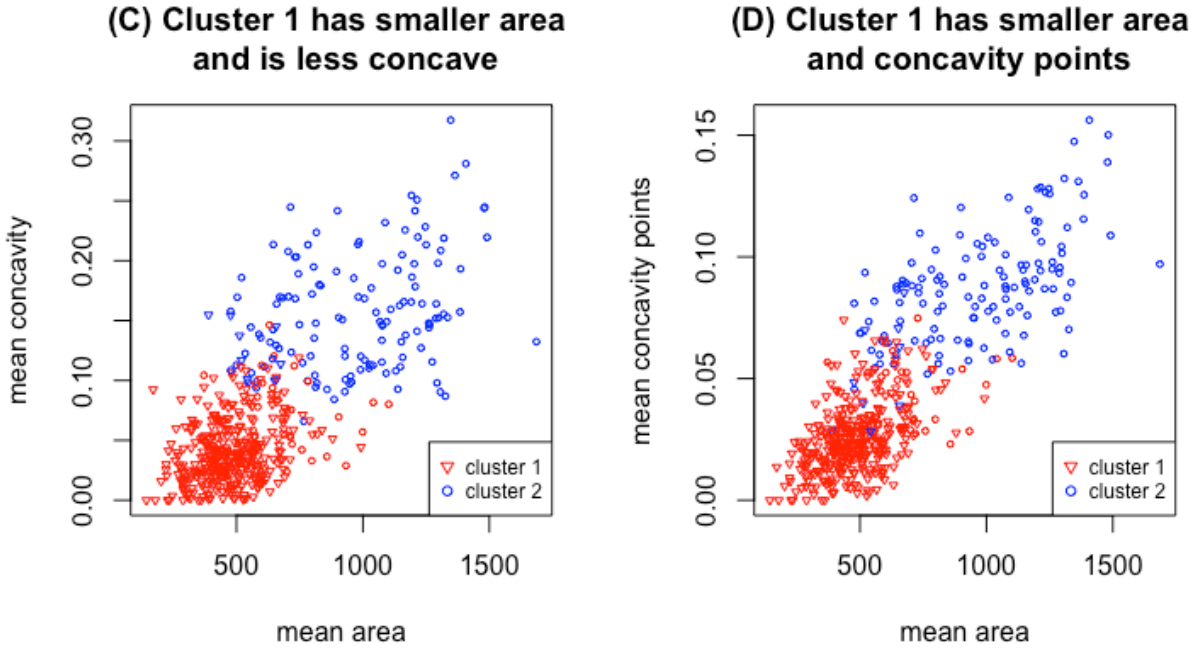

Figure 4. Viewing the clusters through the eyes of different dimensions

Also, upon looking at the various scatterplots, the two plots in Figure 4 further supports that there is a distinct difference that separates the 2 clusters. Cell nuclei in cluster 1 has a smaller average area, is on average less concave and has lesser concavity points. On the other hand, cell nuclei in cluster 2 have higher mean area, is on average more concave and has more concavity points. Cluster 1 generally has a smaller range of area and concavity compared to cluster 2.

From the summaries and scatterplots obtained, I would conclude that there is a cluster of cell nuclei of breast tumor with smaller shape and less structure in the sense that it has less concavity, while another cluster has a generally larger shape and more structure compared to the other cluster. Age and concavity seem to be key variables in separating the groups.

2.3 Labelling by diagnosis

	malignant	benign
Cluster 1	0.910	0.090
Cluster 2	0.064	0.936

Table 4. Proportion of diagnosis in each cluster (rounded to 3dp)

In table 4, we can see that the diagnosis for breast tumor in cluster 1 is approximately 91% malignant (cancerous) and in cluster 2 is approximately 93% benign (non-cancerous). Hence, I can conclude that the diagnosis plays a role in segmenting the cell nuclei in the breast tumor.

2.4 Conclusion

From the cluster analysis using k-means algorithm, I have discovered 2 clusters, differing in area, concavity and diagnosis. One cluster has a set of 355 with smaller area of cell nuclei, less concavity in the cell nuclei and diagnosis of malignant tumor. The other cluster has set of 140 with larger area of cell nuclei, more concavity in the cell nuclei and diagnosis of benign tumor. Based on these two clusters, I would distinguish the type of breast tumor based on the two important characteristics of the cell nuclei – area and concavity.

3 Principal Components analysis

After checking the variance covariance matrix of the scaled and centered data set, I found that there are several high correlations between variables (as high as 0.998). Figure 2 confirms this as well as we can see that several variables are highly correlated. thus, it is worth doing the analysis to learn more. I then conducted a principal components analysis using the program for lesson II-2 of my Stat 102B class [2]. I modified what was needed and began my analysis. In Section 3.1, I will discuss the variance explained by each PC a variable and how many I chose to keep. In Section 3.2, I will discuss the latent variables in the data set and conclude in Section 3.3.

3.1 Principal Components to keep

I summarized the eigenvalues, percentage proportion and cumulative proportion of variability explained by each principal components.

PC	Eigenvalue	Proportion	Cumulative	PC	Eigenvalue	Proportion	Cumulative
1	13.6865	45.6215	45.6215	16	0.0707	0.2358	99.1174
2	5.2190	17.3965	63.0180	17	0.0519	0.1728	99.2903
3	2.8638	9.5461	72.5642	18	0.0407	0.1358	99.4261
4	1.9081	6.3602	78.9243	19	0.0357	0.1190	99.5451
5	1.7700	5.9000	84.8244	20	0.0325	0.1083	99.6534
6	1.2444	4.1480	88.9724	21	0.0236	0.0787	99.7321
7	0.7132	2.3774	91.3498	22	0.0197	0.0656	99.7976
8	0.4997	1.6657	93.0154	23	0.0190	0.0632	99.8608
9	0.3869	1.2897	94.3051	24	0.0142	0.0473	99.9081
10	0.3363	1.1211	95.4262	25	0.0108	0.0359	99.9440
11	0.3051	1.0170	96.4431	26	0.0081	0.0269	99.9709
12	0.2683	0.8945	97.3376	27	0.0067	0.0222	99.9931
13	0.2007	0.6689	98.0065	28	0.0014	0.0047	99.9978
14	0.1586	0.5285	98.5350	29	0.0005	0.0018	99.9996
15	0.1040	0.3466	98.8817	30	0.0001	0.0004	100

Table 5. Eigenvalues and the percentage proportion of variation explained by principal components (rounded to 4 dp)

Taking the sum of all the eigenvalues, we get a total variance of 30. In table 5, the first principal component explains the largest amount of variability in the data, that is approximately 42.6%. the second principal component explains the second largest amount of variability in the data, that is, 17.4%, so on and so forth. The first 7 principal components explain 90% of the variation. So, I chose to keep the first 7 principal components.

3.2 Interpretation of PC

To interpret the principal components, I computed the correlations between the scaled and centered data and each principal components.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
radius_mean	**-0.829**	-0.513	-0.037	-0.062	0.000	-0.014	-0.045
texture_mean	-0.391	-0.074	0.342	**0.781**	-0.088	-0.015	0.039
perimeter_mean	**-0.855**	-0.472	-0.041	-0.061	0.006	-0.013	-0.044
area_mean	**-0.835**	-0.504	0.001	-0.074	-0.019	-0.004	-0.008
smoothness_mean	-0.430	0.539	-0.138	-0.200	-0.475	0.322	-0.104
compactness_mean	**-0.858**	0.385	-0.147	-0.042	0.040	0.015	-0.023
concavity_mean	**-0.944**	0.109	-0.049	-0.017	0.061	0.028	-0.119
concave.points_mean	**-0.947**	-0.037	-0.047	-0.109	-0.103	0.076	-0.111
symmetry_mean	-0.406	0.441	-0.031	-0.129	-0.454	-0.448	-0.033
fractal_dimension_mean	-0.093	**0.890**	-0.098	-0.045	-0.020	0.140	0.252
radius_se	**-0.761**	-0.164	0.449	-0.201	-0.151	0.047	0.273
texture_se	-0.024	0.211	**0.757**	0.355	-0.183	0.071	-0.100
perimeter_se	**-0.784**	-0.117	0.447	-0.185	-0.067	0.008	0.235
area_se	**-0.827**	-0.282	0.322	-0.176	-0.110	0.041	0.233
smoothness_se	0.030	0.558	0.516	-0.147	-0.271	0.289	-0.107
compactness_se	-0.618	0.506	0.204	0.004	0.472	-0.146	-0.038
concavity_se	-0.672	0.358	0.208	0.007	0.464	-0.077	-0.232
concave.points_se	**-0.702**	0.233	0.368	-0.195	0.181	0.063	-0.328
symmetry_se	0.036	0.395	0.531	-0.224	-0.210	-0.541	-0.071
fractal_dimension_se	-0.386	0.666	0.262	-0.064	0.395	0.005	0.260
radius_worst	**-0.872**	-0.457	-0.073	-0.018	-0.063	0.005	0.056
texture_worst	-0.390	-0.047	0.198	**0.873**	-0.165	0.020	0.005
perimeter_worst	**-0.896**	-0.413	-0.081	-0.009	-0.033	-0.006	0.049
area_worst	**-0.868**	-0.446	-0.037	-0.034	-0.080	0.018	0.096
smoothness_worst	-0.423	0.517	-0.329	0.054	-0.453	0.382	-0.076
compactness_worst	**-0.775**	0.344	-0.364	0.175	0.195	-0.086	0.023
concavity_worst	**-0.855**	0.172	-0.289	0.156	0.188	-0.017	-0.125
concave.points_worst	**-0.915**	0.010	-0.258	0.032	-0.015	0.062	-0.166
symmetry_worst	-0.391	0.303	-0.425	0.123	-0.394	-0.579	0.023
fractal_dimension_worst	-0.459	0.658	-0.387	0.174	0.154	0.054	0.316

Table 6. correlations between the principal components and the original variables (rounded to 3dp)

Here, I will consider correlation above 0.7 as important. Referring to table 6, I will discuss, if any, the possible latent variables for each principal components.

The first principal component increases with decreasing radius, perimeter, area, compactness, concavity and concave points. As we have seen in Section 2, where I conducted the cluster analysis, cluster 1 that has mainly malignant diagnosed breast tumor, tend to have cell nuclei with smaller radius, parameter, area, compactness, concavity and concave points. Hence, This latent variable for this component can be the diagnosis of a malignant breast tumor by looking at the cell nuclei.

The second principal component increases with increasing fractal dimension. It is also a contrast between mean smoothness, compactness, concavity, symmetry and fractal dimension, and, mean radius, texture, perimeter, area and concave points. In Section 2, when I conducted the cluster analysis, I found that in general malignant tumors have smaller values in all aspect of the features of the cell nuclei of a breast tumor. The benign tumors, that mainly falls under cluster 2, although it has higher values in all features, but it also has a wider variability in for all features as seen in Table 3. cluster This latent variable for this component can possibly be the cell nuclei of a breast tumor that can either be malignant or benign by looking at the image of the mass. In other words, the features of the cell nuclei in the image of the breast tumor have a mixture of typical features found in both malignant and benign tumor.

The third principal component is associated with high texture only. The same goes for the fourth principal component being is associated with high mean and worst/ largest texture. Recall that the texture measures the standard deviation of gray-scale values. I think that there is no meaningful latent variable for this principal of component. For principal component 5, 6, and 7, I think that there are no latent variables as the correlations are deemed fairly low.

Overall, I think that the first and second principal components are the most meaningful in terms of interpretation of latent variables. Also, they both have the highest variation, so they are more important as well.

3.3 Conclusion

After performing the principal components analysis, I chose 7 principal components to account for 90% variability in the data. I also found several latent variables in the first to fourth principal components that can be helpful in helping us diagnose a tumor based on an image of cell nuclei. I expected that we would be able to reduce the dimension greatly because the main features of the cell nuclei are only radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension while the data set contains the summaries of these 10 features, resulting in 30 variables. Furthermore, as radius, perimeter and area are linearly dependent, I expected that not all 3 features are needed to explain the variation.

4 Model

I think that the most important variable in this data set is mean area and that the log normal probabilistic model applies to this variable.

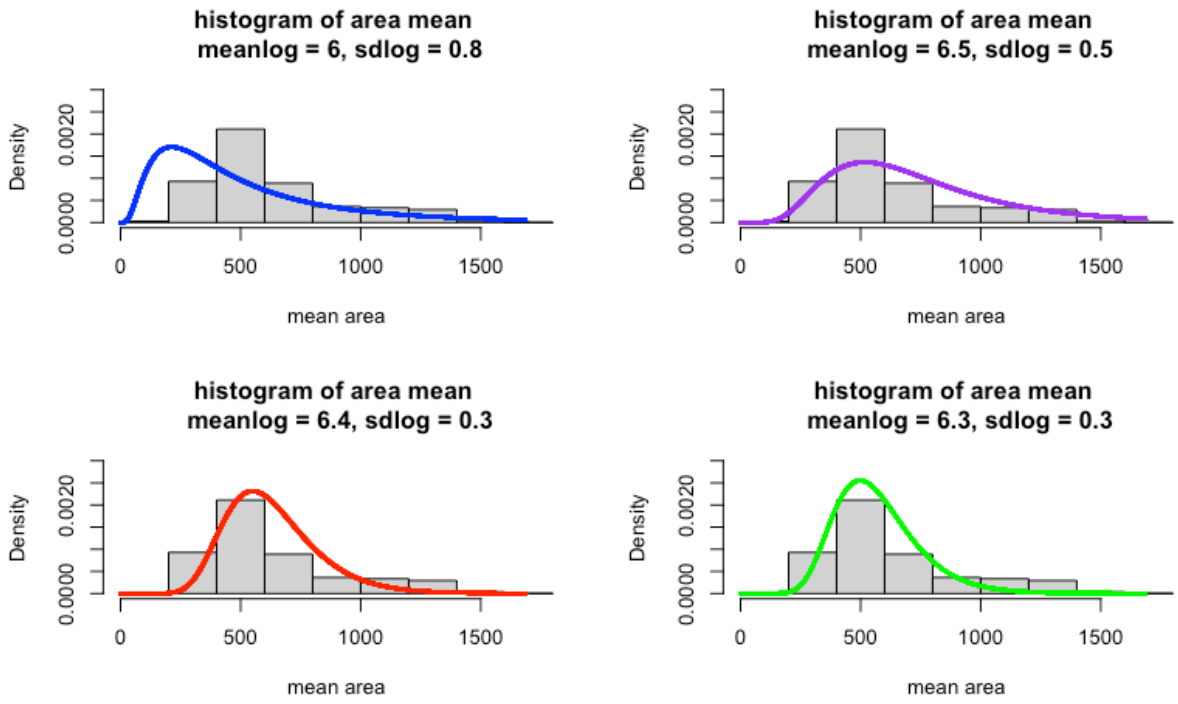

Figure 5. Histogram with different meanlog and sdlog

Based on Figure 5, the log normal probabilistic model with meanlog = 6.3 and sdlog = 0.3 fits the data set best (green color). So, that is used as my initial values for the program. After running the program, I found that the MLE of the parameters μ and σ are 6.3382741 and 0.4329019 respectively.

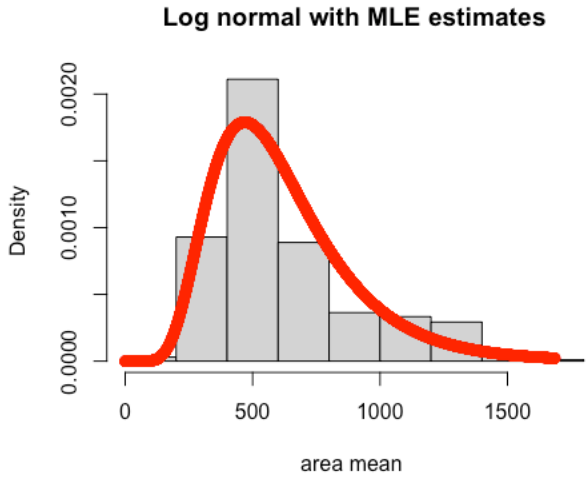

Figure 6. Histogram with MLE estimated

Figure 6 shows the fitted model using the MLE estimates I obtained. I expected the MLE estimates because I fitted the histograms with different values of μ and σ to find the model that has the most similar shape with the histograms of the data. Also, the 95% asymptotic confidence interval for μ is (6.300138, 6.376410) and the 95% asymptotic confidence interval for σ is (0.4059357, 0.4598682). This means that we are 95% confident that the true parameter μ in the population falls between 6.300138 and 6.376410. Furthermore, we are 95% confident that the true parameter σ in the population falls between 0.4059357and 0.4598682.

5 Conclusion

In this paper, I have conducted cluster analysis and principal of components analysis on the Breast Cancer Wisconsin (Diagnostic) data set. Through cluster analysis, I found that the data can be separated into 2 clusters. One with most malignant breast tumors and the other with most benign breast tumors. Furthermore, we can also use area and concavity to characterize the clusters as they are well separated compared to other characteristics of the cell nuclei. From the principal of components analysis, I found that we can reduce the number of variables to 7 principal of components and have observed some latent variables in some of the principal components such as diagnosis of malignant tumors from image and the characteristics of a cell nuclei tumor being either malignant or benign based on the image of the cell nuclei. Last but not least, I have also used a log normal probabilistic model to model the most important variable, area. I think that further analysis can be done in order to extract more information from the data set and possibly use the data set to predict from the images if the tumor is malignant or benign.

6 References

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Sanchez, J. Stat 102B lecture notes and R code