Analysis of Breast Cancer Data

Jamie Lee

I. Introduction

I.a) Explanation of Data

I will analyze the 'Breast Cancer Wisconsin (Diagnostic)' dataset posted on Kaggle[1], which was originally collected and published by the UCI Machine Learning Repository. The data set has 569 observations. Each observation consists of a unique ID number, the diagnosis of breast tissues, and 30 other numerical (interval) variables.

The 30 numerical variables are all interval values, which makes our dataset really nice to work with. These 30 variables are described as below:

For each cell nucleus, ten real-valued features are computed:
a) radius: mean of distances from center to points on the perimeter (unit: millimeters/mm)
b) texture: standard deviation of gray-scale values (unit: none)
c) perimeter: the length of the boundary of the nucleus (unit: mm)
d) area: the surface area (unit: mm^2)
e) smoothness: local variation in radius lengths (unit: none)
f) compactness: perimeter^2 / area - 1.0 (unit: none)
g) concavity: severity of concave portions of the contour (unit: none)
h) concave points: number of concave portions of the contour (unit: none)
i) symmetry: a measure of how similar the shape of the left half of the cell is to the right half (unit: none)
j) fractal dimension: "coastline approximation" - 1 (unit: none)
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.

The diagnosis of breast tissues variable can take one of 2 values - M for malignant or B for benign.

I.b) Explanation of Aim & Motivating Questions

The aim of my analysis is to use these 30 variables to create a method that can accurately predict if a patient's breast tumour should be diagnosed as malignant or benign.
If such a method does exist, a natural question is whether it is possible to finetune a similar, more advanced and accurate, yet still cost-feasible program and implement it in hospitals so female patients can receive quick and correct diagnoses, and doctors can have more time to spend on other cases.

I.c) Overview of Methods used

I first start off the analysis with 2-cluster clustering using the nonprobabilistic k-means clustering algorithm given in Lesson I-10 [2]. I obtained one cluster (cluster 1) with 188 patients and another cluster (cluster 2) with 381 patients. The classification was 91.2% accurate.
I then proceed to perform PCA, using the algorithm given in Lecture II-2 [3], to see if we can reduce the dimensionality of our data. The results of the PCA led me to select only 7 out of the 30 variables. Upon performing nonprobabilistic k-means clustering with these 7 variables only, we obtained 515 observations in cluster 1 and 54 observations in cluster 2. This classification was 87.0% accurate.
Lastly, I distinguished that mean radius is the most important variable in our analysis, and I fit a normal distribution to our sample of mean radius. I used Newton's Method, from Lecture I-7 [4], to find the MLE

II. 2-cluster clustering using non-probabilistic k-means

II.a) Visualization of Clusters

Table II-1: Table of patients' diagnosis with their cluster allocation

```
##                   diagnosis
## cluster_allocation   B   M
##                  1  13 175
##                  2 344  37
```

Our non-probabilistic k-means method classified 188 patients into cluster 1 and 381 patients into cluster 2. Note that there is a pretty clear distinction that benign diagnoses should fall in cluster 2, and malignant diagnoses should fall in cluster 1.
However, there are 50 misclassifications in total. So our classification using non-probabilistic k-means was 91.2% accurate.

Table II-2: Table of cluster percentages of each diagnosis

```
##  cluster     benign malignant
##        1 0.03641457 0.8254717
##        2 0.96358543 0.1745283
```

Table II-2 shows the column-wise percentages of Table II-1.
These percentages confirm that cluster 1 is meant to hold malignant diagnoses and cluster 2 is meant to hold benign diagnoses.
We also see that 17% of malignant diagnoses are misclassified as benign, which is quite troubing.

Table II-3: Table of diagnosis percentages of each cluster

```
##  cluster     benign  malignant
##        1 0.06914894 0.93085106
##        2 0.90288714 0.09711286
```

Table II-3 shows the row-wise percentages of Table II-1.

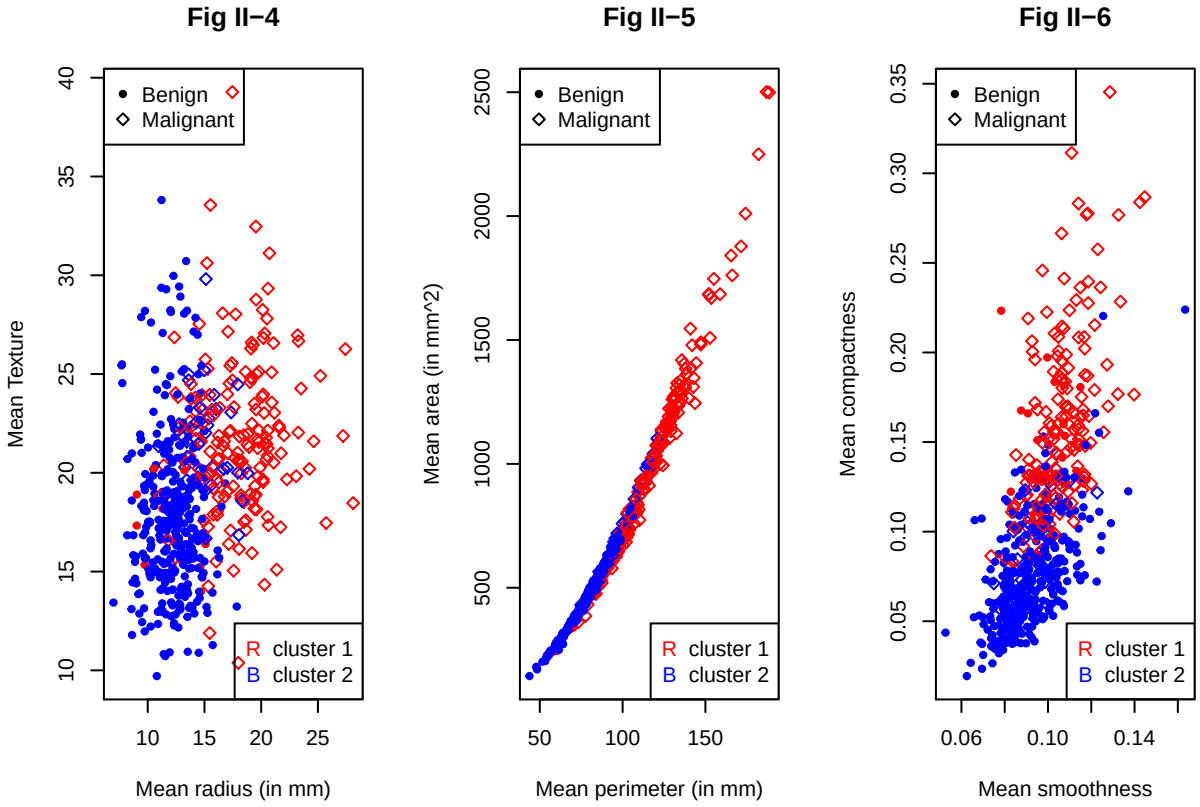
Figure II-4: Nonprobabilistic k-means Clustering: mean radius vs mean texture

Figure II-5: Nonprobabilistic k-means Clustering: mean perimeter vs mean area

Figure II-6: Nonprobabilistic k-means Clustering: mean smoothness vs mean compactness

Figure II-7: Nonprobabilistic k-means Clustering: mean concavity vs mean concave points

Figure II-8: Nonprobabilistic k-means Clustering: mean symmetry vs mean fractal dimension



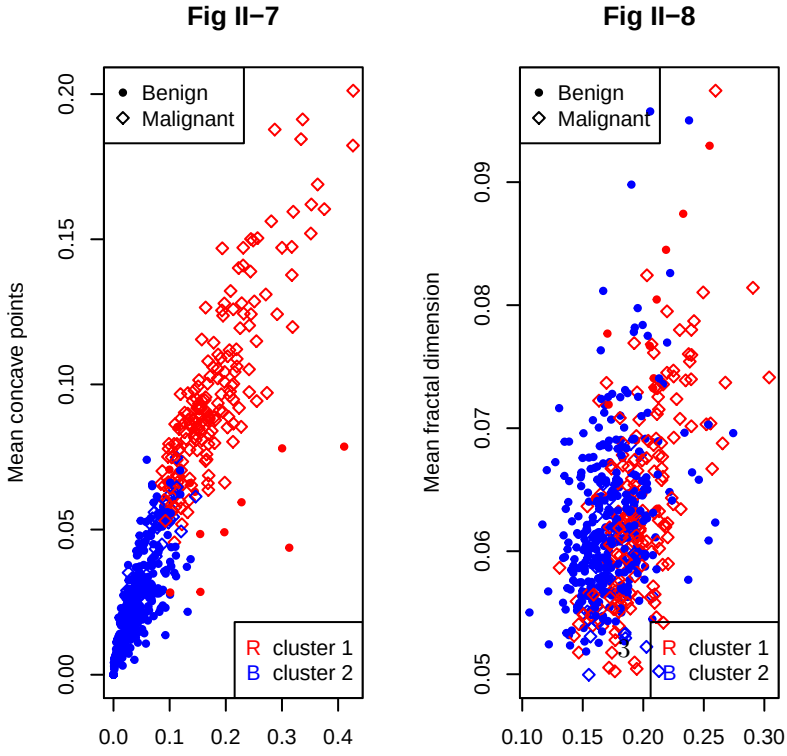


Figure II-4 doesn't show a very clear distinction in the clusters, but it shows that breast cancer cells with smaller radius tend to be benign. However, the spread of texture is pretty similar for both benign and malignant cells.
Figure II-5 doesn't show a very clear distinction in the clusters, but it shows that breast cancer cells with smaller perimeter and smaller area tend to be benign.
Figure II-6 shows a decently clear distinction in the clusters. It shows that breast cancer cells which are less smooth and less compact tend to be benign.
Figure II-7 shows a pretty clear distinction in the clusters, It shows that breast cancer cells which are less concave and have less concave points tend to be benign.
Figure II-8 does not show much distinction in the clusters. There is a large overlap in the 2 clusters, so it is difficult to use these 2 variables to classify our observations. Because of the overlap, it also doesn't really tell us much about whether benign cancer cells typically have low or high symmetry, and low or high fractal dimension.

II.b) Numerical Summaries

Table II-9: Table of mean values of each variable in cluster 1 and cluster 2

```
##                           cluster.1    cluster.2
## radius_mean             0.98563061 -0.486347912
## texture_mean            0.49158876 -0.242568734
## perimeter_mean          1.01777136 -0.502207391
## area_mean               0.97394027 -0.480579452
## smoothness_mean         0.58660145 -0.289451634
## compactness_mean        1.01318150 -0.499942579
## concavity_mean          1.14391592 -0.564451952
## concave.points_mean     1.16925384 -0.576954651
## symmetry_mean           0.60285975 -0.297474104
## fractal_dimension_mean  0.22907278 -0.113033290
## radius_se               0.86235794 -0.425520454
## texture_se              0.04412458 -0.021772759
## perimeter_se            0.86370531 -0.426185298
## area_se                 0.81306079 -0.401195350
## smoothness_se           0.01227864 -0.006058751
## compactness_se          0.69221011 -0.341562996
## concavity_se            0.63920256 -0.315407038
## concave.points_se       0.77098856 -0.380435301
## symmetry_se             0.13786622 -0.068028474
## fractal_dimension_se    0.40349482 -0.199099805
## radius_worst            1.05128810 -0.518745833
## texture_worst           0.51660223 -0.254911339
## perimeter_worst         1.07674731 -0.531308383
## area_worst              1.01302568 -0.499865690
## smoothness_worst        0.59751806 -0.294838308
## compactness_worst       0.95201745 -0.469761894
## concavity_worst         1.05051839 -0.518366032
## concave.points_worst    1.15227453 -0.568576407
## symmetry_worst          0.59888594 -0.295513272
## fractal_dimension_worst 0.61308060 -0.302517460
```

(Note: Recall that X is our scaled and centered data matrix, which explains the negative values for features like mean radius or mean perimeter, that in reality can only take positive values)
Table II-9 is consistent with our results in Figure II-4 through Figure II-8.
I'd like to comment additionally on the other 20 factors that I did not explore in my scatterplots, Fig II-4 through Fig II-8 - Taking a look at row 11 (radius_se), for example, we see that breast tissues with smaller

standard error in their radius tend to get a benign diagnosis. Similar observations can be made for all 20 factors.
Notice that, in fact, all 30 factors have the mean in cluster 1 greater than the mean in cluster 2. Recall that malignant diagnoses typically get classified into cluster 1, and benign into cluster 2. This implies that for all 30 variables, a breast tissue sample with a lower value of any of these 30 variables is more likely to be a benign breast tumour.

The table of medians for each cluster shows the same pattern and gives the same conclusions as Table II-9. So I omit that table, but I move on to show the table of standard deviations for each cluster.

Table II-10: Table of standard deviations of each variable in cluster 1 and cluster 2

```
##                            cluster.1    cluster.2
## radius_mean             3.424401e+00 1.958916e+00
## texture_mean            3.990252e+00 4.053085e+00
## perimeter_mean          2.306993e+01 1.291931e+01
## area_mean               3.871265e+02 1.552721e+02
## smoothness_mean         1.217311e-02 1.309983e-02
## compactness_mean        4.882135e-02 2.953109e-02
## concavity_mean          7.021802e-02 3.021996e-02
## concave.points_mean     3.129896e-02 1.563161e-02
## symmetry_mean           2.739239e-02 2.342818e-02
## fractal_dimension_mean  8.354604e-03 6.160019e-03
## radius_se               3.460642e-01 1.157962e-01
## texture_se              4.830356e-01 5.814732e-01
## perimeter_se            2.559716e+00 7.835579e-01
## area_se                 6.326075e+01 1.011452e+01
## smoothness_se           2.940788e-03 3.028446e-03
## compactness_se          2.012270e-02 1.284772e-02
## concavity_se            3.984509e-02 1.734000e-02
## concave.points_se       6.211718e-03 4.587158e-03
## symmetry_se             1.033386e-02 6.944129e-03
## fractal_dimension_se    3.192778e-03 2.137324e-03
## radius_worst            4.602686e+00 2.312290e+00
## texture_worst           5.776277e+00 5.694254e+00
## perimeter_worst         3.112408e+01 1.551214e+01
## area_worst              6.304860e+02 2.048806e+02
## smoothness_worst        2.184786e-02 2.011215e-02
## compactness_worst       1.625834e-01 8.560296e-02
## concavity_worst         1.835341e-01 1.134403e-01
## concave.points_worst    4.039873e-02 3.753045e-02
## symmetry_worst          7.471750e-02 4.404358e-02
## fractal_dimension_worst 2.171264e-02 1.278032e-02
```

Notice from Table II-10 that in general, across the 30 variables, the variables in cluster 1 and cluster 2 either have very similar standard deviation OR the variables in cluster 1 tend to have much higher standard deviation, and thus higher variability than those in cluster 2.
This shows that patients with malignant breast tumours tend to show more variability in their breast tissue features, which may make it harder to correctly diagnose a breast tumour as malignant.

III. PCA

III.a) Preparation to do PCA - Assess whether data correlation gives good signs to do PCA and choice of number of PCs

Table III-1: Variance-covariance matrix of centered and scaled data

```
##                  radius_mean texture_mean perimeter_mean area_mean
## radius_mean        1.0000000   0.32378189      0.9978553 0.9873572
## texture_mean       0.3237819   1.00000000      0.3295331 0.3210857
## perimeter_mean     0.9978553   0.32953306      1.0000000 0.9865068
## area_mean          0.9873572   0.32108570      0.9865068 1.0000000
## smoothness_mean    0.1705812  -0.02338852      0.2072782 0.1770284
## compactness_mean   0.5061236   0.23670222      0.5569362 0.4985017
##                  smoothness_mean compactness_mean concavity_mean
## radius_mean           0.17058119        0.5061236      0.6767636
## texture_mean         -0.02338852        0.2367022      0.3024178
## perimeter_mean        0.20727816        0.5569362      0.7161357
## area_mean             0.17702838        0.4985017      0.6859828
## smoothness_mean       1.00000000        0.6591232      0.5219838
## compactness_mean      0.65912322        1.0000000      0.8831207
##                  concave.points_mean symmetry_mean fractal_dimension_mean
## radius_mean                0.8225285    0.14774124            -0.31163083
## texture_mean               0.2934641    0.07140098            -0.07643718
## perimeter_mean             0.8509770    0.18302721            -0.26147691
## area_mean                  0.8232689    0.15129308            -0.28310981
## smoothness_mean            0.5536952    0.55777479             0.58479200
## compactness_mean           0.8311350    0.60264105             0.56536866
##                  radius_se  texture_se perimeter_se   area_se
## radius_mean      0.6790904 -0.09731744    0.6741716 0.7358637
## texture_mean     0.2758687  0.38635762    0.2816731 0.2598450
## perimeter_mean   0.6917650 -0.08676108    0.6931349 0.7449827
## area_mean        0.7325622 -0.06628021    0.7266283 0.8000859
## smoothness_mean  0.3014671  0.06840645    0.2960919 0.2465524
## compactness_mean 0.4974734  0.04620483    0.5489053 0.4556529
##                  smoothness_se compactness_se concavity_se
## radius_mean       -0.222600125      0.2060000    0.1942036
## texture_mean       0.006613777      0.1919746    0.1432931
## perimeter_mean    -0.202694026      0.2507437    0.2280823
## area_mean         -0.166776667      0.2125826    0.2076601
## smoothness_mean    0.332375443      0.3189433    0.2483957
## compactness_mean   0.135299268      0.7387218    0.5705169
##                  concave.points_se  symmetry_se fractal_dimension_se
## radius_mean              0.3761690 -0.104320881         -0.042641269
## texture_mean             0.1638510  0.009127168          0.054457520
## perimeter_mean           0.4072169 -0.081629327         -0.005523391
## area_mean                0.3723203 -0.072496588         -0.019886963
## smoothness_mean          0.3806757  0.200774376          0.283606699
## compactness_mean         0.6422619  0.229976591          0.507318127
##                  radius_worst texture_worst perimeter_worst area_worst
## radius_mean         0.9695390     0.2970076       0.9651365  0.9410825
## texture_mean        0.3525729     0.9120446       0.3580396  0.3435459
## perimeter_mean      0.9694764     0.3030384       0.9703869  0.9415498
## area_mean           0.9627461     0.2874886       0.9591196  0.9592133
## smoothness_mean     0.2131201     0.0360718       0.2388526  0.2067184
```

```
## compactness_mean    0.5353154     0.2481328       0.5902104  0.5096038
##                   smoothness_worst compactness_worst concavity_worst
## radius_mean            0.11961614        0.4134628       0.5269115
## texture_mean           0.07750336        0.2778296       0.3010252
## perimeter_mean         0.15054940        0.4557742       0.5638793
## area_mean              0.12352294        0.3904103       0.5126059
## smoothness_mean        0.80532420        0.4724684       0.4349257
## compactness_mean       0.56554117        0.8658090       0.8162752
##                   concave.points_worst symmetry_worst
## radius_mean                  0.7442142      0.1639533
## texture_mean                 0.2953158      0.1050079
## perimeter_mean               0.7712408      0.1891150
## area_mean                    0.7220166      0.1435699
## smoothness_mean              0.5030534      0.3943095
## compactness_mean             0.8155732      0.5102234
##                   fractal_dimension_worst
## radius_mean                   0.007065886
## texture_mean                  0.119205351
## perimeter_mean                0.051018530
## area_mean                     0.003737597
## smoothness_mean               0.499316369
## compactness_mean              0.687382323
```

Looking at the head of the variance covariance matrix, there is a large range in the values of the covariances. Notice that there is a decent proportion of high covariances (ex. > 0.7), so there is high correlation between rows and it is worth it to do PCA.

Table III-2: Individual & Cumulative sum of Proportion of variance of each PC

```
##  PC proportion.of.variance cumulative.proportion
##   1          44.2720256075              44.27203
##   2          18.9711820440              63.24321
##   3           9.3931632574              72.63637
##   4           6.6021349155              79.23851
##   5           5.4957684923              84.73427
##   6           4.0245220399              88.75880
##   7           2.2507337130              91.00953
##   8           1.5887238000              92.59825
##   9           1.3896493746              93.98790
##  10           1.1689781894              95.15688
##  11           0.9797189876              96.13660
##  12           0.8705379007              97.00714
##  13           0.8045249872              97.81166
##  14           0.5233657455              98.33503
##  15           0.3137832168              98.64881
##  16           0.2662093365              98.91502
##  17           0.1979967925              99.11302
##  18           0.1753959450              99.28841
##  19           0.1649253059              99.45334
##  20           0.1038646748              99.55720
##  21           0.0999096464              99.65711
##  22           0.0914646751              99.74858
##  23           0.0811361259              99.82971
##  24           0.0601833567              99.88990
##  25           0.0516042379              99.94150
```

```
##  26          0.0272587995             99.96876
##  27          0.0230015463             99.99176
##  28          0.0052977929             99.99706
##  29          0.0024960103             99.99956
##  30          0.0004434827            100.00000
```

I personally want my PCs to account for at least 80% of the variability, but I want the final PCs to be included to individually account for at least 5% of the total variability.
Thus, based on my own criteria and based on the values in Table III-2, I will choose the first 5 principal components, which account for 84.7% of the variability in the data.

III.b) Analysis of Results of PCA

Table III-4: Correlation matrix of original data with the first 5 PCs

```
##                             [,1]        [,2]         [,3]         [,4]
## radius_mean             -0.79776675 -0.55790267 -0.014321182  0.058276998
## texture_mean            -0.37801323 -0.14243819  0.108358294 -0.848703801
## perimeter_mean          -0.82923555 -0.51334871 -0.015635546  0.059085011
## area_mean               -0.80539280 -0.55126955  0.048177170  0.075200175
## smoothness_mean         -0.51965303  0.44400165 -0.175072188  0.224307700
## compactness_mean        -0.87205011  0.36236113 -0.124375651  0.044746177
## concavity_mean          -0.94171317  0.14353386  0.004589225  0.026912451
## concave.points_mean     -0.95065387 -0.08294330 -0.042912871  0.091950691
## symmetry_mean           -0.50353484  0.45410669 -0.067549766  0.094468500
## fractal_dimension_mean  -0.23456539  0.87452298 -0.037894555  0.068378695
## radius_se               -0.75066782 -0.25181113  0.450692931  0.137837831
## texture_se              -0.06351460  0.21466057  0.628888081 -0.506443503
## perimeter_se            -0.77015490 -0.21341419  0.447610850  0.125243679
## area_se                 -0.73933688 -0.36331782  0.362604709  0.152282609
## smoothness_se           -0.05295834  0.48770074  0.518440202  0.062858236
## compactness_se          -0.62098087  0.55518008  0.259824807 -0.038659071
## concavity_se            -0.55974171  0.47046874  0.296225233  0.001853314
## concave.points_se       -0.66844526  0.31090241  0.377126990  0.104238834
## symmetry_se             -0.15488099  0.43859809  0.484439170  0.062026732
## fractal_dimension_se    -0.37379938  0.66820323  0.355046033  0.021539174
## radius_worst            -0.83090957 -0.52452555 -0.079748785  0.021697488
## texture_worst           -0.38072738 -0.10846933 -0.071004288 -0.890583627
## perimeter_worst         -0.86240823 -0.47684117 -0.081493799  0.019425394
## area_worst              -0.81951682 -0.52329808 -0.019980121  0.036443035
## smoothness_worst        -0.46630955  0.41105891 -0.436115698  0.024842887
## compactness_worst       -0.76567217  0.34256392 -0.396294196 -0.128531255
## concavity_worst         -0.83371903  0.23370868 -0.290506982 -0.104075362
## concave.points_worst    -0.91432733 -0.01969892 -0.285952302  0.008453959
## symmetry_worst          -0.44791263  0.33848486 -0.455445686 -0.051017498
## fractal_dimension_worst -0.48027261  0.65686526 -0.390780903 -0.108441378
##                                [,5]
## radius_mean             0.048518775
## texture_mean           -0.063519440
## perimeter_mean          0.047990153
## area_mean               0.013265627
## smoothness_mean        -0.468784268
## compactness_mean        0.015028239
## concavity_mean          0.110908536
## concave.points_mean    -0.056318830
```

```
## symmetry_mean           -0.392837675
## fractal_dimension_mean  -0.057042168
## radius_se               -0.198326624
## texture_se              -0.246084813
## perimeter_se            -0.155354955
## area_se                 -0.163809274
## smoothness_se           -0.297979065
## compactness_se           0.359487239
## concavity_se             0.454523281
## concave.points_se        0.251089422
## symmetry_se             -0.324690835
## fractal_dimension_se     0.338081553
## radius_worst            -0.005658192
## texture_worst           -0.119264982
## perimeter_worst          0.009571346
## area_worst              -0.035170715
## smoothness_worst        -0.416584530
## compactness_worst        0.156400009
## concavity_worst          0.242063518
## concave.points_worst     0.055639634
## symmetry_worst          -0.314020423
## fractal_dimension_worst  0.121242324
```

I personally don't have the knowledge on how features like radius, fractal dimension, symmetry of a breast tissue can categorize it into a certain type of breast tissue, so I am unable to comment on the presence of latent variables and what latent categorization each PC is describing.

However, I can comment on the independent variables that have the strongest correlation to each PC:
- PC1 has a strong negative correlation to mean concavity, mean concave points, worst concave points, and mean compactness.
- PC2 has a strong positive correlation to mean fractal dimensions
- PC3 doesn't have any strong correlation to any specific factor, but it is most correlated to the standard error of texture
- PC4 has a strong negative correlation to mean texture and worst texture
- PC5 doesn't have even moderate correlation to any factor

Now, because we have identified 7 dimensions that have strong correlation to at least one PC, we will reduce our data to these 7 independent variables and see if our clustering method has changed.

III.c) Checking accuracy of non-probabilistic k-means clustering using only 7 independent variables

Table III-5: Cluster allocation using 7 variables

```
##                      diagnosis
## cluster_allocation_2   B   M
##                    1 333  50
##                    2  24 162
```

From Table III-5, we see that the clustering method classified 383 observations into cluster 1 and 186 observations into cluster 2.
Additionally, benign diagnoses should be put in cluster 1 and malignant diagnoses should be put in cluster 2. We have 74 misclassifications, which means our classification is 87.0% accurate. We see that dropping 23 variables in our analysis has caused us to lose a 4% accuracy rate in our classification.
So, there is a tradeoff between number of variables included and accuracy of the classification method. The

model you choose depends on how accurate you want your classification to be and how many variables you are willing to include.

Table III-6: Table of cluster percentages of each diagnosis

```
##  cluster     benign malignant
##        1 0.93277311 0.2358491
##        2 0.06722689 0.7641509
```

Table III-6 shows the column-wise percentages of Table III-5. These percentages confirm that cluster 1 is meant to hold benign diagnoses and cluster 2 is meant to hold malignant diagnoses.
We also see that 24% of malignant diagnoses are misclassified as benign, which is very troubing.

Table III-7: Table of diagnosis percentages of each cluster

```
##  cluster    benign malignant
##        1 0.8694517 0.1305483
##        2 0.1290323 0.8709677
```

Table III-7 shows the row-wise percentages of Table III-5.

IV. Newton's Method to find MLE

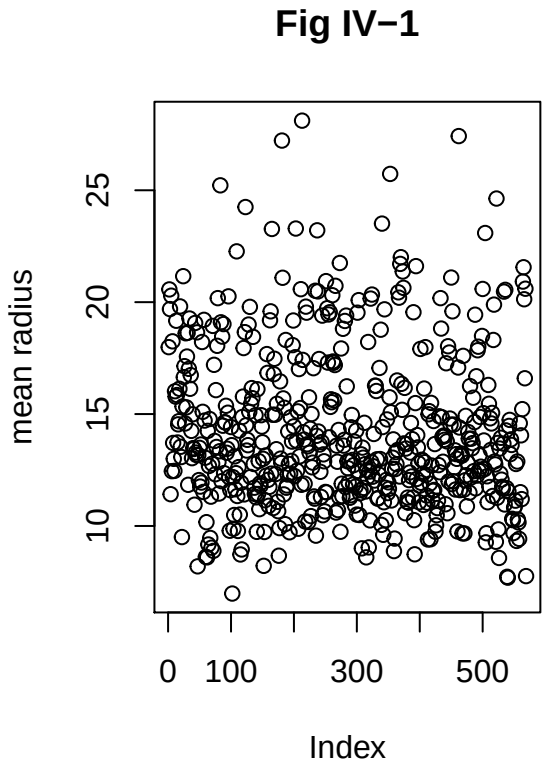
I think the variable mean radius is the most important. This is because we see in Fig II-1 that the clusters are easily distinguishable based on mean radius and in Table III-4 that mean radius has strong correlation with the first 2 PCs.

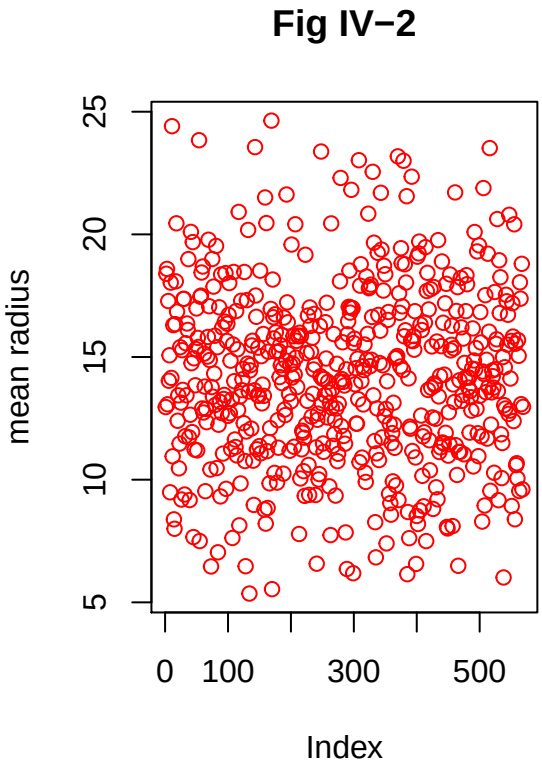
IV.a) Preliminaries of preparing for Newton's Method

We try to fit a normal distribution to the mean radius data.
Using the summary statistics that we can find from this data, we find that the mean radius data has mean=14 and standard deviation=3.5.
We plot a random sample of normally distributed data with this mean and sd and see if it is reasonably similar to our data.

Fig IV-1: Scatterplot of mean radius taken from Breast Cancer Dataset

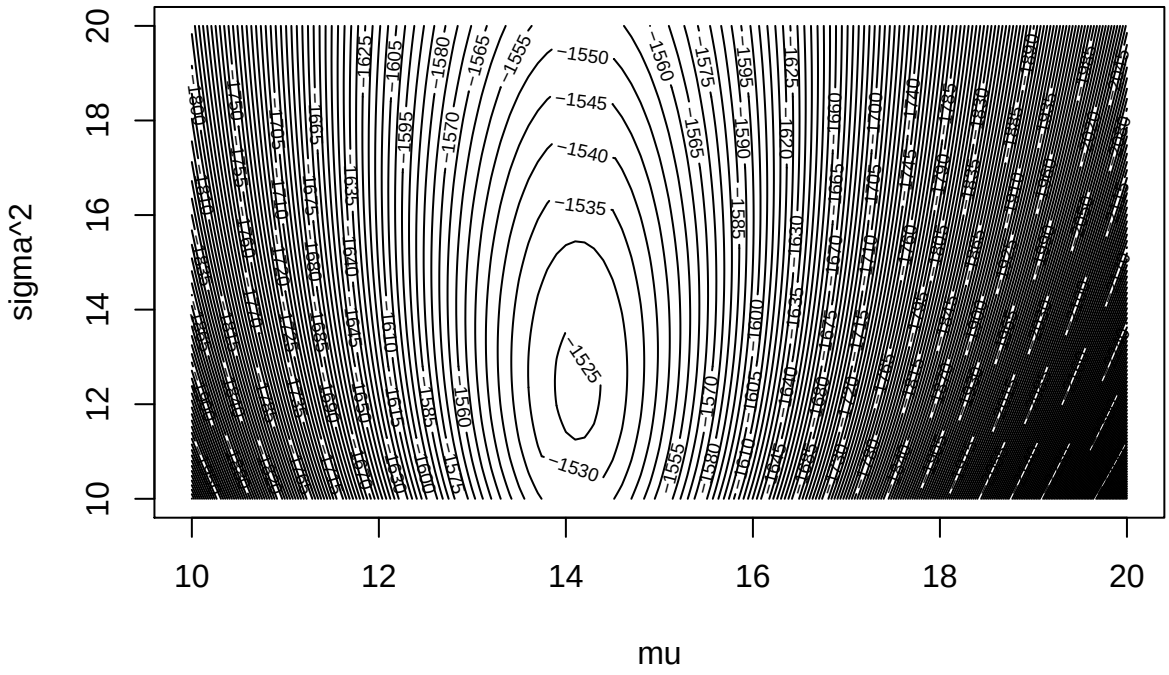
Fig IV-2: Scatterplot of a random sample of normally distributed data with mean=14 and sd=3.5





We see that the 2 figures, Fig IV-1 and Fig IV-2, are quite similar, so a normal distribution is a reasonable model to fit our data to.

Fig IV-3: Contour plot to find initial values of μ and σ^2



From the contour plot in Fig IV-3, we see a critical point around mu=14 and sigmaˆ2=12. So we set those as our inital values.

IV.b) Analyzing the results of Newton's Algorithm

Table IV-4: Table of iterate values for each run of Newton's

```
##  Iterate.number..       mu  sigma.2 log.likelihood
##                 0 14.00000 12.00000      -1524.131
##                 1 14.12350 12.35725      -1523.596
##                 2 14.12728 12.39682      -1523.594
##                 3 14.12729 12.39709      -1523.594
##                 4 14.12729 12.39709      -1523.594
```

From Table IV-4, note that our final estimate for the maximum likelihood estimator is $\hat{\mu} = 14.127$ and $\hat{\sigma}^2 = 12.397$.

The eigenvalues are -1.851154 & -45.897852. Because our eigenvalues are both negative, this means that our critical point (14.127,12.397) is a maximum point.

We are also able to find the standard error for $\hat{\mu} = 0.147$, and standard error for $\hat{\sigma}^2 = 0.734$.
Then we are able to find the 95% confidence intervals for $\hat{\mu} = (13.839, 14.417)$ and $\hat{\sigma}^2 = (10.957, 13.818)$.
Thus,
- We are 95% confident that the true parameter μ in the population is between 13.839 and 14.417.
- We are 95% confident that the true parameter σ^2 in the population is between 10.957 and 13.838.

Conclusions:
- Our final estimate $\hat{\mu} = 14.127$, $\hat{\sigma}^2 = 12.397$ was expected. It was pretty similar to our inital guess of (14,12.25).
- The estimated variance of our population, 12.397, is quite high, which shows that the mean radius of breast cancer tissues differs a lot from patient to patient
- For the asymptotic confidence intervals, the range of the CI for $\hat{\mu}$ is quite narrow, which is good because we are fairly certain of the true parameter mean. But the range of the CI for $\hat{\sigma}^2$ is quite large, so we are not that certain of the true population standard deviation.

V. Final Conclusions

In this analysis, we found that using all 30 variables in a non-probabalistic k-means program to classify whether a breast tumour should be diagnosed as benign or malignant is a pretty reliable method of classification, with a 91.2% accuracy rate. From the cluster visualizations and numerical summaries, we see that tumours that have a smaller value of all 30 variables tend to be classified as benign. This translates to tumours that are smaller, less textured, less smooth, less concave, and less symmetric tend to be classified as benign. We also notice that there is more variabiliy in the observations of cluster 1 than cluster 2 - which shows that malignant tumours show more variability than benign tumours in terms of the 30 physical features included in our analysis.

Then, we used PCA to isolate 7 variables that explain a lot of the variability. Upon performing clustering using non-probabilistic k-means on just those 7 variables, we are able to get a 87% accuracy rate. In my opinion, being able to drop 23 variables for a loss of 4% accuracy rate is pretty good, especially if we are working with really large datasets. The ability to drop so many variables can substantially reduce computation times and machine costs.
On the other hand, we are dealing with diagnoses of breast cancer - an extremely important diagnosis that if misclassified, could have fatal effects.

A possible real-world method to meet a compromise between accuracy and the dimensionality of dataset is as follows:
A hospital technician could run a program that utilize these 7 variables in the non-probabilistic k-means clustering method. When a new patient comes in, the technician inputs her data into the database and runs cluster analysis. This clustering method is just an initial screening for providing a diagnosis. Based on the cluster visualizations, the technician can look specifically at that new data point and how it stands compared to the rest of the data in the dataset. If that new data point is quite clearly in the middle of cluster 1 (malignant) or cluster 2 (benign) , the technician can confidently give the patient her diagnosis. However, if the new data point is somewhere in the boundary of clusters 1 and 2, or it is an outlier that is very far away from the center of either cluster, the technician can consult with another specialist or doctor to determine the patient's diagnosis.

Finally, we selected mean radius as an important variable in our dataset. Using Newton's method, we are able to get an estimate for MLE: $\hat{\mu} = 14.127$, $\hat{\sigma}^2 = 12.397$, and 95% confidence intervals: $\hat{\mu} =$ (13.839,14.417) and $\hat{\sigma}^2 =$ (10.957, 13.818).
Our confidence interval for population standard deviation is wider than for population mean, so we are more certain about the true value of population mean for mean radius of breast tumours.

References

[1] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data?select=data.csv
[2] Sanchez, J. Stats 102B Lesson I-10 Lecture Notes and R code
[3] Sanchez, J. Stats 102B Lesson II-2 Lecture Notes and R code
[4] Sanchez, J. Stats 102B Lesson I-7 Lecture Notes and R code