Analysis of Tissue Specimen for the Identification of Malignant Tumors

Contents

Predictors

Background

Modern diagnosis of breast cancer involves histologic examination of tissue specimen from the suspicious mass. Traditionally, a pathologist would examine a breast biopsy and compose a feature pattern for a particular case and then based on that pattern, associate a probability of malignancy (and a possible diagnosis of breast cancer). Attempts are currently being made to automate this process, using imaging equipment and machine learning (Maglogiannis & Zafiropoulos, 2004).

Origin of Data

Part of the histologic examination is examination of cytologic features which are features of individual cells in the breast biopsy. Three important cytologic features are cell nuclear size, nuclear membrane irregularity, and nuclear chromatin. 569 tissue specimens (from breast biopsies) were collected (Wolberg, 1995). Image analysis looked at 10-20 cells per specimen and classified each on: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. These variables are representations of nuclear size, nuclear membrane irregularity, and nuclear chromatin. The final measurement was an average of each sample's 10-20 cells.

Imputation

Missing data is a serious issue. It can bias parameter estimation, weaken generalizability, and decrease statistical power (Dong & Peng, 2013). Additionally, the techniques used in this section (PCA, Classification, etc.) are not naturally equipped to deal with missing data so imputation is a necessary step to pre-processing the data.

Missing Data Mechanism

Imputation technique is conditional on the missing data mechanism at work in our dataset. If the data are missing completely at random (MCAR), then most imputation techniques won't violate the validity of any inferences we may make. Jamshidian and Jalal (2010) recommend two tests of the null hypothesis that data are MCAR: a normal theory test and a non-parametric test. Testing Mardia's multivariate skew coefficients, we found that the data are not multivariate normal, $p<0.05$. Applying the non-parametric MCAR test, we cannot reject the null hypothesis that our data are MCAR, $p>0.05$.

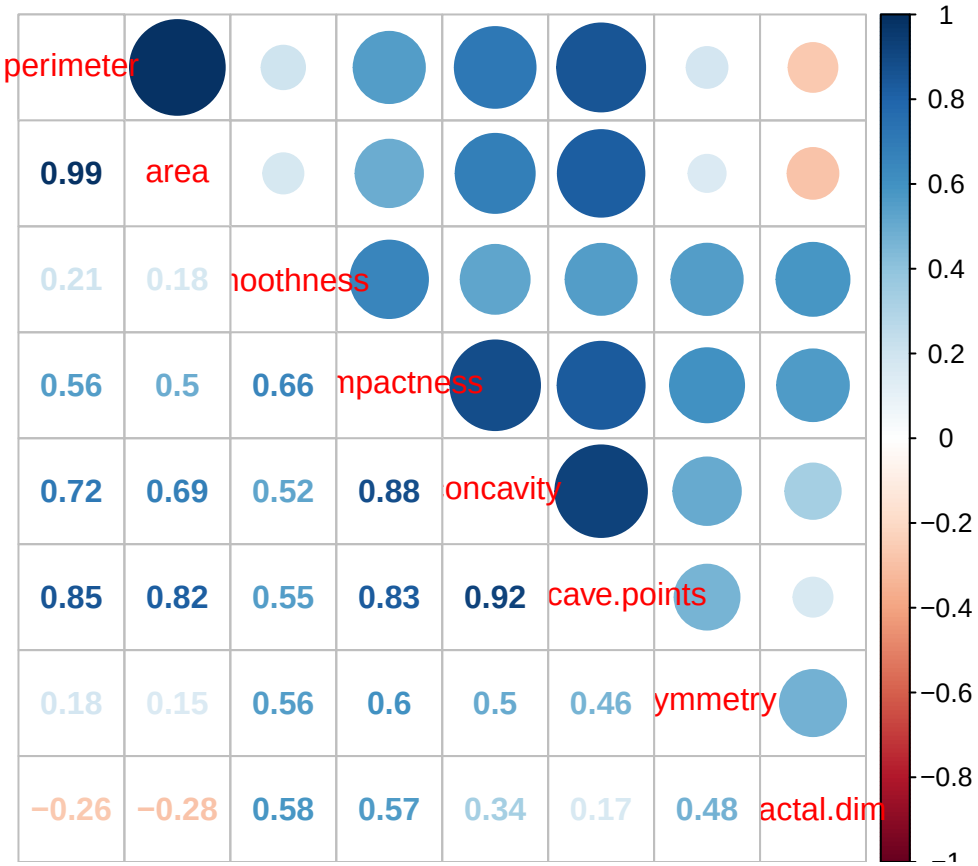
Non-Parameteric Imputation

Another important factor is the prevalence of missingness – this also determines what fix to apply. There is one pattern of missingness (Compactness, Concavity, and Concave Points) in the data for two observations: clients with IDs 843786 and 844359. Since the proportion is so low (0.4%), the go-to method would usually be list-wise deletion (Izenman, 2013). Additionally, since the proportion is so low, if we elect for some form of imputation, the difference between list-wise deletion and between different forms of imputation will be minimal in terms of analyses we run. In our case, the choice of how we deal with the missing data is very low-impact/risk. Simple imputation has the easiest implementation but it may bias standard error estimates to make our models look more precise than they actually are (Izenman, 2013). Thus I elected to use the distribution-free imputation method of Srivastava and Dolatabadi (2009; implemented using the MissMech R package) for the two cases with missing data.

Dimension Reduction through PCA

Are the predictors related?

Let's investigate the relationship between the predictors. Below is a correlogram, a visualization of the correlation matrix of the predictors. We can see strong positive relationships between the variables including a near perfect correlation between perimeter and area ($r = 0.99$).



	Squared Multiple Correlation
perimeter	0.99
area	0.98
smoothness	0.65
compactness	0.93
concavity	0.91
concave.points	0.95
symmetry	0.44
fractal.dim	0.84

Further, the squared multiple correlations (SMCs; tabulated above) between the predictors are incredibly high. For example, 98.7% of the variance in Perimeter is accounted for by the other predictors; four other predictors have SMC's above 90% There is a lot of redundancy in these predictors. A lot of the variance that exists in the Perimeter dimension also exists in the Area dimension also exists in the Concave Points dimension and so on. We don't need to keep a full, 8-dimensional predictor space because the data don't occupy the whole space. Geometrically, there may be an opportunity to reduce the predictor space in such a way as to preserve much of the original variance of the predictors.

Is PCA necessary?

From a statistical point-of-view we can take two stances. On the one hand, high inter-predictor associations often result in estimation problems (which may propagate to interpretation problems) so we may want to perform PCA. The phenomenon of such high correlation between predictors is multicollinearity, which interferes with matrix inversion and inflates parameter standard errors. In addition to the high SMCs, the

correlation matrix of the predictors has a condition number of $\kappa(R)$=446, indicating ill-conditioning and multicollinearity. The inflated standard errors ruin parameter interpretations as we can't tell the contribution of individual predictors. On the other hand, as long as we can get our matrix to invert and we don't care about our high standard errors (essentially making our method a black box), we may not want to perform a PCA as preprocessing to our main analysis. From a practical point-of-view we don't really have a large number of predictors to project to a smaller subspace and even if we were to do so, does it make sense to interpret composite indices of these predictors? For example, what does it mean to have a weighted difference between smoothness and area, and perimeter and fractal dimension? The typical application of PCA involves many, many predictors whose variance can be effectively reduced down to only a handful of linear combinations. In this case with small p, perhaps a better solution would be some penalized glm like a lasso which would shrink some collinear predictors to zero.

Should we standardize for PCA?

PCA of unstandardized variables (i.e. PCA of Σ) is not equivalent to PCA of standardized variables (i.e. PCA of R). In fact, the principal components derived in one situation are not a simple function of those derived in the other. When there is a disparity in the scales of the variables, the variables with larger variances dominate (have larger coefficients for) the principal components. It is apparent from the various measures of dispersion in the table below that the eight predictors have wildly different scales – to be exact Area (and to a lesser extent, Perimeter) have much larger variances and so would dominate the PCA. Therefore, the PCA should be performed on R.

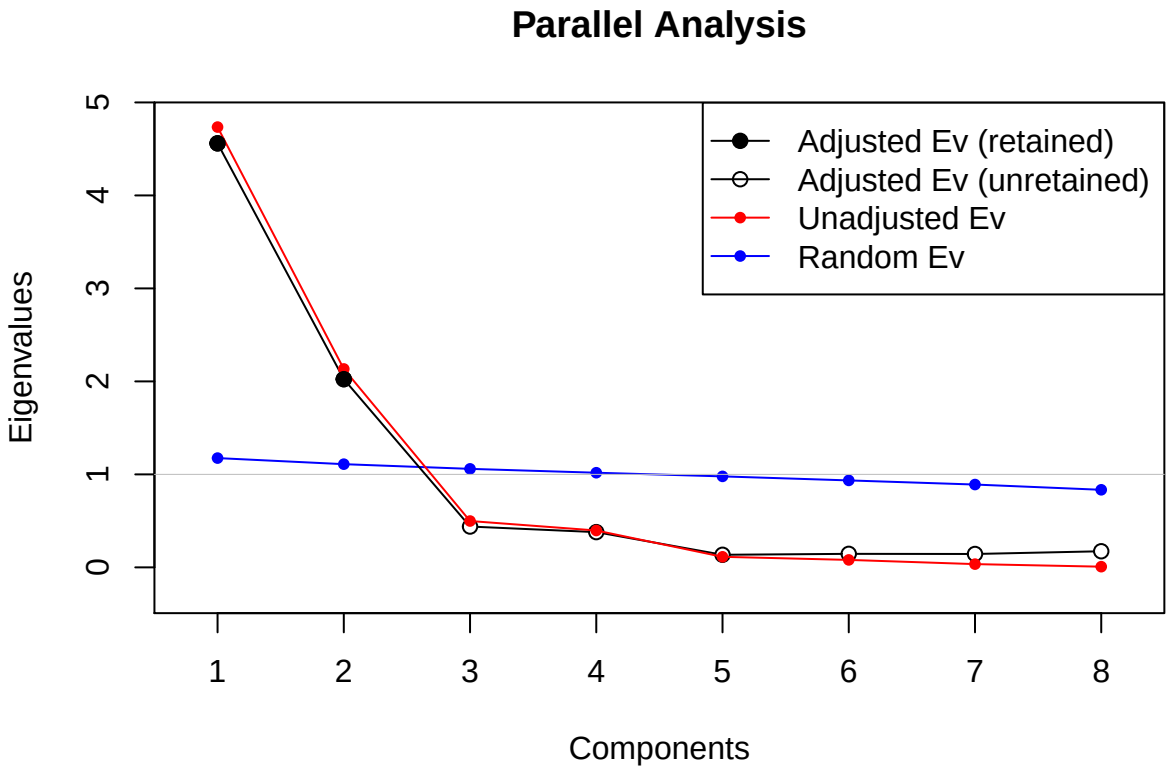
	SD	IQR	Min	Max
perimeter	24.30	28.93	43.79	188.50
area	351.91	362.40	143.50	2501.00
smoothness	0.01	0.02	0.05	0.16
compactness	0.05	0.07	0.02	0.35
concavity	0.08	0.10	0.00	0.43
concave.points	0.04	0.05	0.00	0.20
symmetry	0.03	0.03	0.11	0.30
fractal.dim	0.01	0.01	0.05	0.10

To drive this point home, we can look at the difference in coefficients between PCA of standardized and unstandardized predictors. The coefficients are shown below. Notice that the PCA of unstandardized predictors results in a first principal component totally dominated by Area.

	PC1 S Loadings	PC1 R Loadings
perimeter	0.07	0.36
area	1.00	0.34
smoothness	0.00	0.30
compactness	0.00	0.42
concavity	0.00	0.43
concave.points	0.00	0.44
symmetry	0.00	0.28
fractal.dim	0.00	0.15

Perform PCA

To select the optimal number of principal components we will use Horn's Parallel Analysis and the amount of variation explained:

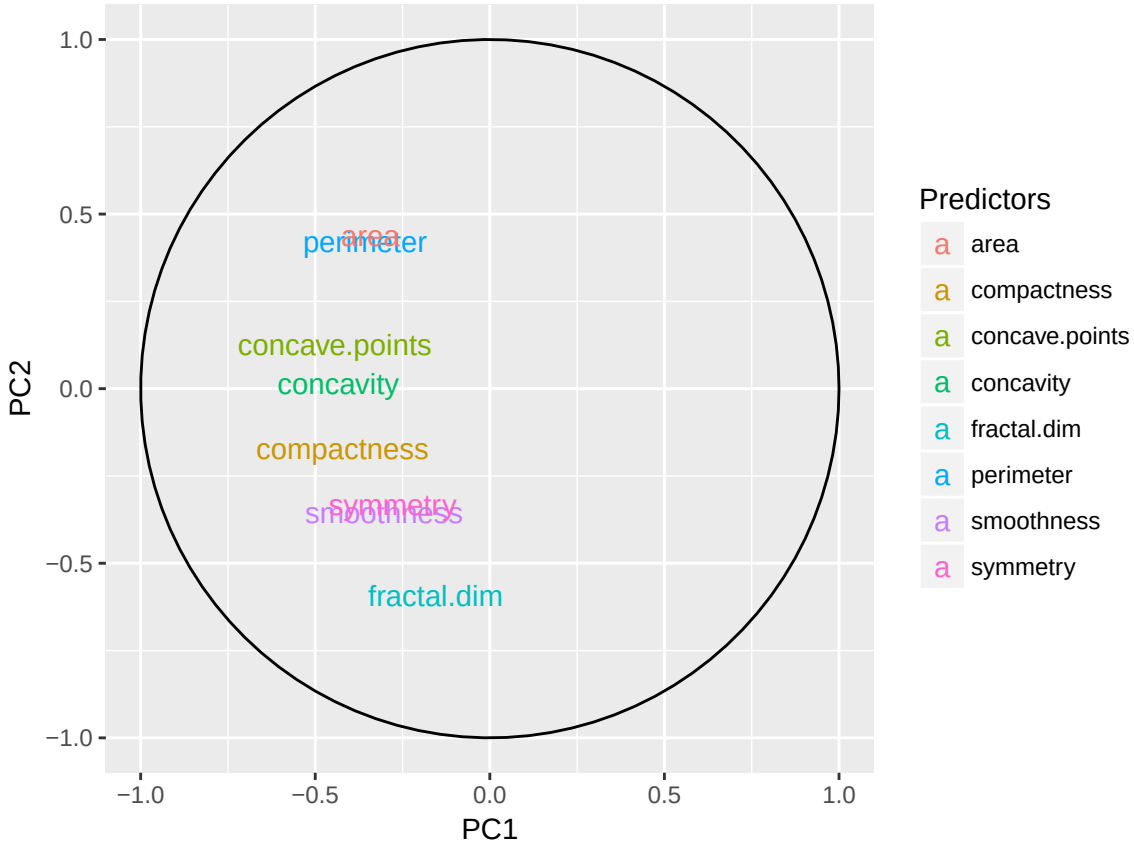


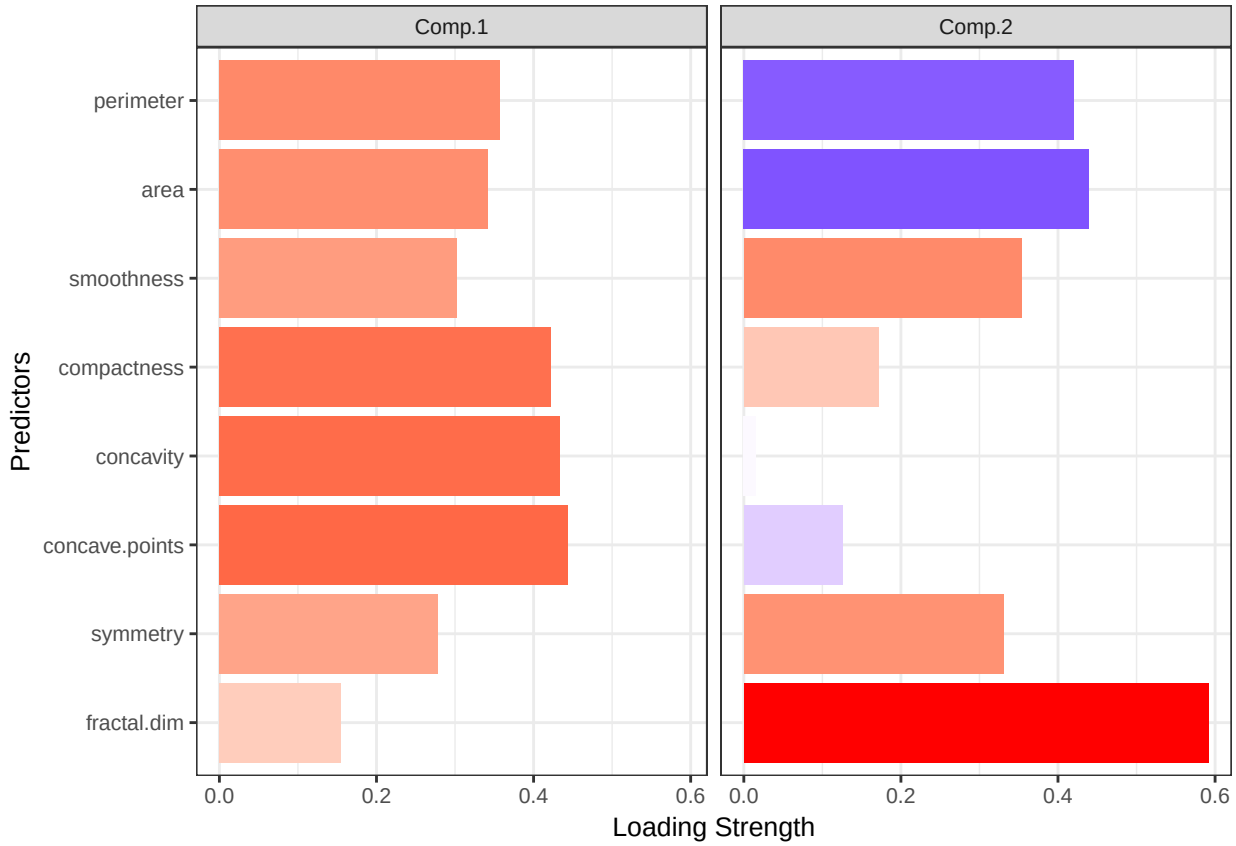

```
##
## Using eigendecomposition of correlation matrix.
## Computing: 10%  20%  30%  40%  50%  60%  70%  80%  90%  100%
```

HPA tells us that we should stick with two principal components. As the table below shows, these correspond to 86% of the total variance in the predictors! Including an additional principal component only increases the total variance accounted for by 6%. Thus we stick with two principal components.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Prop of Var.	0.59	0.27	0.06	0.05	0.01	0.01	0.00	0.00
Cum. Prop.	0.59	0.86	0.92	0.97	0.98	0.99	0.99	0.99

The two figures below aid us in interpreting the principal components as composite indices of our predictors:

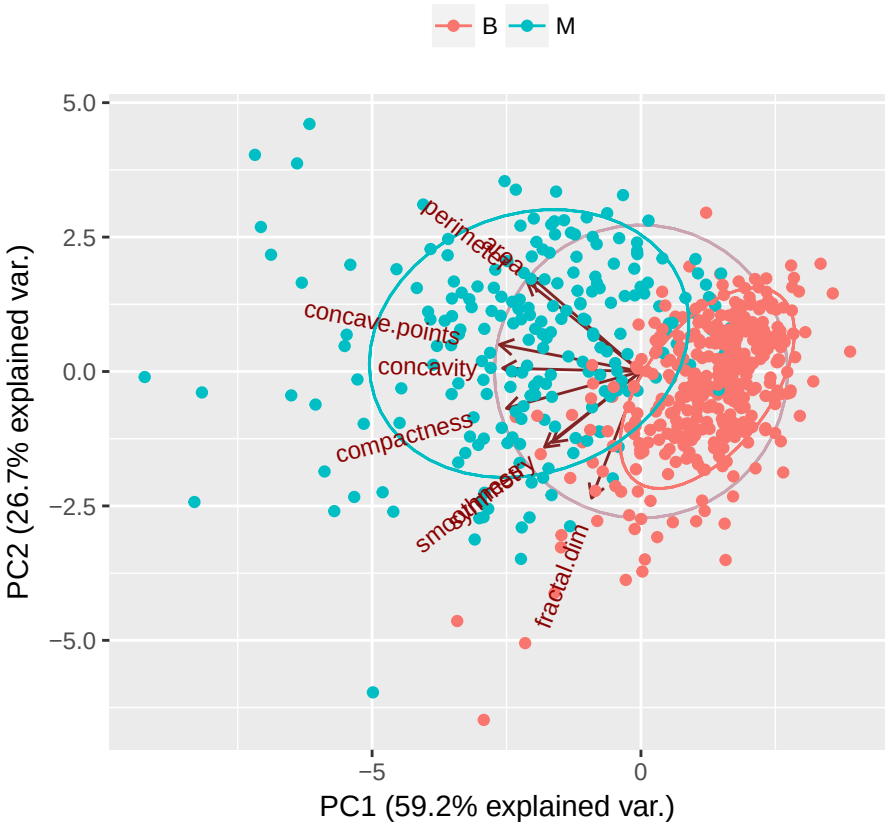

1.0
0.5
0.0
−0.5
−1.0
PC2
−1.0
−0.5
0.0
0.5
1.0
PC1
area
perimeter
concave.points
concavity
compactness
symmetry
smoothness
fractal.dim
Predictors
a area
a compactness
a concave.points
a concavity
a fractal.dim
a perimeter
a smoothness
a symmetry



Our first principal component is a (nearly equal) weighted mean of all predictors; Fractal Dimension has the least weight. The second principal component is a weighted difference between Fractal Dimension, Symmetry, and Smoothness; and Perimeter and Area.

PCA and Classification

Can we see separation between the two diagnoses in our reduced subspace? In other words, is the 14% loss in fidelity of predictor variance worth it in terms of our ultimate goal of classification? The biplot below answers the question. The first principal component and the 59.2% variance it accounts for actually does a decent job of separating the diagnoses; incorporate the second principal component (86% cumulate variance) and the separation is even better. The question now is how much better this separation is than if we had used the original data. This will be answered in the next section.



Classification

Now we want to apply a selection of classifiers to the data to classify cases as benign vs. malignant. Building on our PCA, we must ask, can we perform classification on the reduced space formed by the first two principal components? In other words, is the variance we lost when performing data reduction very important in the prediction of diagnosis? If it is, we will see a large difference in classification metrics; if it isn't, the difference should be minimal. The advantage of a 2-component PCA is that we can visually assess if a classifier is performing well by looking at the decision boundary superimposed on a scatterplot of the first two principal components. Note that using PCA to deal with multicollinearity isn't really an advantage here because in this classification context, we don't care about the effects of multicollinearity, only about prediction precision. We may not be able to figure out how important each predictor is in terms of the prediction, but the prediction itself should not be affected. We use a validation set approach, randomly splitting the observations 50-50 into a training set and a test set. We train the classifiers on the training set and evaluate their performance with the test set.

Logistic Regression

As mentioned above, we don't bother with subset selection as our focus is on prediction only. We fit all predictors. Logistic Regression fit very well with an overall classification rate of 93.3%, a sensitivity of 89.3% and a specificity of 96.7%. Individuals with benign masses were nearly classified perfectly. About 1 out of every 10 individuals with malignant were misdiagnosed. Sensitivity takes high precedence since the error of making a negative diagnosis when the client is positive is worse than making a positive diagnosis when the client is negative. It is better to administer cancer medication to someone without cancer (even if the

medication has terrible side effects) than to give someone a clean bill of health and let someone slip past the point of no return.

Logistic Regression (Reduced Space)

With the reduced information, logistic regression performance fell across overall classification rate (1% drop), sensitivity (~4% drop), and specificity (1.6% drop). While this isn't a big deal in a statistical sense, from a practical sense any increase in error can come at the cost of human lives.

Linear Discriminant Analysis

LDA performed better than logistic regression overall (94.7% overall classification rate). It outperformed logistic regression in specificity (99.4%) but at the cost of worse sensitivity (86.3%) which is the error rate that we care about. Therefore logistic regression is still the superior classifier.

Linear Discriminant Analysis (Reduced Space)

Like logistic regression, LDA on the reduced data decreased performance across the board; unlike logistic regression, the decrease was marked. This is very unacceptable from a practical standpoint.

Quadratic Discriminant Analysis

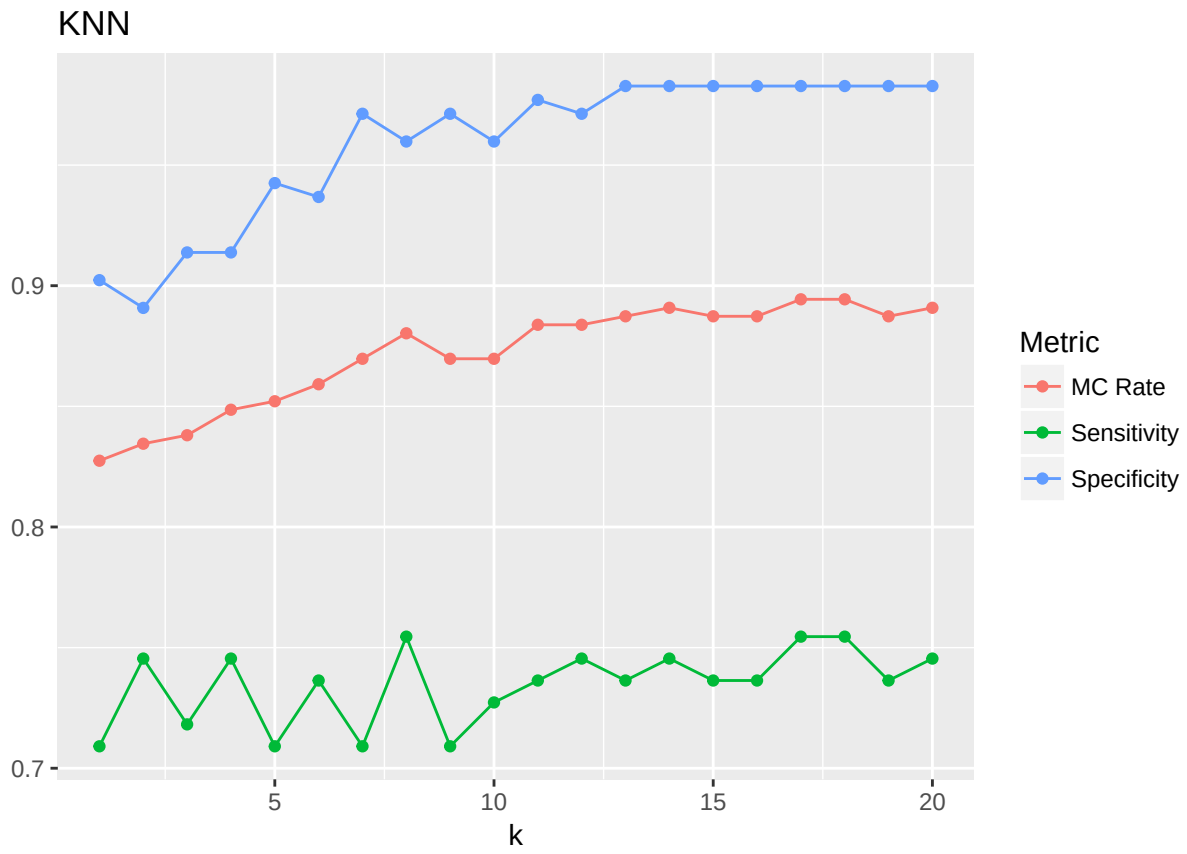
QDA performed the same as LDA in terms of sensitivity (86.3%) but was 2.7% worse in terms of specificity. So far it is the worst classifier of the three.

Quadratic Discriminant Analysis (Reduced Space)

Curiously, QDA on the reduced space saw a huge decrease in sensitivity (drop of 4.4%) and a small increase in specificity (plus 1.6%). Overall, it's still unimpressive.

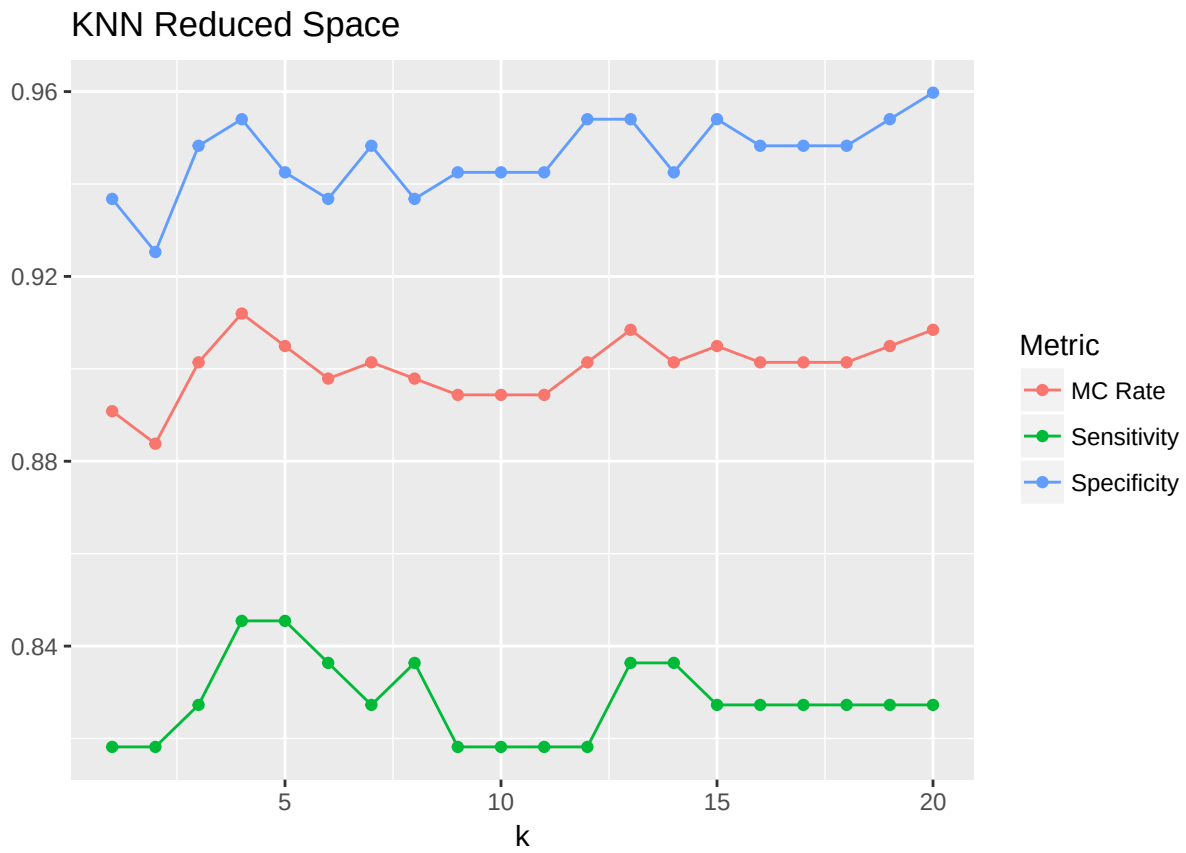
K-Nearest-Neighbors

To pick the optimal K we tune it with the total misclassification rate (TCR), sensitivity, and specificity:



Total classification rate is highest for k=14. The k with best tradeoff between optimum sensitivity and total classification rate is k=3. So, we will use the 3-Nearest-Neighbors approach when comparing to other classifiers. K-Nearest-Neighbors appears to be the worst so far.

K-Nearest-Neighbors (Reduced Space)



	mcr	sens	spec
1	0.891	0.818	0.937
2	0.884	0.818	0.925
3	0.901	0.827	0.948
4	0.912	0.845	0.954
5	0.905	0.845	0.943
6	0.898	0.836	0.937
7	0.901	0.827	0.948
8	0.898	0.836	0.937
9	0.894	0.818	0.943
10	0.894	0.818	0.943
11	0.894	0.818	0.943
12	0.901	0.818	0.954
13	0.908	0.836	0.954
14	0.901	0.836	0.943
15	0.905	0.827	0.954
16	0.901	0.827	0.948
17	0.901	0.827	0.948
18	0.901	0.827	0.948
19	0.905	0.827	0.954
20	0.908	0.827	0.960

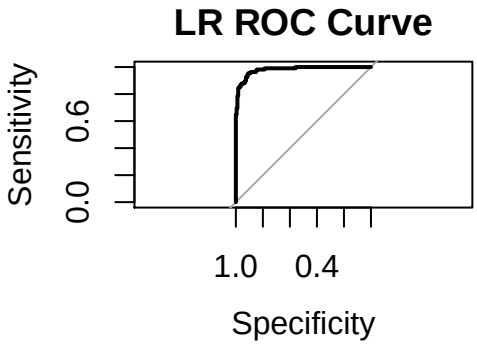
Prioritizing sensitivity, the optimal k for the reduced space is k=4. What's interesting here is that classification for KNN in the reduced space is better in every way than it is for the full space. Why should this be if we lost information (variation) when we reduced dimensions? My guess is that the nature of this algorithm is to classify based on neighbors and with reduced dimensions, there is less noise in the distances between neighbors. An observation may have been close to another observation with a different diagnosis in the full space but once the space was reduced, the new projected observation is now closer to a projected observation with the same diagnosis.

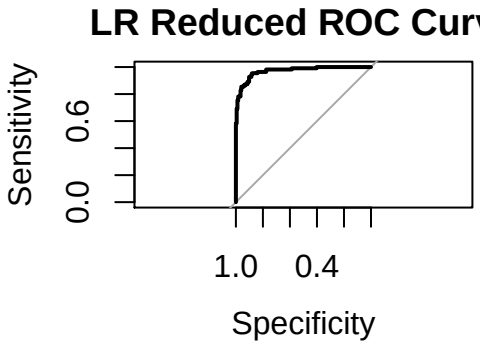
Compare Classifiers

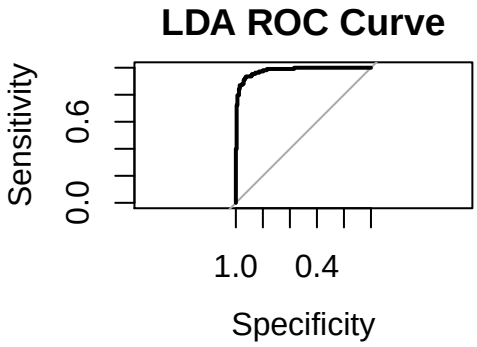
To compare the performance of classifiers we will use five measures. The first is the misclassification rate which is simply, "when did the classifier get it right?" Out of all of the cases, how many cases were correctly diagnosed as benign or malignant? Sensitivity is our true positive rate – out of those cases with malignant tumors, how many were correctly diagnosed as malignant? Specificity is our true negative rate – out of those cases with benign tumors, how many were correctly diagnosed as benign? Finally, we look at the ROC curve for each. The ROC curve tells us the relationship between sensitivity and specificity for a classifier as we vary the threshold. For example, for a poor classifier, decreasing the threshold of classification to increase the sensitivity will tank the specificity. A great classifier will have a threshold at which both specificity and sensitivity are maximized at a high value (close to 1). The area under the ROC curve is a measure of the discrimination power of the test. It is bounded between 1 and 0.5 with the former being a perfect classifier and the latter being random chance (a coin flip for a diagnosis). Note that KNN in its original form is a non-probabilistic classifier so it does not have ROC curve or corresponding AUC.

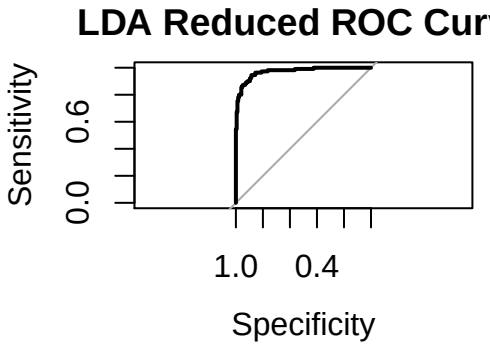
	Accuracy	Sensitivity	Specificity	AUC
LR	91.2%	87.3%	93.7%	98.1%
LR reduced	91.9%	85.5%	96%	97.1%
LDA	92.6%	85.2%	97.1%	97.8%
LDA reduced	89.8%	76.4%	98.3%	97.2%
QDA	92.2%	86.1%	96%	97.4%
QDA reduced	91.9%	85.5%	96%	96.7%
KNN	84.9%	74.5%	91.4%	NA
KNN reduced	91.2%	84.5%	95.4%	NA

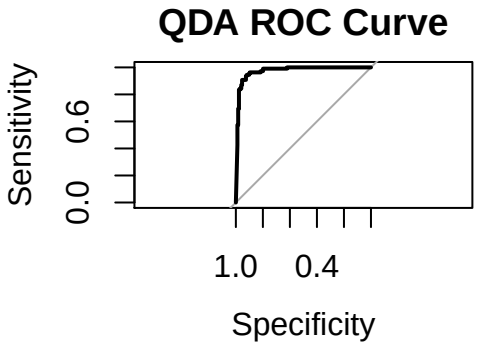
In terms of overall classification, LDA is the best, correctly diagnosing nearly 19 out of 20 patients. It also has the highest specificity, with nearly perfect classification of patients with benign tumors. The problem is its sensitivity is a full 3 percentage points lower than the two classifiers with the highest sensitivity: KNN (reduced) and logistic regression. KNN is a surprise because when applied to the original data it is easily the worst classifier. However, when applied to the reduced data, it's sensitivity improves markedly. Unfortunately it doesn't have a ROC curve because it's a nonprobabilistic classifier. Logistic regression is the best overall performer (besides LDA) with high sensitivity, decent specificity, and large AUC. If specificity was important, LDA would be the top classifier; but in this context, sensitivity is mostly what matters and that makes Logistic Regression the best classifier.

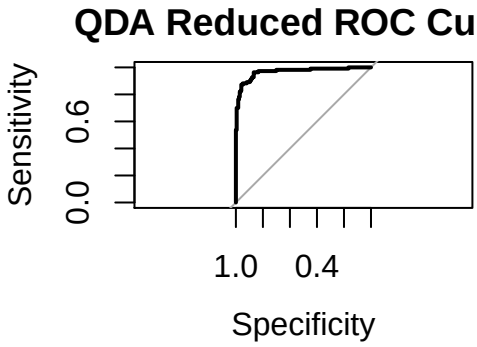












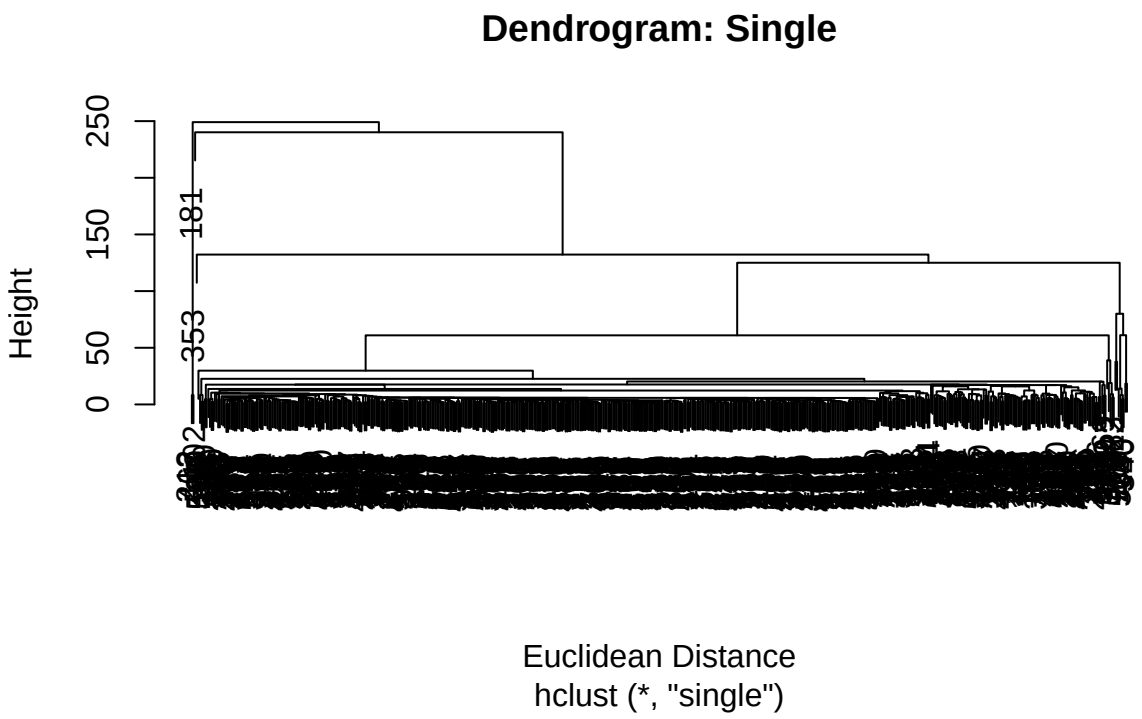
Cluster Analysis

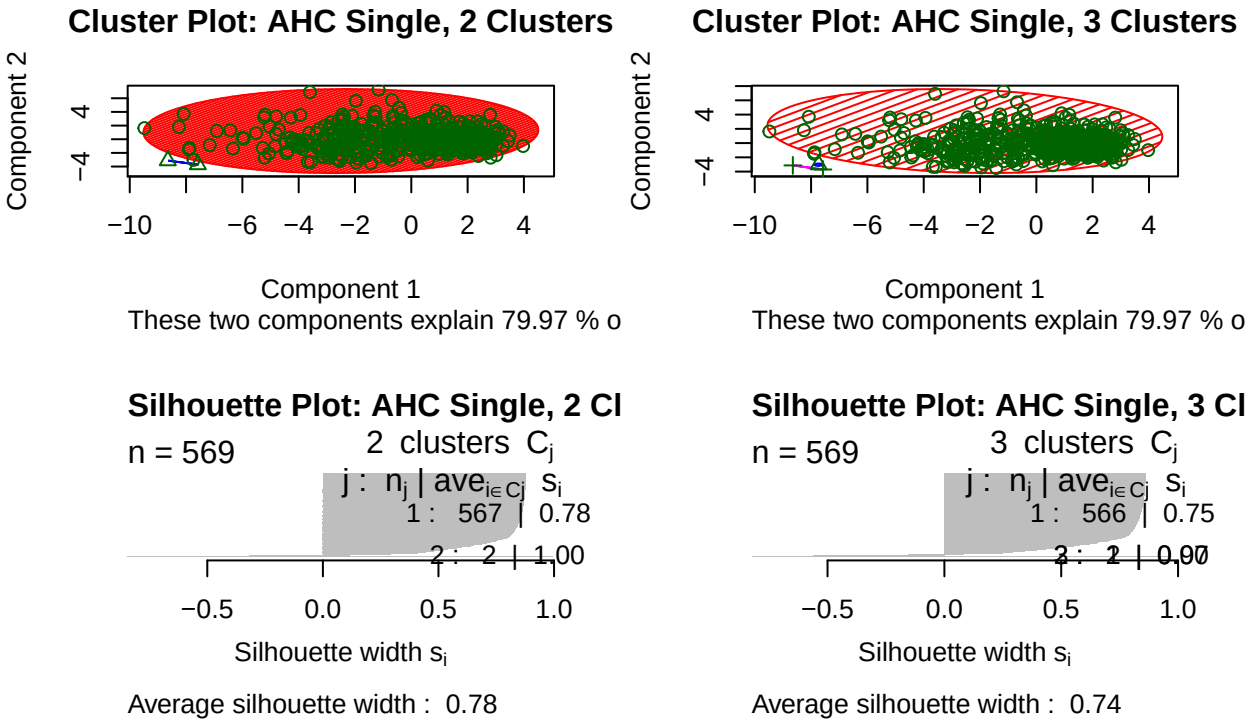
Agglomerative Hierarchical Clustering (AHC)

There are a number of different linkage methods and the process of cluster analysis necessitates application of multiple linkage methods to the data and to compare results. We apply single linkage, complete linkage, average linkage, centroid linkage, and Ward's linkage and examine two- and three-cluster solutions using three graphical methods: dendrograms, cluster plots, and silhouette plots. A cluster plot plots the clustering solution over the first two principal components. It is very similar to the PCA biplot from earlier. We can evaluate the performance of a linkage method by looking at the size, separation, and shape of clusters (while keeping in mind that this is a cluster solution projected into a reduced space, necessarily losing some variance). A silhouette plot allows interpretation and validation of consistency within clusters of data (Rousseuw, 1987). It displays, for each observation, a silhouette value which is a measure of cohesion (distance to all other points in cluster) vs. separation (distance to points in closest other cluster). The value is bounded between -1 (high separation) and 1 (high cohesion). Silhouette values are grouped into the corresponding clusters

allowing you to evaluate how well certain clusters are performing. If a cluster has a large number of very negative silhouette values, the cluster solution is poor.

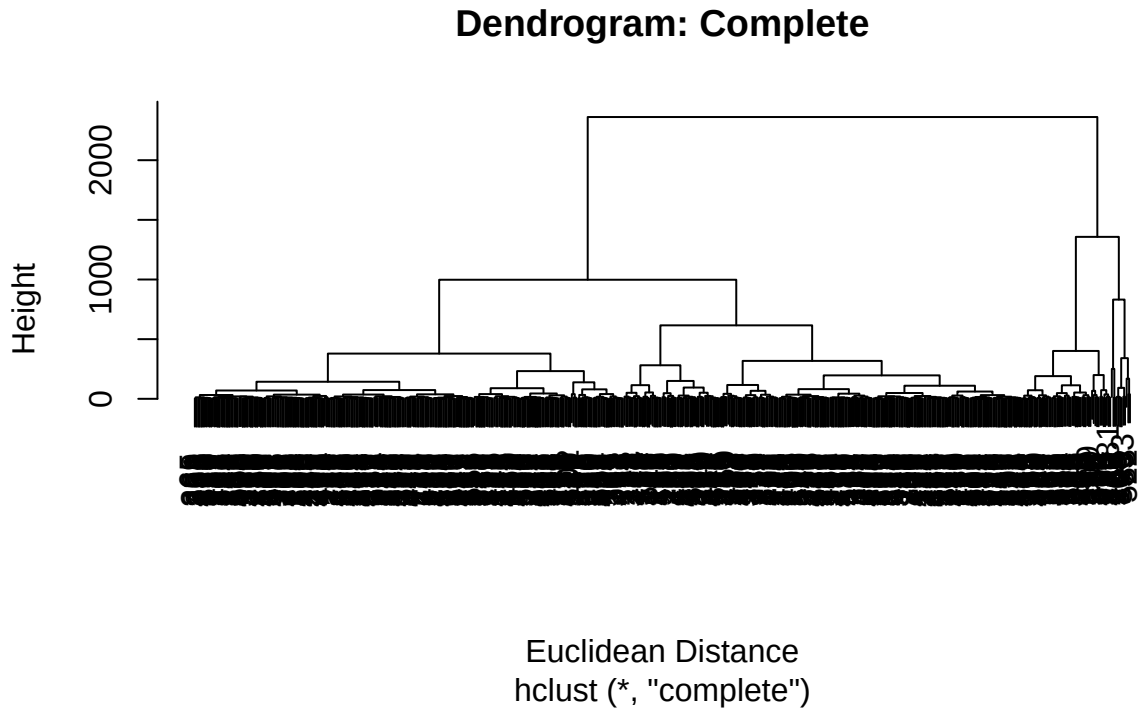
Single Linkage

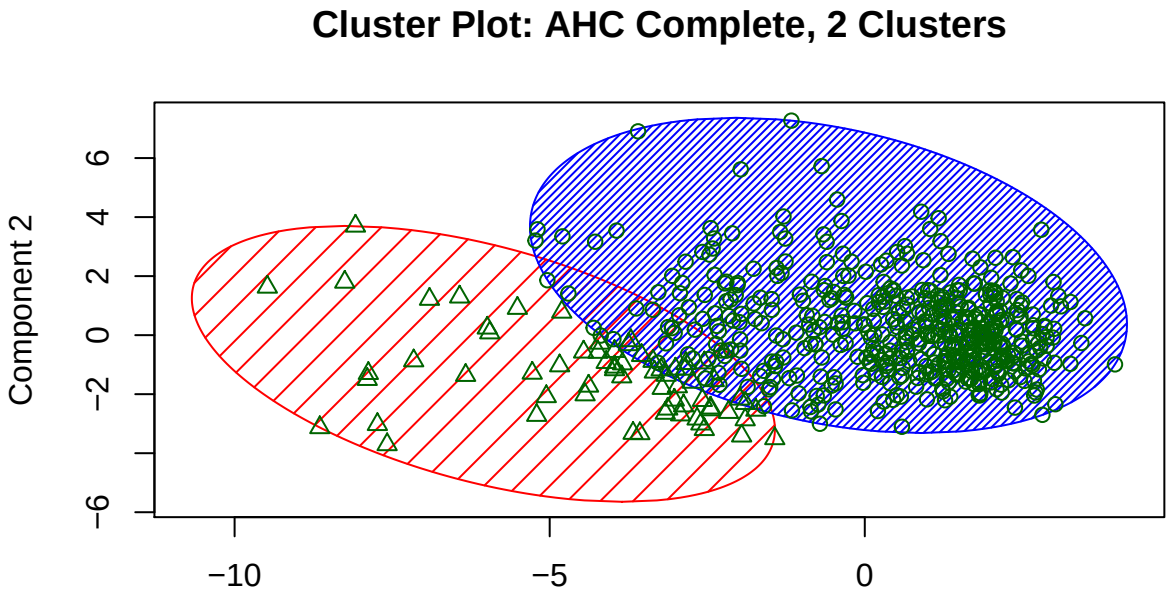


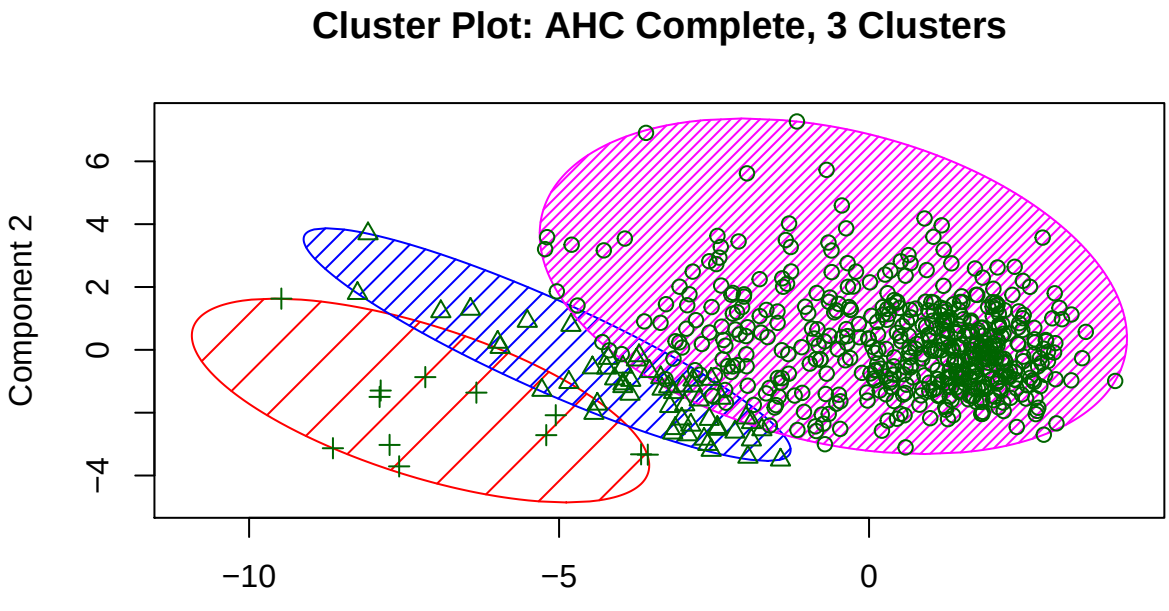


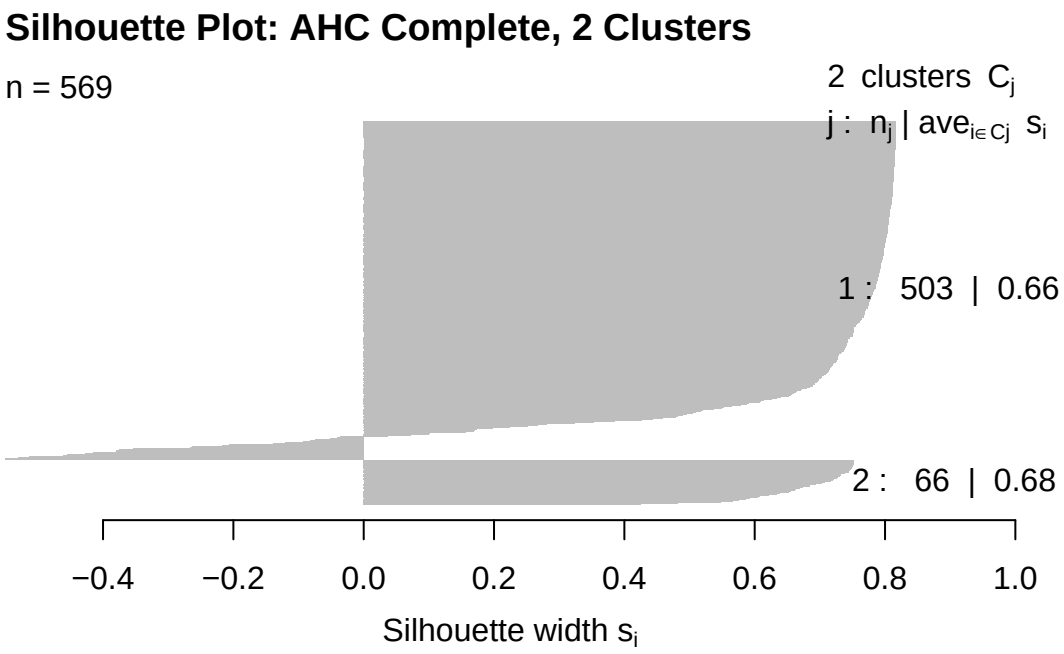
The dendrogram for single linkage doesn't look very good. The first cluster is two observations versus the rest. The third and fourth clusters split off individual observations from the majority cluster. The clusters are tiny! Obviously in a diagnosis application this is useless. Examining the cluster plot on the first two principal components, we see a nice spatial representation of what the dendrogram was telling us – the two and three cluster solutions just pick out a couple outlying observations as their own cluster. The silhouette plots aren't much use here because of the uneven size of the clusters.

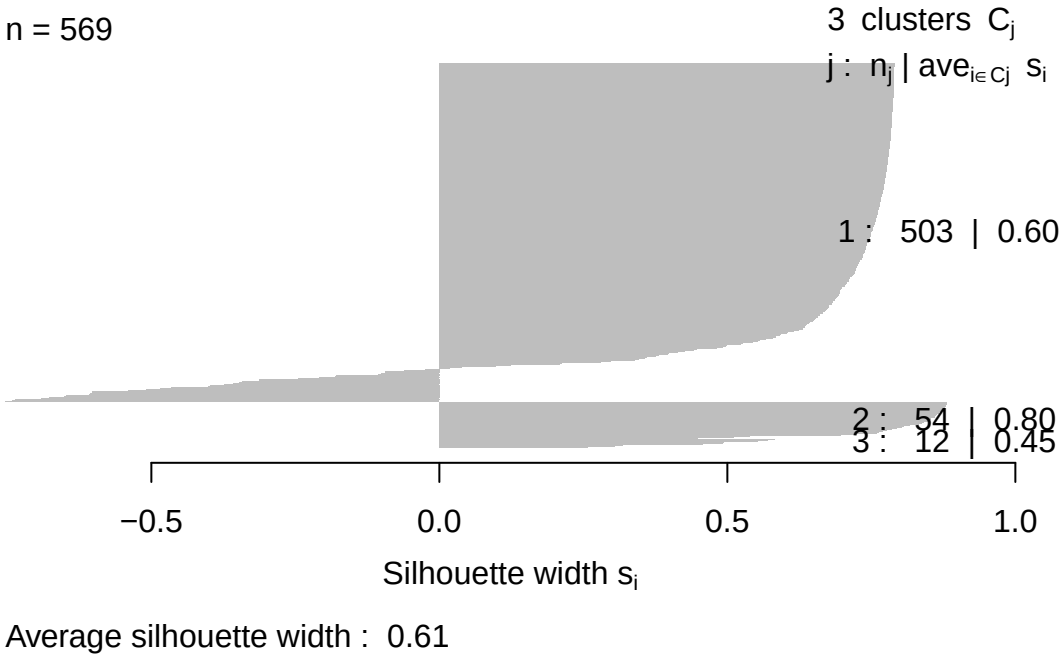
Complete Linkage




Cluster Plot: AHC Complete, 2 Clusters
Component 2
6
4
2
0
−2
−6
−10
−5
0
Component 1
These two components explain 79.97 % of the point variability.

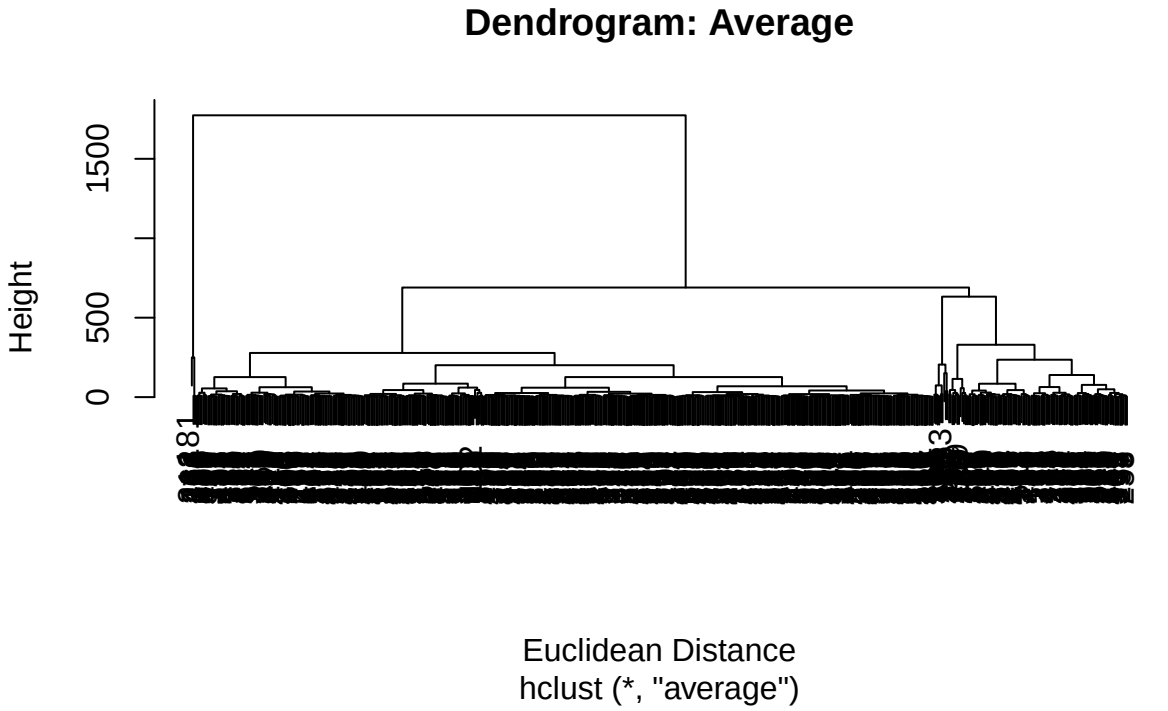

Cluster Plot: AHC Complete, 3 Clusters
Component 2
6
4
2
0
−4
−10
−5
0
Component 1
These two components explain 79.97 % of the point variability.

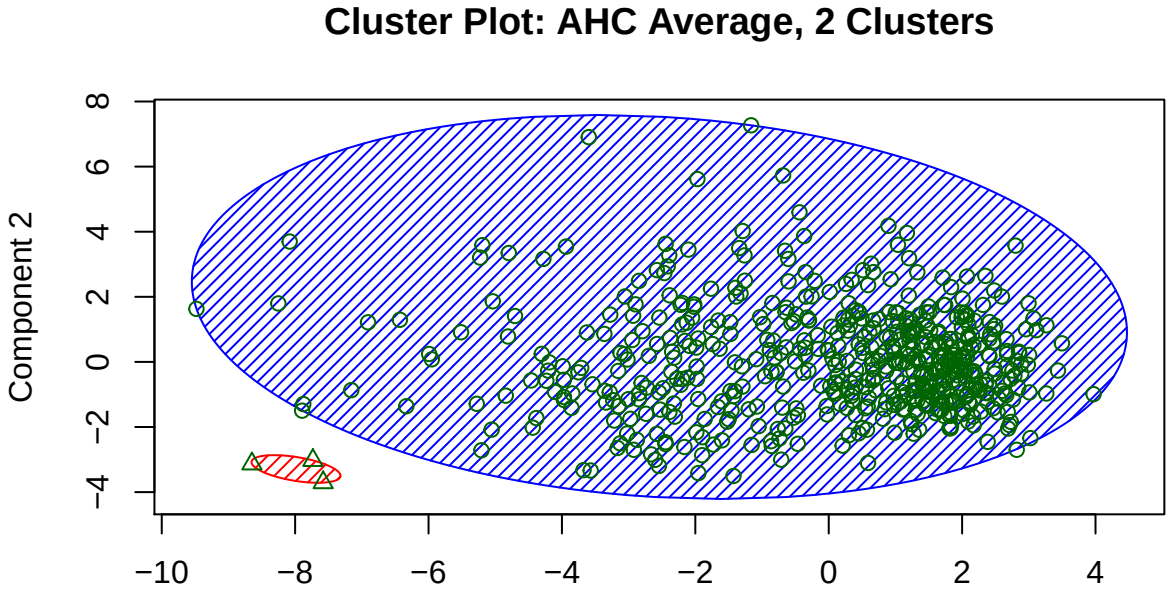




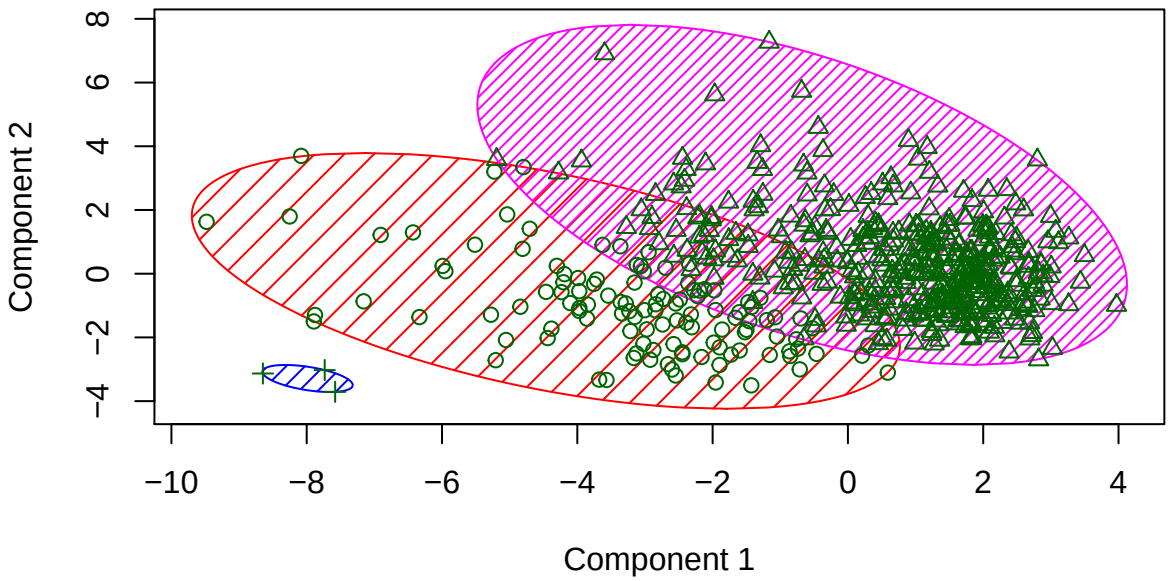
The dendrogram looks a lot better here; the first and second split produce large clusters. When we look at the cluster plot there is a nice separation between the clusters in the two-cluster solution (just like the true diagnoses). The poor silhouette values of the larger cluster are reflective of the overlap between the two clusters (i.e. the two diagnoses).

Average Linkage



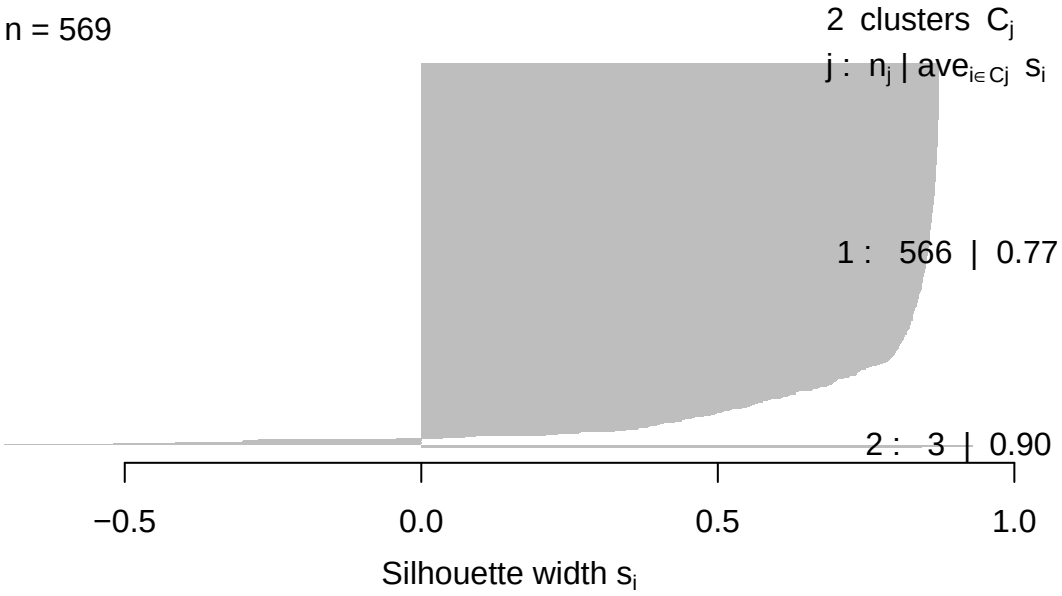


Cluster Plot: AHC Average, 3 Clusters

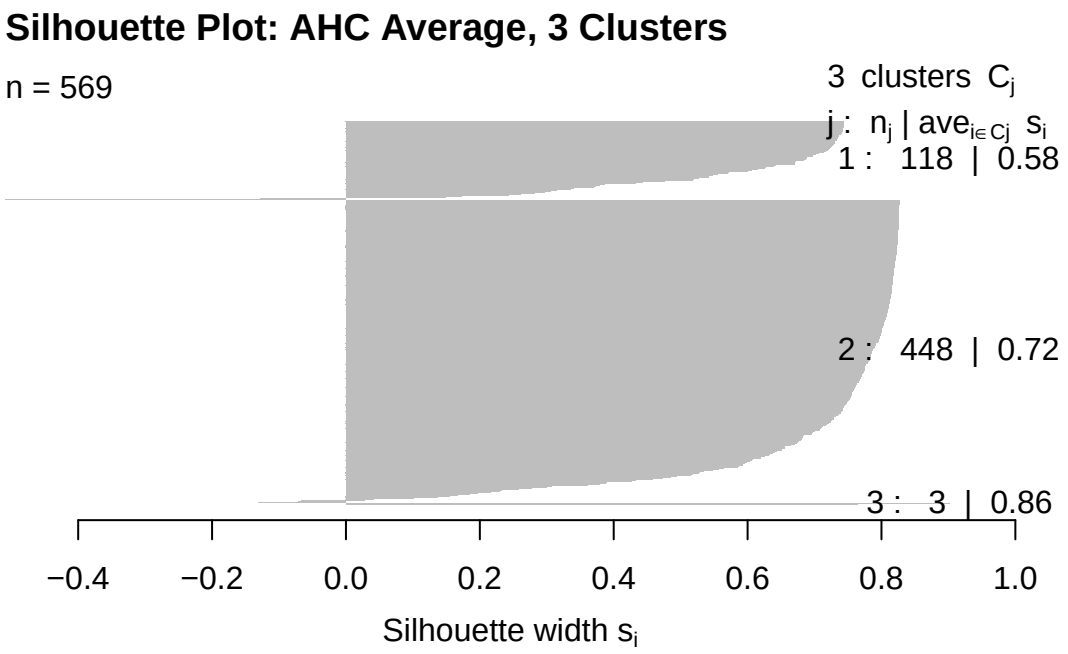


These two components explain 79.97 % of the point variability.

Silhouette Plot: AHC Average, 2 Clusters

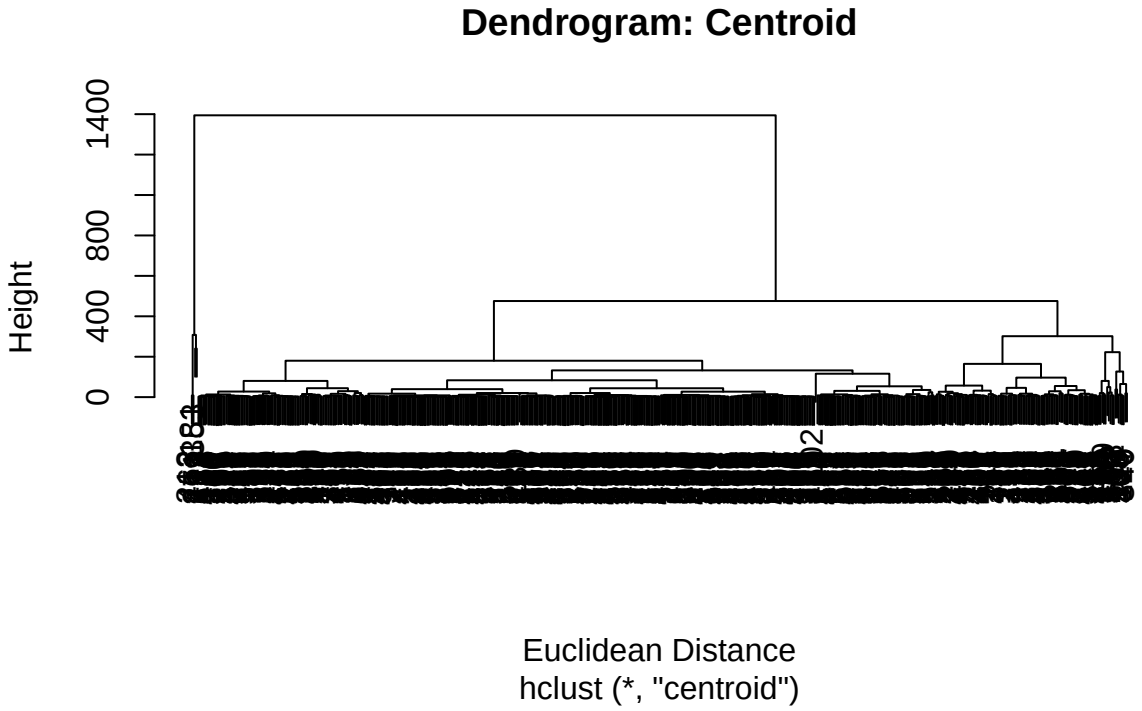


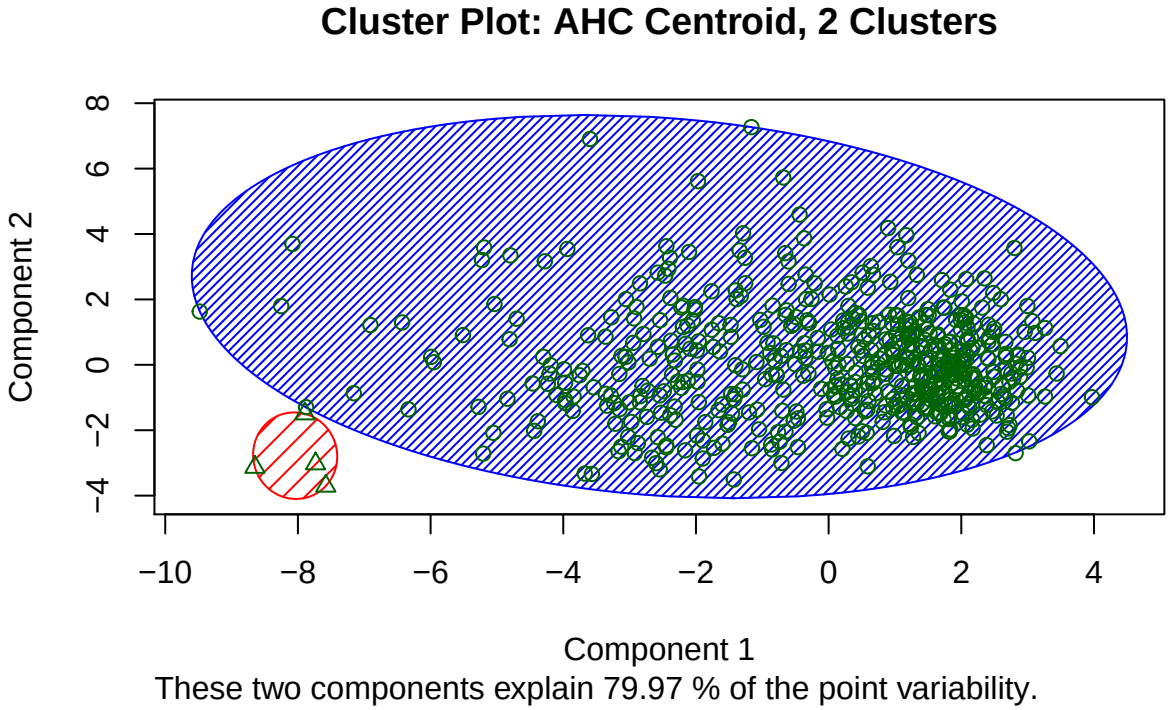
Average silhouette width : 0.77

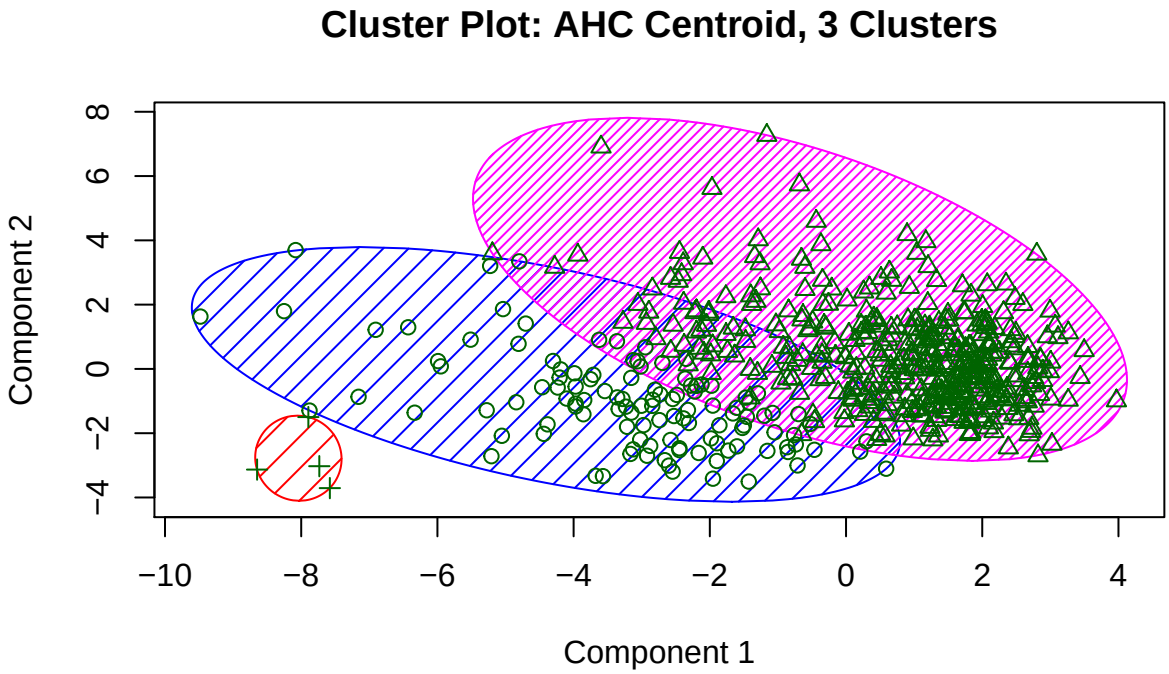


Reminiscent of single linkage, the two-cluster solution creates a tiny cluster on the first split. Interestingly, the next split recovers what looks like the two-cluster complete linkage solution; it's too bad that the agglomerative algorithm has already created an erroneous small cluster. The silhouette plots look fine but that's because although the cluster shapes and sizes are wrong, membership within the clusters looks good.

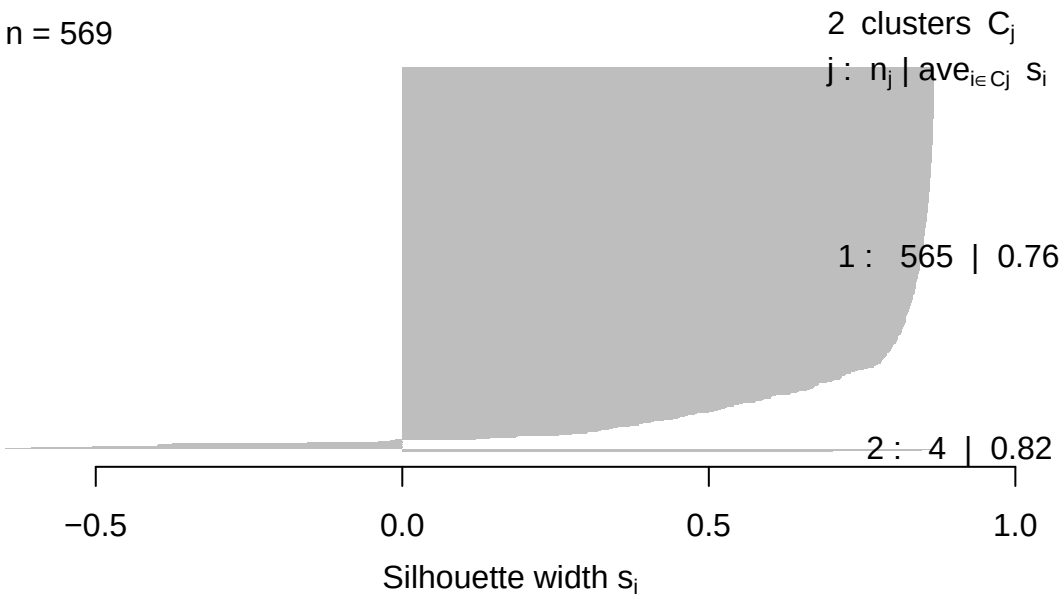
AHC with Centroid Linkage



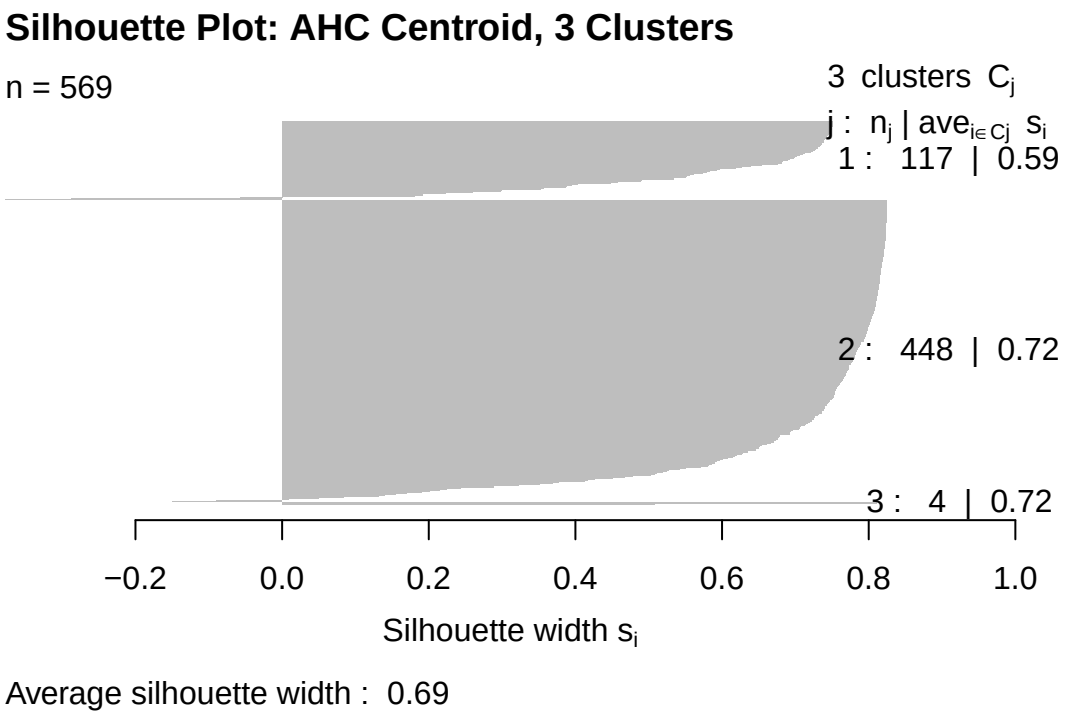




These two components explain 79.97 % of the point variability.

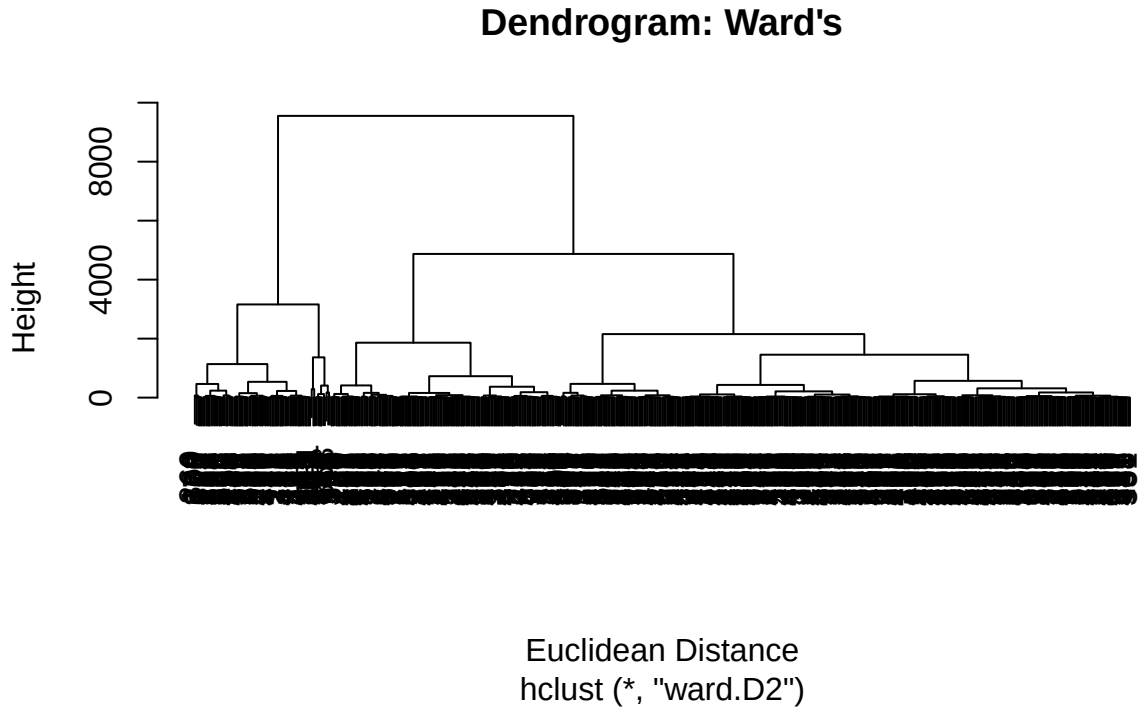


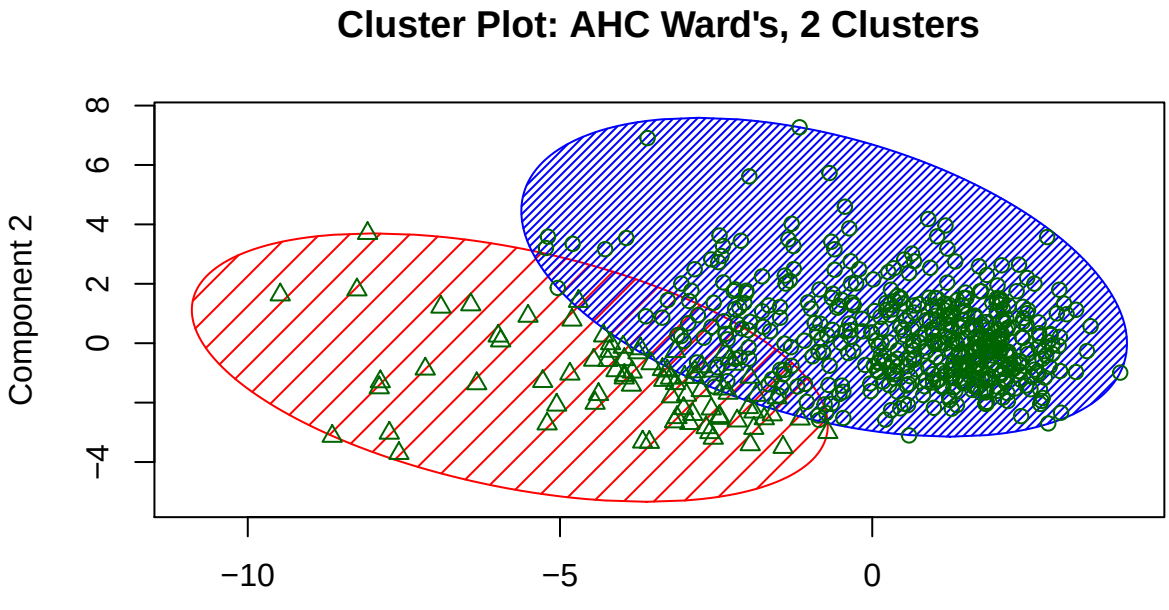
Average silhouette width : 0.76

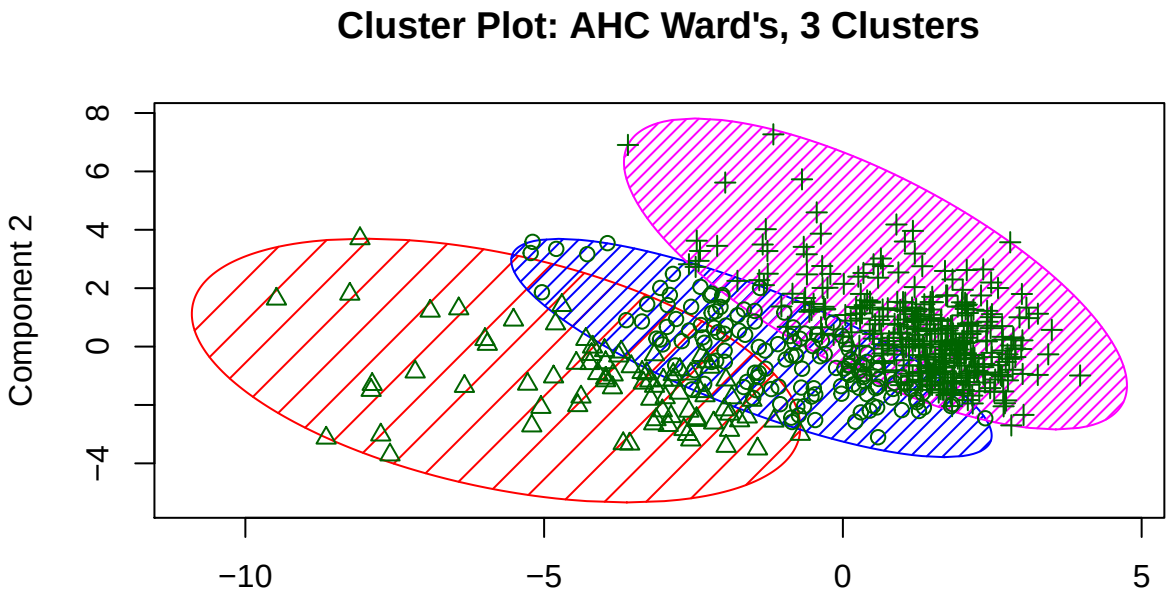


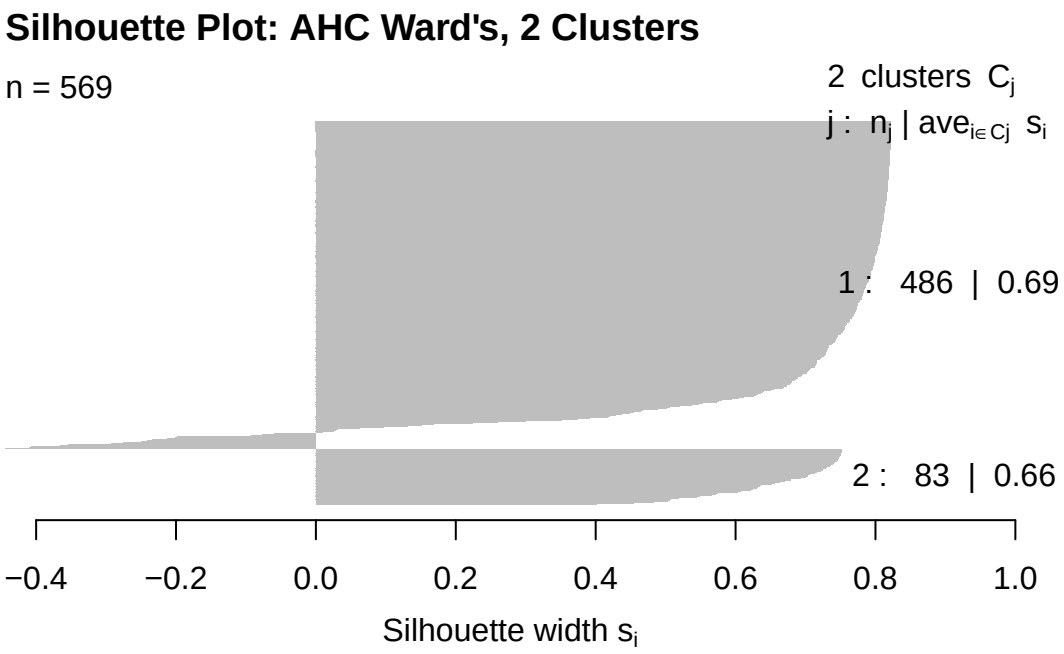
Results are extremely similar to the average linkage method. . . poor.

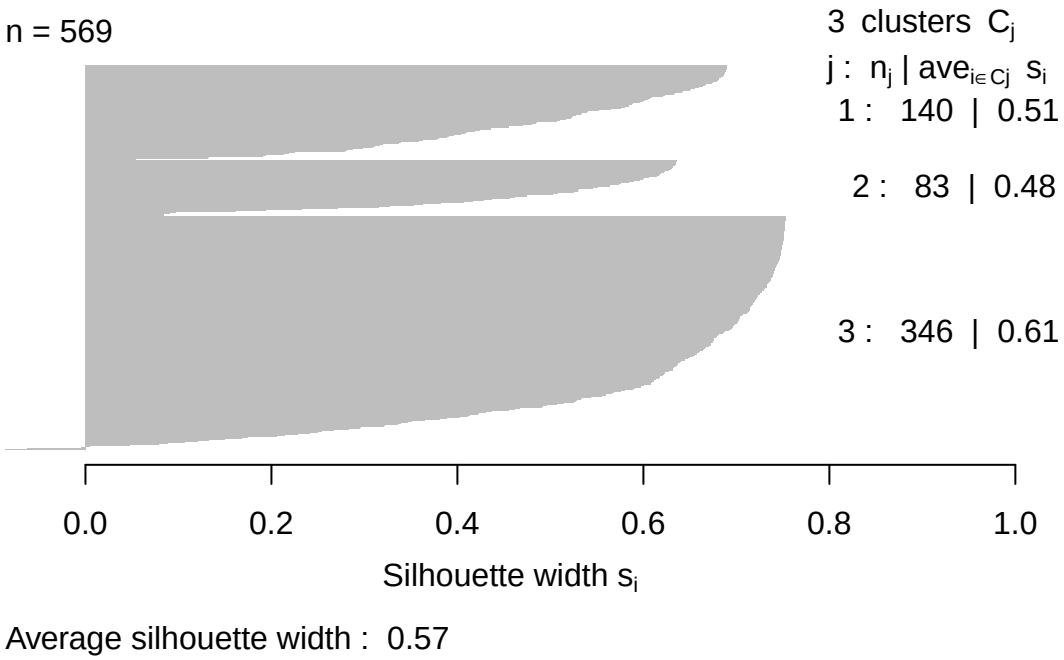
AHC Ward's Linkage




Cluster Plot: AHC Ward's, 2 Clusters
Component 2
8
6
4
2
0
−4
−10
−5
0
Component 1
These two components explain 79.97 % of the point variability.


Cluster Plot: AHC Ward's, 3 Clusters
Component 2
8
6
4
2
0
−4
−10
−5
0
5
Component 1
These two components explain 79.97 % of the point variability.





Ward's performs just like Complete Linkage. It separates two large clusters which look like our two diagnoses. Interestingly, the 3-cluster solution separates out the "boundary group" consisting of individuals with cytologic features that could just as easily lead to a benign diagnosis than a malignant one. This makes the silhouette plot for the 3-cluster solution look a lot better than that of the 2-cluster solution.

Determine the Optimal Number of Clusters

Using Ward's Linkage (the superior method), we will utilize the C-statistic and the d-statistic to determine the optimal number of clusters (K).

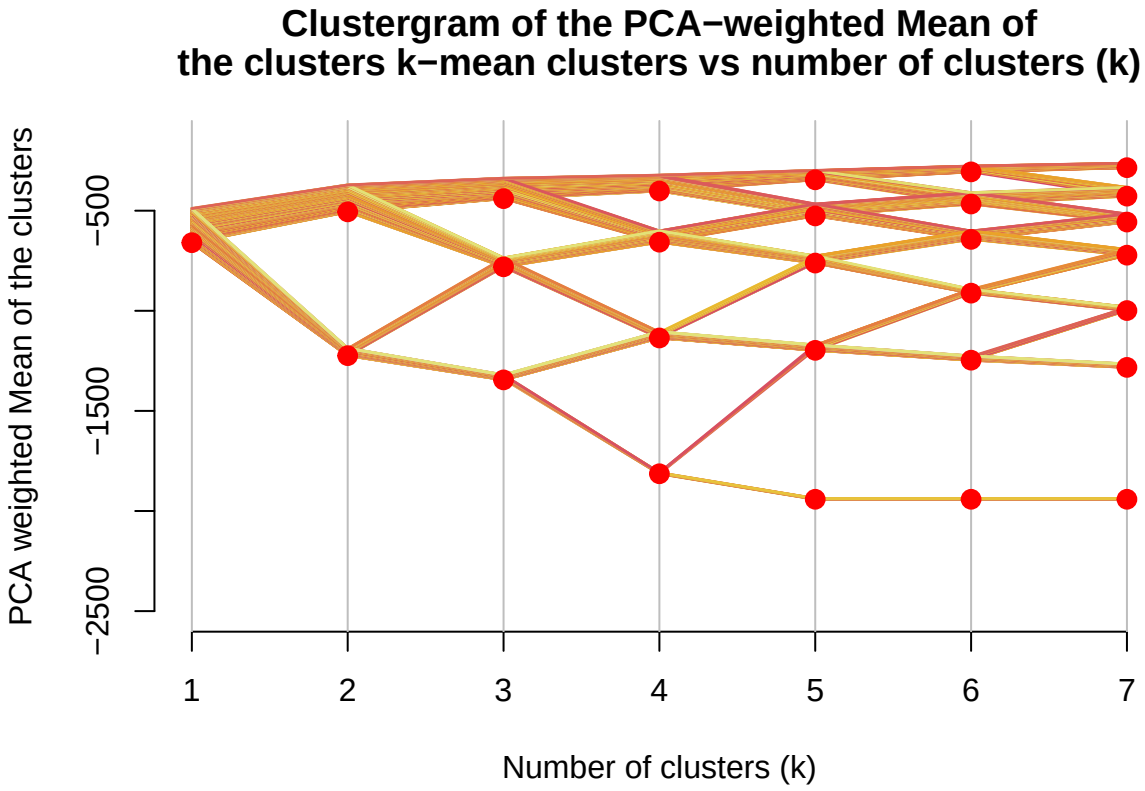
$$C = \frac{\text{tr}(B)/(K-1)}{\text{tr}(W)/(n-K)}$$

$$d = K^2|W|$$

C	d
0.00	0.000000e+00
1031.81	3.603018e+10
119.37	1.370742e+11
26.62	2.581346e+11
19.54	2.682608e+11
11.95	6.438436e+11
8.65	1.043928e+12

Based on both statistics, the two-cluster solution is optimal. This makes sense given what we know about the diagnoses! A visual method of determining the optimal number of clusters is the clustergram (Schonlau, 2002) which utilizes K-Means clustering:

K-Means clustering



The clustergram shows that using k-means for a two-cluster solution appears to produce a separation of groups (in terms of mean separation and cluster size) similar to Complete linkage and Ward's linkage. A further split results in a boundary group (again like Complete and Ward) that pulls a little from the benign group and the malignant group. The splits begin to make less and less sense after this. The clustergram suggests that a 2- or 3-cluster solution is best. Combined with our C- and d-statistics, a two cluster solution is optimal.