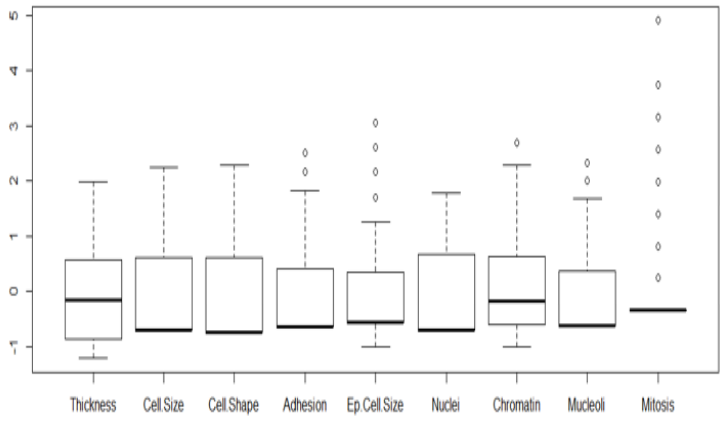
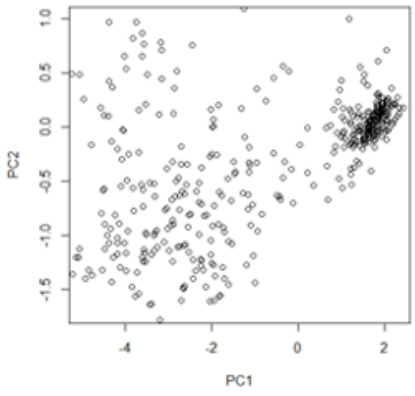
Data: The data set chosen for this analysis was sourced from the UCI machine learning repository and contains 699 observations of breast tumors which are classified as either benign or malignant. Each observation has nine independent variables that represent physical measurements of each tumor. Measurements relate to the symmetry, thickness and texture of the tumor and each measurement is given an integer value from 1 to 10 by the medical examiner. All 699 observations in the data set have been labelled as either benign or malignant and there is no missing data.[i]





The above boxplot shows the dispersion of all nine independent variables after having been scaled as well as a scatterplot representing the dispersion of the data when plotted against the first two principal components. This could be interesting later in the analysis as there clearly are two distinct regions that represent the data with the data on the left arguably having more than one distinct cluster from which we can model the data using a mixture of normal distributions.

Problem/Purpose: We have many observations which explain the relationship between the physical measurements of a tumor and its diagnosis as either benign or malignant. Our goal is to create a model that if successful, would be able to classify observations accurately into either of the two classes. If the results are accurate enough this could have significant value to the healthcare system as it could reduce human error in the diagnosis stage of a medical examination. The goal is to use mixture discriminant analysis to classify tumors as either benign or malignant using the nine independent variables described above.

Techniques: Mixture discriminant analysis can be thought of as an extension of linear and quadratic discriminant analysis. Instead of using a single multivariate normal distribution to explain the density of a given class, mixture discriminant analysis allows for the use of multiple normal distributions to model the data in each class and takes the following general form:

$$\sum_{k=1}^{G_j} \tau_{jk} \phi(\boldsymbol{y} \mid \mu_{jk}, \Sigma_{jk})$$ [ii]

Traditionally, mixture discriminant analysis requires that the covariance matrices (Σ_{jk}) be the same for each class and that the number of mixture components (G_j) for each class is known in advance. However, the use of MclustDA allows for the relaxation of these requirements and lets the component covariance matrices differ between classes and subclasses.[iii] When using MclustDA, the training data chooses the number of components that should correspond to each class as well as the corresponding covariance matrices. To determine the number of components to use within each class, each possible number of components is given a value corresponding to its Bayes information criteria (BIC).[iv] The component within each class that has the lowest BIC is used. Broadly speaking, mixture discriminant analysis allows for more closely fitted models than linear or quadratic discriminant analysis by allowing there to be a non-linear combination of normal models fitted to each class.

However, caution should be exercised when using mixture discriminant analysis as is explained by McNicholas who cites an example of where misclassification is amplified on classes that are fitted using several components instead of one.[v]

The R (v3.5.1)[vi] Mclust package (v5.4.1)[vii] was used to carry out mixture discriminant analysis on the breast tumor dataset.

Results: To account for variation in the training of the model, the algorithm was run ten times with each iteration selecting a randomly generated split between unlabeled (25%) and labelled (75%) observations. Running the model ten times produced a misclassification error rate on the training data that ranged from a minimum of 3.20% to a maximum of 6.90% with a standard deviation of 1.3%. The train and test outcomes of the model with the most accurate testing classification rate is represented by the following confusion matrices:

```
Training classification summary:
           Predicted
Class       Benign Malignant
  Benign       323        19
  Malignant      5       178

Training error = 0.04571429
```

```
Test classification summary:
           Predicted
Class       Benign Malignant
  Benign       113         3
  Malignant      2        56

Test error = 0.02873563
```

Here we can see that the training model had a classification error of 4.57% while the result of the model applied to the unlabeled testing data was 2.87%. This model had a Log-likelihood of -6305.085 which is the smallest value from all ten trials as well as a BIC of -13180.14 which is the optimal value that determines the number of components for each class. More information on BIC and Log-likelihood can be found in Hastie (2001).[viii] Furthermore, the model for the benign class was fitted with two components while the model for malignant tumors was fitted for five. This could be of interest for further analysis

since malignant tumors are not represented by a simple unimodal or bimodal distribution which could indicate further subgroups within the data. This could have a relationship with what we saw in the PCA graph above.

In order to contrast the results of the mixture discriminant analysis model produced above, a random forest was also generated. The random forest was created using the 'randomForest' (v4.6-14)[ix] package in R with the 'mtry' parameter set to 4 which produced the following confusion matrix.

	tumor.pred	
tumor.test	Benign	Malignant
Benign	224	4
Malignant	2	120

The misclassification rate for this random forest was 1.7% with an adjusted rand index of 0.93. This result is more accurate than any of the ten mixture discriminant models produced via their respective random samples and simply shows that a mixture discriminant model did not produce the most accurate classification model given the data as well as other popular classification methods available.

Conclusion: Our initial goal was to create a mixture discriminant analysis model that would be accurate enough to suggest there could be real-world value in applying such classification models to the diagnosis stage of a medical examination. Although a misclassification rate of even 2.87%, as was produced with our most accurate mixture model, would not successfully solve the issue of human error, it along with our random forest classification model shows that there could be value in conducting further research on the development of models that aid medical examiners in the classification of potentially dangerous tumors.

---

[i] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[ii] Hastie, T., Tibshirani, R.,, Friedman, J. (2001). The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc

[iii] Fraley, R. and Raftery, A. Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association; Jun 2002; 97,458

[iv] Ibid.

[v] McNicholas, Paul D.. Mixture Model-Based Classification, Chapman and Hall/CRC, 2016. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/mcmu/detail.action?docID=4709751

[vi] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[vii] Scrucca L., Fop M., Murphy T. B. and Raftery A. E. (2017) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models The R Journal 8/1, pp. 205-233

[viii] Hastie, T., Tibshirani, R.,, Friedman, J. (2001). The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc

[ix] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.