Clustering Analysis of Surgeons from Surgical Data

Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science: Combined Honours in Statistics and Mathematics

Jingyu Li

Supervisor: Dr. Edward Susko

Co-Supervisor: Dr. Huaichun Wang

Dalhousie University

April 23th, 2021

Table of Contents

1. Introduction

Whether people like it or not, hospitals are essential. In hospitals, many major or minor procedures are performed every day. Within the same field, a doctor can perform different surgeries. For example, a heart surgeon can do heart bypass surgery, and heart bypass surgery can be performed on a different vein or artery. Although these surgical procedures are different, they share some similarities. The purpose of this thesis is to cluster surgeons from two surgical datasets, determine the optimal number of clusters and try to analyze the relationships between the surgeries that doctors undertake. The processing of the data will be implemented using R.

These two datasets were provided by Dr. Huaichun Wang, who works for the Nova Scotia Health Authority (NSHA). Each dataset contains six types of surgical procedures, which are IIa-7, IIb-21, III-42, IV-91, V-182 and VI-365. These types are priority classification levels and each type of surgery has a target time corresponding to it.

Priority Classification Level	Target Time for Surgery
Priority IIa	Within 1 week
Priority IIb	Within 3 weeks
Priority III	Within 6 weeks
Priority IV	Within 3 months
Priority V	Within 6 months
Priority VI	Within 12 months

Table 1: Surgery priority classification scheme.

Surgeon ID	**IIa-7**	**IIb-21**	**III-42**	**IV-91**	**V-182**	**VI-365**
1	0	2	1	0	0	0
2	0	1	0	4	0	0
3	0	1	3	0	6	474
4	0	2	2	1094	0	0
5	0	1	9	124	0	0
6	2	0	2	2204	393	0
7	0	0	0	1899	1	0

Table 2: A few rows of data on cataract extraction surgical priority.

Surgeon ID	IIa-7	IIb-21	III-42	IV-91	V-182	VI-365
1	0	0	0	4	54	0
2	0	8	19	37	19	20
3	3	1	4	526	6	0
4	0	1	2	3	534	0
5	0	1	2	14	461	1
6	0	0	0	0	38	0
7	0	0	0	1	0	0

Table 3: A few rows of data on knee, hip, shoulder replacement surgical priority.

The first dataset is about cataract extraction surgery (hereafter referred to as 'cataract data') and there are 42 surgeon samples in it. The second dataset is about knee, hip, and shoulder replacement surgery (hereafter referred to as 'knee data'), it has 38 surgeon samples. *Table 2* and *Table 3* show some data extracted from them, each row represents a surgeon and the number of times he performed each surgery. The basic computational analysis of the data is first performed. The mean of the cataract data is 6276.833 with a standard deviation of 8516.237. The mean of the knee data was 2360.667 with a standard deviation of 2677.832. Cataract extraction surgery is being done more often than knee replacement surgery and cataract data is more discrete. *Table 4* lists the overall proportion of each type of surgery performed in both datasets. Surgeons conduct the most IV-91 and the least IIa-7.

	IIa-7	IIb-21	III-42	IV-91	V-182	VI-365
Cataract data	0.00079658	0.01696715	0.09168636	0.60205518	0.21337723	0.07511750
Knee data	0.00098842	0.00691895	0.09149958	0.38830839	0.42191471	0.09036995

Table 4: Overall proportions of types of surgery in cataract data and knee data.

Going through each doctor's data, some doctors such as ID 1 and ID 2 in the cataract data and ID 7 in the knee data, performed only one or two surgeries and had a very low total number of surgeries. Some doctors such as ID 4, ID 6, and ID 7 in the cataract data, repeated a surgery up to thousands of times. Some doctors, such as ID 2, ID 3, and ID 5 in the knee data, performed many types of surgeries. Based on the above summary of the data, different physicians may have a propensity to perform the procedure. We became interested in the relationship between the data, that is, whether there is a similar frequency of doing surgery between doctors then these doctors can be seen as a group. That is the reason why we wanted to cluster them.

Clustering is done by determining the similarity or dissimilarity between samples and then putting the similar samples together. Each group of objects divided in this way is called a 'cluster' and the samples should be sufficiently dissimilar from one cluster to another (Hastie, et al., n.d.). Clustering and classification are not the same. Classifying data is grouping them according to a

division criterion, while clustering data is needed because the specific division criterion is not known. Clustering algorithms can be divided into three types, which are combinatorial algorithms, mixed modeling and mode seeking (Hastie et al., n.d.). The most widely used of these are hierarchical clustering and K-means clustering, which are part of the methods used in this paper. Also, because of the sparsity of the data caused by some doctors doing low number of surgeries, the mixture modeling would be a better clustering method as well. In the meantime, how to determine the optimal number of clusters is also a question worth exploring. Because there is no division criterion, it is difficult to define precisely how many clusters a set of data needs to be divided into. Based on K-means clustering, the curvature method is the most common one to estimate the optimal number of clusters. It determines the number of clusters by using the elbow point where the decreasing trend of the sum of square within a cluster slows down. For the mixture modeling, the likelihood is the main consideration. The optimal number of components is usually determined by information criteria or testing.

2. Hierarchical Clustering

Hierarchical clustering can be seen as generating clustering results by determining the measure of dissimilarity (distance is often used) between two clusters based on the dissimilarities of the objects within the two clusters, which can be represented as a dendrogram (Hastie et al., n.d.). The way the dendrogram is formed can classify hierarchical clustering into two types: one is agglomerative, where each of the *n* objects is treated as a cluster by merging the two clusters with the least dissimilarity until the *n* clusters finally merge into a single cluster; another one is divisive, where the whole sample is treated as a cluster and a cluster is split into the two clusters with the greatest dissimilarity step by step until it splits into *n* clusters. Because agglomerative hierarchical clustering is relatively simpler, it is more commonly used. Based on the dissimilarities of objects within two clusters, the method with merging minimal dissimilarity is called single linkage, and the method with merging maximal dissimilarity is called complete linkage. However, these two methods are susceptible to extreme anomalous objects. Average linkage is merging mean dissimilarity, so it can get relatively suitable results.

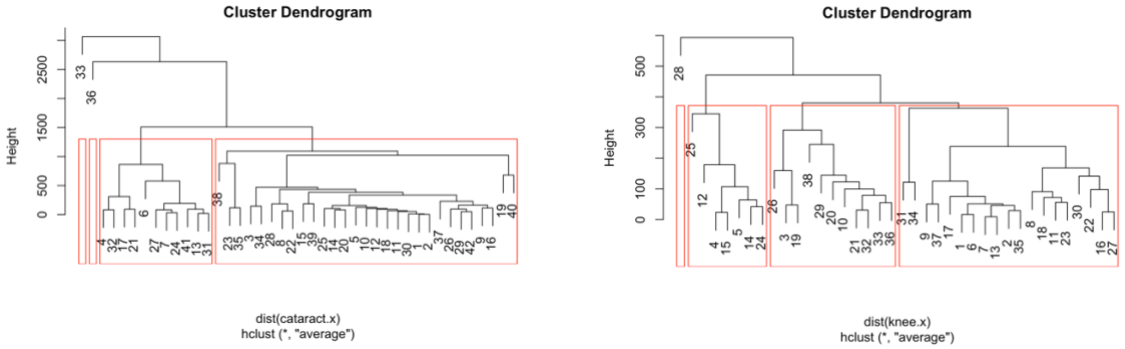

*Figure 1:*Dendrograms for cataract data (left) and knee data (right) computed by average linkage.

In R, the function *hclust* is used to implement hierarchical clustering and visualize it with *plot()*. This requires a subjective determination of how many clusters the data are divided into. In *Figure 1*, each leaf of the tree represents a surgeon. The leaves from low to high represent the dissimilarity

between clusters from small to large. The dendrogram on the left-hand side is a hierarchical clustering of the cataract data, with the number of clusters roughly divided into four. Point 33 and point 36 are two clusters that are well separated from the other clusters because these two points are at the top of the tree. The dendrogram on the right-hand side is a hierarchical clustering of the knee data, which can also be divided into four clusters based on the merging of the clusters. It can be visualized that point 28, as a cluster, has a large dissimilarity with another large cluster. In the figure, the red borders divide each tree into four groups, making it is easy to observe clusters. The size of the clusters from left to right is getting larger. The clusters are merging upwards in sequence, which indicates that the dissimilarity between clusters from right to left is increasing from small to large.

3. K-Means Clustering

The K-means clustering algorithm (Hastie et al., n.d.) is not difficult. Similar to hierarchical clustering, K-means clustering is also based on the dissimilarity of objects within two clusters and uses squared Euclidean distance as a measure of dissimilarity.

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = \|x_i - x_{i'}\|^2$$

The process of the K-means algorithm is a continuous iteration of the following steps:

a) Choose a value of *k* as the number of clusters derived after clustering.

b) Let the value of *k* be the centers $(m_1, m_2, ..., m_k)$ of the clusters, and calculate the distances of other objects to each center using the Euclidean distance formula. By determining which center the object is closer to, then let this object be in the same cluster as that center. This is the first clustering.

c) Calculate the mean value of each cluster and set it as the new clustering center, again using the Euclidean distance formula to determine the distance and grouping to get the new clustering.

d) Repeat the previous step until there is no more change in data grouping, which indicates that the clustering centers finally converge.

Since it is difficult to determine the value of *k* in advance, it is necessary to try to set multiple *k* values. The *kmeans()* function is used to cluster *k* = 2:6. Since there are not many observations in each data set, the value of *k* does not need to be large. According to *Figure 1*, we can separate both datasets into four clusters. In this case, 4 might be the best number of clusters. Because the principles of hierarchical clustering and K-means clustering are similar, we prefer to compare the results from the two methods when *k* = 4.

Table 5 shows four clusters with 3, 28, 10, and 1 observations respectively, and the clustering vector provides the specific surgeons included in each cluster. Knee data has clustering with sizes of clusters are 1, 15, 10 and 12. Both cluster 2 have the largest proportions, which means that these doctors perform procedures with a high degree of similarity. Define y_{ijg} be the count for category

j for each surgeon *i* in cluster *g* with size *l*. Define m_{jg} be the mean of category *j* in cluster *g*. Then the 'between_SS / total_SS' item

$$\frac{between\ SS}{total\ SS} = \frac{(total\ sum\ of\ squares) - (total\ within\ cluster\ sum\ of\ squares)}{(total\ sum\ of\ squares)}$$

$$= \frac{\sum_{i=1}^{42}\sum_{j=1}^{6}(y_{ij} - m_j)^2 - \sum_{i=1}^{l}\sum_{j=1}^{6}\sum_{g=1}^{k}(y_{ijg} - m_{jg})^2}{\sum_{i=1}^{42}\sum_{j=1}^{6}(y_{ij} - m_j)^2}$$

indicates the ratio of the sum of squares of the between group distances to the sum of squares of the overall distances. Since we want the distance within cluster to be as small as possible and the distance between clusters to be as large as possible, this value is better to close with 1.

Clusters Sizes	3, 28,10, 1					
Cluster means	**IIa-7**	**IIb-21**	**III-42**	**IV-91**	**V-182**	**VI-365**
1	0.6666667	6.00	11.0000	15.66667	868.00000	566.6667
2	0.6071429	20.75	118.0357	156.35714	68.14286	39.85714
3	1.0000000	3.40	6.6000	1819.20000	48.60000	0.50000
4	1.0000000	6.00	49.0000	57.00000	3038.00000	8.00000
between_SS / total_SS = 80.6 %						

Table 5: Extraction of output of K-means on cataract data.

Compare cluster vectors of hierarchical clustering and K-means clustering. For the left confusion table in *Table 6*, the true label that corresponds to cluster 1 is 26 and they have high correlation, the true label that corresponds to cluster 3 is 2. There is no same observation in cluster 2 and cluster 4. For right confusion table, only in cluster 2 has one same observation and cluster1 of hierarchical clustering has high correlation with cluster 2 of K-means clustering.

	1	2	3	4
1	26	0	1	10
2	1	0	0	0
3	0	1	2	0
4	1	0	0	0

	1	2	3	4
1	0	14	0	5
2	0	1	10	0
3	0	0	0	7
4	1	0	0	0

Table 6: Left table gives the result of hierarchical clustering (rows) versus the results of K-means clustering (columns) for cataract data. Right table gives the result of hierarchical clustering (rows) versus the results of K-means clustering (columns) for knee data.

Visualizing the clusters can be an intuitive way to determine the distribution of clusters. When the data were generated as images using principal component analysis (PCA) to reduce the

dimensionality, so the horizontal axis represents the dimensionality of the first principal component and the vertical axis represents the dimensionality of the second principal component. The purpose of PCA is to project high-dimensional data into a low-dimensional space through a few linear combinations of variables (Johnson & Wichern, 2019). Each convex region or each line is a cluster, each point with number in the cluster represents a doctor and the point in the cluster without number is the mean of cluster.

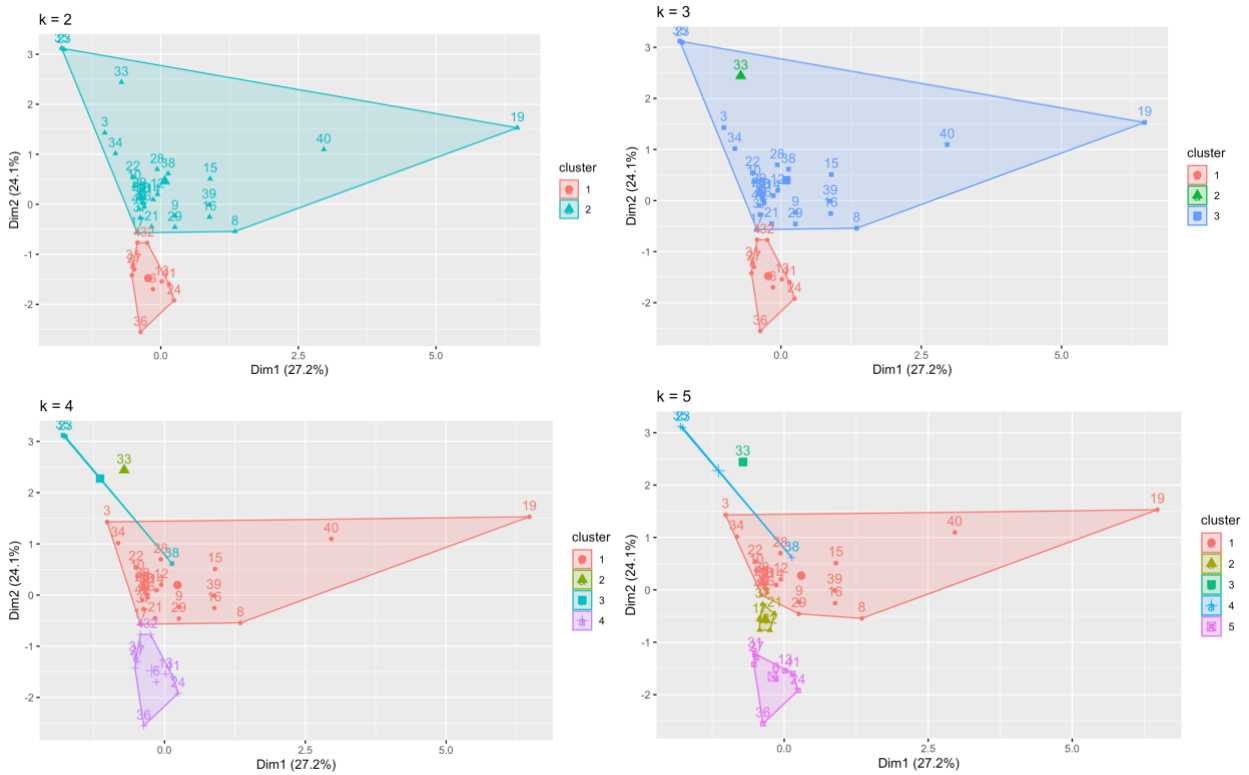

Figure 2: Scatterplots of K-means clustering on cataract data when k = 2:5.

By looking at *Figure 4*, the largest cluster will be divided when k becomes larger, so we may conclude that K-means clustering tends to generate clusters that have similar size. The cluster in the upper left part only has one observations point 33. This cluster does not change as the number of clusters increases, it might be the outlier. Because there are not many observations in the dataset, this data may be outlying behavior for the doctor, which suggests that having more observations may be better for clustering.

The use of K-means clustering depends on the number of clusters specified to be generated, so we will expect to be able to determine the optimal number of clusters. Based on the idea of K-means clustering, the curvature method (elbow method) uses wss (within sum of square) value to measure the optimal number of clusters. As the number of clusters increases, each cluster has less observations and the wss value decreases. When the wss value decreases slowly, it is considered useless to increase the number of clusters, so the "elbow point" is the optimal number of clusters. In *Figure 5* below, the value of wss decreases slowly when k = 5 on both plots, so we use 5 as the optimal cluster number for both cataract da and knee data. This is completely different from the number of clusters we judged by hierarchical clustering before, which indicates that hierarchical

clustering cannot give an optimal number of clusters when the dendritic separation of the dendrogram is not obvious.

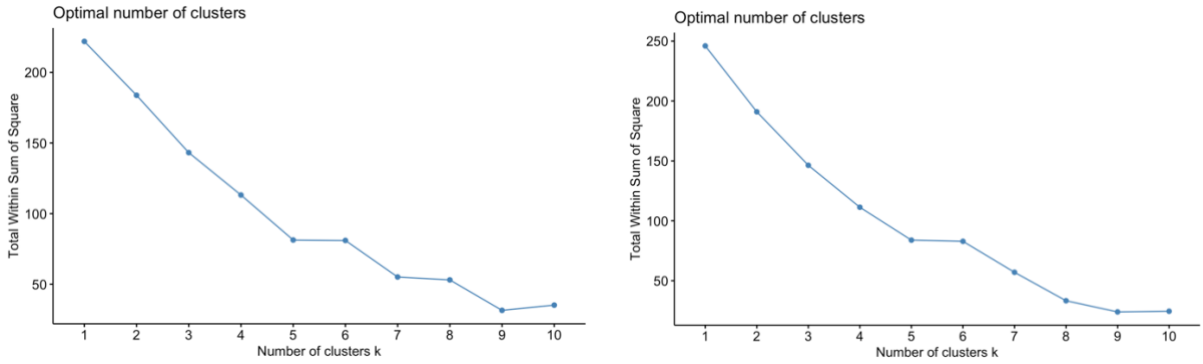

Figure 5: Both knee data (left) and cataract data (right) show the optimal number of clusters is 5.

4. Mixture Modeling

4.1 Clustering

For the mixing model, here we are interested in is multinomial mixtures because there is always one outcome that each surgeon is going to do one of these types of procedures. Suppose $y_i = [y_{i1}, y_{i2}, \ldots, y_{i6}]$ = [0, 2, 1, 0, 0, 0] where the list of numbers is the first row of *Table 2*. Each element in vector represents the number of procedure that every surgeon undertakes. Let each k_i be the total number of procedures and the distribution with probability $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_6]$ where $\boldsymbol{\theta}$ being the probability of these different procedures.

$$f(y_i; \theta) = \frac{k_i}{y_{i1}!\, y_{i2}! \ldots\ y_{i6}!} \{\theta_1^{y_{i1}} \ldots \theta_6^{y_{i6}}\}$$

So it will be $Mult(k_i, \theta)$ where k_i is size parameter and θ is probability parameter. Then suppose surgeons in one cluster prefer to do procedure IIa-7 and surgeon in another cluster prefer to do procedure V-182. In this case, probability of different clusters will depend upon the probability of individual categories. Let x_i be the indicator of cluster, so we end up with y_i from cluster j that is $y_i | x_i = j \sim Mult(k_i, \theta_j)$. But we do not know which cluster each surgeon corresponds to, so the probability of data that we actually observe is

$$f(y_i) = \sum_{x_i} f(y_i | x_i) f(x_i)$$

$$= f(y_i | x_i = 1) p(x_i = 1) \ldots f(y_i | x_i = 6) p(x_i = 6)$$

$$= f(y_i;\theta_1)\lambda_1 + \cdots + f(y_i;\theta_6)\lambda_6$$

$$= \sum_{j=1}^{k} f(y_i;\theta_j)\lambda_j$$

λ_j is the frequency with which surgeon appears in the cluster. The number of components *k* is considered as the number of clusters. We would like to use the maximum likelihood method to estimate the probability parameter θ, and to facilitate the calculation, we expect to maximize the log-likelihood function, which is

$$\prod log \sum \lambda_j f(y_i;\theta_j)$$

To determine the cluster, derive the posterior probability of cluster *j* by Bayes rule. If

$$p(j|y_i) = \frac{\lambda_i f(y_i;\theta_j)}{f(y_i)}$$

is maximal, then *i*th observation will in cluster *j*.

To implement the mixture modeling in R, we need to use the EM algorithm (Expectation–Maximization Algorithm) of Dempster et al. (1977). We use function multmixEM() in *mixtools* package in R (Benaglia et al., 2009). The value 'posterior' that multmixEM() returns gives posterior probabilities for observations, this is the basis we use to clustering. Take a cataract model with the two components as an example, 'posterior' is a 42× 2 matrix, each row is the posterior probability of component 1 and component 2. The larger probability of component in each row determines the surgeon is an observation in this component. Our goal is to determine the clusters, so let *k* = 2:5, λ and θ are NULL. Since λ and θ are randomly selected, in order not to be affected by the local maximizer, a better way is to create a loop and run the function a large number of times like 1000 times. This way will give the model with the largest log-likelihood for each *k*.

How about setting the initial values of λ and θ?

When λ and θ are NULL, then λ and each row of θ are random from uniform Dirichlet (McLachlan & Peel, 2000). By hierarchical clustering and K-means clustering, we have obtained clusters for *k* = 2:5. We can assign values to λ and θ based on these results, so that we can compare their log-likelihood values. Take K-means clustering with 4 clusters on cataract data as an example, λ is a vector of length 4 and θ is a 4 × 6 matrix. As shown in *Table 5*, the sizes of clusters are 3, 28, 10 and 1, then $\lambda = [\frac{3}{42}, \frac{28}{42}, \frac{10}{42}, \frac{1}{42}]$. To obtain θ, consider each cluster as a new data matrix. The values of each row are the probabilities of total number of corresponding procedures in a component and sum of each row should be 1. That will be initial value we would like to set.

Surprisingly, the function with no initial value setting has a larger log-likelihood value, which means that the initial value we set is a local maximizer. We can find in running the cataract data model that the best log-likelihood value of cataract data -13127.02 occurs 17 times out of 20 runs

when k = 2. When k = 3, the best log-likelihood value -6719.798 appears 10 times out of 20 runs, the best value -4152.523 appears twice when k = 4, and the best value does not occur in 20 runs when k becomes larger. Similar with knee data. This shows that when k is larger, it is more susceptible to the local maximizer. Thus, the large number of runs is necessary.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
Cataract data	0.2494911	0.2714365	0.2142931	0.1457316	0.1190476
Knee data	0.26436351	0.05267198	0.33509651	0.18857456	0.15929344

Table 7: Estimated λ of cataract data model and knee data model.

	theta.1	theta.2	theta.3	theta.4	theta.5	theta.6
comp.1	1.214224e-03	0.0068690410	0.0844501123	8.202160e-01	0.086556793	0.0006938420
comp.2	4.553781e-04	0.0019581448	0.0042724224	9.871665e-01	0.005578381	0.0005692226
comp.3	1.679902e-03	0.0105376603	0.0264203229	4.948076e-02	0.880268656	0.0316127016
comp.4	1.116016e-03	0.1378277335	0.7575204530	7.786742e-02	0.022041321	0.0036270528
comp.5	1.000000e-100	0.0004766444	0.0007149666	1.000000e-100	0.380362250	0.6184461392

Table 8: Estimated θ of cataract data model.

	theta.1	theta.2	theta.3	theta.4	theta.5	theta.6
comp.1	2.631549e-04	0.0007894647	0.003684166	0.06659013	0.926831012	0.0018420753
comp.2	1.865666e-03	0.0055969987	0.973877808	0.01119686	0.007462665	0.0000000000
comp.3	2.088235e-03	0.0144088197	0.117567616	0.84108534	0.024014699	0.0008352939
comp.4	3.819126e-84	0.0014781045	0.036213565	0.33780833	0.583852108	0.0406478874
comp.5	8.572458e-04	0.0081438350	0.042433667	0.12818103	0.323610287	0.4967739375

Table 9: Estimated θ of knee data model.

Combined with K-means clustering, we can treat the division of doctors into 5 clusters as relatively appropriate clusters. The models give the parameter λ and θ, which helps us to determine the frequency and propensity of doctors within a group to perform the procedure. λ of cataract data model shows that cluster 2 has the more frequency of arising surgeon and cluster 3 in knee data has the most frequent appearance of surgeons. For cataract data, doctors in cluster 1 and cluster 3 are most likely to perform surgery V-182, both with more than 80% chance. But doctors in cluster 1 have the next highest preference for surgery III-42, and doctors in cluster 3 have the next highest preference for surgery IV-91. Doctors in cluster 2 have the highest preference for surgery IV-91, with a 98.71% probability. In the knee data, the most interesting result is that the surgeons in both cluster 1 and cluster 4 do surgery V-182 the most and then both tend to do surgery IV-365. The reason for this clustering may be that there is this tendency within the same group but not similar between clusters.

4.2 Number of Components

McLachlan and Peel (2000) believed it is not easy to determine the number of components in the mixture distribution, and they noted that “there are two main ways: One way is based on a

penalized form of the log likelihood, the other main way is to carry out a hypothesis test, using a likelihood ratio test (LRT)" (p. 370).

4.2.1 Akaike Information Criterion and Bayesian Information Criterion

For the mixed model, a higher log-likelihood value means that this is a better model, so we usually choose the component of this model to be the optimal number of components. However, the model set in R will give higher log-likelihood values as the number of parameters increases. We need to penalize the log-likelihood value with AIC and BIC, then choose the number of components for the model which gives the largest log-likelihood value. In AIC and BIC, use the largest log-likelihood value we have got.

$$AIC = l(\hat{\lambda}, \hat{\theta}, k) - \{k - 1 + 5k\}$$

$$BIC = l(\hat{\lambda}, \hat{\theta}, k) - \frac{log(n)\{k - 1 + 5k\}}{2}$$

	AIC	BIC
k = 1	-31341.12	-31345.46
k = 2	-13138.02	-13147.58
k = 3	-6736.798	-6751.569
k = 4	-4175.523	-4195.507
k = 5	-3383.459	-3408.655

Table 10: AIC and BIC of cataract data model.

	AIC	BIC
k = 1	-9657.646	-9661.99
k = 2	-4751.157	-4760.714
k = 3	-3134.469	-3149.239
k = 4	-2237.227	-2257.261
k = 5	-1759.427	-1784.623

Table 11: AIC and BIC of knee data model.

After summarizing the results, it appears that we need 5 or more components.

4.2.2 Likelihood Ratio Tests

Take the test of null hypothesis H_o: $k = k_o$ versus alternative hypothesis H_A: $k = k_o + 1$.

$$Test\ Statistic = 2\{l(\hat{\lambda}, \hat{\theta}, k_o + 1) - l(\hat{\lambda}, \hat{\theta}, k_o)\}$$

We can consider that the test statistic is going to be χ^2 distribution with degree of freedom (1+5), where 1 is additional weight parameter λ, 5 is additional θ. But this theory does not apply here since we have less data. After each resampling run with the model, the likelihood ratio statistic was used to calculate -2logλ, but "the asymptotic null distribution of −2 log λ will not necessarily be chi-squared, as regularity conditions still do not hold" (McLachlan GJ & Peel D, 2000). Then we would like to use parametric bootstrap to achieve likelihood ratio test because we need to generate a large number of samples from model. For each sample, there exists a test statistic (likelihood ratio statistic). We would like to use the largest log-likelihood value we obtained before

to form the observed likelihood ratio statistic. The p-value can be considered as the proportion of likelihood ratio statistic of each sample greater than observed likelihood ratio statistic.

It may be implemented in R by the function *boot.comp*. There is one difference from what we do is that this function uses the data processed by the cut-point method. Therefore, creating a new function to compare with *boot.comp* is reasonable. We set the maximum number of components is 5. However, there exists a difficulty with convergence as k_o becomes larger when creating the new function. Testing the null hypothesis and alternative hypothesis of cataract data with 2 versus 3 by two functions, they have same p-value which is 0 and similar likelihood ratio statistics. The observed likelihood ratio statistics from *boot.comp* is 12396.871, which almost twice the observed likelihood ratio statistic of the new function, 6407.22. We can refer to the p-value of *boot.comp* on cataract data, the first k_o for which we cannot reject is the maximum number of components. All the p-values of *boot.comp* on knee data are 0, which suggests we may need more components.

4.2.3 Mixturegram

Mixturegram (Young et al., 2019) is a graphical representation of the number of components of the mixture model. Use the best model with k = 2:5 to create mixturegram. The first step is choosing a K as the maximum number of components, here K is 5. The second step is running these models to get the posterior probability matrices for k = 2:5. Then dimensionality reduction of this list of matrices is performed using principal component analysis and kernel principal component analysis, and the first vector from each method is transformed separately. The transformed data are then plotted on a coordinate plot and clustered by K-means clustering. The objects are grouped into clusters and assigned colors, and k-component mixture settings are performed. The number of colors for the profile is the value of k. Finally, the observation profiles that locate together are treated as a cluster to judge the value of k.

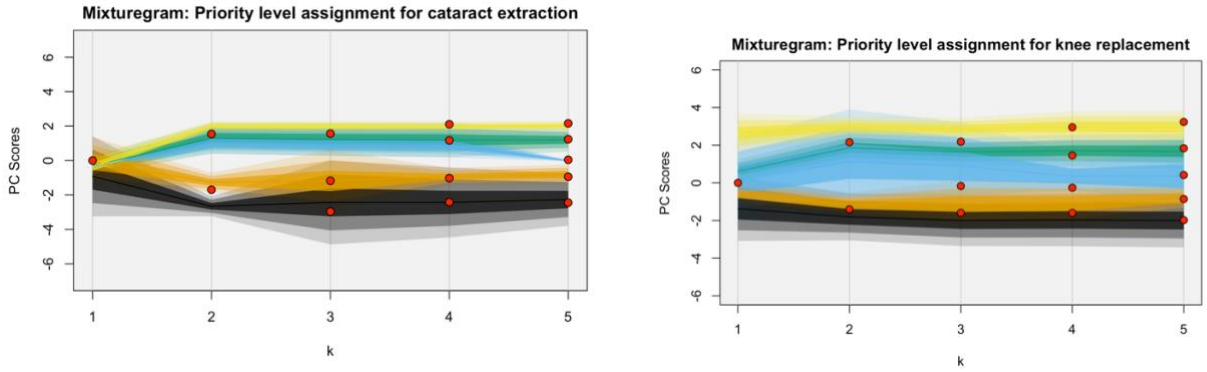

Figure 6: Mixturegrams of cataract data (left) and knee data (right) using PCA.

Figure 6 shows the mixturegrams of cataract data and knee data. The number of components on the horizontal axis and the pc scores on the vertical axis. Since we set 5 components as maximum, then there are 5 colors for profiles. Each red dot represents mean of each component group. The number of profile groups might be the number of components that we would like to choose, but when we look at the graph on the left, determining the number of profile groups can be very subjective. It can be seen as being divided into two groups: yellow, green and blue are in one,

orange and black are in another one. But it also seems to be able to be divided into three groups: yellow, green and blue are in one group, orange is a group and black is a group. So we also consider the location of the red dots. When *k* becomes larger, the red dots can be separated well and aligned well, *k* becomes better.
The profiles in mixturegram of knee data lie together, they cannot be separated into groups. It can be treated as the limitation of mixturegram. This method does not accurately determine the number of components, it only serves as an auxiliary way.

5. Conclusion

In this thesis, hierarchical clustering, K-means clustering and multinomial mixture models are used to cluster the two datasets. The clusters generated by all three methods are somewhat different. But the generated clusters will always have a small number of outliers, possibly because there are only a few dozen samples, and there will be a large dissimilarity in the propensity to perform the procedures between doctors. Because hierarchical clustering and K-means clustering are based on distance and are more susceptible to outliers, the results of the mixture model with larger log-likelihood values are relatively better. According to the results of the best mixture models of two datasets, doctors are most inclined to do the procedure among the III-42, IV-91, V-182 and VI-365. On the contrary, surgical IIa-7 and IIb-21 are rarely conducted.

At the same time, we also want to identify the optimal number of clusters. The elbow method used for K-means clustering shows both optimal number of clusters to be 5. The AIC and BIC used to penalize log-likelihood values, resampling for likelihood ratio test, and the mixturegram all show that more clusters are better. But if the optimal number of clusters is set to a large number, there will be many clusters with only one or two observations inside, which makes no sense. Overall, we prefer to separate these two datasets into 5 clusters. When we have more samples we can set the optimal number of clusters larger. So how to be able to accurately determine the best number of clusters still requires working.

References

Benaglia, T., Chauveau, D., Hunter, D., & Young, D. (2009). *mixtools: An R Package for Analyzing Mixture Models.* Journal of Statistical Software, 32(6), 1 - 29. DOI: http://dx.doi.org/10.18637/jss.v032.i06

Derek S. Young, Chenlu Ke & Xiaoxue Zeng (2018) *The Mixturegram: A Visualization Tool for Assessing the Number of Components in Finite Mixture Models,* Journal of Computational and Graphical Statistics, 27:3, 564-575, DOI: 10.1080/10618600.2017.1398093

Hastie, T., Tibshirani, R., & Friedman, J. (2nd ed.). *The Elements of Statistical Learning.* Springer. DOI: 10.1007/b94608

Johnson, R., & Wichern, D. (2019). *Applied multivariate statistical analysis*. Uttar Pardesh, India: Pearson India Education Services.

McLachlan GJ & Peel D (2000). *Finite Mixture Models.* John Wiley & Sons, New York.