Clustering Final - Group Portion
Orange Team 2: Qing Feng, Julie Huang, Carlos Chavez, Andrew Moolenaar, Bill Jenista

a) Perform a principal components analysis on columns 2 through 65. List the standard deviations for the first 5 components.

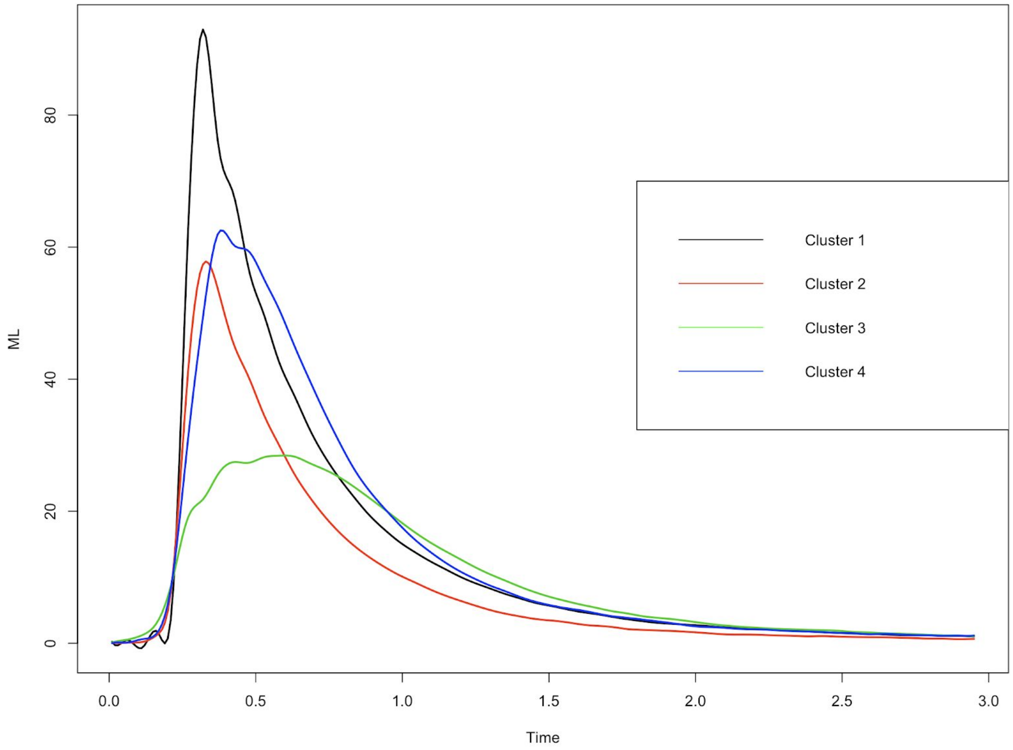
| Principal Components (PC) | Standard Deviation |
|---|---|
| PC1 | 45.21175 |
| PC2 | 31.29008 |
| PC3 | 22.37538 |
| PC4 | 17.32995 |
| PC5 | 13.04712 |

b) Using all pca scores compute the optimal number of clusters using kmeans using both "wss" and the "silhouette" method. What is the optimal number of components using each method. Why may this number be different?

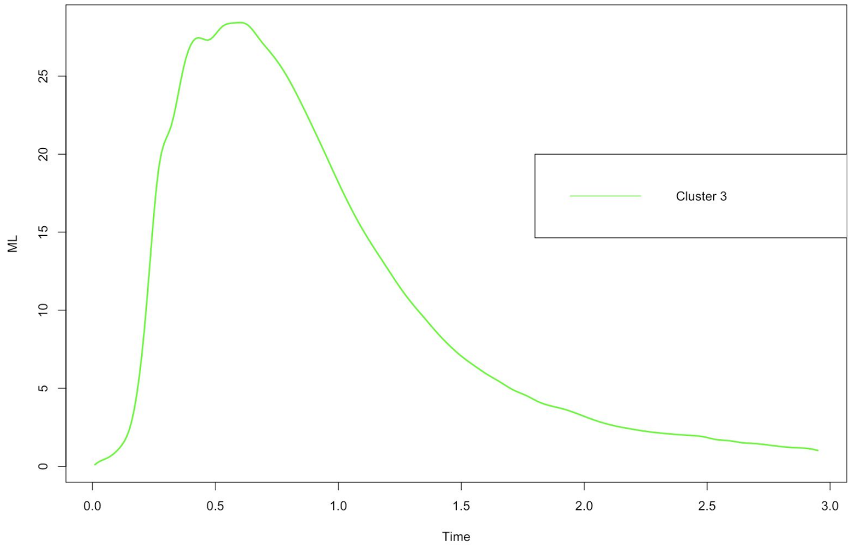
The optimal number of clusters using the “wss” method ranged between 2-5 clusters based on where the elbow is located. The optimal number of clusters using the “silhouette” method was 2. This can differ because these are computing the clusters using different algorithms. The “wss” method tries to minimize the within cluster sum of square, whereas the “silhouette” method tries to maximize the distance between ‘average within cluster distance’ and ‘average between cluster distance’

c) Run the command "set.seed(12345)" and run a k-means clustering algorithm using the pca scores.

a) Compute the graph of mean spirometry for the 4 clusters (all 4 on one graph).

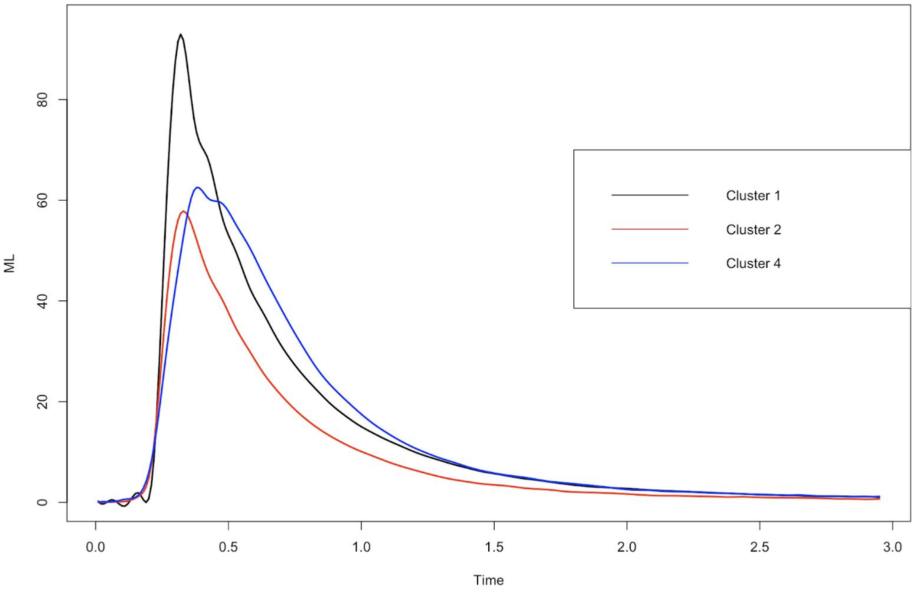


b) Look at cluster 3. Plot the graph of this cluster and give the mean values (on the original scale) for columns 2-65. What makes this cluster different from the other clusters? Describe this cluster so a physician can better understand important characteristics of these clusters.

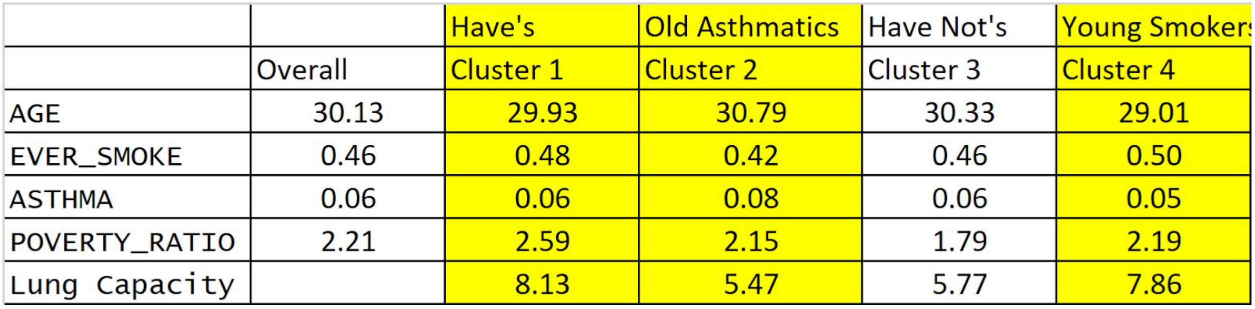


Cluster 3 is much flatter than the other clusters. This means that the initial intake of air is not very large, but it is maintained just as long if not longer than the other clusters. People in this cluster are on average 30 years old and match up with the sample average for people that smoke and have asthma, 46% and 6% respectively. These people are poorer than the average within the sample (1.79 poverty ratio). Overall, people in this cluster do not have the lowest lung capacity but are near the bottom.

c) Looking at clusters 1,2, and 4 which clusters has the largest lung capacity? which one has the least lung capacity? Describe these three groups in terms of the curves as well as the additional variables that are available in the data frame cdata. Provide figures with your descriptions.



Cluster 1 (black) has the greatest area under the curve and thus has the largest lung capacity. Cluster 2 (red) has the least area under the curve and thus has the least lung capacity. Cluster 4 (blue) has a somewhat large area under the curve, but not the largest lung capacity when compared to Cluster 1.

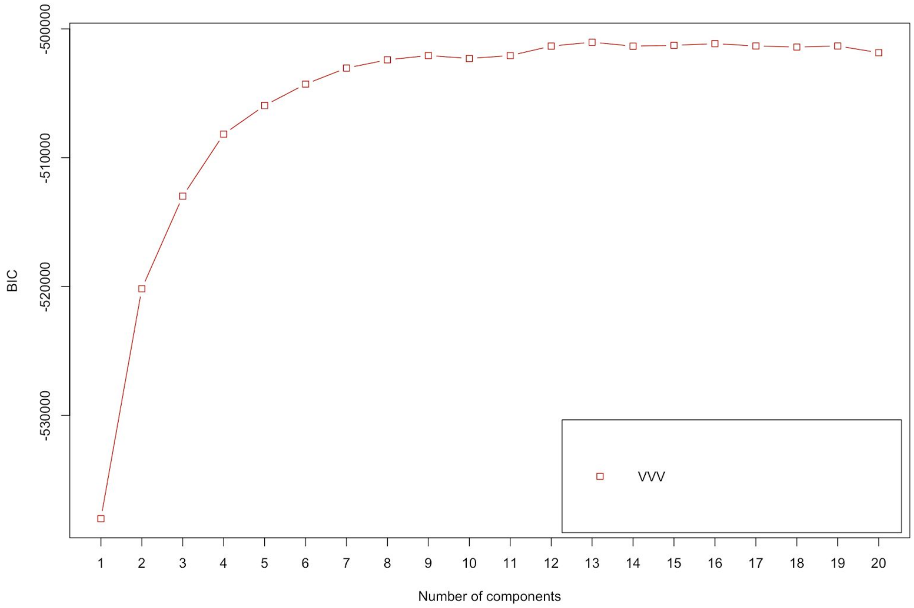


| | | Have's | Old Asthmatics | Have Not's | Young Smokers |
|---|---|---|---|---|---|
| | Overall | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
| AGE | 30.13 | 29.93 | 30.79 | 30.33 | 29.01 |
| EVER_SMOKE | 0.46 | 0.48 | 0.42 | 0.46 | 0.50 |
| ASTHMA | 0.06 | 0.06 | 0.08 | 0.06 | 0.05 |
| POVERTY_RATIO | 2.21 | 2.59 | 2.15 | 1.79 | 2.19 |
| Lung Capacity | | 8.13 | 5.47 | 5.77 | 7.86 |

Cluster 1 has the highest poverty ratio, so it is made of “rich” people.
Cluster 2 is made up of older people who have a relatively higher asthma rate.
Cluster 4 is made up of younger people who smoke.

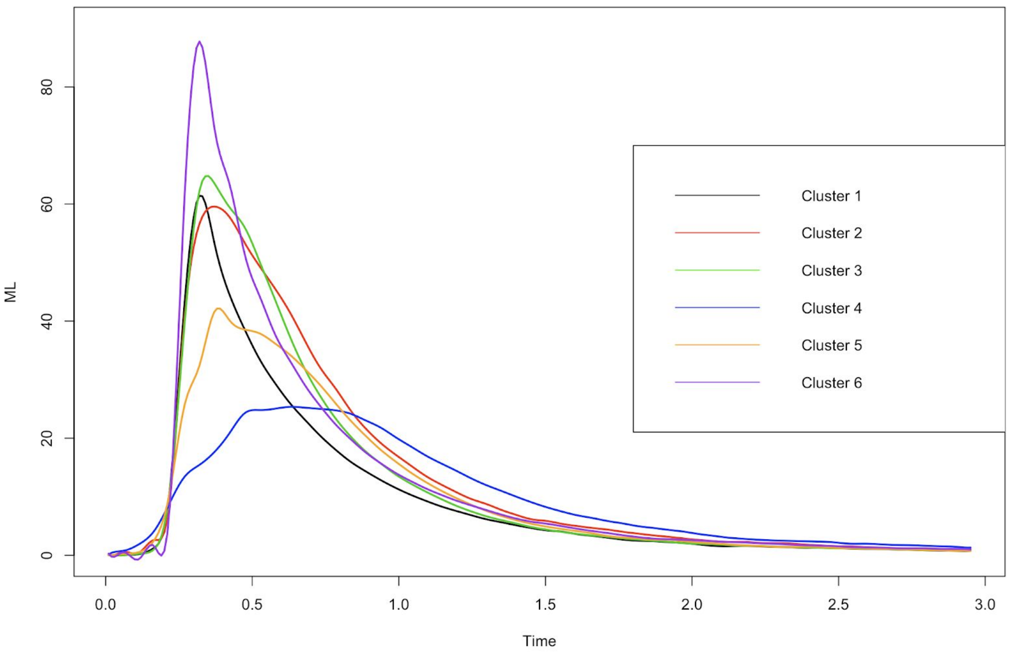
NOW look at the data using MCLUST type 'set.seed(12345)':

a) Using mclustbic() and columns 10-20 of cdata (NOT the principal component values). estimate the optimal number of cluster components using the BIC and only with modelNames='VVV' and G = 1:20. Show a graph of the estimate. Is this number different than the ones given above, why? (This will take a while).



When we use mclustbic() on just columns 10-20 of cdata, we found that the optimal number of clusters was 13. The differs notably from kmeans. The most obvious reason for this difference is because we are only using 11 betas (which are probably among the most significant) in mclustbic(), whereas we were using all the the betas in kmeans. The other reason for this difference is that kmeans and mclustbic() are two different algorithms. Kmeans is a Hard Assignment i.e. We are certain that particular points belong to particular centroid and then based on the least squares distance method, we will optimize the place of the centroid; mclustbic(), on the other hand, uses the probability of a sample to determine the feasibility of it belonging to a cluster.

b) Now using G = 6 and modelNames='VVV' and the same columns, provide a graph of each cluster's mean curve (USING ALL OF THE DATA COLUMNS). Put all plots on one graph.

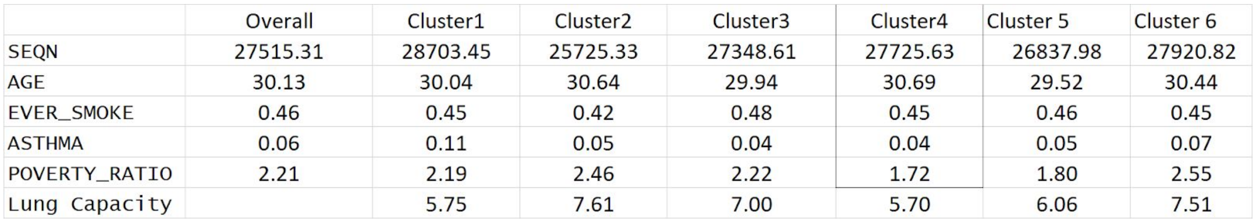


c) Using all of the data compare cluster 4 with cluster 3 from the kmeans() cluster what can you say about the similarities between these two clusters, what are the differences? Which estimate makes more sense? What do you trust more? What are the benefits of using mixture modeling over kmeans, what are the issues?

Both Cluster 4 from mclust and Cluster 3 from kmeans seem to have similar lung capacity, and they are both the lowest of their respective clusters. They both peak and taper of at similar times. Additionally, both of these clusters have similar ages, smokers, and poverty ratios. They differ in that Cluster 4 from mclust has fewer people with asthma. Because of this difference, we think that the estimates from kmeans' Cluster 3 make more sense since we would expect people with low lung capacity to have a greater probability of having asthma. In Cluster 4 of mclust, the estimate for asthma seems too low. Thus, we trust Cluster 3 from kmeans slightly more.

The mixture model works better when two (or more) clusters are embedded within another, in such case, observations are hard to be clustered. Instead of saying something belonging to one cluster, mixture modeling makes 'soft' clusters because there is never 100% probability an observation falls into one cluster or another.

d) Are there any clusters similar to the k-means clusters? Describe each cluster.



| | Overall | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster 5 | Cluster 6 |
|---|---|---|---|---|---|---|---|
| SEQN | 27515.31 | 28703.45 | 25725.33 | 27348.61 | 27725.63 | 26837.98 | 27920.82 |
| AGE | 30.13 | 30.04 | 30.64 | 29.94 | 30.69 | 29.52 | 30.44 |
| EVER_SMOKE | 0.46 | 0.45 | 0.42 | 0.48 | 0.45 | 0.46 | 0.45 |
| ASTHMA | 0.06 | 0.11 | 0.05 | 0.04 | 0.04 | 0.05 | 0.07 |
| POVERTY_RATIO | 2.21 | 2.19 | 2.46 | 2.22 | 1.72 | 1.80 | 2.55 |
| Lung Capacity | | 5.75 | 7.61 | 7.00 | 5.70 | 6.06 | 7.51 |

Cluster 1 (“Asthmatics”) is average except for a higher proportion of asthmatics. Their lung capacity is close to the lowest with an average peak that tapers off quickly.
Cluster 2 (“Monied non-Smokers”) is made up of more “richer” people with fewer smokers. They have the highest lung capacity with an average peak that tapers off the slowest.
Cluster 3 (“Smoking, Non-Asthmatics”) is made up more than average smokers and less than average asthmatics. Their lung capacity is above average.
Cluster 4 (“Older Poorish People who are not Asthmatic”) is made of older people closer to the poverty limit and contains fewer asthmatics. This group has the lowest lung capacity that has a low peak capacity the tapers off the slowest.
Cluster 5 (“Young Poorish People”) is younger than average people who are closer to the poverty line. They have an average lung capacity.
Cluster 6 (“Just Monied”) is average except they are the highest above the poverty line and have the second highest lung capacity with a high peak that tapers off quickly.

Cluster 1 from k-means and Cluster 6 from mixture clustering are similar.
Cluster 3 from k-means is similar to Clusters 4 and 5 from mixture clustering.
Cluster 4 from k-means is somewhat similar to Cluster 3 from mixture clustering