Final Test (Team Portion)

**BEFORE EACH Mclust COMMAND kmeans COMMAND or mclustBIC COMMAND use**

**THE 'set.seed(12345)' COMMAND to get the same answer as me.**

**We are looking at the spirometry data shown in the final lecture.**

**The code below will prepare the data.**

**For the file 'final\_data.Rdata' run the following code:**  
library(splines)  
times <- seq(1,295)/100 # Observations in 1/100th of a second  
X <- bs(times,intercept=TRUE,df=60) #create a spline to   
                                    #model the data

betas <- matrix(0,ncol=60,nrow = 6792)  
###########################################################  
# run a linear regression on each data set  
# here I am manipulating my data you I can cluster  
###########################################################  
for (ii in 1:6792){  
  temp <- lm(as.numeric(final\_data[ii,6:300])~X-1) #-1 removes the natural intercept  
  betas[ii,]  <- coefficients(temp)  
}

cdata <- cbind(final\_data[,1:5],betas)

#CONVERT EVERTYING TO 'numbers'  
cdata$AGE <- as.numeric(cdata$AGE)  
cdata$EVER\_SMOKE <- as.numeric(cdata$EVER\_SMOKE)  
cdata$ASTHMA <- as.numeric(cdata$EVER\_SMOKE)  
cdata$POVERTY\_RATIO <- as.numeric(cdata$POVERTY\_RATIO)

**Now:**

**a) Perform a principal components analysis on columns 2 through 65. List the standard**

**deviations for the first 5 components.**

**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?**

**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.**

**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.**

**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).**

**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?**

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