Homework 2

46-923, Fall 2017

Due Thursday, November 9 at 3:00 PM

You should submit the Rmd and pdf file for Question 1. You should also submit a pdf file with your responses to Questions 2 through 4. There is nothing wrong with handwritten solutions; I am not asking you to learn Latex to complete the homework.

Please do not submit photos of your homework. Scanners are available for your use.

Question 1:

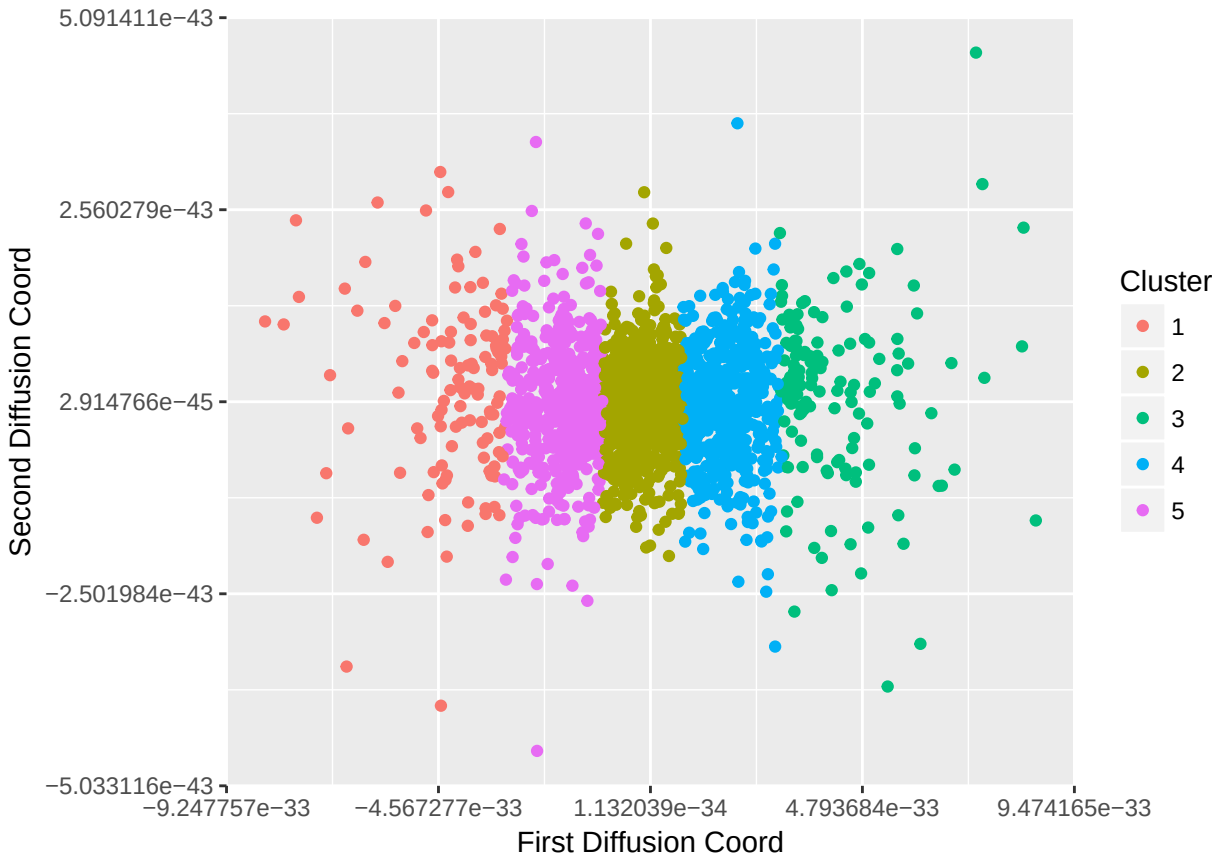
Run a **both** K-means and hierarchical clustering algorithm on the yield curve shift data that was considered back when PCA was introduced. Discuss anything interesting that you find. You are free to make decisions regarding settings to the algorithms as you see fit.

```
## Scrape the data
library(ggplot2)
library(xml2)
library(rvest)
library(diffusionMap)
library(reshape2)

fullYCweb =read_html("https://goo.gl/j97141")
tvdnodes =html_nodes(fullYCweb, ".text_view_data")
tableelements =html_text(tvdnodes)

#Change N/A in NA
tableelements[grep("N/A", tableelements)] = NA
YCdata =matrix(tableelements, ncol=12,byrow=TRUE)
YCdata =data.frame(YCdata, stringsAsFactors=FALSE)
names(YCdata) =c("Date","1mo","3mo","6mo","1yr",
                 "2yr","3yr","5yr","7yr","10yr",
                 "20yr","30yr")
YCdata$Date =as.Date(YCdata$Date,format="%m/%d/%y")
YCdata[,2:12] =apply(YCdata[,2:12],2,as.numeric)
YCrates = YCdata[YCdata$Date > "2010-01-01", -1]
YCrates = YCrates[-which(apply(YCrates,1,sum) == 0),]
YCshifts = apply(YCrates,2,diff)

## K-Means against dimension reduction
km_out = kmeans(YCshifts[complete.cases(YCshifts),], centers=5, nstart=10)
YCshifts_cpy = YCshifts[complete.cases(YCshifts),]
YCshifts$km_clust = factor(km_out$cluster)
YCshifts = data.frame(YCshifts)
```

```
dist_Map = dist(YCshifts_cpy[complete.cases(YCshifts_cpy),])
diff_map = diffuse(dist_Map, eps.val=50, t=10)
```

```
## Performing eigendecomposition
## Computing Diffusion Coordinates
## Used default value: 2 dimensions
## Elapsed time: 3.818 seconds
```

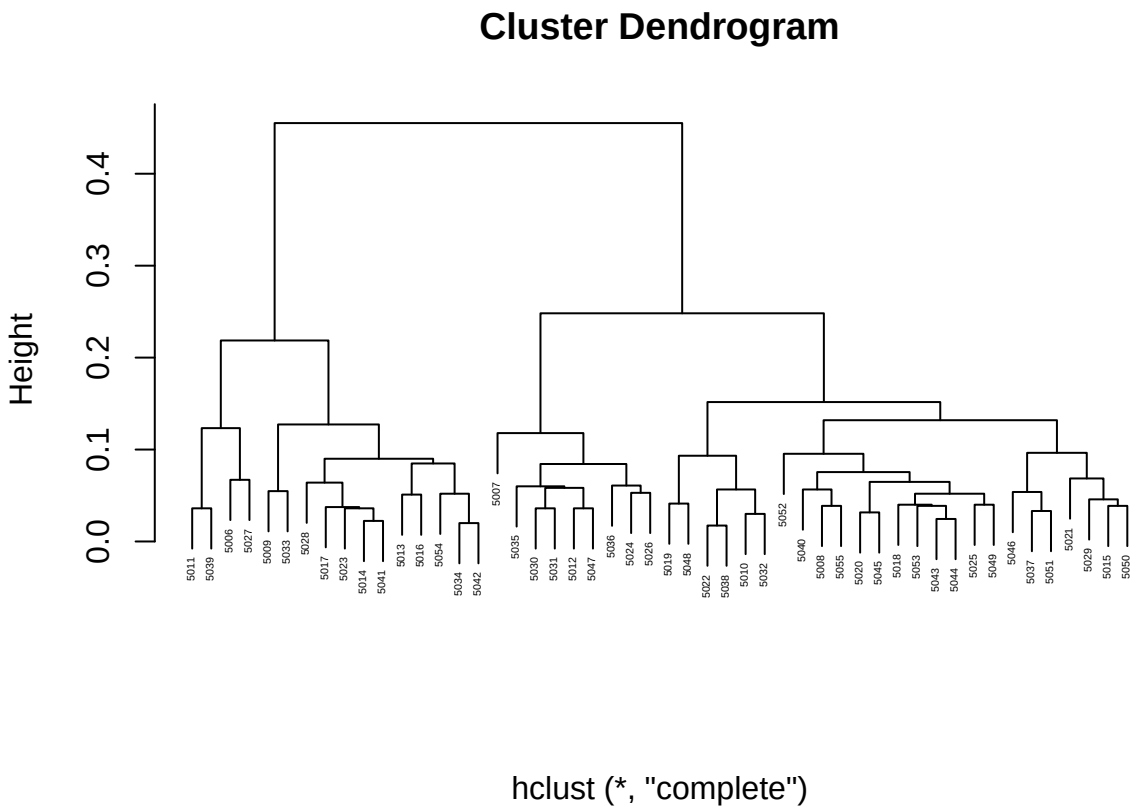
```
YCshifts$diff_1 = diff_map$X[,1]
YCshifts$diff_2 = diff_map$X[,2]
YCshifts$clust = km_out$cluster

## convert to df for ggplot
YCshifts = data.frame(YCshifts)

plt = ggplot(YCshifts, aes(x=diff_1, y=diff_2, colour=km_clust)) +
  geom_point() +
  labs(x="First Diffusion Coord", y="Second Diffusion Coord", color="Cluster")

plt
```



In this lower dimensional representation of the data, there is clearly some clustering displayed. The k-means algorithm was run with 5 dof; cluster two is much more contained while clusters

five and three are more spread out. There seems to be some symmetry in this representation of the data, specifically around the point (-1.132x10^{-34},-2.9147x10^{-45}). As one moves away from this point, the variability in the data (i.e. how spread out it is) seems to increase.

```
## Hierarchical
hcout = hclust(dist(YCshifts_cpy[1:50,]), method="complete")
plot(hcout, cex=0.35, xlab="")
```



“‘