Functional Data Analysis (AMS Spring 498B)

Homework 1

This assignment should be submitted to Gradescope in pdf foramt. You can work together in groups but submit the homework separately. .

◇ For the parts that involve computing submit a copy of your source code and output and/or plots. Use the rMarkdown format in R studio to include written answers along with your analysis. Be thrifty in what you include in your output and avoid listing extraneous matrices and vectors.

◇ All figures should be well-crafted and include labels for the axes and a title.

◇ Ten (10) points are given for each separate item following a ● so some problems count for more than others.

For these problems recall the definition of *depth* of an observation is given in the second lecture. These problems will also lead you through developing skill in R . If you get stuck on the coding use the web for searching examples, look at the code posted for the plots from lectures, come into office hours, or send email questions to `nychka@mines.edu`.

1. Suppose you have a data set where $y_1, \ldots, y_n$ are just numbers and to make this problem simpler assume n is odd and there are no ties. Consider the $\binom{n}{2}$ intervals formed from taking all the pairs of values of these data. For example if the data set was $\{1, 4, 5\}$ there would be three intervals: [1,4] [1,5] and [4,5].

   Now given a value of α for a data set count the number of intervals that contain this value, call this the *depth* of α. In the example above if $\alpha = 2$ then 2 intervals cover 2 and one does not.

   ● Show (prove or give a convincing picture) that the usual median statistic of a data set is also the value of α that has maximum depth.

2. The motivation for problem is to understand how the depth of a sample stays the same under a data transformation such as the *log*

   ● Suppose you have a data set where $y_1, \ldots, y_n$ are just numbers, n odd and all values different. From the first problem we know that the deepest observation is also the median. Now take the log of these data $z_1, \ldots, z_n$ with $z_i = log(y_i)$ and show that the deepest observation stays the same. (and equivalently $median(\{z_i\})$ is $log(median(\{y_i\}))$

   ● Given a data set where $y_1(t), \ldots, y_n(t)$ are now curves. Consider the new transformed data set by taking logs $z_1(t), \ldots, z_n(t)$ with $z_i = log(y_i)$.

For each i show that the depth of z_i is equal to the depth of y_i.

3. The Bolder Boulder race results for 2013 are in the R data set BB10K.rda. `head( BB10K)` is a handy R function to list the first 5 rows of these data. Note that not all columns are numbers; some are text strings. To simplfy this homework just look at the mile split times (columns 5 through 10) for the 30 year old age category. To read and then pare down these data into a handy subset, in R

```
# edit this next line so that the folder path
# points to where you have out the data set.
   dataLocation<- "/Users/nychka/Home/Teaching/FDA/data/BB10K.rda"
   load( dataLocation)
#  or if the data set is in your working directory:  load("BB10K.rda")
# just grab the spit times and switch rows and columns
   split<-  as.matrix(BB10K[,4 + 1:6])
   split <- t( split)
#Now rows index the splits,  columns index the runners
  indM<- BB10K$DIV == "M30"
  indF<- BB10K$DIV == "F30"
# select only female 30 years old
  split30F<- split[,indF]
# total race time
  timeF30<- BB10K$TIME[indF]
# the guys ...
split30M<- split[,indM]
```

   - I think `split30F` is an example of "functional data". Argue why it is.

   - Argue why `split30F` is not really a good example of "functional data".

   - Create a figure that gives boxplots for each F30 split time plotted against the mile. Note that to use the `boxplot` function you will have to take the transpose of splitF30 (`t(splitF30)`) because it will need the columns indexing the split times and rows being the runners. Based on the classic boxplot criterion are there outliers?

   - Now create a figure for the F30 based on a functional boxplot. This is the function `fbplot` from the `fda` R package. Does this more sophisticated plot give you different insight in to the distribution of the data set? Based on the functional boxplot criterion are there outliers?

4. Working with centered data. Subtract the mean split time from each runner's performance :

```
split30FCentered <- scale(split30F, center=TRUE, scale=FALSE)
```

- Create an fBoxplot of these centered data and comment on the differences with the plot of the raw split times.

- For the centered data is there a relationship between the *depth* of a runner's times and their total time? How about if you just focus on the outliers?