**Math 537, Summer 2020, Exam 2.**

Exam 2 consists of two data analysis problems for you to complete. The exam is due Friday, June 19th at 11:59 PM. Please submit a single file on Titanium that includes your written response to the exam along with any relevant code/output.

The instructor may schedule an interview with you to discuss your exam where you may be asked to elaborate on your solutions.

**Problem 1: 40 points**

Download the Apple.txt dataset, which includes the data which compares apple trees from different rootstocks. Eight trees (observations) are measured from six different rootstocks. The variables of interest are:

* = Trunk girth at 4 years of age (m/10)
* = Extension growth at 4 years (m)
* = Trunk girth at 15 years (m/10)
* = Weight of the tree above ground at 15 years (lb\*1000)

1. (10 points) Do you think if the mean vector of the 4 variables above is the same across the 6 rootstocks? Write an appropriate hypothesis to test this and provide a detailed statistical analysis to test your null hypothesis. Interpret your findings in the context of the problem.
2. (10 points) Analyze the residuals from (a). Do the usual MANOVA assumptions appear to be satisfied. You don’t need to do a formal test for Independence, but casually discuss.
3. (20 points) Calculate the p-value from parts a.) and b.) using a bootstrap approach. Briefly discuss the strengths and weaknesses of the three approaches you’ve considered (MANOVA, trace test, Bootstrap).

**Problem 2: 60 points**

The DrivFace database contains image sequences of subjects while driving in real scenarios. It is composed of 606 samples of 640,480 pixels each, acquired over different days from 4 drivers (2 women and 2 men) with several facial features like glasses and beard.

A set of labels assigning each image into 3 possible gaze direction classes are given (variable 4 in the dataset: lr- looking right, f – looking front, and lf-looking left). Variable 5 gives the angle associated with the head direction (lr: -45 to 30, f: -15 to 15, lf: 30-45). Consider these response variables, the first three variables are just identifying information.

Variables 6-9 contain information on face position (x,y,w,h) are features for location, width and height.

Variables 10-11 are x and y coordinates for the right eye position

Variables 12-13 are x and y coordinates for the left eye position

Variables 14-15 are x and y coordinates for the nose position

Variables 15-17 are x and y positions for the right corner of the mouth

Variables 18-19 are x and y positions for the left corner of the mouth

1. (40 points) Perform a full PCA analysis of the data using variables 6-19 to predict either variable 4 or 5. How many principal components did you settle on and why? Please interpret some of the loading coefficients for a few of your most prominent components.
2. (20 points) Perform a Factor analysis of the data using variables 6-19. How many underlying features do you think account for the observed variables, which features are loaded from the same factors? Why can’t you use Factor Analysis to predict variables 4 or 5 like Principal Component Analysis?