# Homework 1 – Due 12 am CST, 18 February 2021

The total points on this homework is 100.

1. The exercise here is to provide us with some facility with R but mainly to highlight the value of using multivariate statistics. The application comes from fracture mechanics as applied to forensics. I will first explain the application and then the problem (and illustration).

   We can consider an example of a crime scene where investigators found the tip of a knife or other tool which broke off from the rest of the object. Later, investigators recover a base which appears to match and they wish to show the two pieces are from the same knife in order to use that evidence later at trial. We evaluate 9 knives that were broken in a controlled environment and then the edges of both the base and the tip for each knife were imaged at nine overlapping intervals. Therefore, each knife yields nine pairs of images. Each image is decomposed into the frequency domain and we will focus on the two frequencies of 5-10 and 10-20 Hz. The correlation of these frequency bands of images between any two images of all corresponding pairs is calculated. That is, we calculate the correlation between the 5-10 Hz frequency bands of the $k$th image ($k = 1, 2, \ldots, 9$) from the different tip-tip, base-tip and base-base pairs. Simularly, for the 10-20 Hz frequency band. Thus, of all the 729 pairs of correlations that we have for the (5-10, 10-20) Hz bands, only 81 of them are from matches. The rest are more-or-less non-matches. These 81 are really nine 9-vector measurement pairs.

   The file `Knife-matching.csv` contains the correlations between different pairs at the two frequency ranges. The columns are in the following order:

   - `first column (unnamed)`: 1-729, has no value in our exercise
   - `5-10`: The correlation at the 5-10 Hz frequency band between the image pairs
   - `10-20`: The correlation at the 10-20 Hz frequency band between the image pairs
   - `knife`: the knife pair (3 alphanumeric numbers for each specimen)
   - `img`: the image number of the calculated pair (1-9)
   - `match`: match or nonmatch
   - `set`: all 1s, may be ignored

   Answer the following questions. For each part, you may place all the figures in the sub-parts on the same plot, or on different plots.

   (a) In this part of the problem, we will analyze the raw data.

      i. For each frequency band, make histograms of the correlations of the match and non-match pairs in one figure. Comment on the separation of the matches and non-matches. [*5+5 points*]

      ii. For each pair, and using color to denote match/non-match, plot the correlation between the 5-10 and 10-20 band pairs. Comment on the separation of the matches from the non-matches in the bivariate plot vis-a-vis the univariate plots. [*10 points*]

      iii. The bivariate plot clearly shows some minor overlap that comes from one set of images. But there are actually 9 correlation pairs from that set (T10:T10). Connect all the 9 correlation pairs from this set by means of a line, and comment on its separation when all nine correlation pairs are considered together. [*10 points*]

   (b) The inverse hyperbolic tangent transformation is often used on correlations to stabilize its variance. Repeat the exercise in the previous part, but use the inverse hyperbolic tangent transformed correlations. Comment. [*20 points*]

2. Two different visual stimuli (S1 and S2) produced responses in both the left ($L$) and right ($R$) eyes of subjects with multiple sclerosis and others. The file `sclerosis.dat` having no header contains measurements on

the age, total response of both eyes to stimulus S1, absolute difference between response of the two eyes to stimulus S1, total response of both eyes to stimulus S2 and absolute difference between response of the two eyes to stimulus S2 respectively. The first 69 records contain responses for the normal subjects while the last 29 records contain responses for the subjects with multiple sclerosis.

(a) Calculate the means for each group. [*5 points*]

(b) Use the `plotcorr` function to display the correlation matrix for each group. Comment. [*10 points*]

3. Each pixel in an image is represented in terms of its primary components namely Red, Green and Blue. Therefore, each pixel has a certain amount of Red, a certain amount of Green and a certain amount of Blue. This way of representing color is known as *RGB* format.

(a) *Reading in a TIFF image* The file provided `jubabrinda.tif` in the usual places provides a digital file in the Tagged Image File Format (TIFF) of bamboo-handiwork decorations at a street-side exhibit during Kolkata's (formerly, Calcutta) famed fall festival. The R package `rtiff` can read in this file as follows:

```
library(rtiff)
juba <- readTiff(fn="jubabrinda.tif")
plot(juba)
```

The amount of the primary colors at each pixel can be obtained from the read TIFF file by using: `owlet@red`, `owlet@green`, `owlet@blue`. Note that these are all matrices of the same dimension as the image. (Make sure that you look at the help on `pixmap-class` to get an idea of what the components of pixmap are). Note also that these values are all between 0 and 1.

Read in the TIFF image and convert to a dataset of 375000 3-dimensional observations. [*10 points*]

(b) Use the `plotcorr` function to display the correlation matrix between the three colors. Comment. [*10 points*]

4. For any $p$-dimensional random vector $\boldsymbol{X}$, let $\boldsymbol{X}^\circ = (\boldsymbol{X} - \boldsymbol{X}'\mathbf{1}1/p)/\left\|(\boldsymbol{X} - \boldsymbol{X}'\mathbf{1}1/p)\right\|$, where $\mathbf{1} = (1, 1, \ldots, 1)'$ is the $p$-dimensional vector of all ones and $\|\cdot\|$ is the Euclidean norm. Let another $p$-dimensional vector $\boldsymbol{Y}$ and $\boldsymbol{Y}^\circ = (\boldsymbol{Y} - \boldsymbol{Y}'\mathbf{1}1/p)/\left\|(\boldsymbol{Y} - \boldsymbol{Y}'\mathbf{1}1/p)\right\|$ be similarly defined. Then show that the square of the Euclidean distance between $\boldsymbol{X}^\circ$ and $\boldsymbol{Y}^\circ$ is equivalent to $2 - 2\text{Corr}(\boldsymbol{X}, \boldsymbol{Y})$ where $\text{Corr}(\boldsymbol{X}, \boldsymbol{Y})$ is the correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$. [*15 points*]