Quantitative Methods in Neuroscience
(NEU 466M)
Homework 6
Due: Tuesday April 7 by 12:00 pm (to upload in Canvas)

General guidelines: Read through each complete problem carefully before attempting any parts. Feel free to collaborate in groups of size 2-3, but always note the names of your collaborators on your submitted homework. For graphs: clearly label your axes and use good color and symbol choices. Print out your matlab code (in the form of a script file). For derivations you're asked to do 'by hand' (in other words, analytically, using paper and pencil) feel free to turn in handwritten or typed-out work.

Applying change-of-basis to find structure in data. In this problem we're going to look for structure in data that might not be immediately apparent. We're going to re-plot a dataset along its "important" directions, where we use a specific definition of "important". We'll see that this procedure allows us to find interesting structure in the data. This problem is a motivating setup for PCA, which we'll be studying next.

a. Download the Matlab file `clusters_generate_fake_3d.m` and run it. This file generates a matrix `D` of data. The (three) columns represent the three variables, and the rows consist of many ($N = 800$) different measurement trials or samples of these variables. Hidden in the big scatter of data are four clusters. The file produces a 3-dimensional scatterplot of the data. Do you see separate, clear clusters in the scatterplot? Now, for each variable separately, plot the histogram of values the variable takes (use `hist`). Do you see separate, clear clusters in the data histograms? Next, use the rotate tool in the toolbar above the plot to rotate the data-cloud and hunt visually for structure. Feel free to explore: can you find some rotation at which you see the clusters? (Looking at the data from any given angle is really projecting the data cloud from three dimensions onto the two dimensions of your screen, along the particular angular direction. As you have already verified from the 3 separate histograms above, some standard projections e.g. onto the x,z plane do not clearly show any clusters.)

b. Now we'll answer what are good axes along which to examine the data. From [a.] above, you will have noticed that at certain angles you can view the clustered structure of the data. Let us work with the hypothesis that the directions of maximum covariation between the three variables are important. To interpret: use `cov` to construct the covariance matrix `covD` of the data `D`. Next, use the command `[v,d] =`

`eig(covD)`. The first (second) column in the resulting matrix `v` is the vector corresponding to the direction of (second) maximum covariance in the data. Use `plotv` to plot the three vectors (maybe multiplied in amplitude by a factor of 10) onto the original scatterplot of the data. Can you see which of the vector(s) cut across the clusters in the data?

c. Let us now try to visualize the projection of the data onto the first column vector in `v`: Each row of the data is a sample, given in the original basis of the 3 measurement variables. We now want to project each row onto the first column vector in `v`: recall that projection is an inner-product. The resulting N-dimensional vector are the coefficients of the data samples along the first column vector. Histogram these coefficients. Contrast with your results from a. above. Take note that `eig` allowed for automatic discovery of cluster structure that you had to locate by eye in a. In > 3 dimensions, it is impossible to search for structure by eye, since we can at most make 3-dimensional plots. Thus, automated techniques are critical.