ECE 3502 / CS 3501: Foundations of Data Analysis **Due Th 4/21**

Homework 4: PCA

Instructions: Submit a single Jupyter notebook (.ipynb) of your work to Collab by 11:59pm on the due date. All code should be written in Python. **Be sure to show all the work involved in deriving your answers! If you just give a final answer without explanation, you may not receive credit for that question.**

You may discuss the concepts with your classmates, but write up the answers entirely on your own. Do not look at another student's answers, do not use answers from the internet or other sources, and do not show your answers to anyone. **Cite any sources you used outside of the class material (webpages, etc.), and list any fellow students with whom you discussed the homework concepts.**

1. In this exercise, you are going to use Principal Component Analysis to find the most relevant dimensions of variance in a set of hand shapes. Load the matrix `all-hands.dat`. Each row of this matrix is an entire set of hand points as a list of x, y coordinates: $(x_1, y_1, x_2, y_2, \ldots, x_{72}, y_{72})$.

   Do the following:

   (a) Compute the mean hand (this should be a $72 \times 2 = 144$ vector, consisting of the means of each column in your matrix). Plot this mean as a hand shape.

   (b) Compute the covariance matrix Σ for this data. Use the formulas we covered in class, not a covariance function! What is the total variance of the data?

   (c) What is the covariance between the x_1 coordinate and the x_2 coordinate? What is the correlation between these two coordinates? These are adjacent points on the hand, can you explain why the correlation comes out to this value?

   (d) Compute the PCA of the hands. (Use the Python function `numpy.linalg.eigh` to get eigenvalues and eigenvectors of Σ).

   (e) Plot a scree plot of the eigenvalues. How many eigenvalues are nonzero? What does this tell you about the dimensionality of your data?

   (f) Plot a sequence (as a strip of 5 side-by-side figures) of hand shapes along the first principal component at $s = -3\sqrt{\lambda_1}, -1.5\sqrt{\lambda_1}, 0, 1.5\sqrt{\lambda_1}, 3\sqrt{\lambda_1}$, where λ_1 is the first (largest) eigenvalue. So, you will plot hands corresponding to:

   $$\mu + s v_1,$$

   where v_1 is the first eigenvector. What does this dimension in the data correspond to, in terms of hand shape changes? Repeat this process for the second and third principal component.

   (g) How many dimensions do you need to represent 95% of the variance in the hand data?

   (h) Using your PCA results with the reduced number of dimensions you found in the previous answer, project the first hand (row 1 of the matrix) onto this reduced dimensional subspace. What is the vector of weights needed to represent this hand? Plot the reconstructed hand shape on top of the original hand shape (again, use two different colors). Is the reconstructed hand similar to the original?