**ENGR 2013: Engineering Analysis and Applications**

**Laboratory No. 13: What Do Eigenvalues Tell Us About Data?**

**Goal:** The goal of this lab is to introduce you to the Gram-Schmidt Orthogonalization algorithm and discuss some of the implications of this algorithm.

**Preliminary Work:**

The data you will use for this lab is located here:

https://isip.piconepress.com/courses/temple/engr\_2011/resources/data/lab\_13/

There are three sets of data (set\_00.csv, set\_01.csv, set\_02.csv). All three sets contain data tagged with one of two labels: cat (blue) and dog (black). We will refer to these as ‘classes’.

There are several ways to load this data into your program. I would suggest you spend a moment to learn how to use data frames and csv files in Python using the Pandas library:

https://www.geeksforgeeks.org/python-pandas-dataframe/

Load the data into your Python program and produce the scatter plots shown in the corresponding png files (set\_00.png, set\_01.png, set\_02.png). The data consists of 2D vectors.

You will also need to know how to compute the covariance matrix for a set of 2D vectors:

https://www.geeksforgeeks.org/python-numpy-cov-function/

and how to compute the percentage of the variance accounted for by each eigenvalue:

$$\% variance\_{i}={λ\_{i}}/{\sum\_{j=0}^{p}λ\_{j}}$$

where $p$ is the number of eigenvalues (in this case, $p=2$).

**Tasks:**

For each of the three sets, do the following analysis and produce one plot for each set:

1. Compute the mean for each class (e.g., compute the mean for all the data tagged with ‘cat’ in set\_00.csv, then compute the mean for data tagged with ‘dog’).
2. Compute a covariance matrix for each class.
3. Find the eigenvalues and eigenvectors.
4. Plot the eigenvectors in the same 2D space as the data. Plot these as vectors extending from the mean of the class. Plot a vector of length 1 in a high contrast color like yellow extending from the mean for each class. The angle of this eigenvector should reflect the actual angle of the eigenvector. The reasons for this will become clear shortly.
5. Compute the percentage of the variance accounted for by each eigenvalue.

Justify/explain your results. What relationships in the data do the eigenvectors reveal?

**Summary:**

Principal Component Analysis (PCA) was one of the first extremely powerful techniques introduced to give us insight into the relationships in data. You can review the history of PCA here:

https://en.wikipedia.org/wiki/Principal\_component\_analysis

This technique was first introduced by Karl Pearson in 1909. It became popular in the 1960’s when computing technology made it feasible to apply these techniques to real world problems. It remains one of the most popular techniques for gaining insight into data.