# Homework 3

October 16, 2018

## Due: October 23, 2018, 11:59 PM EST

## Instructions

Your homework submission must cite any references used (including articles, books, code, websites, and personal communications). All solutions must be written in your own words, and you must program the algorithms yourself. If you do work with others, you must list the people you worked with. *Please type your solution.*

Your programs must be written in Python 3+. All code must be able to compile and run for full credit. Comment all code following proper coding conventions. Remember, if we can't read it, we can't grade it! (For more information on python coding standards, refer to: *https://www.python.org/dev/peps/pep-0008/*)

You should submit your assignment via Github. Submit your solutions as a PDF named "hw(hw #).pdf". For example, homework 3 should be submitted as hw03.pdf. If the assignment requires coding, submit your working code as a .py file with the same name, i.e., hw03.py.

If you have any questions address them to:

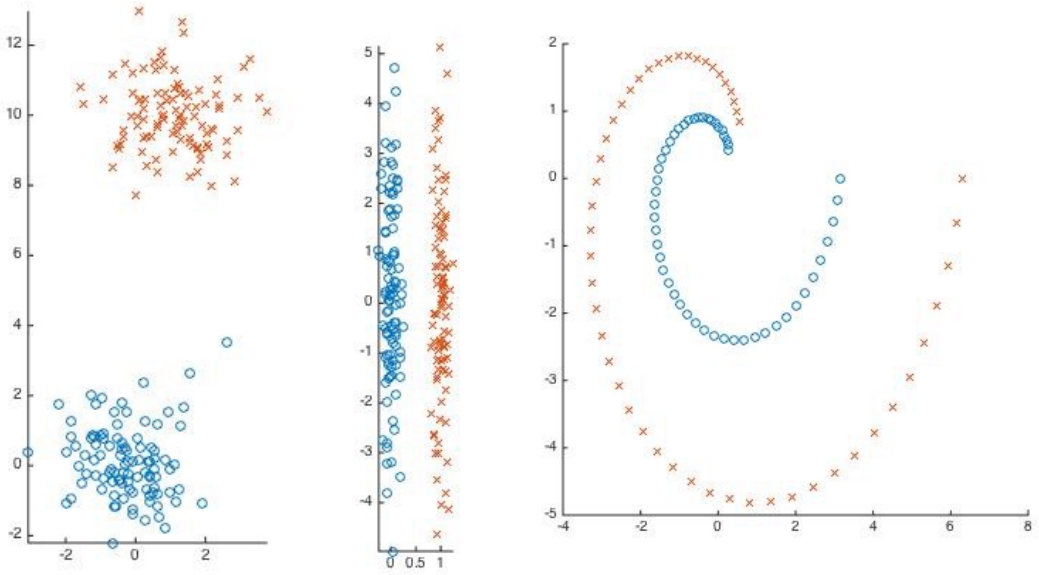
- Daniel Wells (TA) – dwells@ufl.edu
- Xiaolei Guo (TA) – suninth@ufl.edu
- Connor McCurley (TA) – cmccurley@ufl.edu

## Question 1 - 5 points

Consider the following three two-dimensional data sets each containing two clusters of data points (shown with 'circles' and 'crosses'). Suppose you would like to apply either Principal Components Analysis (PCA) or Linear Discriminant Analysis (LDA) to reduce the dimensionality of each of these data sets from 2-D to 1-D where the two clusters remain separated in the reduced dimensional data set. For each data set, address each of the following two questions:

1. Will PCA be effective and keeping the two clusters separated in the reduced dimensionality data? Why or why not? If yes, state what characteristics of the data set allow PCA to be effective. If no, state what characteristics of the data set cause PCA to fail.

2. Will LDA be effective and keeping the two clusters separated in the reduced dimensionality data? Why or why not? If yes, state what characteristics of the data set allow LDA to be effective. If no, state what characteristics of the data set cause LDA to fail.



## Question 2 - 5 points

This question focuses on Principal Components Analysis and Data Whitening.

1. Suppose $\mathbf{z}$ is the *whitened* version of $\mathbf{x}$ where $\Sigma$ is the covariance for the original data set $\mathbf{X}$ and $\mu$ is the mean ($\mathbf{x}$ is one data point in $\mathbf{X}$). Write the formula for computing $\mathbf{z}$ from $\mathbf{x}$.

2. Suppose the original data set, $\mathbf{X}$, has features with widly varying range and variance. Could this data be whitened? Why would you or would you not want to perform whitening on data of this type? Generate two dimensional data with these characteristics (i.e., the two features have very different range and variances from each other). Implement python code that applies principal components analysis (without dimensionality reduction, only decorrelation) and data whitening to your generated data set. Scatter plot the original data, the data after application of the PCA transformation and the data after data whitening. What is the covariance of the original data, after PCA and after data whitening? Use these results to motivate your discussion of why you might or might not want to apply data whitening to data with features of widely varying range and variance.