Homework 3

Discussion of Part 1, Question 1:

PCA is not effective for all of these three data sets because PCA doesn't label data points. It considers the whole bunch of data as one data set, thus PCA is not able to separate different clusters of points no matter how the data sets are distributed.

Discussion of Part 2, Question 1:

LDA is effective for data set 1 and data set 2, but not for data set 3 because LDA is effective only for linearly separable data sets. If the two clusters of data points can be separated by a straight line as decision boundary in 2-D coordinate system, then LDA could be used to separate two clusters in the reduced dimensionality data. It is apparent that a straight decision boundary could be found in data set 1 and data set 2, but not in data set 3, thus LDA is only effective for data set 1 and data set 2.

Discussion of Part 1, Question 2:

Compute the correlation matrix of data set X:

$$R_x = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T$$

Compute the max eigenvalues λ and corresponding eigenvector v of R_x

Set A to be equal to v and compute the PCA transformed data:

$$Y = vX$$

Compute the covariance matrix of data set Y and we have $\sum_Y$:

We define λ_{ij} as each element of $\sum_Y$, where i and j are the numbers of row and column, respectively.

We multiply $1/\sqrt{\lambda_{ii}}$ with y_i, where y_i is the ith feature of data point y, thus we have:

$$z_i = \frac{y_i}{\sqrt{\lambda_{ii}}} = \frac{vx_i}{\sqrt{\lambda_{ii}}}$$

Where x_i is the ith feature of data point x.

Discussion of Part 2, Question 2:

The data set could be whitened. The covariance matrices of original data set, after PCA and after whitening could be found after running my code. After PCA and after whitening, data sets are decorrelated in each feature. Typically, the covariance matrix after PCA is a diagonal matrix, where the elements are variances of different features, respectively. The covariance matrix after whitening is an identity matrix, so it means that variance of each feature is equal to 1. Thus, sometimes whitening data set is not needed because the original data set has different variances in each feature. If it is whitened, then the variances will be the same and some information contained in each feature might be lost. However, whitening data is needed if the object is the noise. Some noise filters may only be effective when the noise has the same variances in different dimensions. Therefore, whitening data set might be necessary in some situations but not in other situations.