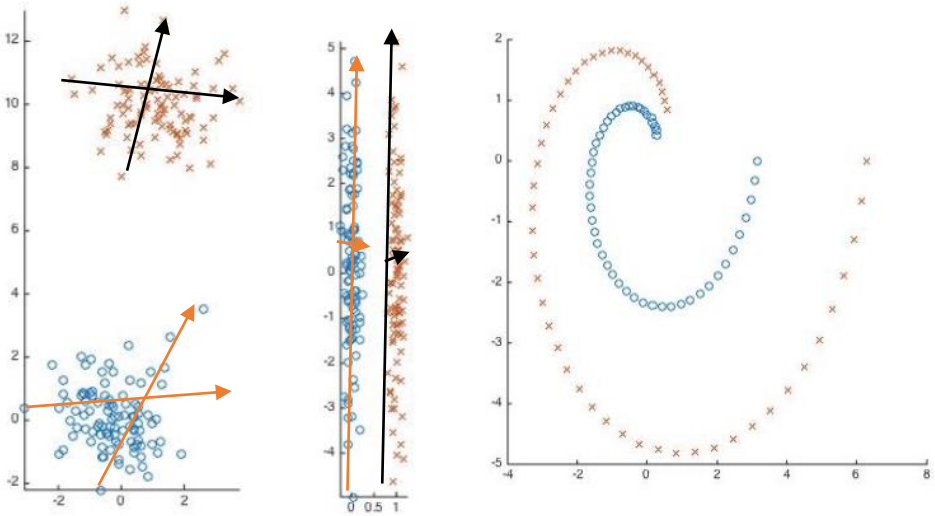
**Note: I have discussed the problem with Diandra Prioleau generally, but everything is %100 my own work without getting help from anyone. **

# Question 1

Our main purpose in this question is to differentiate Linear Discriminative Algorithm (LDA) and Principal Component Analysis (PCA) by considering which one is applicable to each plot. The applicability of PCA for each plot would be considered and later observe each plot to decide if LDA is applicable or not.



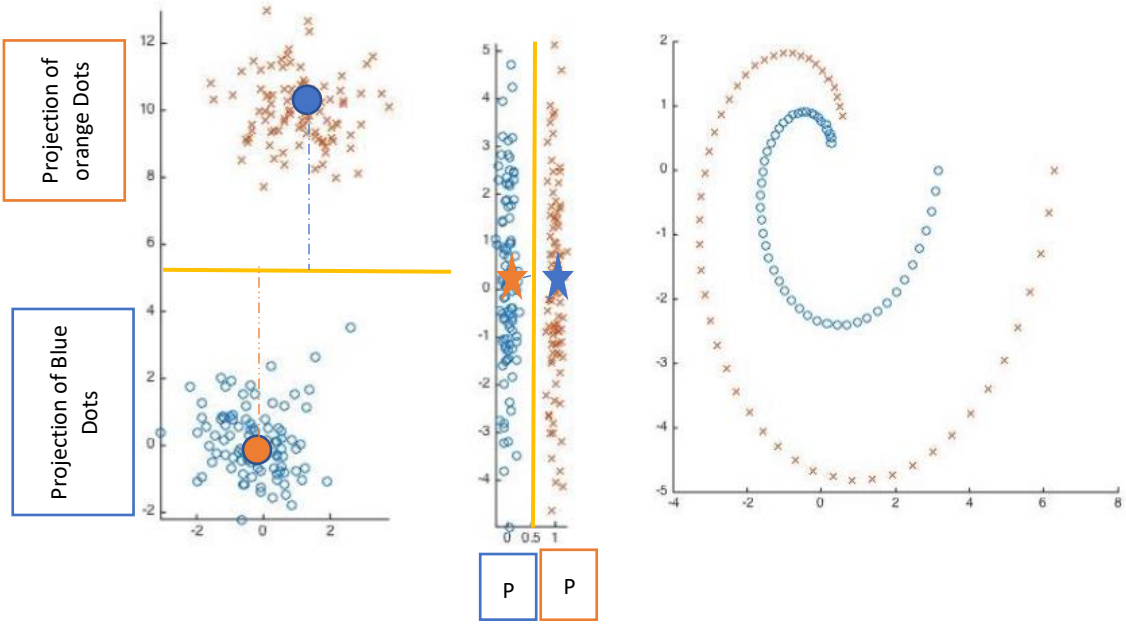
A B C

As we discussed in the class and lecture notes, PCA is a procedure to decorrelate the data points into the principal components linearly. To do this, we are trying to find the orthogonal eigenvectors of the covariance matrix where it will show the variance distribution through the data. If we are reducing the dimensionality from 2D into 1D, then we only consider the largest eigenvector of the data which shows the highest variance.

For Figure A, the eigenvectors' direction has been shown for each data separately. For both datasets, if the eigenvector which is parallel to the x-axis considered as the largest one. These vectors would be selected for projecting the data points for each dataset. After projecting the data to the dimension with highest variance the points would be projected on y-axis and we would be able to discriminate between them. However, if the other vector is chosen for the projection, the points will be overlapped with each other which makes the distinction between them impossible. Therefore, it depends on the selection of the eigenvector for projecting the class points.

In Figure B, due to the same reason as Figure A we would see the principle direction with small variance is the direction that discriminate between these two sets and if we use dimension reduction, then there is no way to distinct between projected points. Therefore, we are not

able to clearly distinguish the two datasets after applying the PCA and reducing the dimension into 1D.
One of the first assumptions about the ordinary PCA is the linearity and falling into the normal gaussian distribution. In case of Figure C, none of these assumptions are valid. Therefore, we cannot reduce the dimensionality using the ordinary PCA and we may need another method which reduce the dimensionality by considering the non-linear nature of the data.




A B C

Before discussing the LDA algorithm for each case separately, the intuition behind this algorithm will be discussed. First step in the LDA algorithm includes spotting the mean of each class which has been shown in the above figure. Once the mean is calculated, the line which maximizes the distances between both means is recognized. At the same time, the line should minimize the variance of projection of data points onto the orthogonal subspace to the hyper plane which in this case is a line. These constraints will be beneficial in our ability to discriminate two classes.
As it is shown in Figure A, the projection of data points belonging to each class is not overlapping. Therefore, we would be able to distinguish the data points and LDA would be a viable method to user in this case.
The same arguments about Figure A apply to Figure B, if the projection to the orthogonal line has no overlap between two classes. Due to the small range of values on orthogonal hyperplane(y=0), it is possible to overlap in some degree between the two classes. However, I think the probability is very low, so I confirm that LDA is applicable to second figure.

In case of the last Figure, the non-linear nature of the two classes would hinder us from separating these two sets with a line. Therefore, we are not able to use LDA algorithm for discrimination between the points.

# Question 2

## Part 1

Suppose we have X= {$x_1$,……, $x_n$} as the actual point. We want to derive the formula for whitened version of each $x_i$ after applying PCA and whitening methods. Covariance of X is equal to product of X and $X^T$. After standardizing the data, the eigenvectors and values of covariance would be computed. If the matrix of all Eigenvectors is shown with U (we are not reducing dimensionality, so we are including all the eigenvectors), the decorrelated version of X using the PCA would be:

$$X_{Standard} = x_i - \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$X_{PCA} = X_{Standard}.U^T$$

Following the PCA, the covariance becomes almost diagonal which confirms the fact that features are decorrelated. However, they are not normalized yet and that's the whole purpose of performing data whitening. For Data Whitening or Sphering, we compute the linear transformation which is shown by W. According to textbook [4] (pp. 586), W is computed with the following formula:

$$W = L^{-\frac{1}{2}}\,U^T$$

By considering the above equation we could compute "z" the whitened version of each data point such as $x_i$ as follow:

$$z_i = W.x(i)_{standard}$$

## Part 2

The data with highly varying variance for features has been generated using multivariate normal by considering covariance in Table 1.

| Name | Covariance |
|---|---|
| **Original** | [0.20822577 -1.55612962<br>-1.55612962 20.56663065] |
| **PCA** | [2.06848891e+01 -6.71028586e-16≈ 0<br>-6.71028586e-16≈ 0 8.99672796e-02] |
| **Whitening** | [1 -5.37534835e-16 ≈ 0<br>-5.37534835e-16≈ 0 1] |

Table 1: Covariance Matrix of Original Data, PCA transformation and data whitening

As it is shown in Figure 1, there is a correlation between the two features in the original data. When two features are highly correlated, they could have same effects on the machine learning model [1].

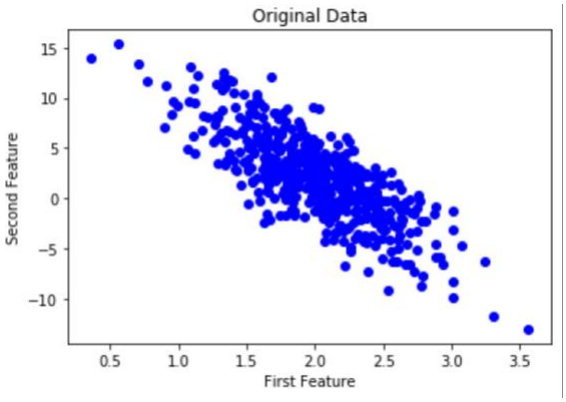



Figure 1: Original Data Scatterplot

Therefore, it would be more beneficial to decorrelate the features or apply feature selection. PCA algorithm would help us in decorrelating the data based on the principal directions of the data. To find the principal direction of the data, we computed the eigen vectors of the data after subtracting the mean. The scatterplot of our PCA method and PCA method from Scikit-Learn library [2] have been shown in Figure 2. According to these plots, our method was able to decorrelate and center the data around the origin without reducing the dimensionality. The almost diagonal covariance after PCA, is shown in Table 1, validates that our features are decorrelated. It is worthy to mention that the off-diagonal elements are not exactly zero due to the round off error.

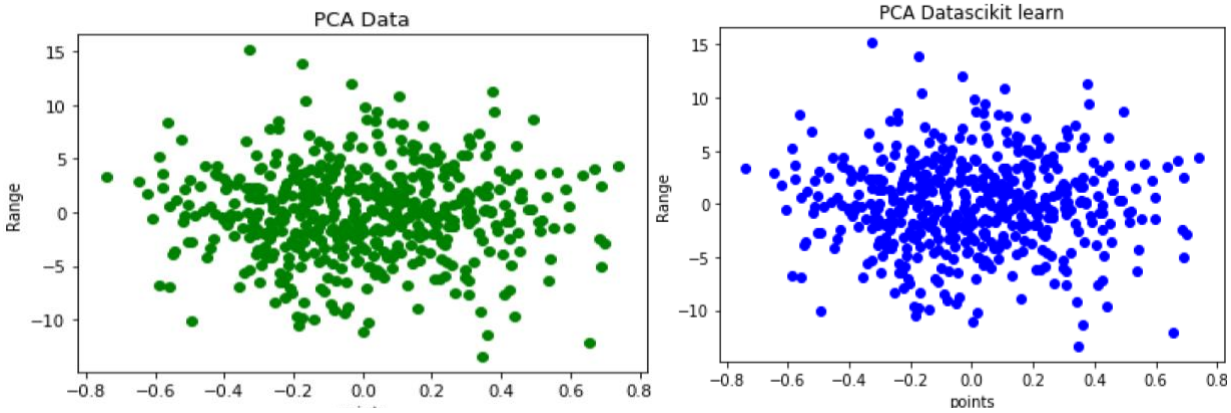



Figure 2: Left is our implementation of PCA and right plot is Scikit Learn method [2]

While dealing with different features over the same range, we could end up having redundancy in the data which lower down the performance of our models. In this project we apply data whitening or sphering onto the data to make the features normalized. The whitening data scatter plot, is shown in Figure 3, validates that our data are normalized. The covariance is almost identity matric as it is shown in Table 1.

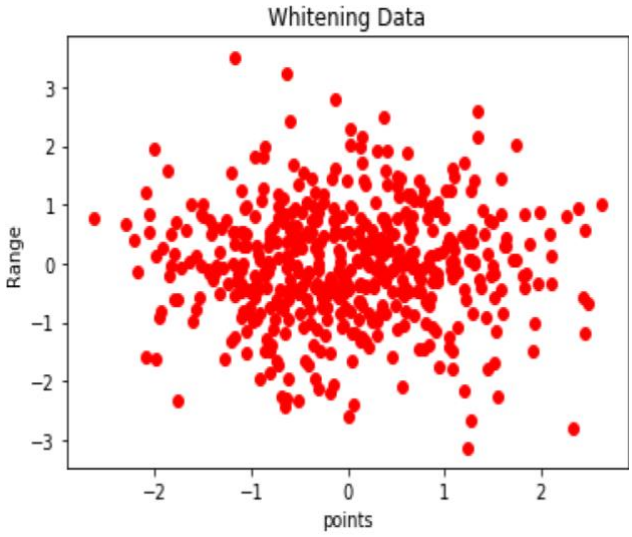



Figure 3: Data Whitening scatter plot

By considering all the results we gathered through our implementation, applying PCA and data whitening for normalizing the data would help us to lower than the correlation and redundancy in our feature space which could affect our predictive models' performance significantly.

References:

1. https://towardsdatascience.com/feature-selection-correlation-and-p-value-da8921bfb3cf
2. http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
3. http://courses.media.mit.edu/2010fall/mas622j/whiten.pdf
4. Bishop, C.M., 2006. Pattern recognition and machine learning (information science and statistics) springer-verlag new york. *Inc. Secaucus, NJ, USA*.
5.