Dry Bean Data Visualization

Math 250 Final Project

Spring, 2021

Team Members

Kaylyn Vo

Kevin Pham

1. Background

“Multiclass classification of dry beans using computer vision and machine learning techniques” by Murat Koklu and Ilker Ali Ozkan wanted to contrast various methodologies to achieve principles of sustainable agricultural systems. A basis for such agricultural systems is through the understanding of seed classification. Through seed classification, more specifically seed quality is pivotal for the influence of crop production. Therefore, for optimal crop production, a uniform seed variety is desired. The methodologies through computer vision and machine learning techniques aim to obtain uniform seed varieties based on seed classification. The dataset they compiled for experimenting consists of numerous images of 7 various bean types with a set of attributes for machine learning techniques to contrast between them. The best technique between them will yield the highest accuracy for meeting the correct seed classification.

The 7 various beans used for this experiment were Barbunya, Bombay, Cali, Dermason, Horoz, Seker, and Sira. Images of these dry beans were obtained by processing bags of the beans from certified seed producers and fed through a computer vision system they designed and developed to capture the images. After the images go through segmentation and processing to prevent altering classification results, they undergo feature extraction to retrieve each bean attribute. In total, it yielded 12 dimensional and 4 shape effective features from the dry beans using pixel count as the unit measurement for each respective image. Those attributes were then fed to the following techniques to contrast accuracy, Multilayer perceptron (MLP), Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Decision Tree (DT) classification models with 10-fold cross validation and performance metrics. Overall, the SVM classification model yielded the highest accuracy for uniformity.

The following 12 effective dimensional features for the dataset are:

1. Area (A) - The area of the bean zone and the number of pixels within its boundaries.
2. Perimeter (P) - The bean circumference defined as the length of its border.
3. Major axis length (L) - The distance between the ends of the longest line that can be drawn from the bean.
4. Minor axis length (l) - The longest line that can be drawn from the bean while standing perpendicular to the main axis.
5. Aspect ratio (K) - Defines the relationship between ‘L’ and ‘l’.

$$\mathrm{K} = \frac{\text{Major Axis Length}}{\text{Minor Axis Length}} = \frac{\mathrm{L}}{\mathrm{l}}$$

6. Eccentricity (Ec) - Eccentricity of the ellipse having the same moments as the region.
7. Convex Area (C) - The number of pixels in the smallest convex polygon that can contain the area of a bean seed.
8. Equivalent diameter (Ed) - The diameter of a circle having the same area as a bean seed area.

$$(\text{diameter})\ \mathrm{d}\ =\ \sqrt{\frac{4*\mathrm{A}}{\pi}}$$

9. Extent (Ex) - The ratio of the pixels in the bounding box to the bean area.

$$\mathrm{Ex}\ =\ \frac{\mathrm{A}}{\mathrm{A_B}}, \mathrm{A_B}\ =\ \text{Area of bounding rectangle}$$

10. Solidity (S) - i.e. convexity. The ratio of the pixels in the convex shell to those found in beans.

$$\mathrm{S}\ =\ \frac{\mathrm{A}}{\mathrm{C}}$$

11. Roundness (R) - Calculated with the following:

$$\mathrm{R}\ =\ \frac{4\pi\mathrm{A}}{\mathrm{P}^2}$$

12. Compactness (CO) - Measures the roundness of an object:

$$\mathrm{CO}\ =\ \frac{\mathrm{Ed}}{\mathrm{L}}$$

The following effective shape features for the dataset are:

1. ShapeFactor1 (SF1) = $\frac{\mathrm{L}}{\mathrm{A}}$
2. Shape Factor2 (SF2) = $\frac{\mathrm{l}}{\mathrm{A}}$
3. ShapeFactor3 (SF3) = $\frac{\mathrm{A}}{\frac{\mathrm{L}}{2}*\frac{\mathrm{L}}{2}*\pi}$
4. ShapeFactor4 (SF4) = $\frac{\mathrm{A}}{\frac{\mathrm{L}}{2}*\frac{\mathrm{l}}{2}*\pi}$

2. Introduction

With the attributes of the dataset above, it can be explored further such as the visualization of a single or small subset of the dimensional features or shape features and any preprocessing. Dimensional reduction techniques such as PCA, LDA, MDS, and Laplacian Eigenmaps will be utilized to visualize the data.

3. Low-Dimensional Data Exploration

The data set is highly dimensional, and thus only a subset of the explanatory variables can be visualized simultaneously. The descriptions of the variables suggest that they can be approximately separated into groups that describe similar features about the types of beans. The first group of variables contains information about the dimensions of a bean with features such as the area, perimeter, and lengths. The second group of measurements describe the shape of a bean with features such as eccentricity, roundness, and compactness. The third group contains variables which are called ShapeFactors that were calculated from the area and major and minor axis lengths. The separation of variables into groups is not absolute and was done with the purpose of exploring related variables together in low dimensional spaces.

We begin exploring the data with a scatter plot of the variables that convey information about the dimensions of a bean. A plot between the lengths of minor and major axes along with the area as the size of the scatter points reveals two distinct elliptical-shape clusters and a positive correlation trend between the lengths as seen in Figure 1. With labels adding to the plot presents further information that there are seven groups which coincide with the seven types of beans in the data set. Although the six types overlapped; however, the observations within each group stay closely in its own group. Additionally, the size of each cluster and the amounts of variations of the lengths depend on the type of bean which prompted us to examine the total observations for each type with a bar plot. We discovered that the counts of observations between the bean types are unevenly distributed with the largest count is almost six times larger than the smallest count as seen in Figure 3. The total count for Bombay is the smallest, yet the bean has the largest amount of variation in length while Dermason has the highest total number of observations, yet it doesn't have much variation in lengths. This informs us that the size of Dermason is about equally distributed while it might not be unexpected to see different sizes of type Bombay. The differences in dimensions between each type can be summarized in the boxplots which can be used to informally compared the averages of dimensions across the seven bean types as seen in Figure 2. Unarguably, Bombay is the largest type out of the seven types while Dermasion and Seker beans are some of the smallest beans.

To distinguish the types of beans, beside the dimensions, we also want information about the shape of a bean. Plot of the variables - compactness, roundness, and eccentricity - reveals information about the shape of a bean. Figure 5a shows that the higher the compactness and roundness measurements the rounder the beans while beans with large values of eccentricity have elongated shape. From the plot, we see that the Seker has the roundest shape while Horoz is elongated and the other types are in between as seen in Figure 5b.

Finally, there are four ShapeFactors variables that were calculated from the area and the major and minor axes measurements. These four measurements were derived from the other variables and thus, they don't have a direct interpretation, but each one of them tell a different story and is helpful in classifying beans. Figure 6 shows that large dimensions beans correspond to small ShapeFactors values.

Lastly, we will look at the correlation plot to study the linear relationship of all the variables. The correlation plot in Figure 4 shows that there are two blocks of variables with equal number of features in each group. The variables within a group are positively correlated to one another while negatively correlated to the variables that don't belong to its group.

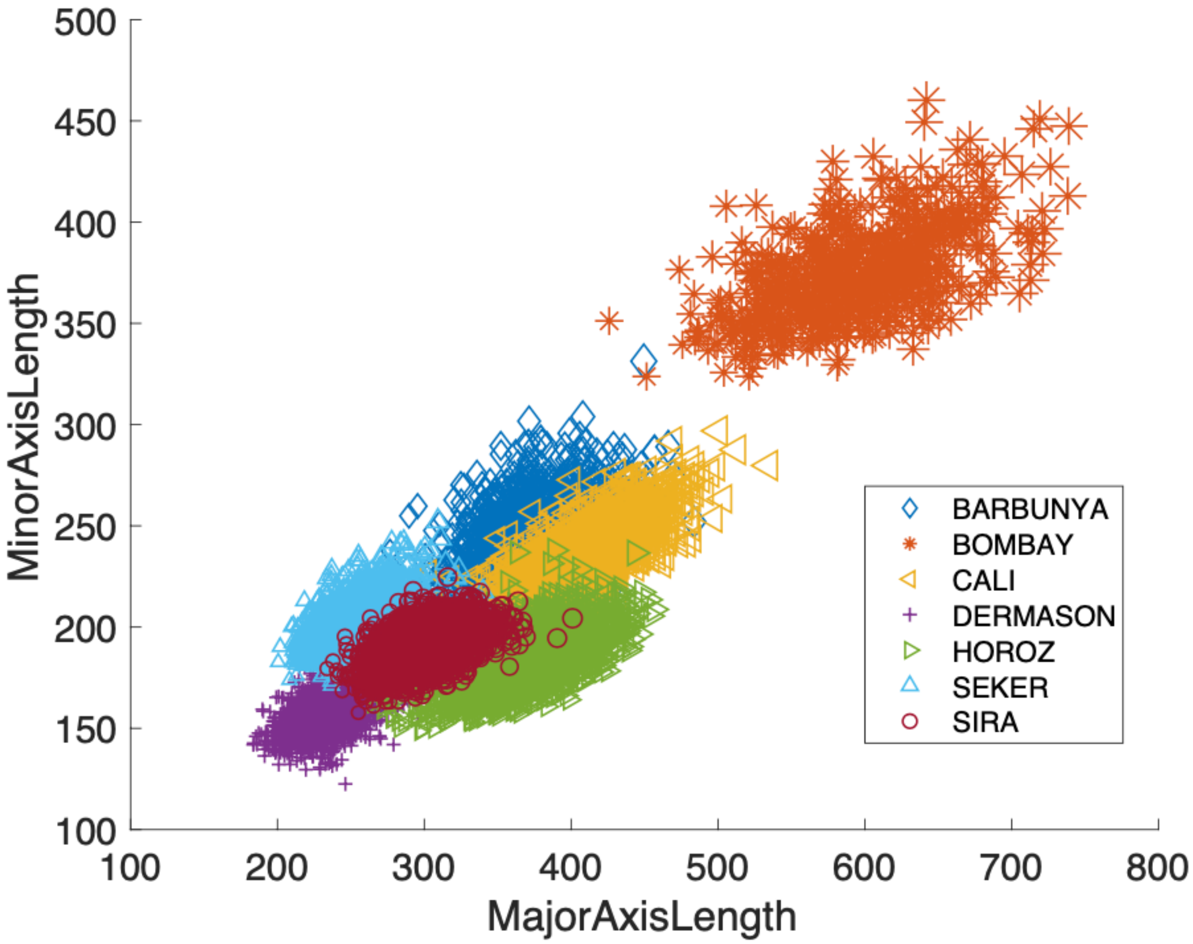


Figure 1: Scatter plot of major and minor axes using area as the size of each scatter

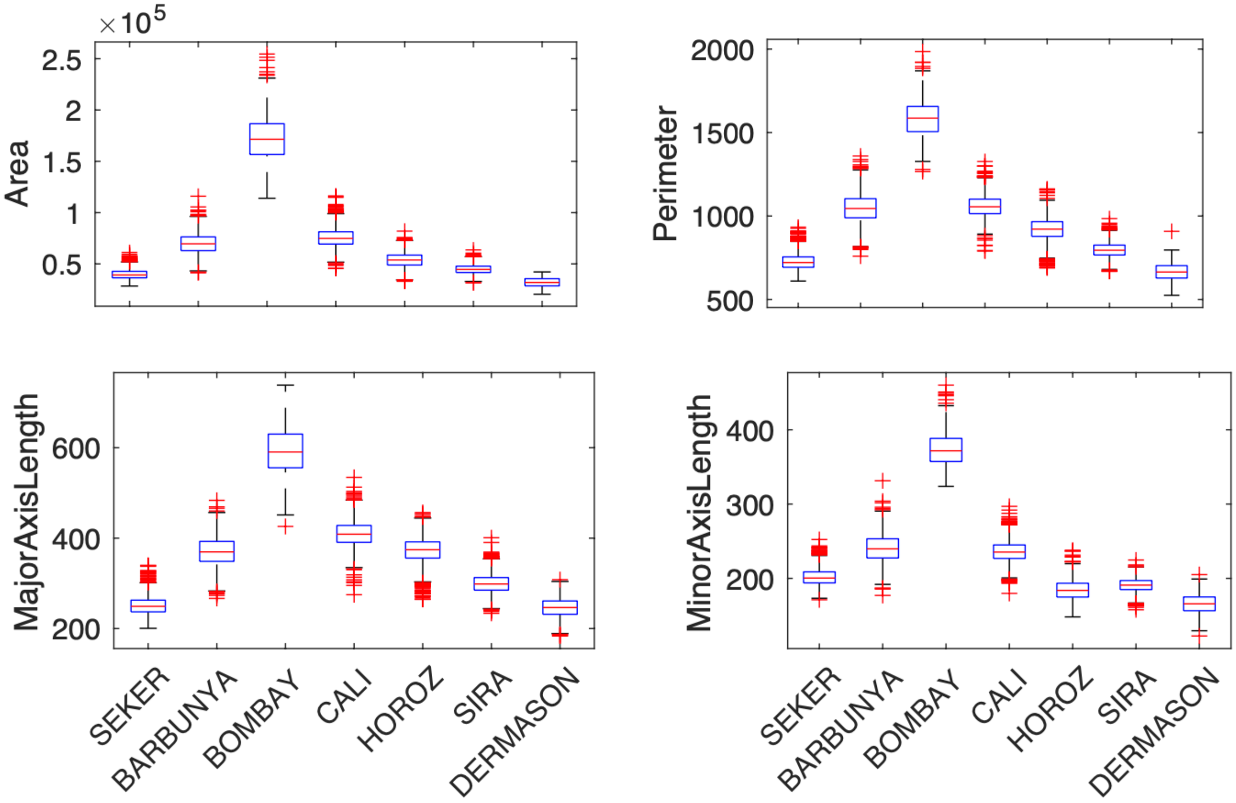


Figure 2: Boxplots of the four dimensions variables

Figure 3: Counts of Bean Types

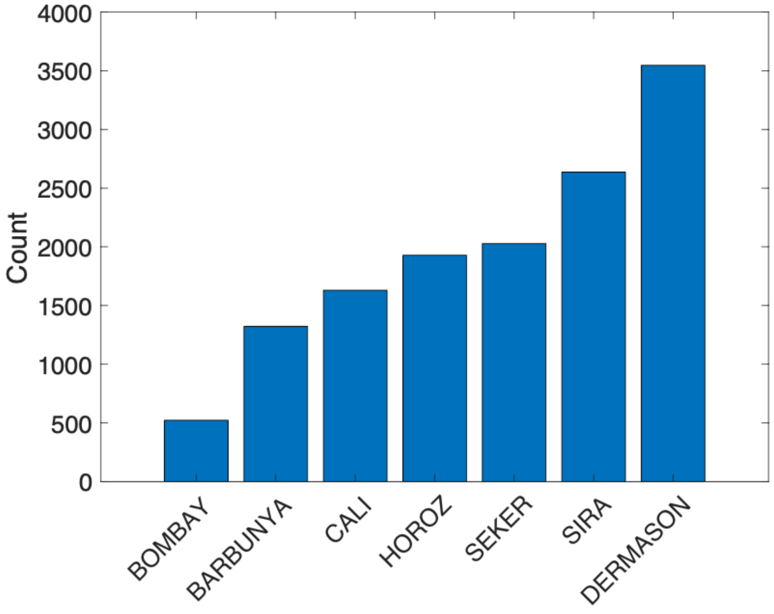
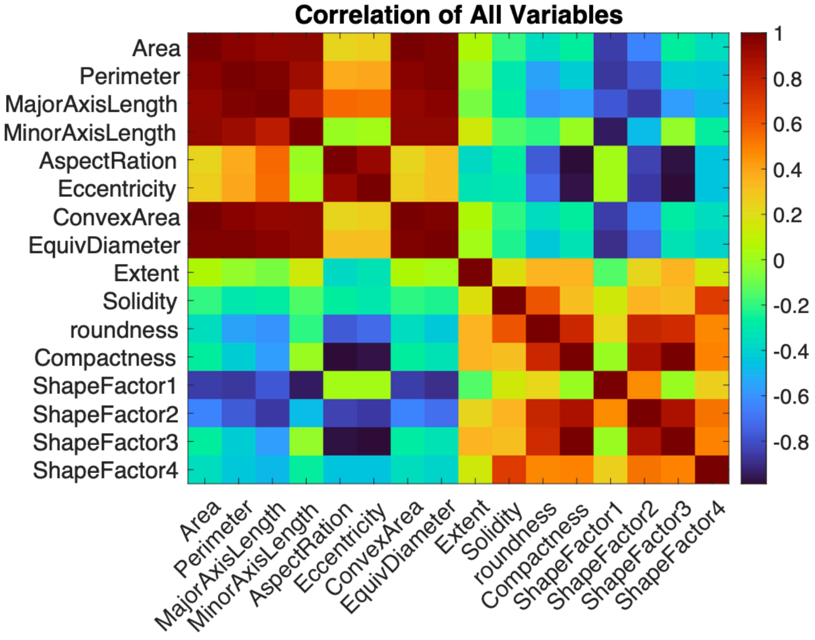


Figure 4: Correlation of all variable



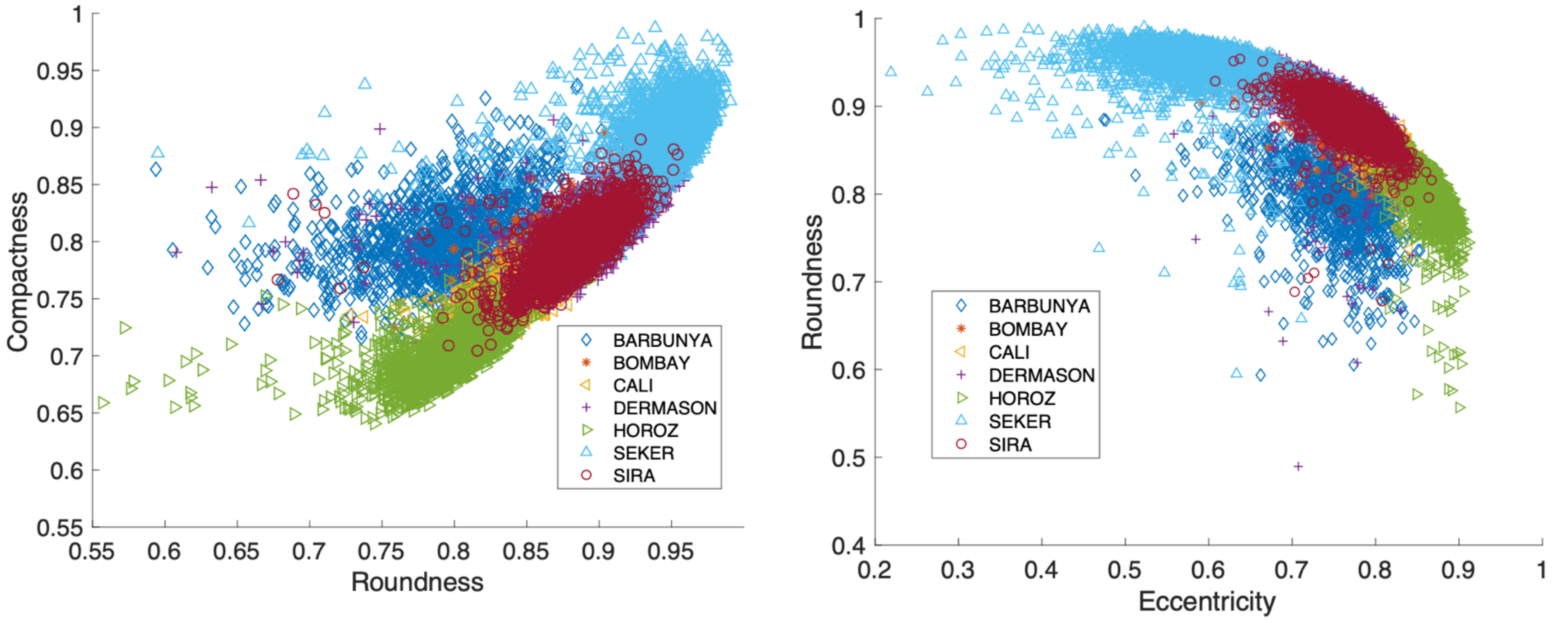


Figure 5a and 5b: Scatter plots of various shape variables

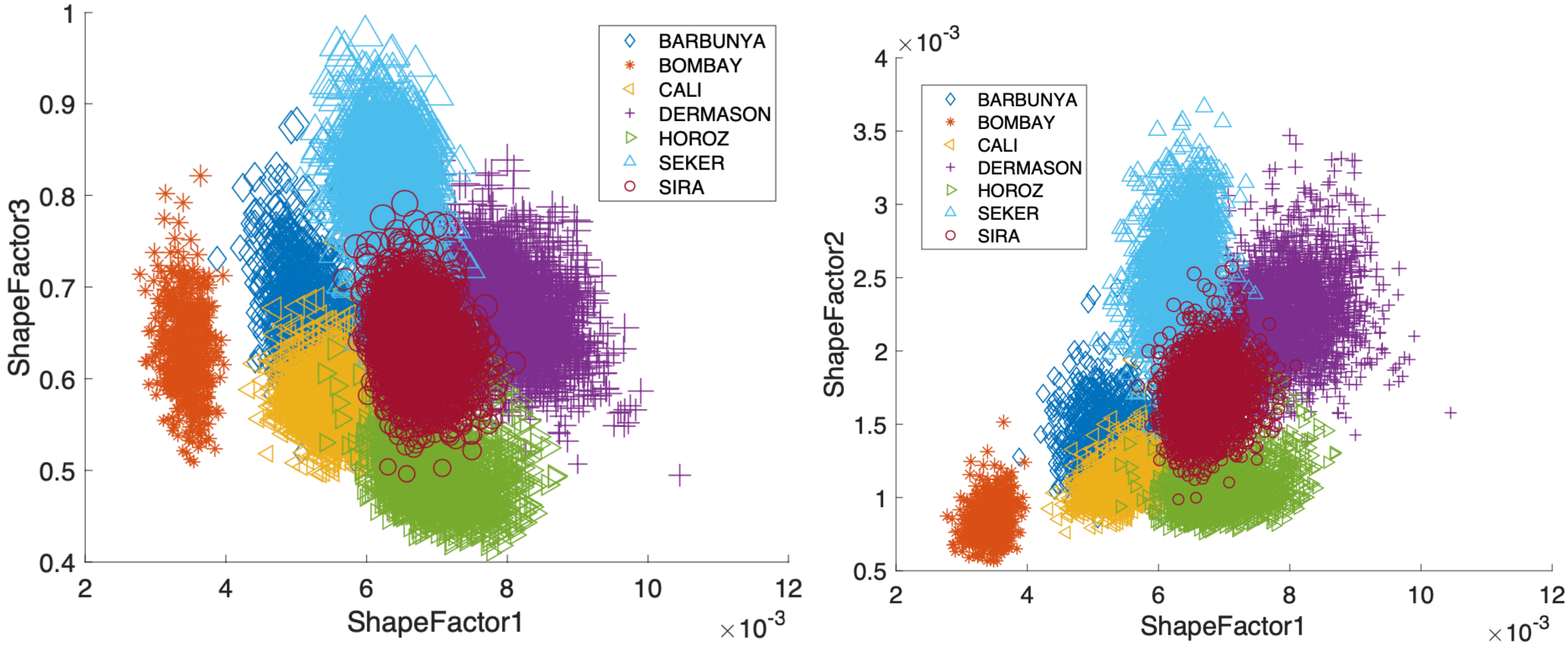


Figure 6: Scatterplot of ShapeFactors variables

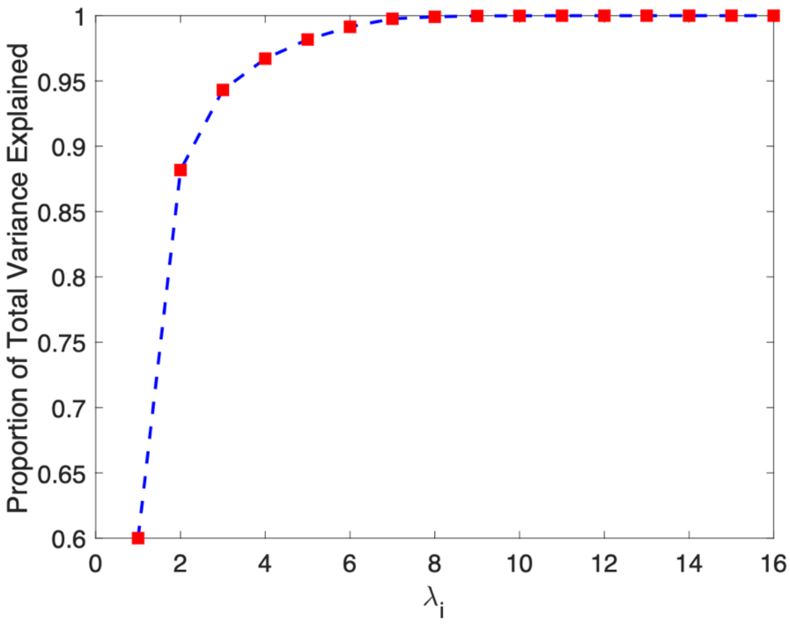
4. Principal Component Analysis

This first method that we used to explore the data is Principal Component Analysis (PCA). It is a linear dimensionality reduction technique with the main objective is to maximize the explained variance while reducing the number of dimensions and minimizing as much loss of the real structure of the data. The magnitudes of the measurements of the features are different, and we want to avoid having one feature dominating the principal components; thus, we scaled the data to be between the range of [-1,1] before we proceed with PCA. We used the top three principal components to analyze and visualize the data because approximately 94% of the total variation in the data accounts for by the first three principal components, with the top two components explain 88% of the total variation as seen in Figure 7.

The first principal component is a contrast between the dimension variables and ShapeFactors variables with eight negative coefficients in the first principal direction belong to the ShapeFactors features and eight positive coefficients belong to the dimension features. This division of features is what we have seen in the correlation matrix. Figure 9 emphasizes that the types of beans with the large dimensions and small ShapeFactors values correspond to large principal component values while the beans with small dimensions have small values. The plot also reveals an inverse relationship between the principal component one and two for types Seker, Sira, and Dermason showing that large scores of the second component correspond to small scores for the first component.

The coefficients of the second and the third principal directions are not straight forward to interpret, but the scatter plots of the two components against the first component reveal an interesting pattern that one is almost a mirror image of the other as seen in Figure 10. We also visualize the top three components using a 3-D to check for more patterns; however, the pattern is similar to the 2-D scatter plots as seen in Figure 11.

We were also interested to find out which variables are the most influential by analyzing the weight of each variable of the first principal direction. The analysis reveals that ShapeFactor2 and Compactness are the top two dominant features. The rank of each variable can be seen in a bar plot as seen in Figure 8.



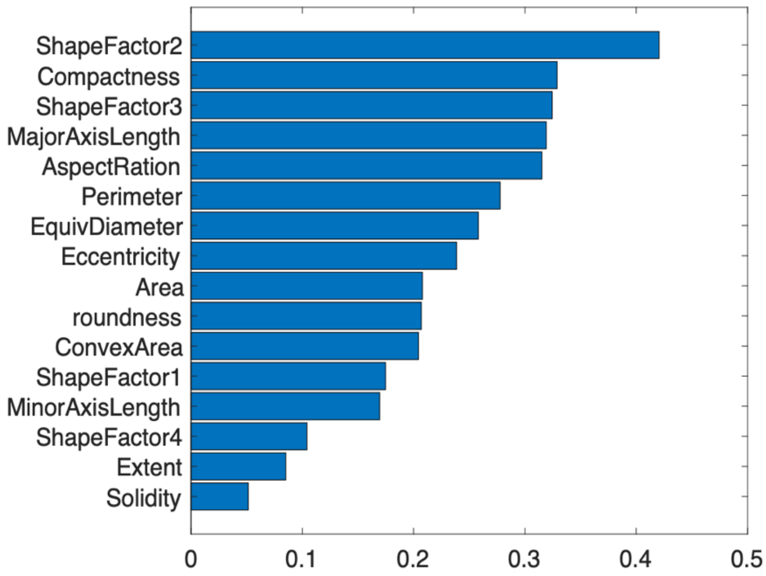


Figure 7

Figure 8

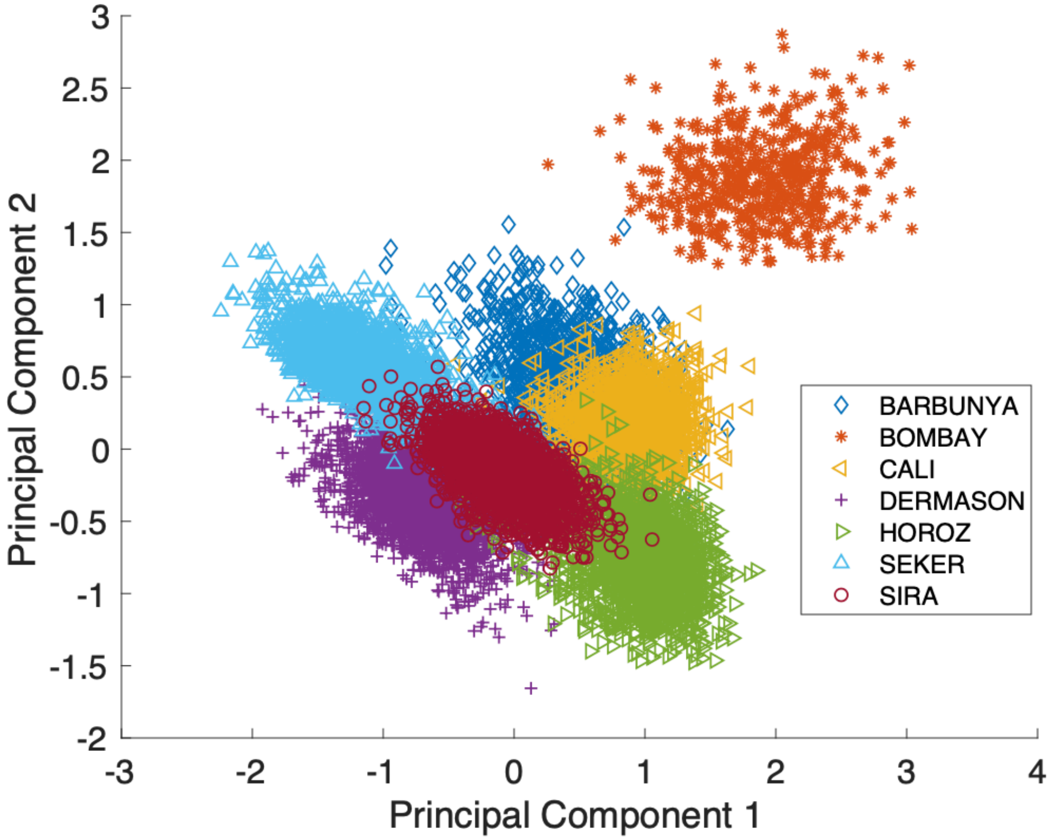


Figure 9: Scatter plot of the top two components

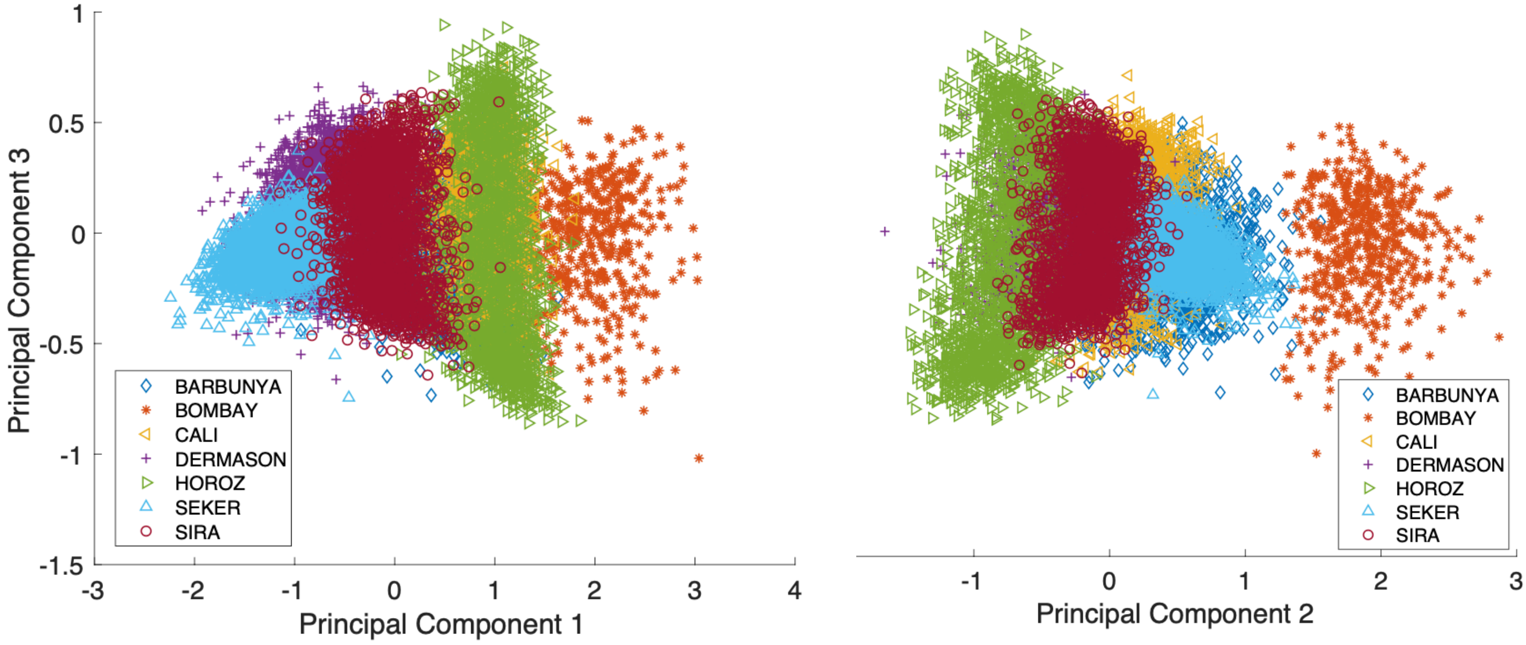


Figure 10: 2-D scatter plots of the top 3 principal components

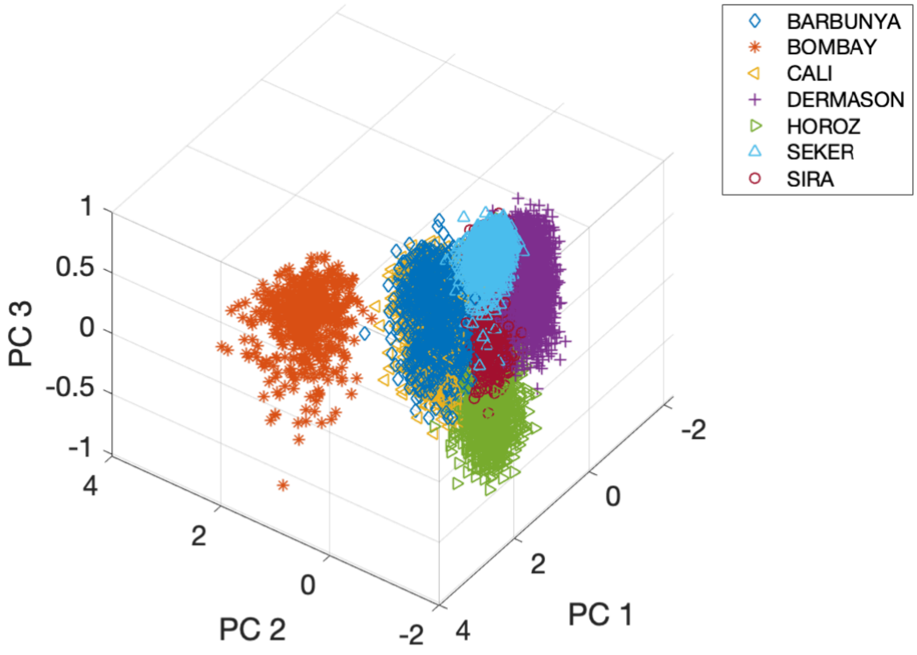


Figure 11: 3-D plot of the top 3 principal components

5. Linear Discrimination Analysis

Linear Discrimination Analysis (LDA) is another linear technique for dimensionality reduction that is different from PCA in two main aspects. The method requires knowing in advance the labels of the data and its sole objective is to maximize the variance between groups while minimizing the variance of the data points within each group.

There are seven types of beans in the data set, and we can only visualize four or fewer types simultaneously with LDA. The three types that we want to visualize together are Cali, Dermason, and Sira. We purposely chose these three because by inspection, they are similar to each other in terms of shape and dimension, and we wanted to examine how well LDA can detect the differences with this subset. Since the three types are quite similar to each other, we expected that the separation between the three groups would be negligible. Figure 12 shows the three clusters representing the three groups which are greatly overlapped. The second group that we want to visualize consists of types Bombay, Cali, and Seker. We chose to visualize these three types together because they have different dimensions and shape characteristics, and thus we expect that there would be significant differences between them. Figure 13 shows that the separation between type Bombay and the other two types is quite substantial while the separation between Cali and Seker is adequate with overlapping data points. We also analyze the types Barbunya, Cali, and Horoz together and we don’t expect these three types to be different from each other as they are very similar in shape. Figure 14 shows the visualization of these three types, and as expected, the three clusters greatly overlapped.

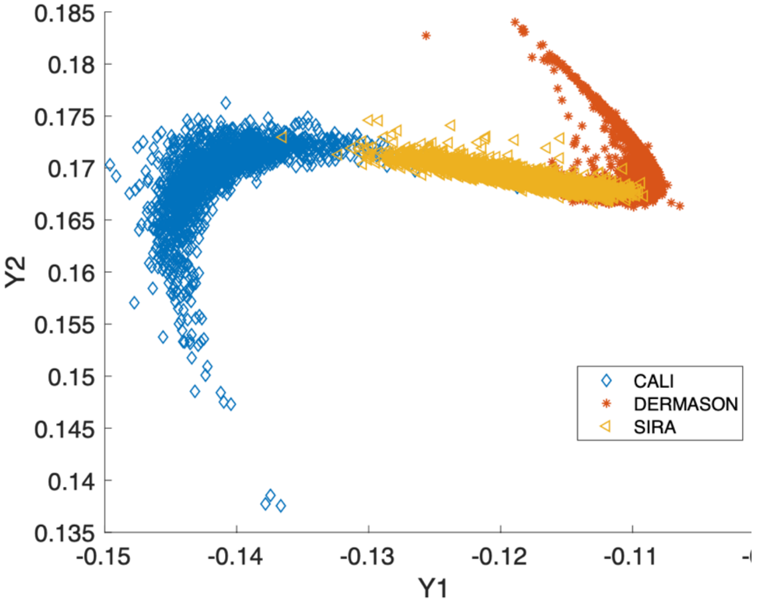


Figure 12

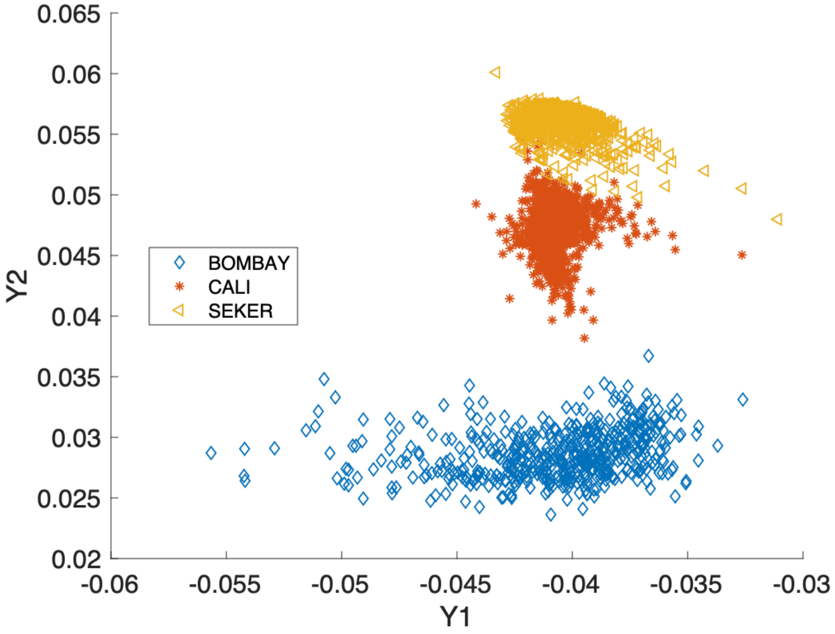


Figure 13

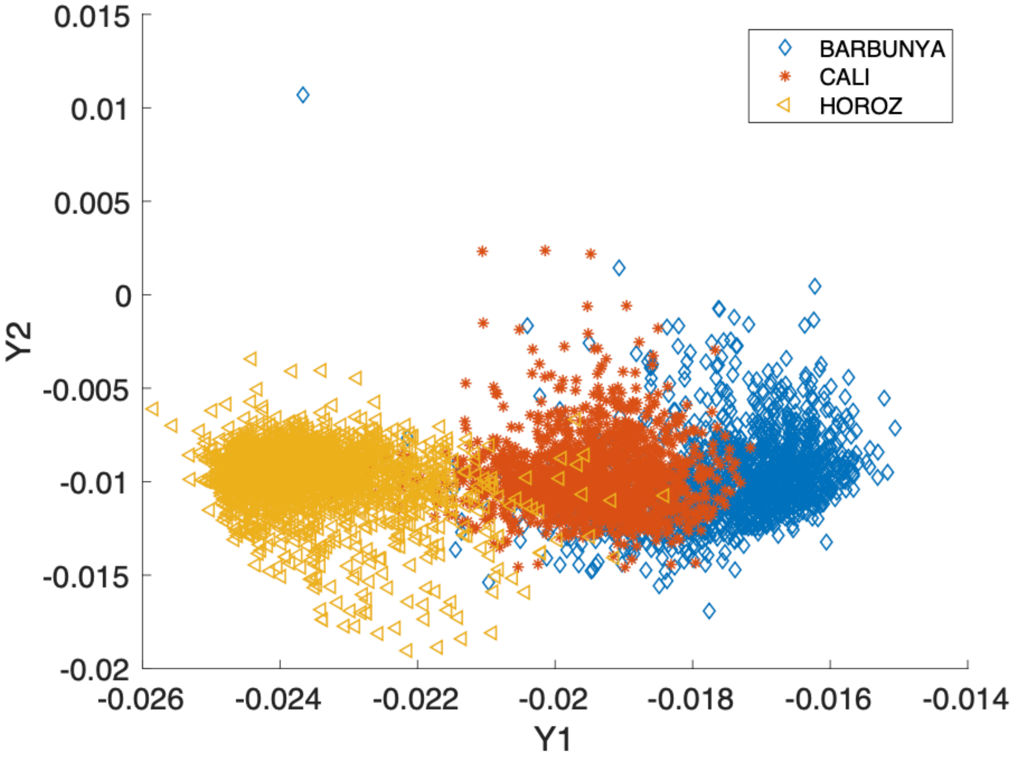


Figure 14

6. Classical Multidimensional Scaling

Classical Multidimensional Scaling (MDS) is a nonlinear dimensionality reduction technique that works by measuring information in terms of pairwise distances between each observation and preserving the global structure of the data. We used four distant metrics - Euclidean, Correlation, Cosine, and City block - to measure the distance between each data point. Furthermore, we used the Kruskal stress to measure how well MDS can represent the data in the low dimensional space.

We used the Euclidean metric as a baseline to compare whether using MDS will give us new insights or a new representation of the data. Figure 15 shows four scatter plots using the four metrics. Beside the magnitudes of Y1 and Y2, the differences between these four plots are quite subtle. All four plots show that the six types of beans are tightly clustered together when projected into the 2-D space. Furthermore, the locations of each type are about the same on all four plots. We use the Kruskal stress score to evaluate the performance of each distance metric. The Euclidean and City block performed quite well with the stress scores of 0.1184 and 0.1225, respectively while the performance of MDS using Cosine and Correlation metrics with the Kruskal stress score of 0.3393 and 0.3567, respectively.

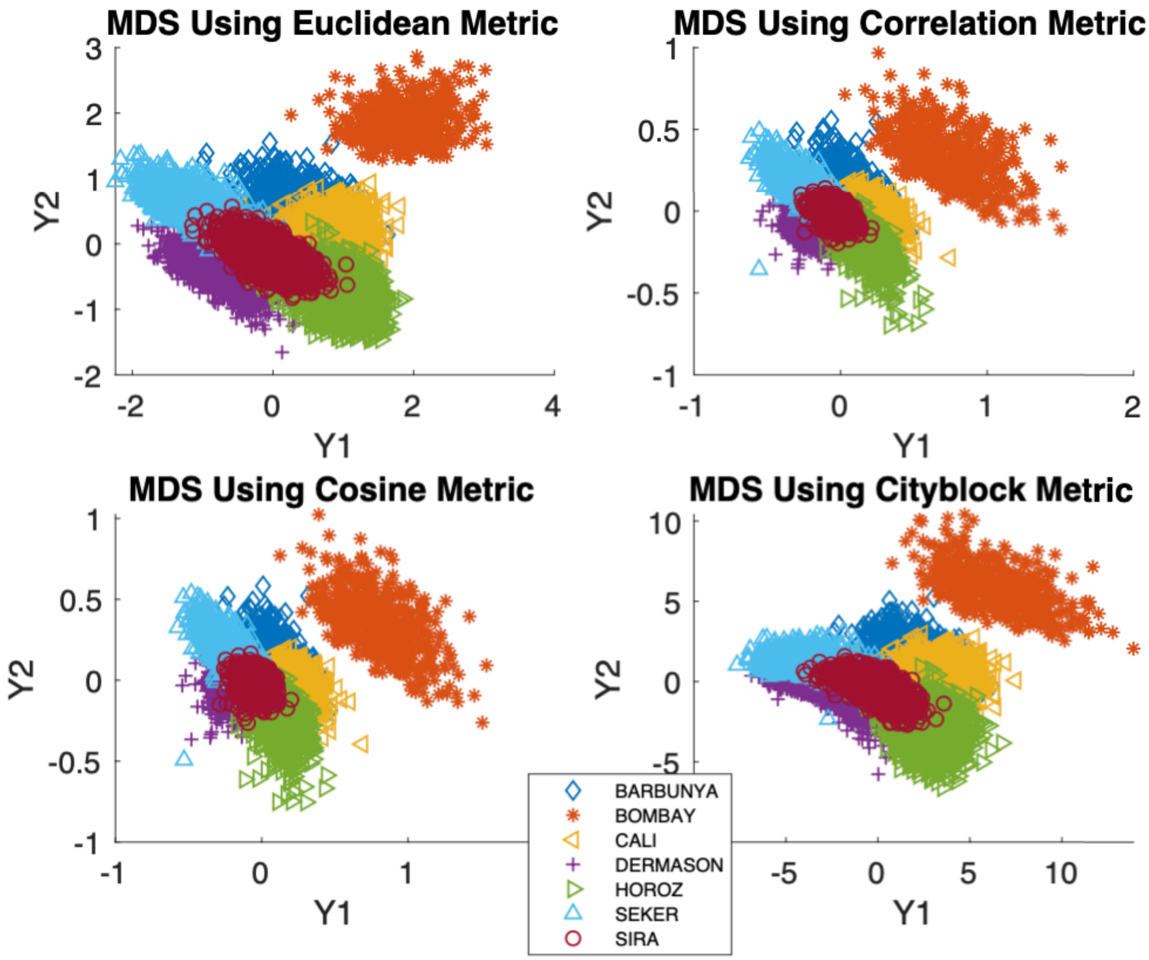


Figure 15

7. Laplacian Eigenmaps

Laplacian eigenmaps is another nonlinear dimensionality reduction method that projects the highly dimensional data to low dimensions while preserving its local structure. Laplacian eigenmaps works by building neighborhood graphs from the data with the assumption that nearby data points have the same linear structure. We measured the nearby points using the kNN approach with k=10 – the nearest 10 neighbors - and then calculated the weights using the Gaussian density function. We proceed by reducing the dimensions of the data by applying PCA to the scaled data before using Laplacian Eigenmaps because the weight matrix W was highly ill-conditioned. We keep six principal components which account for 99% of the total variance as our new data and we seek to reduce the number of dimensions of the transformed data.

Figure 16 shows the weight matrix which suggests that there are about six clusters where each cluster represents a type of beans in our data. The sizes of the clusters are uneven which we suspect is caused by the unbalanced sizes of the bean types in the data. We use the embedding matrix to visualize the data in 2-D and 3-D. Figure 17 shows that the embedding values of type Bombay are quite far from the rest of the six types. As for other types, they are very similar locally and thus Laplacian Eigenmaps accurately show the close relationship between them. We also visualize the three embedding components to learn more about the data as seen in Figure 18. It seems that types Horoz and Seker are quite different from each other as their embedding values are on the opposite ends albeit tightly clustered together. Overall, the insights that we got from Laplacian Eigenmaps matched with what we have in the three previous techniques.

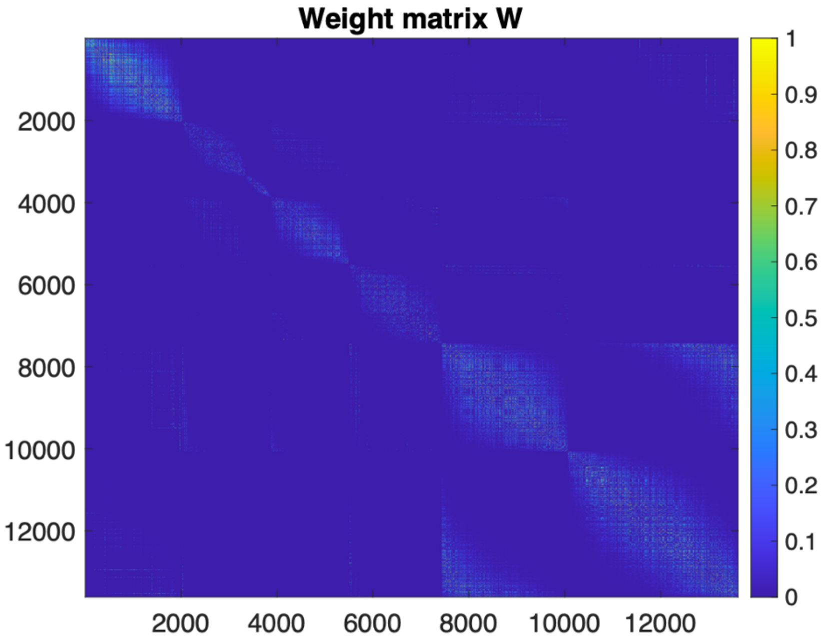


Figure 16

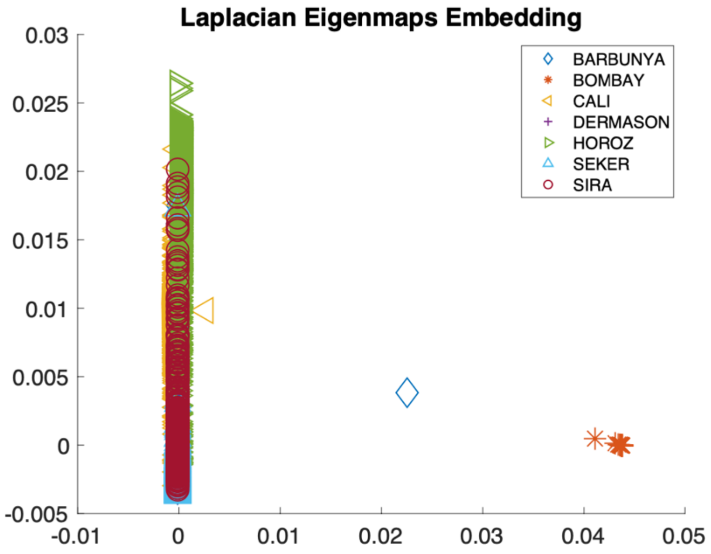


Figure 17

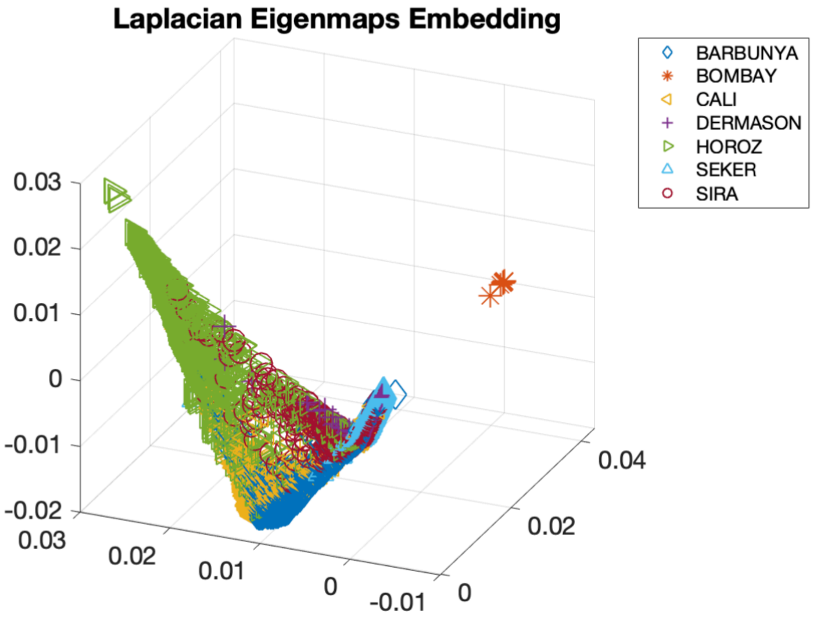


Figure 18

8. Conclusion

We have utilized low-dimensional data visualization and four dimensionality reduction techniques to visualize the dry bean data set. Low dimensional data exploration is useful for a subset of variables and can reveal patterns and relationships in low dimensional space, but it has its limitations because we cannot gain insights into relationship of all the variables simultaneously. Thus, we proceed to reduce the dimensions while keeping the intrinsic structure of our data using four dimensionality reduction techniques - PCA, LDA, MDS, and Laplacian Eigenmaps. The four methods have different objectives, and it depends on what one is trying to achieve and how much resources in terms of computing power and time one has to tuned parameters that comes with some of the methods.

From the four techniques, we learn that the data is intrinsically linear because the four methods produce similar results. Furthermore, it is not necessary to have sixteen variables to describe the shape and the dimensions of a bean. We also learn the type Bombay is consistently different from the rest regardless of what dimensionality reduction method we used which suggests that we don’t need a lot of information about Bombay to correctly classify it. However, for the remaining six types, they are similar in shape and dimension, so we need to collect more information to accurately describe and classify them.

9. References

Koklu, Murat, and Ilker Ali Ozkan. “Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques.” *Computers and Electronics in Agriculture*, Elsevier, 30 May 2020, www.sciencedirect.com/science/article/pii/S0168169919311573?via%3Dihub.