Shuang Lin

CSE 347 - Individual Project - Clustering Analysis

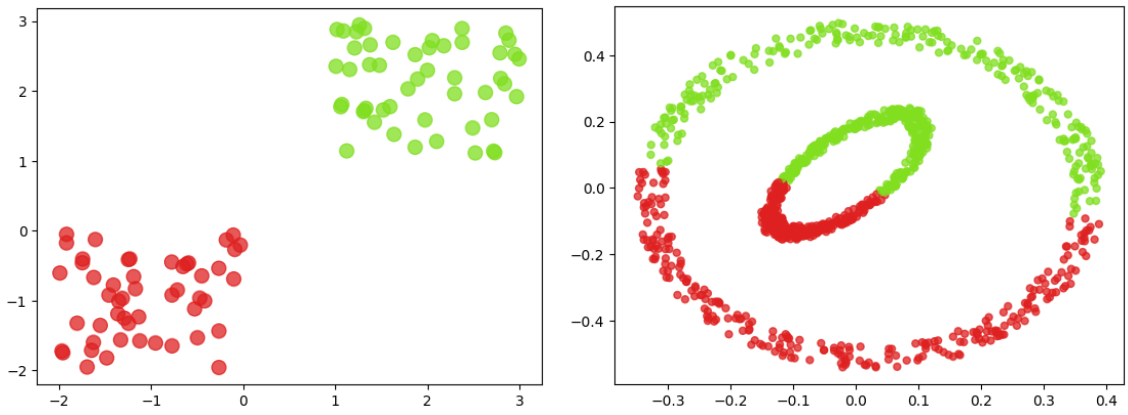
Simulation

Square and elliptical datasets with given X and Y points.

Two distinct clusters can be formed by the coordinate points.

1. *Compare the two methods and discuss their pros and cons.*

    a. **K-means**



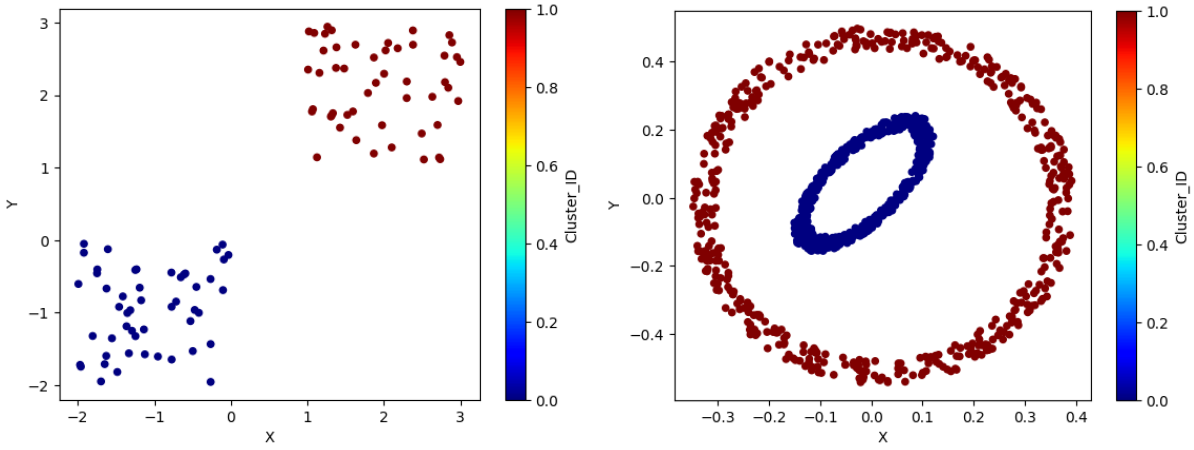
i. *Pro*:

1. Linear time complexity; only 1 second run time when calculating the elliptical clusters on my PC.
2. Easy implementation and understanding, so it's the most commonly used method.

ii. *Con*:

1. Need to manually choose a k-number to find the correct clusters.
2. Generally finds spherical clusters, so it's weak to other shapes.
3. Sensitive to centroid initialization, outliers, data density, and data size.
4. Due to the elliptical data being non-convex, it could not be clustered correctly.

b. Spectral



i. Pro

1. Can separate non-convex clusters.

ii. Con

1. Exponential time complexity with a runtime of 2 seconds when calculating the elliptical clusters on my PC.
2. Need to manually tune the kernel parameter (gamma) to find the correct clusters, and higher values lead to greater time complexity.

2. *Centroid initialization*

a. A technique of randomly choosing k-data-points, or seeds, as the initial centroids. The initial centroids are important because it then repeats through iterations to choose the best centroids

3. *Performance Analysis - Similarity Distance Measures*

a. Cosine similarity

$$Cosine\ Similarity = 1 - \frac{A \cdot B}{||A||\,||B||} = 1 - \frac{\sum_i A_i B_i}{\sqrt{\sum_i A_i^2}\sqrt{\sum_i B_i^2}}$$

i. The cosine similarity ranges from -1 to 1, where -1 means the vectors are opposites, 0 means the vectors are perpendicular, and 1 means they are identical. In the square data, the cosine similarity between the two clusters is *-0.9119*, which tells us that the two square clusters are opposites and two separate clusters. In the elliptical data, the measure is *0.0041*, which shows that the two clusters have no correlation.

b. Gaussian kernel similarity

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

i. The Gaussian kernel is a nonlinear function of Euclidean distance, ranging from 0 to 1. Higher values mean that the two points are similar. The square data measured *0.002844*, which means that the two clusters are not similar. The elliptical data showed a measure of *0.9014*, which means that the two clusters are similar.

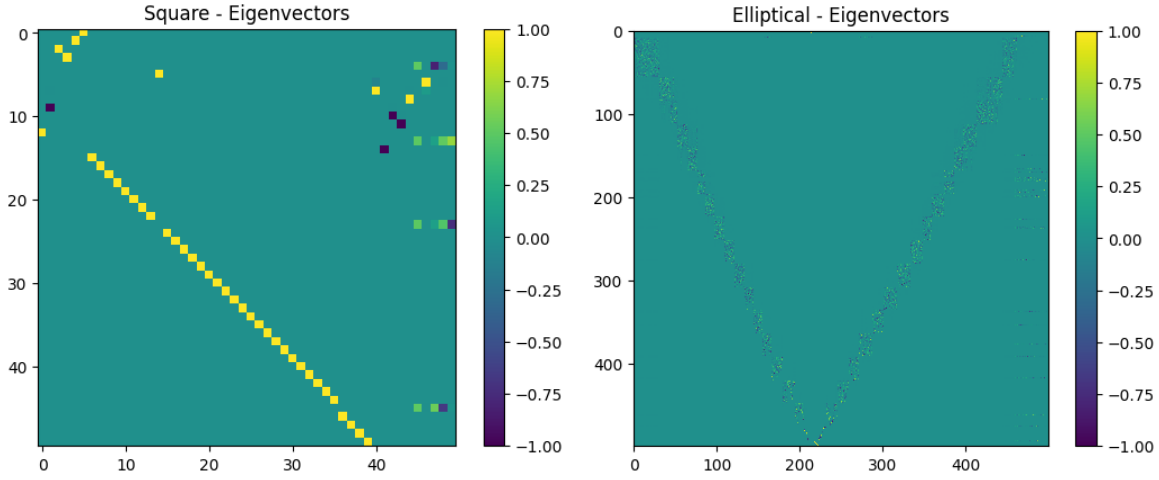
ii. The bandwidth parameter decides how much the similarity between two data points decreases as their distance increases. I set my sigma to 1 since it is the common approach.

4. Performance Analysis - Laplacian Matrices

a. Unnormalized Laplacian Matrix

$$L = D - W$$

D: degree matrix W: weighted adjacency matrix

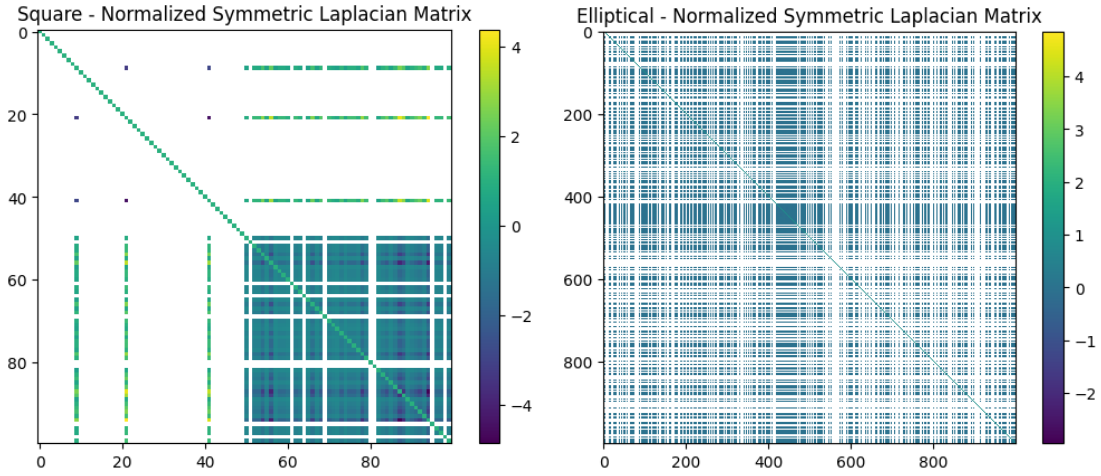


i. In the square data, the second smallest eigenvalue of the unnormalized Laplacian matrix is 7.3255^{-17}. This is a large spectral gap, compared across all eigenvalues, meaning that there are at least 2 distinct clusters. The Fiedler vector has values ranging from -1 to 1 which also shows clusters at two different points.

ii. In the elliptical data, the second smallest eigenvalue of the unnormalized Laplacian matrix is 1.9702^{-288}, also showing two clusters. The Fiedler vector also shows values from -1 to 1, stating two clusters.

b. Normalized Symmetric Laplacian Matrix

$L_N = D^{-1/2}LD^{-1/2}$ Normalized cuts tend to avoid unbalanced cuts.



i. The diagonal elements are 1, which means that each node is fully similar to itself. The off-diagonal elements are between -1 and 0, which means that the nodes are dissimilar. They are closer to -1, meaning that clusters of dissimilar nodes exist.

ii. Both square and elliptical data show the diagonal and off-diagonal elements obeying the rules, indicating that there are clusters available.

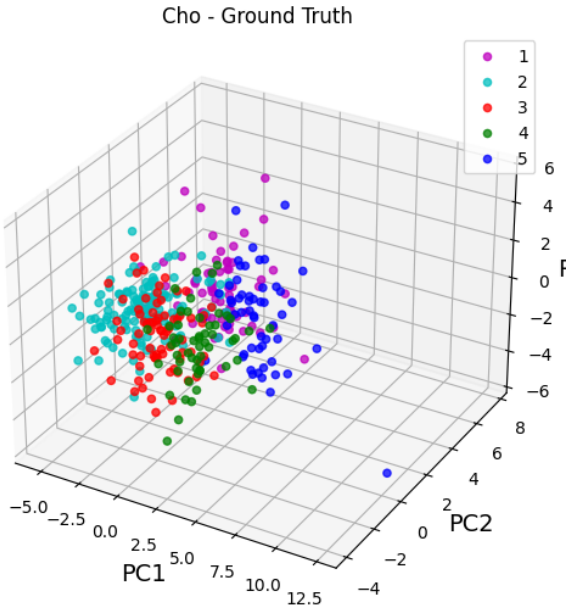
Real-World

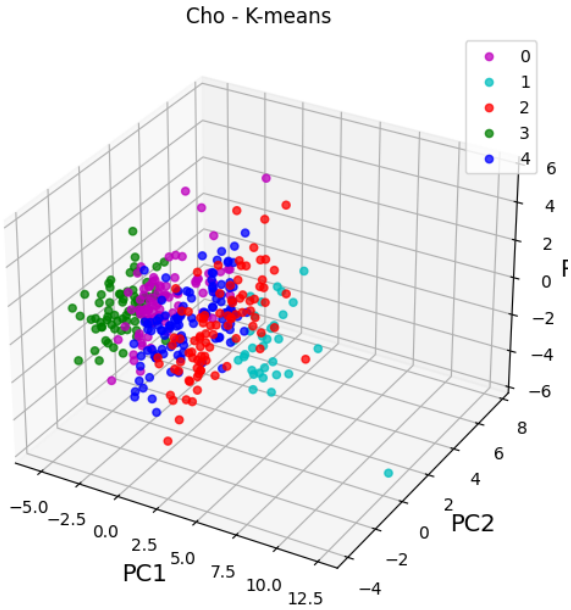
Cho and Iyer datasets are collected from the UCI machine learning repository.

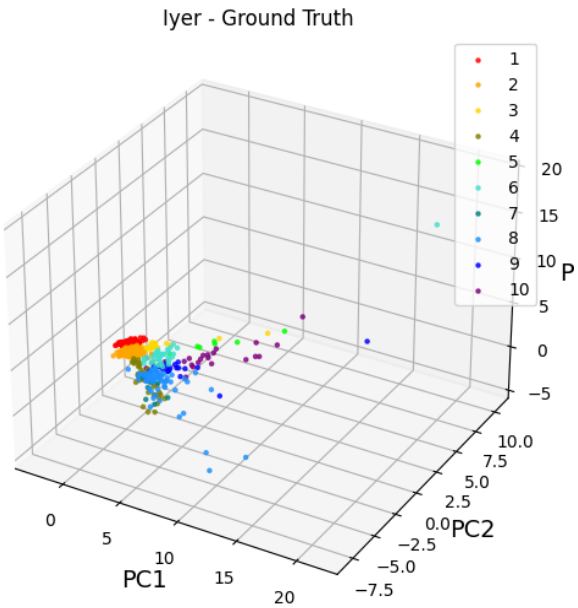
The first two columns are Gene_ID and ground truth clusters (-1 is outliers). The rest of the columns are the gene's expression values (attributes).

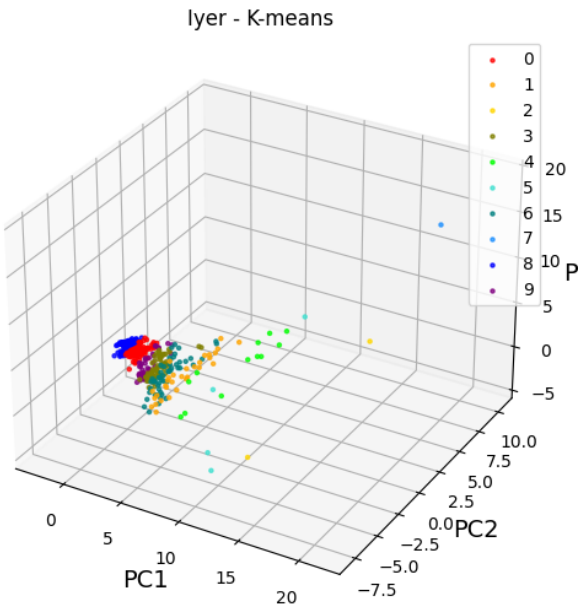
1. *Clustering Methods*

K-means

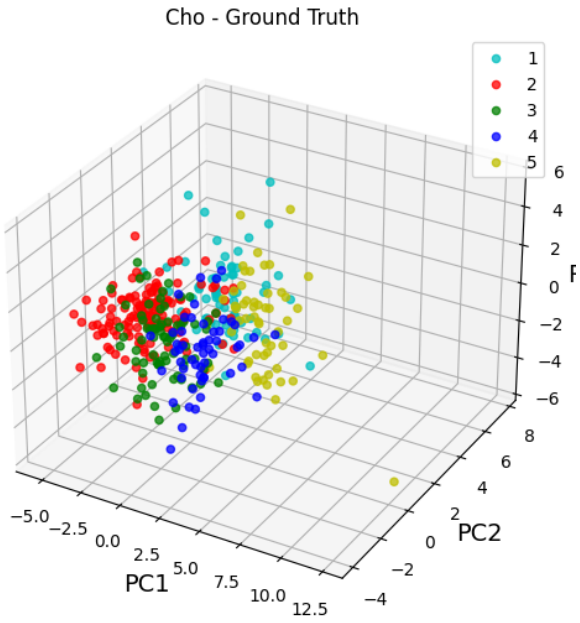


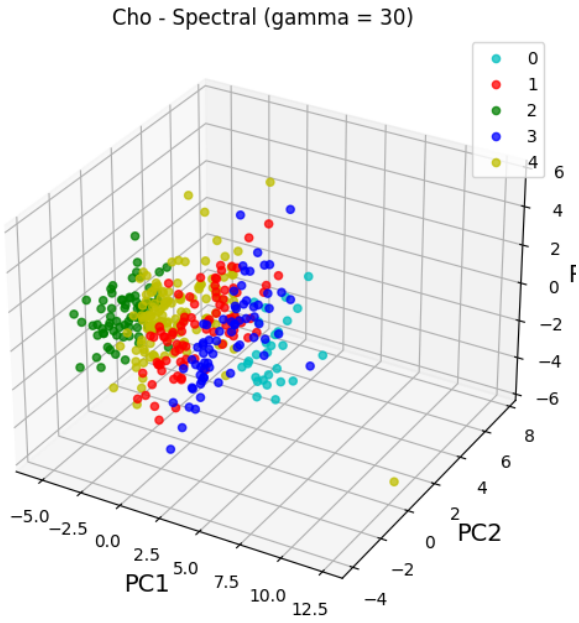


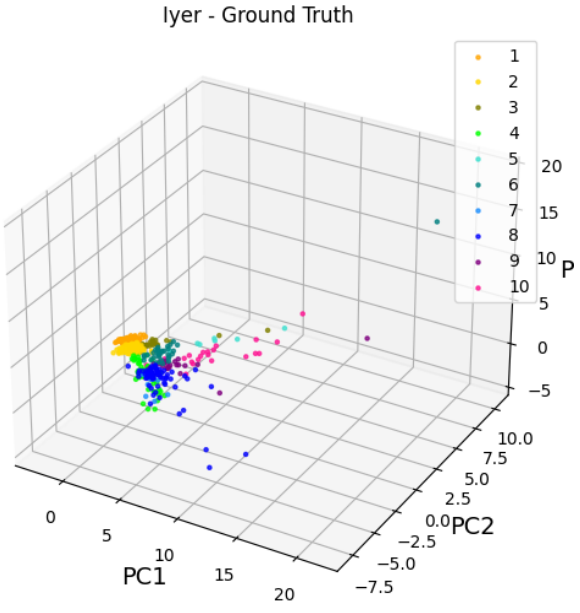


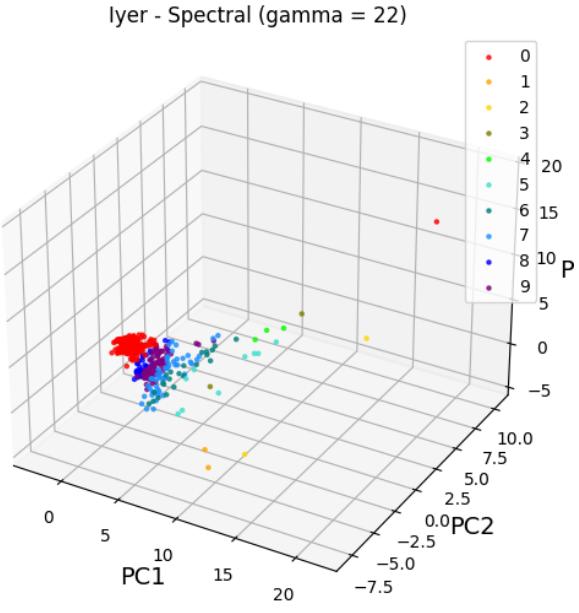


Spectral


Cho - Ground Truth
1
2
3
4
5
PC1
PC2


Cho - Spectral (gamma = 30)
0
1
2
3
4
PC1
PC2


Iyer - Ground Truth
1
2
3
4
5
6
7
8
9
10
PC1
PC2


Iyer - Spectral (gamma = 22)
0
1
2
3
4
5
6
7
8
9
PC1
PC2

2. *Clustering Validation*

a. **External Index**

```
Cho - K-means
Adjusted Rand Index: 0.2516074828815705
Purity: 0.004759386567953464
Normalized Mutual Information: 0.3863242288241127
```

```
Iyer - K-means
Adjusted Rand Index: 0.363279766892468
Purity: 0.004510921177587844
Normalized Mutual Information: 0.43590377295244015
```

```
Cho - Spectral
Adjusted Rand Index: 0.25207374796201043
Purity: 0.0044598612487611496
Normalized Mutual Information: 0.38497848131060664
```

```
Iyer - Spectral
Adjusted Rand Index: 0.36890801794919287
Purity: 0.005492917028042787
Normalized Mutual Information: 0.41705520580456856
```

i. The results are similar with spectral clustering proving more accurate.

ii. The Rand Index score ranges from 0 (no agreement) to 1 (perfect agreement) and is affected by chance and may not be normalized. Therefore, the Adjusted Rand Index is preferred. The Adjusted Rand Index measures similarity by counting pairs of points that are either in the same cluster or in different clusters in both inputs. This is adjusted for chance; 0 for random partitions and 1 for identical partitions.

iii. Purity is the measure of clusters that were classified correctly, ranging from 0 (no agreement) to 1 (perfect agreement).

iv. Mutual Information measures the amount of information shared by two clusterings. Normalized mutual information is the normalization of the mutual information to scale the results between 0 (no mutual information) and 1 (perfect correlation).

b. Internal Index

```
Cho - K-means
Clusterwise SSE: [1200.62529054289, 2043.206970309856, 1874.139628668478, 1741.7458367857907, 817.7983656478506]
Total SSE: 208.57237804096195
```

```
Iyer - K-means
Clusterwise SSE: [447.87563453203313, 980.3121181421941, 954.3795738708596, 423.46121814880087, 961.5668848948111,
Total SSE: 45.07600725643452
```

```
Cho - Spectral
Clusterwise SSE: [84.72043563659211, 674.1528964843112, 210.17776716716864, 486.1643607733016, 1035.890246701838]
Total SSE: 2491.1057067632114
```

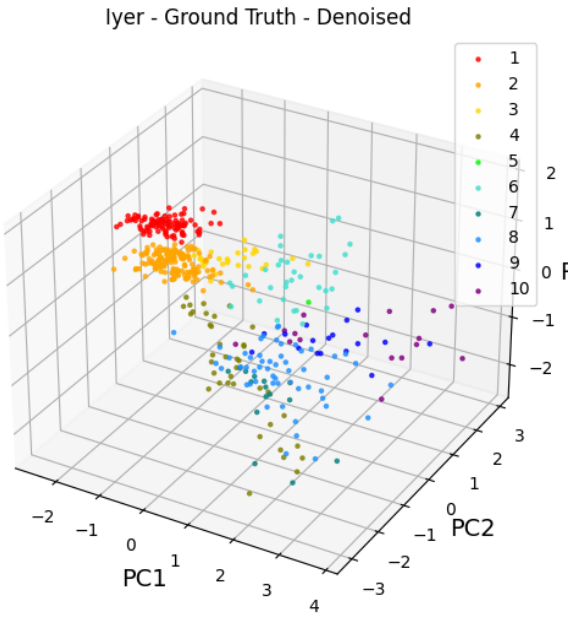
```
Iyer - Spectral
Clusterwise SSE: [1042.1027301196189, 1.343953584121469, 83.39263384399693, 61.09259147514411, 9.822546438960016,
Total SSE: 1997.2588516235157
```

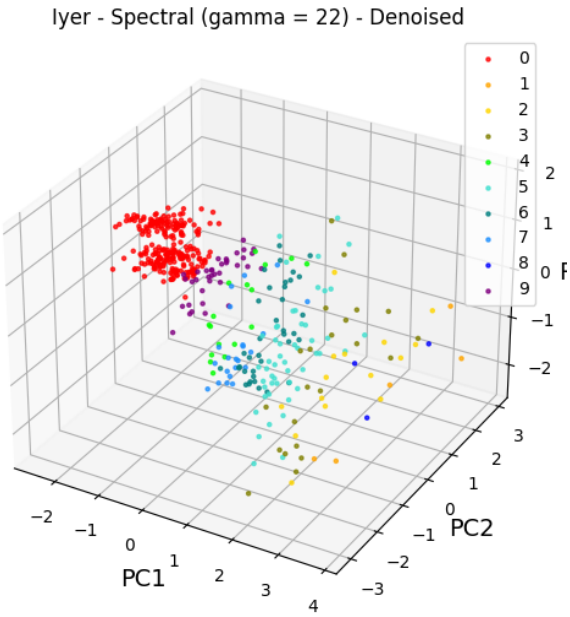
i. The Sum of Squared Error (SSE) measures how much variation is within a cluster. The lower the Total SSE, the more compact and homogeneous your clusters are. The lower the clusterwise SSE, the more similar the data points are to their cluster center. The Iyer clusterwise SSEs are cut off due to their length.

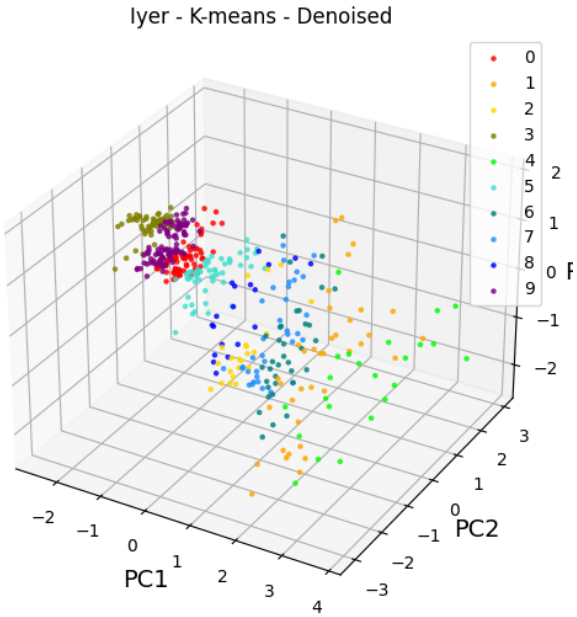
3. *Impact of Data Normalization*

   Data normalization transforms the data into a standard scale or range. Normalizing the data into three components reduced the influence of large-scaled variables and allowed easier visualization on an XYZ graph. However, the first three dimensions had a summed variance of 0.7267 in Cho and 0.7842 in Iyer. Over 20% of the weights by other components weren't loaded by PCA. This could have played a role in decreasing the accuracy of the clusterings.

4. *Impact of Noise*







```
Iyer - K-means
Adjusted Rand Index: 0.3004724478308412
Purity: 0.005054535780792764
Normalized Mutual Information: 0.44640450986330277
```

```
Iyer - K-means
Clusterwise SSE: [380.84221537215336, 642.0337104671293, 82.81545817603234, 743.8731486649194, 701.9959615429112,
Total SSE: 11.993281534010375
```

```
Iyer - Spectral
Adjusted Rand Index: 0.4194453636289485
Purity: 0.006690140845070422
Normalized Mutual Information: 0.454809745423741
```

```
Iyer - Spectral
Clusterwise SSE: [90.2749514016293, 27.811862661300292, 56.62977665265869, 77.03359575757511, 25.68572946422401,
Total SSE: 444.6352206623011
```

After the standard scaler, z-score normalization was applied with a threshold of 2. It resulted in 33 rows of noise data being removed. This cleaned the outlying data points. Turns out the K-means accuracy decreased by 6% while the Spectral accuracy increased by 5%. Both methods total Sum of Squares Error also decreased.