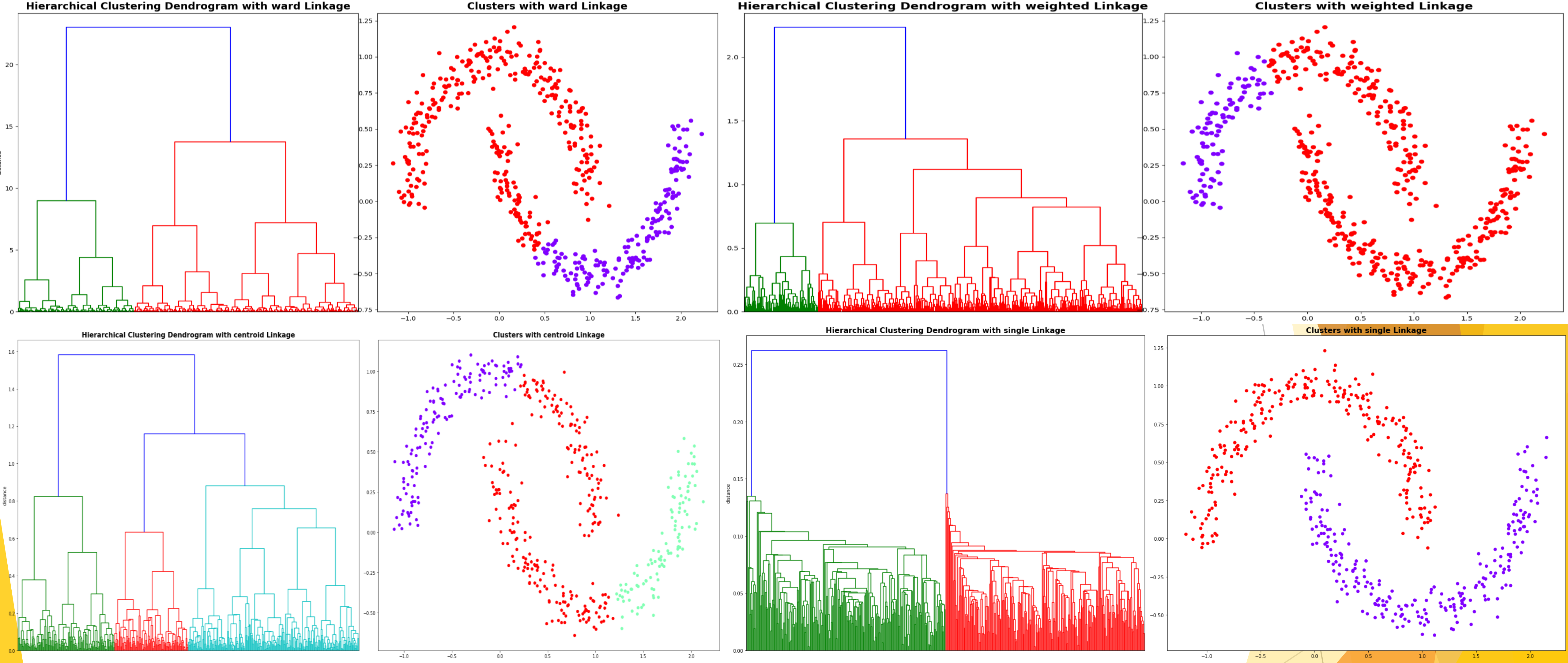
Impact of linkage on hierarchical clustering and covariance type on GMM clustering

Data Description And methods

- 2 datasets were used, one a simulated dataset, the moons dataset with noise =0.08
- And the other a real dataset of RNA-Seq samples for tumor tissues. This dataset had 800 samples and 20531features. It has 5 classes and 5 clusters can be detected given the sufficiently informative dimension reduction.
- The linkage method used were, ‘centroid’, ‘weighted’ and ‘ward’ along ‘single’ as a comparison
- The covariance matrices used for GMM were ‘full’, ‘diagonal’, ‘spherical’ and ‘tied’.



Hierarchical Clustering Dendrogram with ward Linkage

Clusters with ward Linkage

Hierarchical Clustering Dendrogram with weighted Linkage

Clusters with weighted Linkage

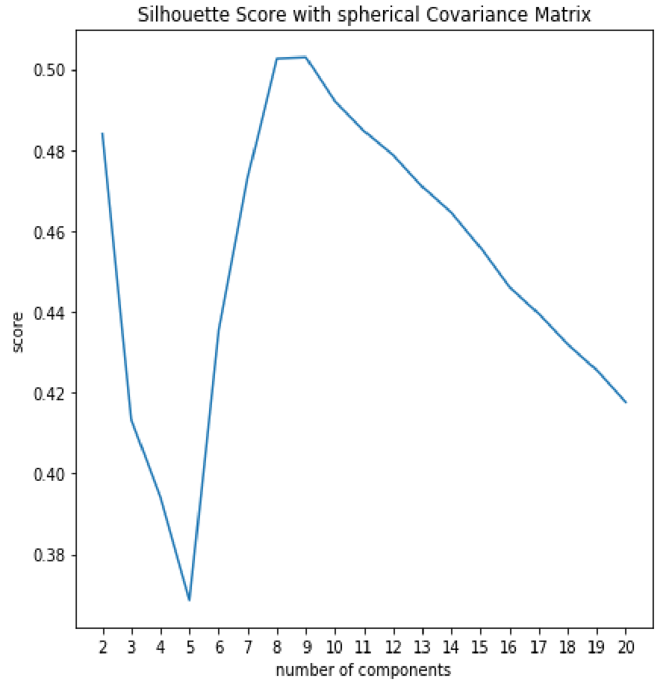
Hierarchical Clustering Dendrogram with centroid Linkage

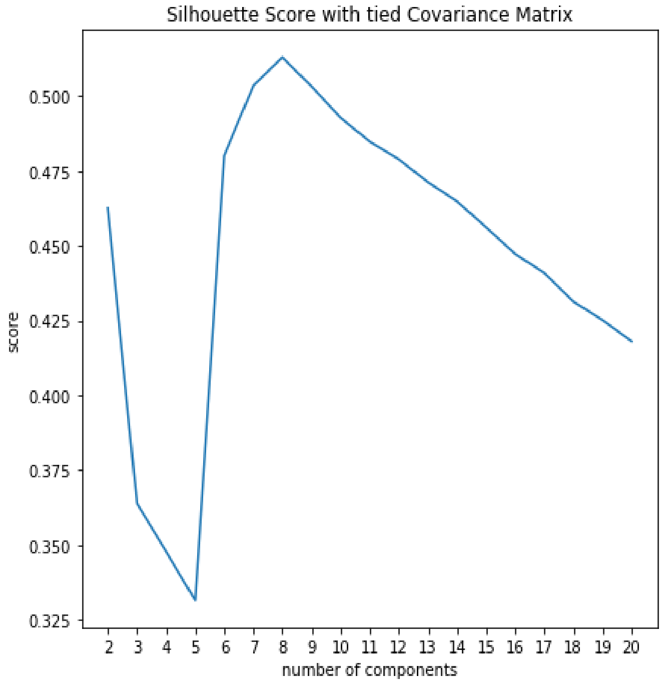
Clusters with centroid Linkage

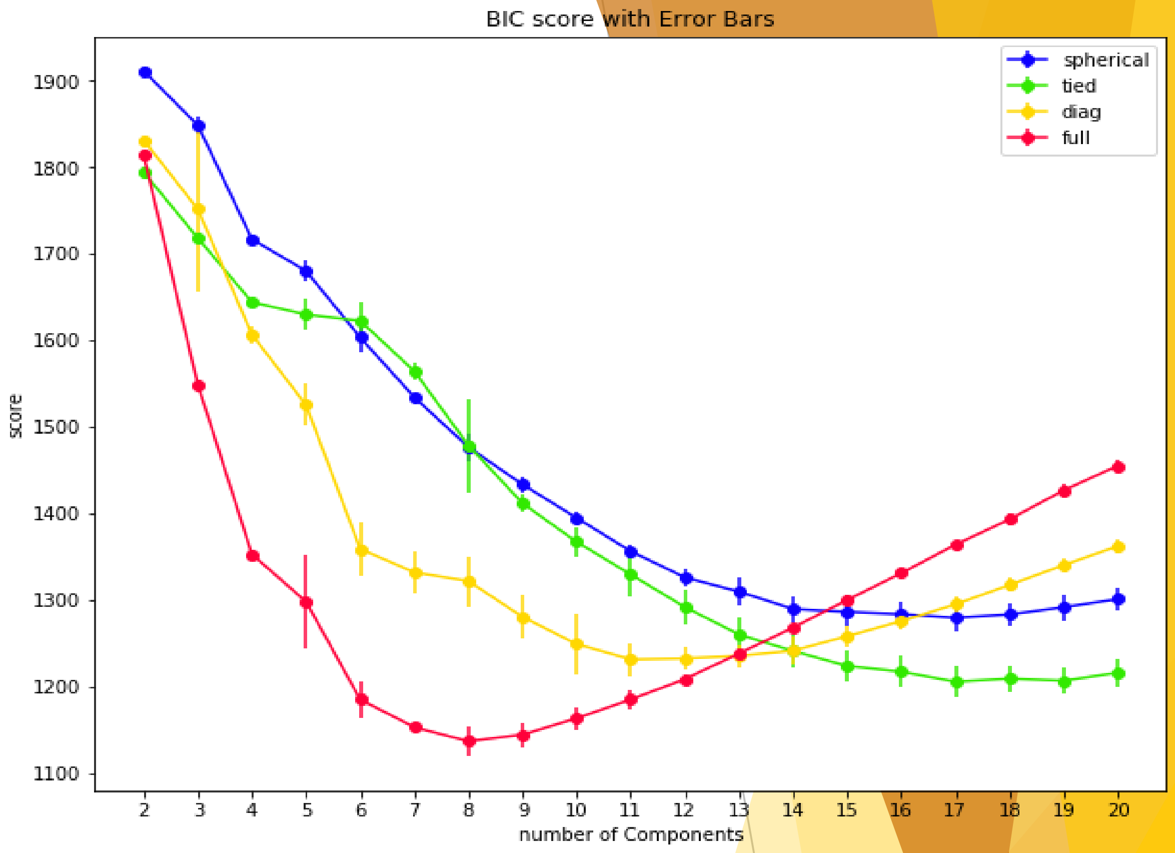
Hierarchical Clustering Dendrogram with single Linkage

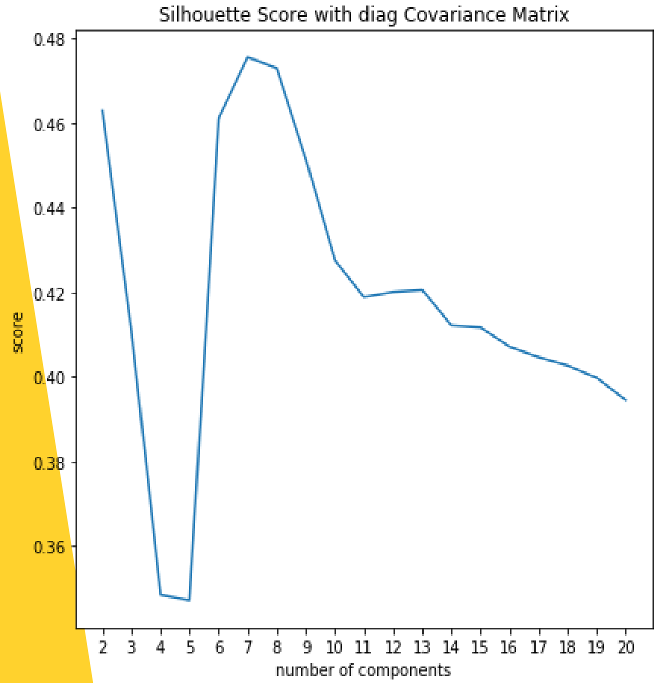
Clusters with single Linkage

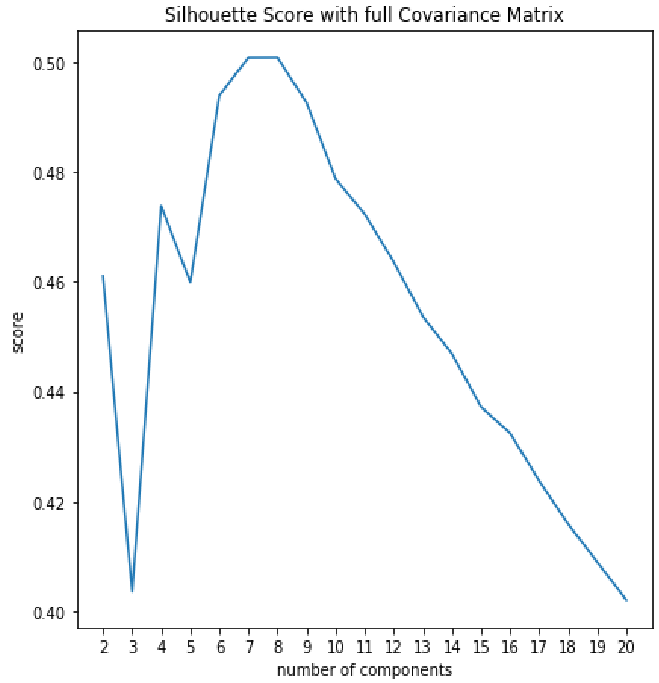
Here the best linkage method is the single method, the number of clusters were chosen based on the dendrogram trees. Since single method only looks at nearest neighbors, it forms chains along the data points thus capturing curvilinear shapes. The centroid method measures the Euclidean distance between two clusters' geometric centroids. In this case the outer tails of the curves are counted as distinct clusters as they are farther from the inner tails, and the inner tails are closer to each other than the outer tails, in other words it's as if they have gravitated towards each other and thus form their own cluster with their centroid about the center of the figure. In Ward method, proximity between two clusters is the magnitude by which the summed square in their joint cluster will be greater than the combined summed square in these two clusters. Ward linkage captures shapes ('clouds') that are more concentrated towards the center and less dense in their outer rims so it is reasonable that it does not perform well here. In the weighted method or WPGMA, proximity between two clusters is the arithmetic mean of all the proximities between the objects of one, on one side, and the objects of the other, on the other side. Here ward and weighted have somewhat similar results perhaps as only two clusters were found.


Silhouette Score with spherical Covariance Matrix

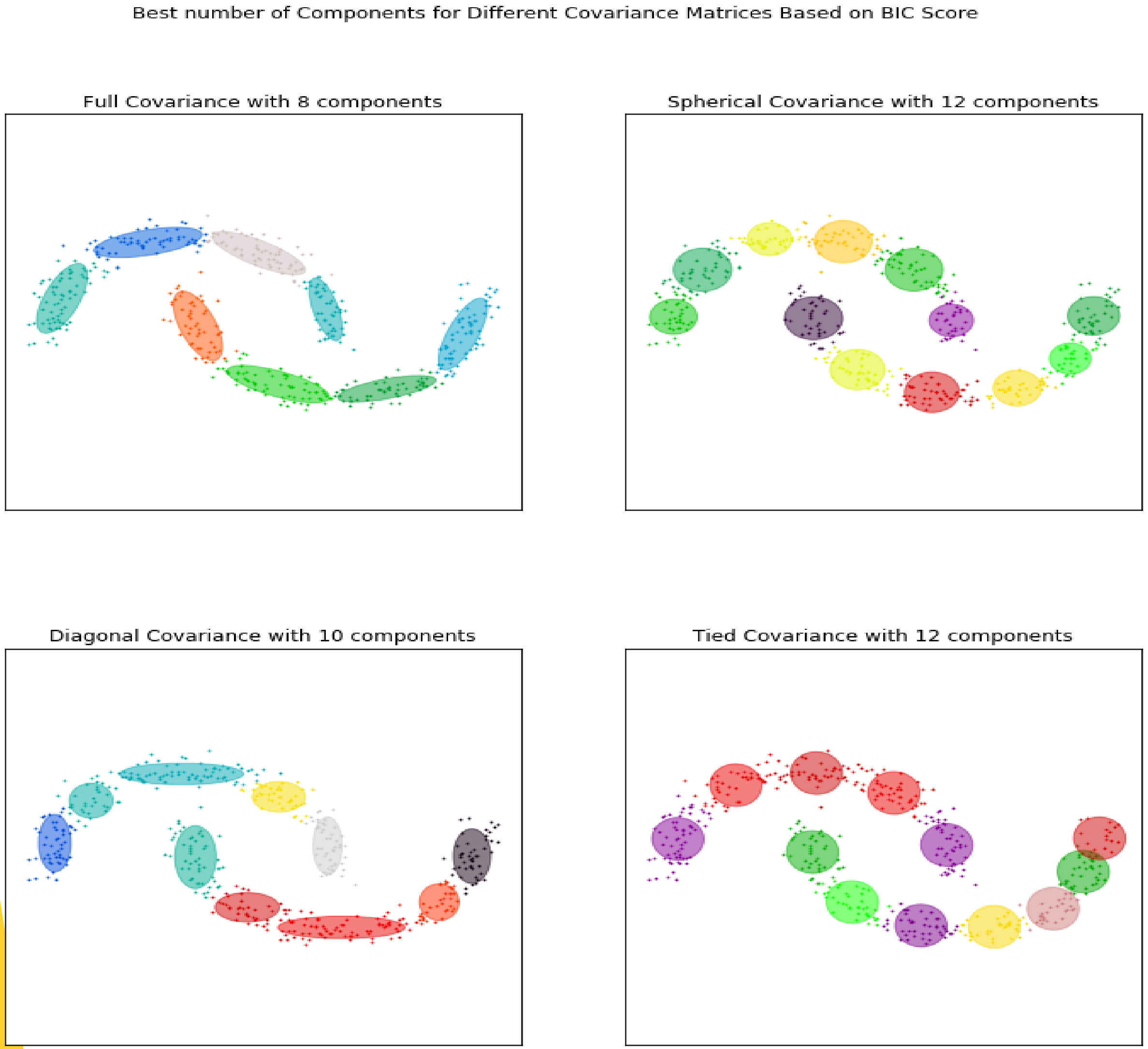

Silhouette Score with tied Covariance Matrix

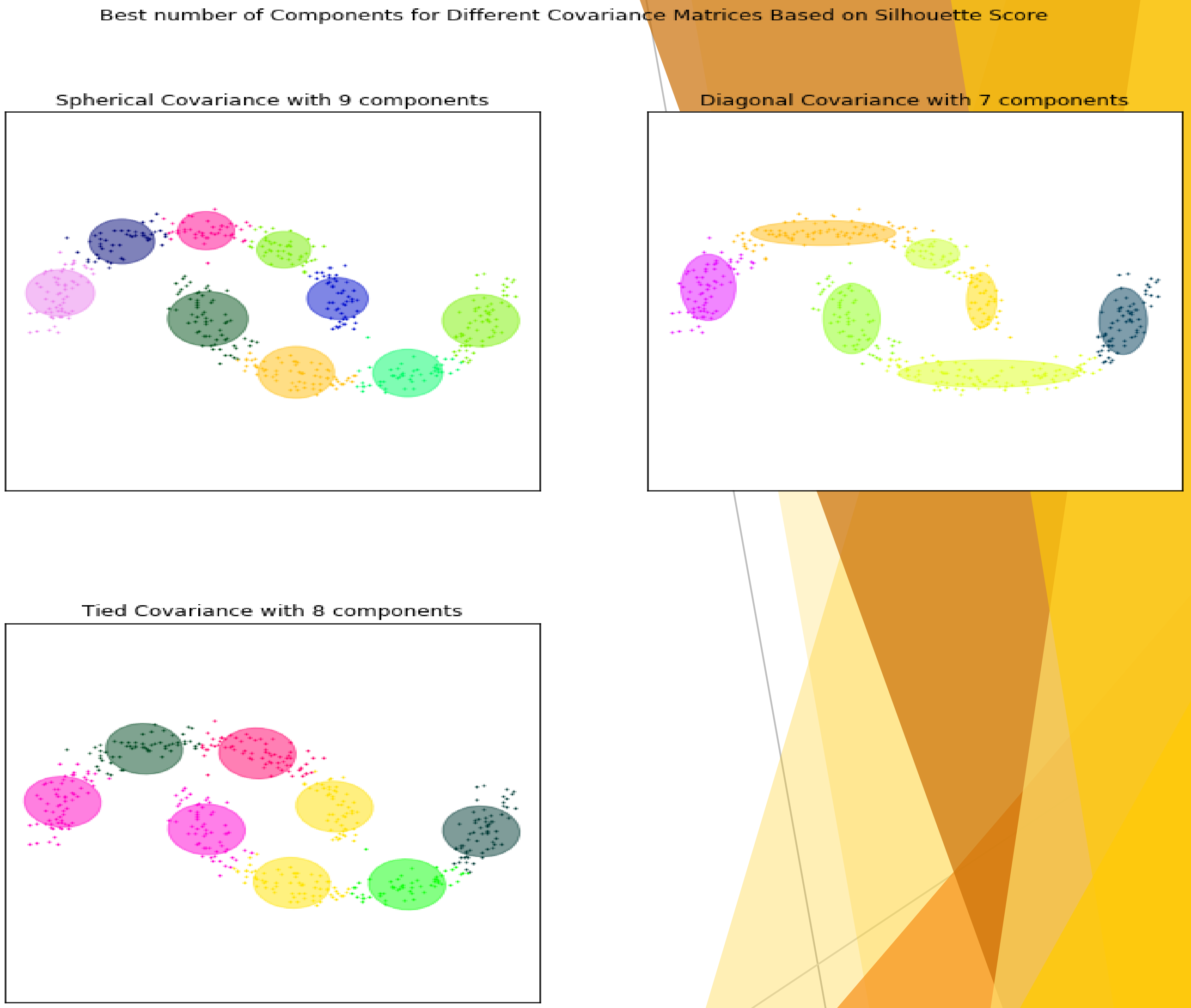

BIC score with Error Bars


Silhouette Score with diag Covariance Matrix


Silhouette Score with full Covariance Matrix

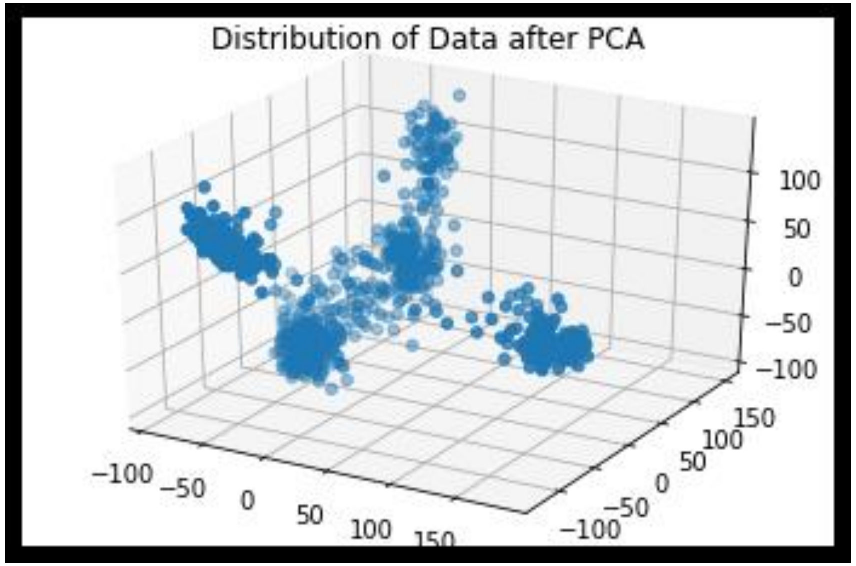
Two different measures, the silhouette score and the BIC score were used to deduce the number of clusters for the GMM method for different types of covariance matrices. As expected full covariance has best performance according to BIC score but according to silhouette score tied covariance is better in this case. For choosing the number of components based on the BIC score the elbow rule was used. The results shown for both scores are the average of 100 trials. The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. Here, the score shown is mean for all values in the dataset. As expected there is not a significant difference between the two scores.


Best number of Components for Different Covariance Matrices Based on BIC Score

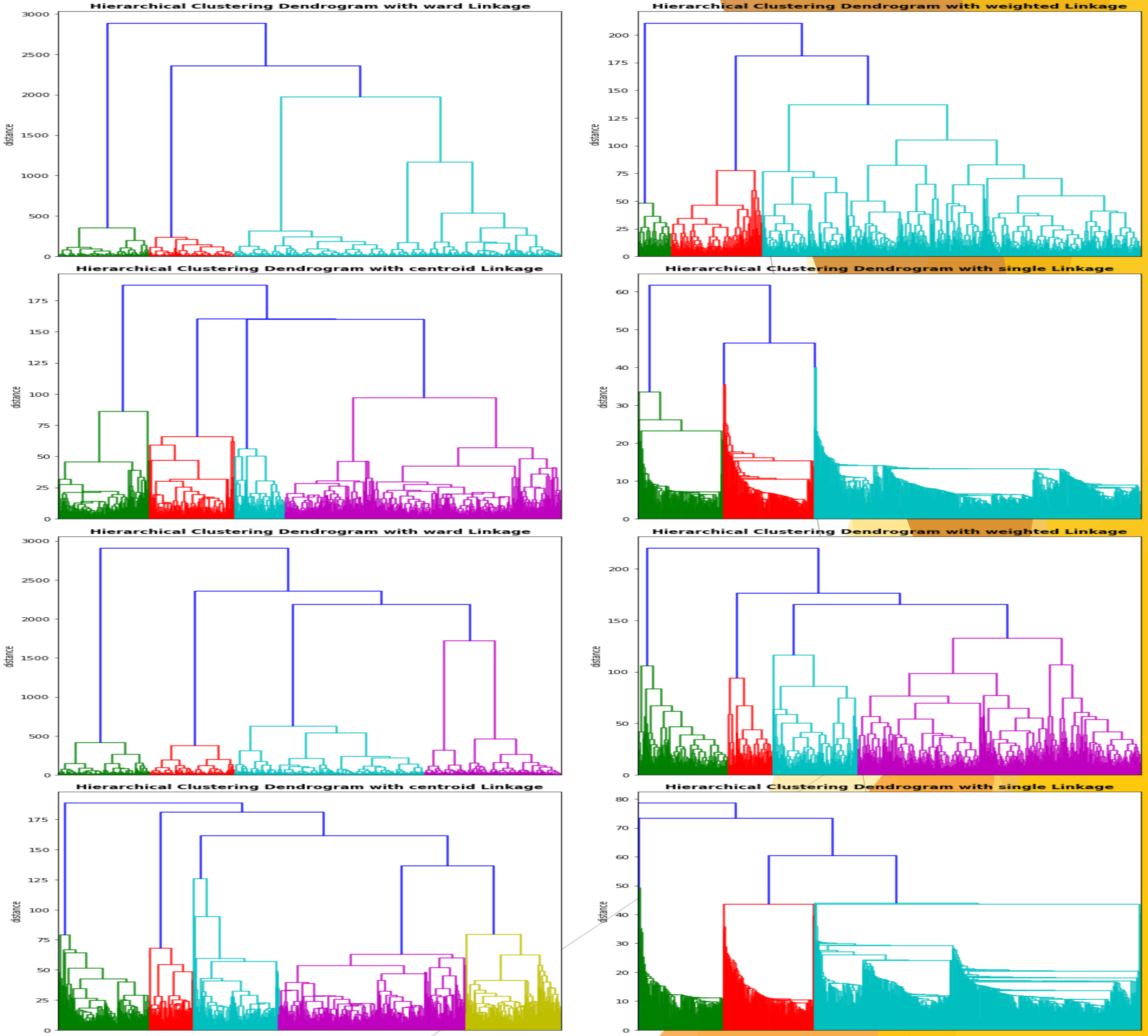

Best number of Components for Different Covariance Matrices Based on Silhouette Score

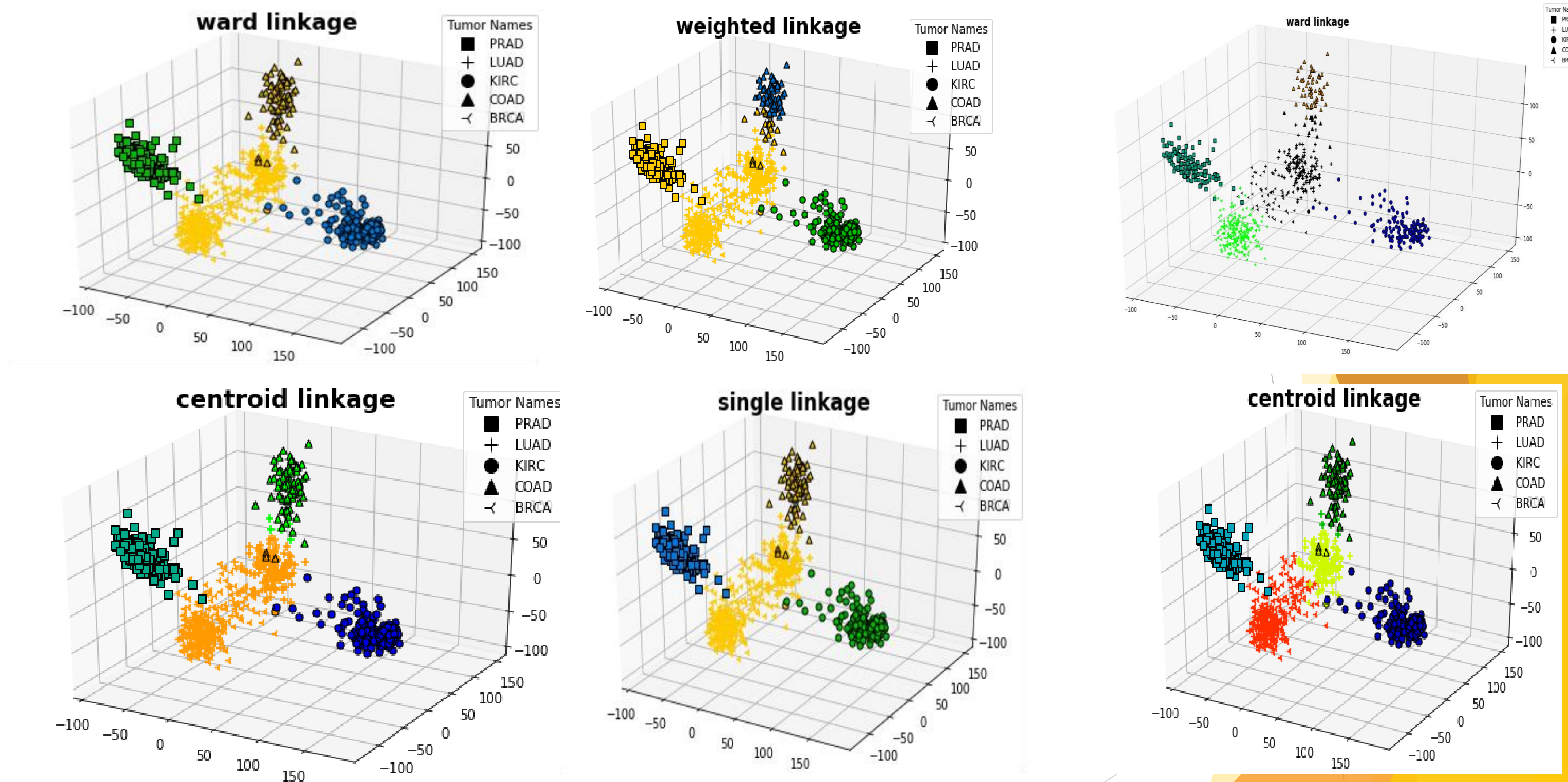
The full covariance matrix attained best performance at 8 components for both silhouette score and BIC. As expected full covariance performs best here, as the components may adopt any shape. In the tied form they have the same shape but this shape can be anything. Since the data is a bit noisy the tied covariance is not performing as well as the full matrix. The spherical covariance has circular (or spherical in higher dimensions) contours.

The RNA-Seq Dataset

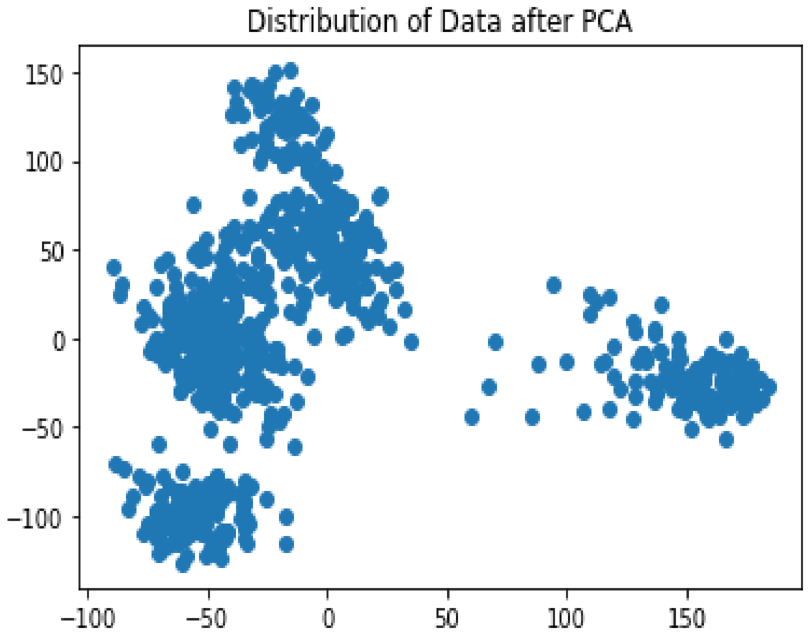


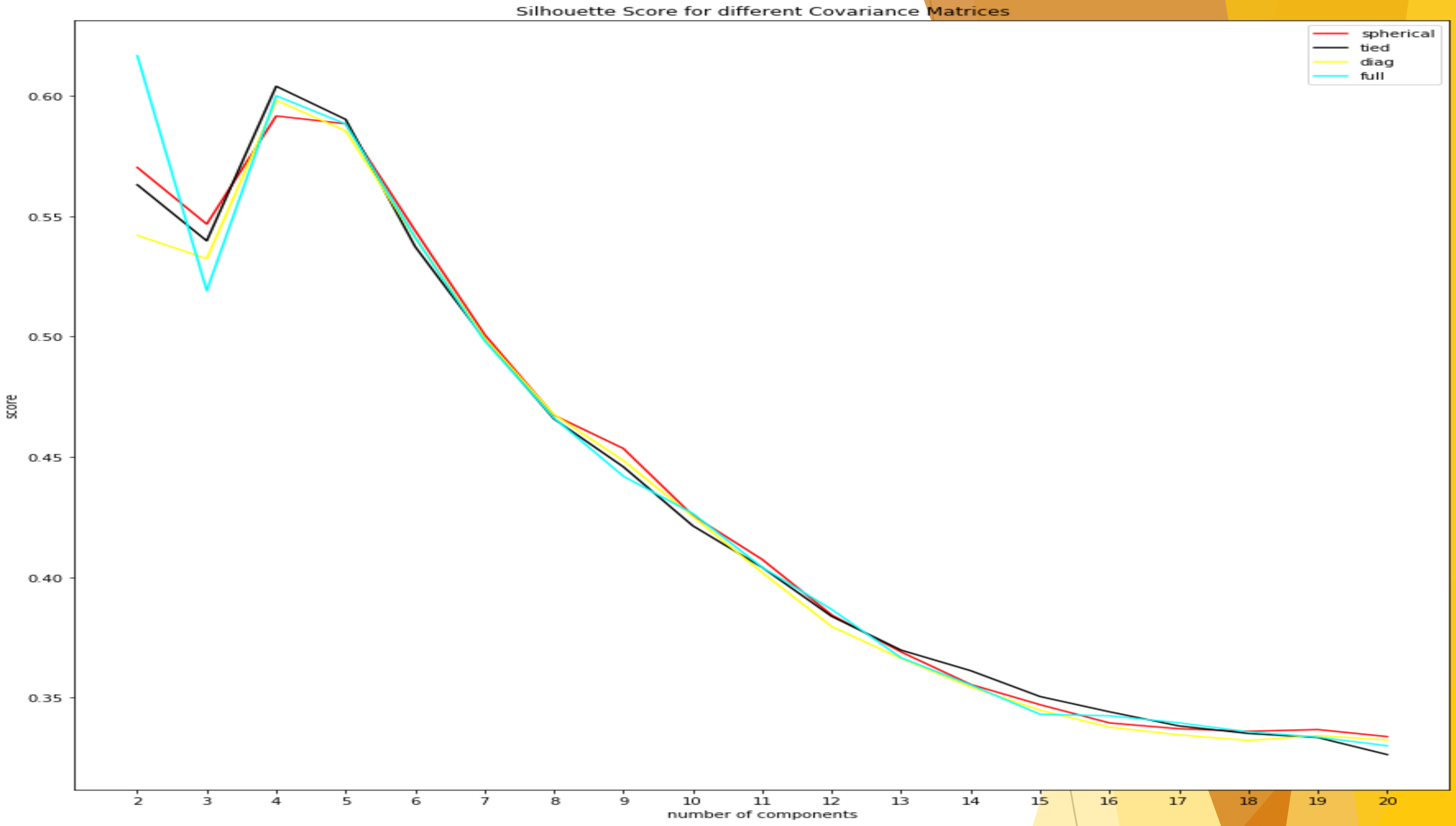
The 4 upper dendrograms were implemented for PCA with 3 components and the centroid linkage performs best here and this is expected, the clusters are compact so the ‘cloud’ form with more concentrated center than rims is not very apparent and this makes ward linkage less useful. The 4 lower dendrograms correspond to PCA with 4 components. Here the centroid linkage find 5 clusters and it has more consistent results up to 10 principal components than the other methods. (more than 10 were not checked).

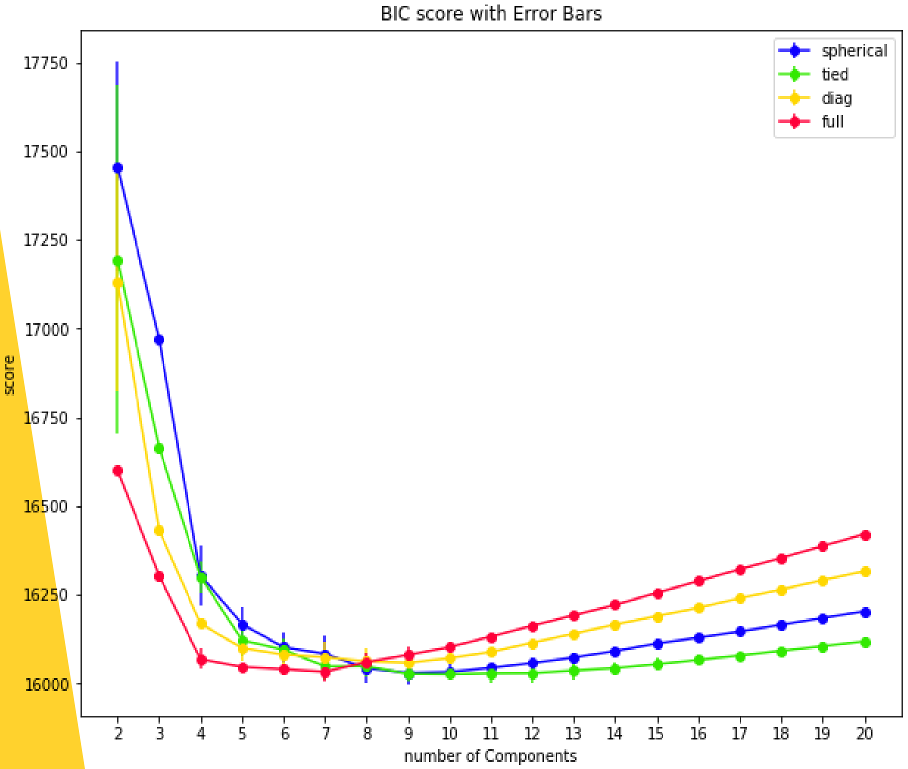




The 4 figures to the left use the number of clusters as given by the dendrograms, the different shapes as mentioned in the legend are various tumor types as given by the labels in the actual dataset. Interestingly when we force the algorithm to find 5 clusters, the centroid linkage performs very well and outperforms ward linkage as expected. Since the clusters are mostly uniform in density the ward linkage performs worse and the centroid measure is best here as the clusters form distinct uniform sphere-like shapes. The weighted linkage is not able to capture the significant cluster to the upper left. As this linkage method combines two close enough clusters into a supergroup it cannot distinguish them as seen here. It might be that this configuration maximizes the distance between clusters

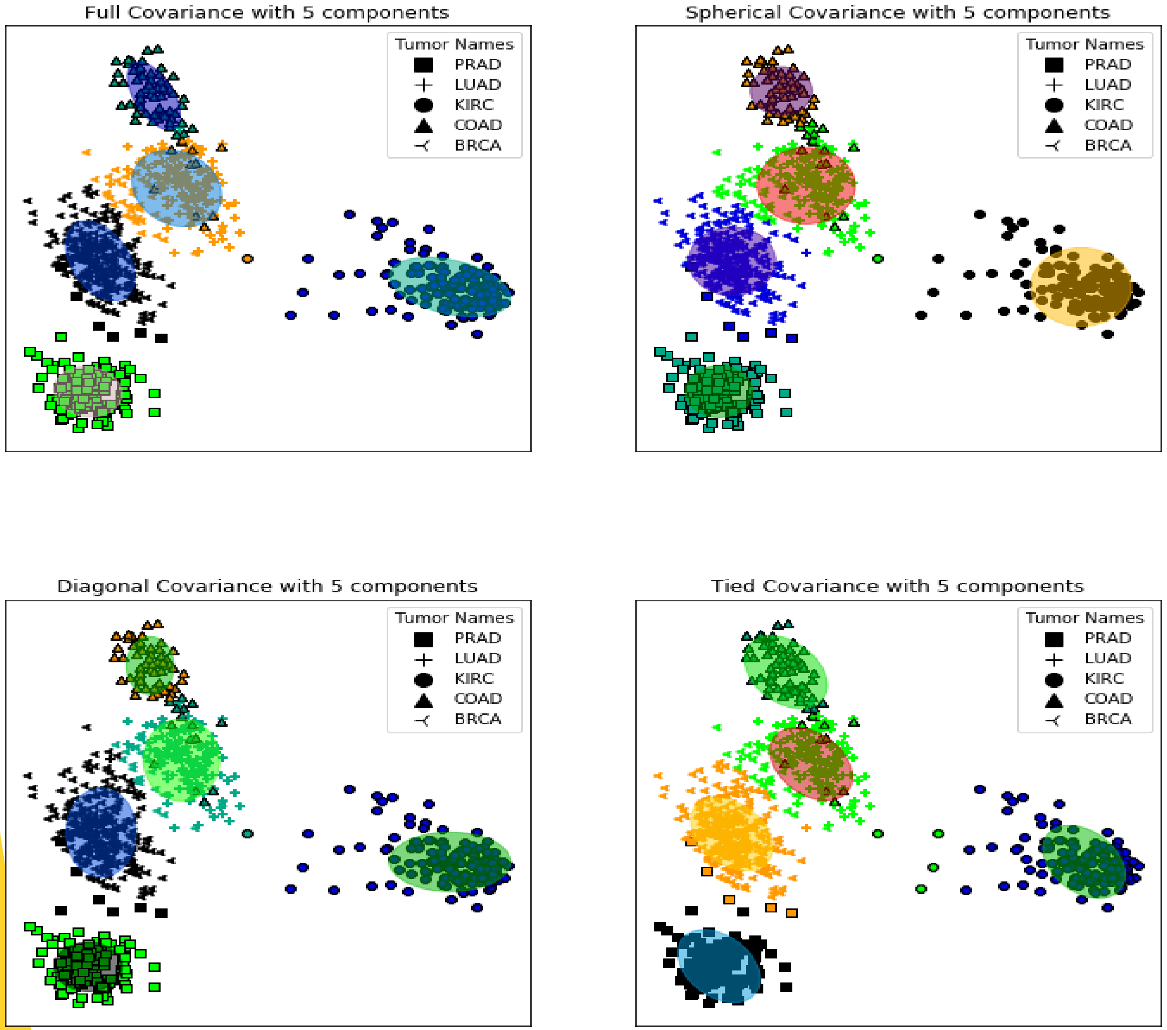




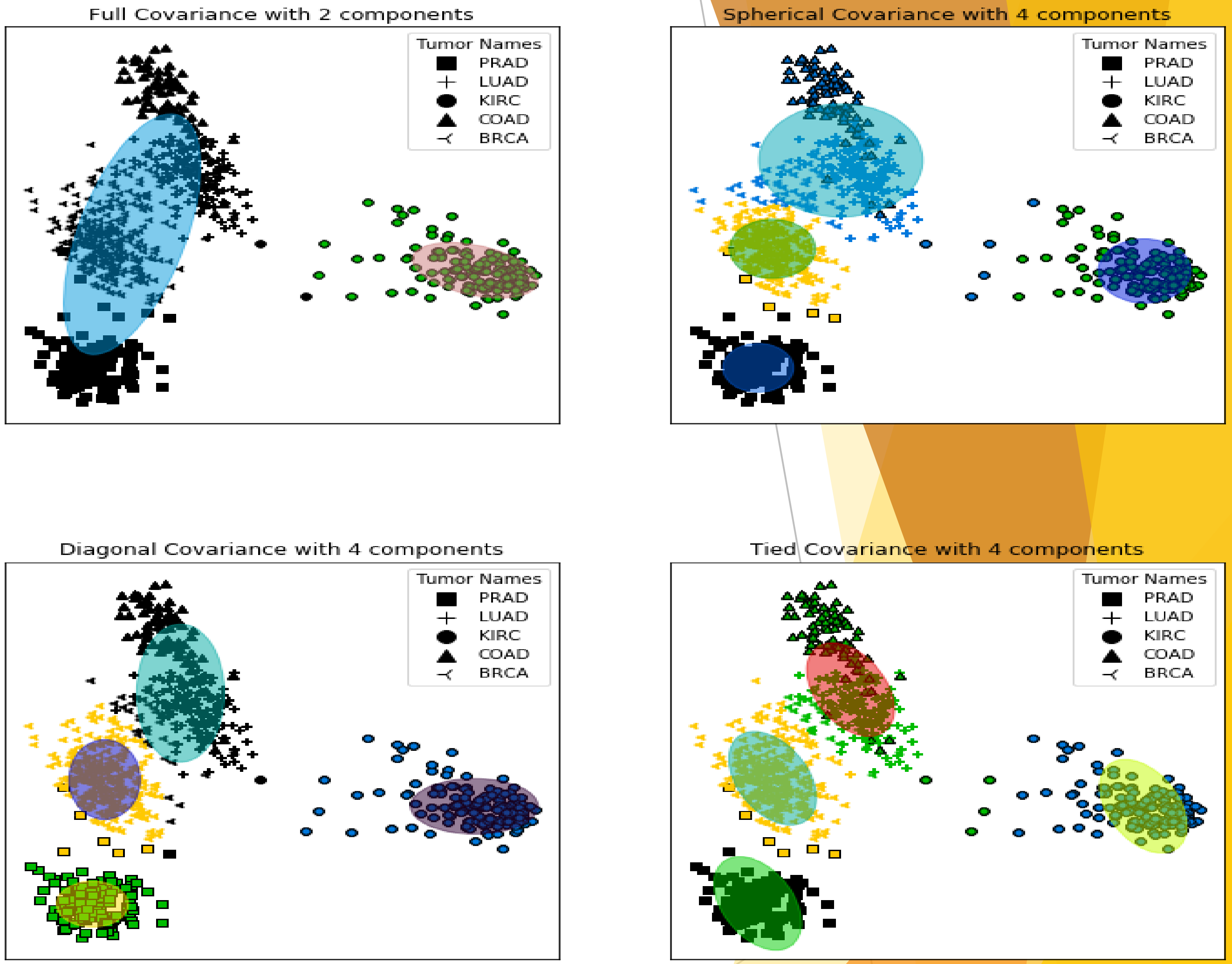


For the GMM, the dimensionality of the data was reduced to two for better visualization. Here the silhouette and BIC scores are significantly more consistent than the moons dataset. As this case pertains to a real data it is reasonable that it can be better explained by Gaussian distributions than a simulated data. Interestingly, in the silhouette score case full covariance captures less clusters than the other methods which guess closer to reality. However, the bic score shows better correspondence to reality and using the elbow rule we can safely assume 5 clusters based on the plot.

Best number of Components for Different Covariance Matrices Based on BIC Score



Best number of Components for Different Covariance Matrices Based on Silhouette Score



A comparison of best cluster count based on two different criteria. When we assign 5 components the tied covariance has best performance after full covariance. Although in this particular case as also evident from the silhouette score and BIC score matrices all covariance matrices perform close to each other. This might be due to somewhat similar density and shapes of the clusters.

Final Remarks

- GMMs perform better on real datasets than simulated non-Gaussian datasets.
- For real datasets single linkage is not a good method most of the time as we rarely observe such curvilinear or ‘chained’ data. However, it can capture certain shapes best as illustrated by the moons dataset.
- Based on cluster shapes, density and distanceeither centroid, ward or average (WPGMA, UPGMA) linkage types can have better performance compared to the others for most real datasets.
- BIC score gave better results for choosing the number of components compared to silhouette score.