Title: Case Study 5, Model Selection:

Ronit Ganguly: 2487190G:

Introduction

Clustering is a way of identifying the unlabeled data in groups. It is based on the understanding that points in a cluster are similar to each other and they are different from the other points in the other clusters [1]. For example, in our dataset of digits (total 1797 in number), each and every digit is 32X32 pixels hand-written picture of 0 to 9 digits. When transformed to 2 dimensional PCA coordinates, we get many scattered data-points which can be grouped and analysed.

This case study revolves around finding clusters in our digits dataset taken from SKlearn's repository [2].

Objective:

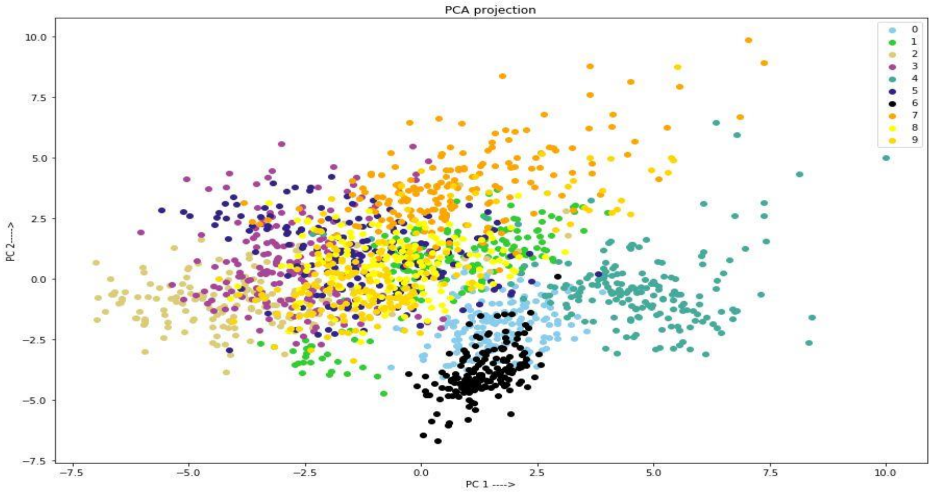
1. To use Gaussian Mixture model as clustering algorithm and find out which digit is getting grouped or split between clusters.

2. To explore several model selection methods:
    - AIC (Akaike information criterion)
    - BIC (Bayesian information criterion)
    - Silhouette Score
    - Cross-validation

   And analyse what each model is saying about the optimum number of clusters in our data.

Methods:

1) **PCA (Principal Component Analysis)**: Our digit dataset is in higher dimension (1797, 64). So, in order to visualize the clusters, we are taking its first two principal components and projecting on them. A visualization of the original data:

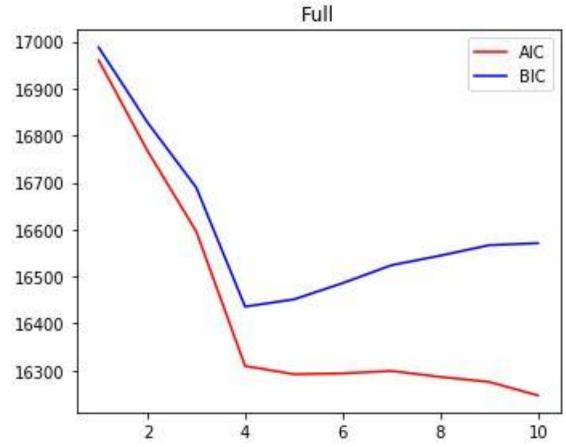


We can see that the digits are all over the space and therefore, a clustering algorithm is needed to identify similar groups of data.

2) **AIC (Akaike information criterion)** : This is used to select a model which is giving the lowest value of this formula: 2K - ln (L). Where K is the number of parameters and L is the maximum of likelihood.

3) **BIC (Bayesian information criterion)** : This is quite similar to AIC formula with only difference that it puts more penalty with increasing dataset. The penalty term is ln(n) where n is the number of data points in the space. Thus the formula is : K ln(n) - ln(L). Again the lowest BIC indicates that the model is the optimum one.

A visualization of AIC and BIC done on a Gaussian Mixture Model :



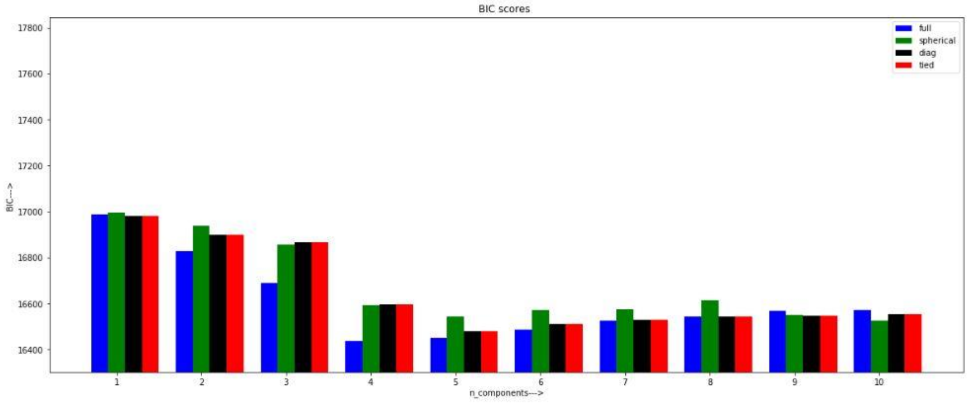
We can consider Elbow method and depict that cluster number 4 is the optimum choice of clusters for our dataset. However, AIC also indicates that cluster 10 has the lowest score and we know that we do have 10 groups of data in our dataset. So why not go with 10 clusters with confidence? Results of analysis is in the results and discussion sections.

4) **Cross-Validation and maximum likelihood on held out set**: Stratified KFold [3] has been used so that our testing dataset gets exposure to all groups in each fold. The maximum likelihood is given by gmm_clf.score(X_test) function [4].

Results:

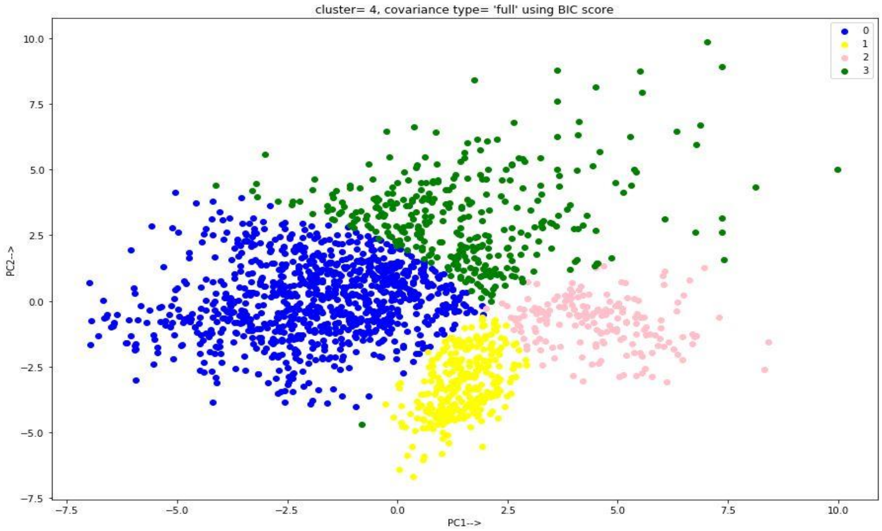
1) BIC:

The BIC value was taken for every cluster K and for every covariance matrix type (Full, Spherical, Diagonal and Tied). Below is the bar plot:

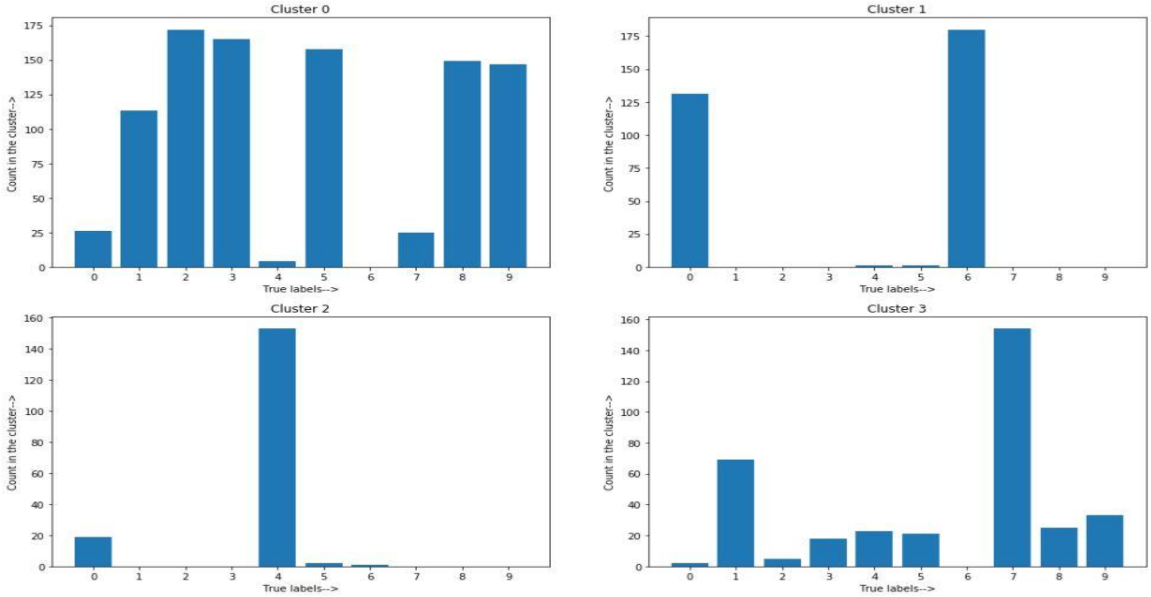


We can see that cluster K=4 i.e. 4 clusters and covariance matrix type: Full, when chosen, give the lowest BIC value.

Outcome of choosing cluster K=4 and covariance matrix type: Full

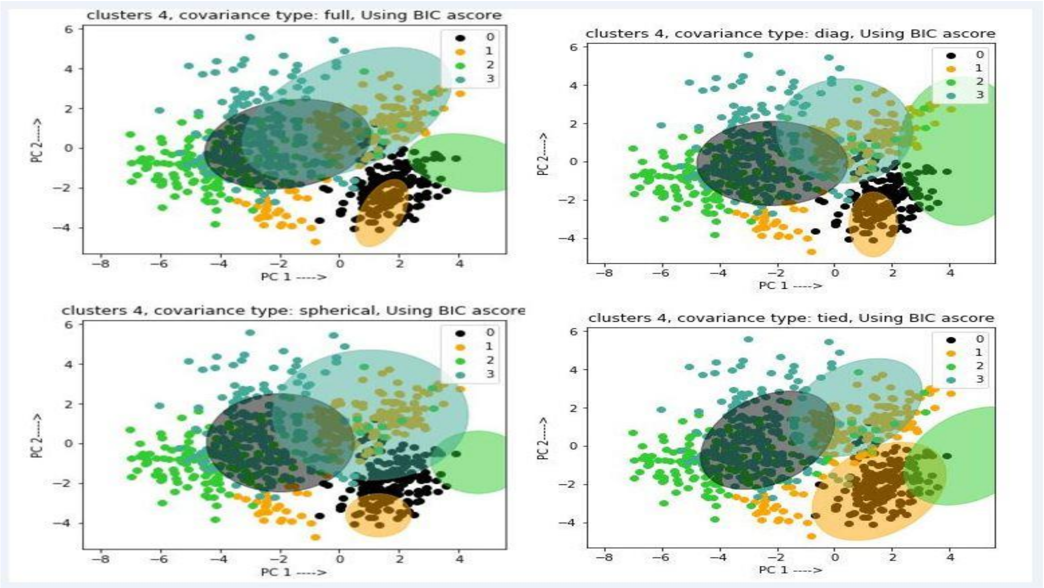


Checking which true label got split and grouped due to BIC model:



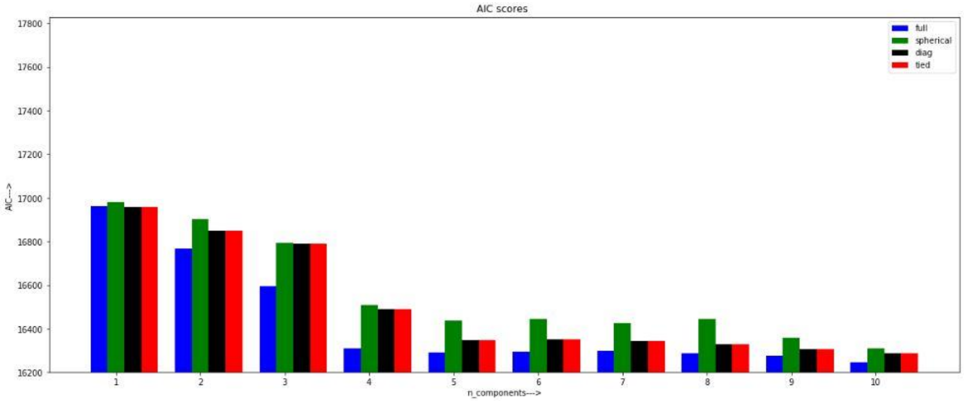
For example: cluster 1 is highly confident about containing approximately all the 6 digits while cluster 2 says it has a majority of 4s.

Covariance matrix wise visualization of 4 clusters:



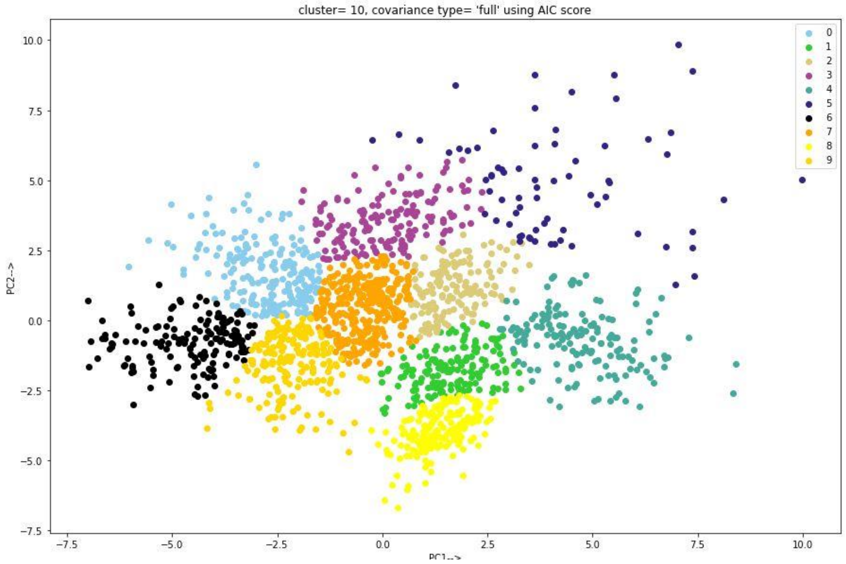
2) **AIC**:

The AIC value was taken for every cluster K and for every covariance matrix type (Full, Spherical, Diagonal and Tied). Below is the bar plot:

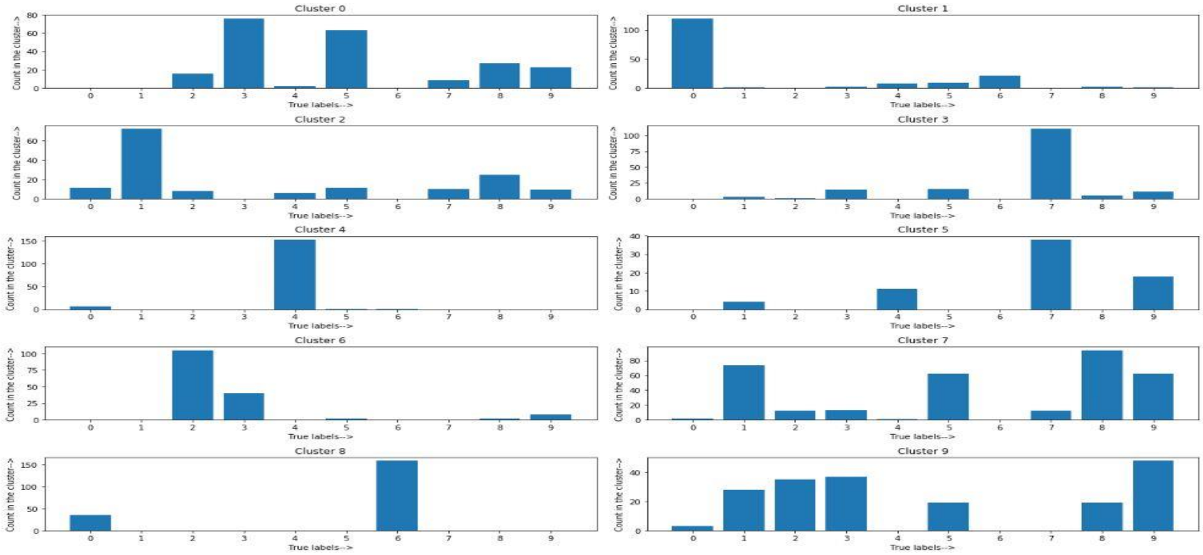


We can see that cluster K=10 i.e. 10 clusters and covariance matrix type: Full, when chosen, give the lowest AIC value.

Outcome of choosing cluster K=10 and covariance matrix type: Full:

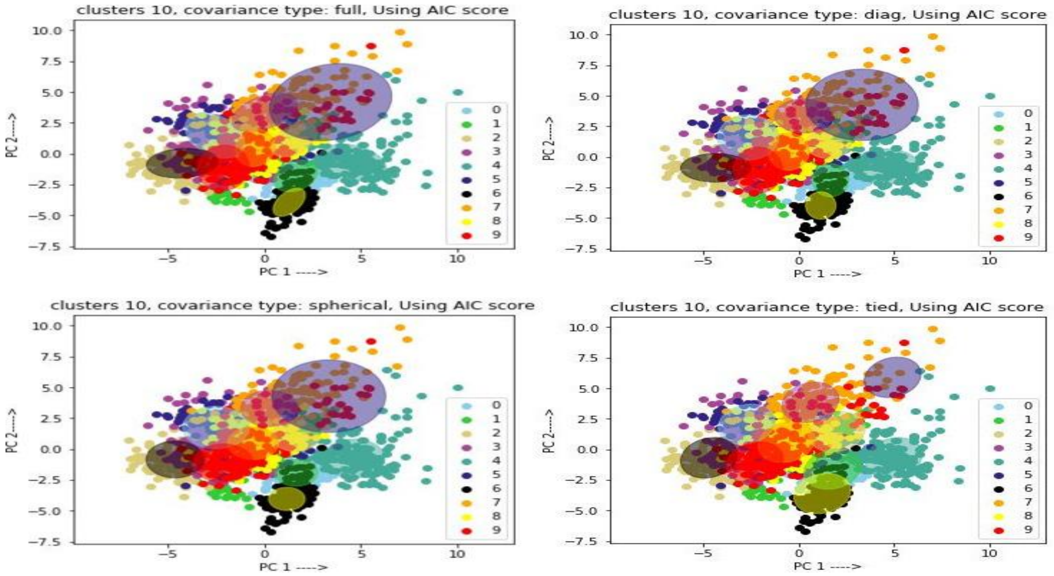


Checking which true label got split and grouped due to AIC model:



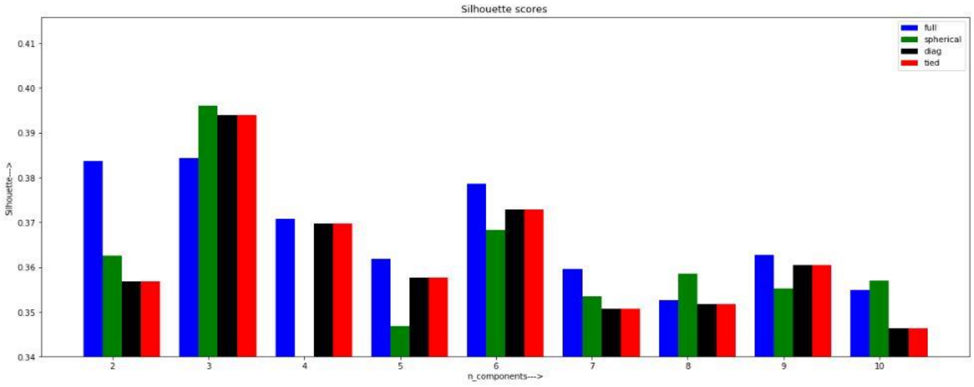
We can see that cluster 1 is pretty confident on 0s and cluster 4 is confident on 4s while cluster 8 is confident on 6s, similarly cluster 3 is confident on 7s etc.

Covariance matrix-wise visualization of 10 clusters:



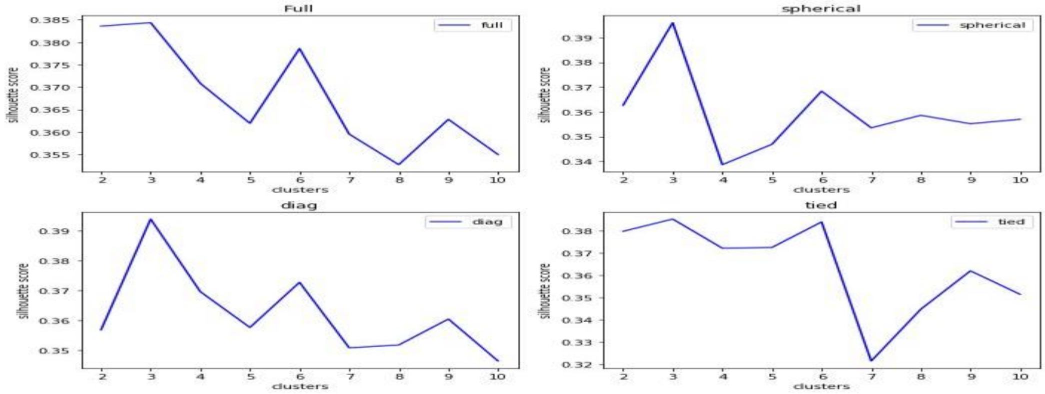
3) **Silhouette Score** : [5]

The Silhouette value was taken for every cluster K and for every covariance matrix type (Full, Spherical, Diagonal and Tied). Below is the bar plot:

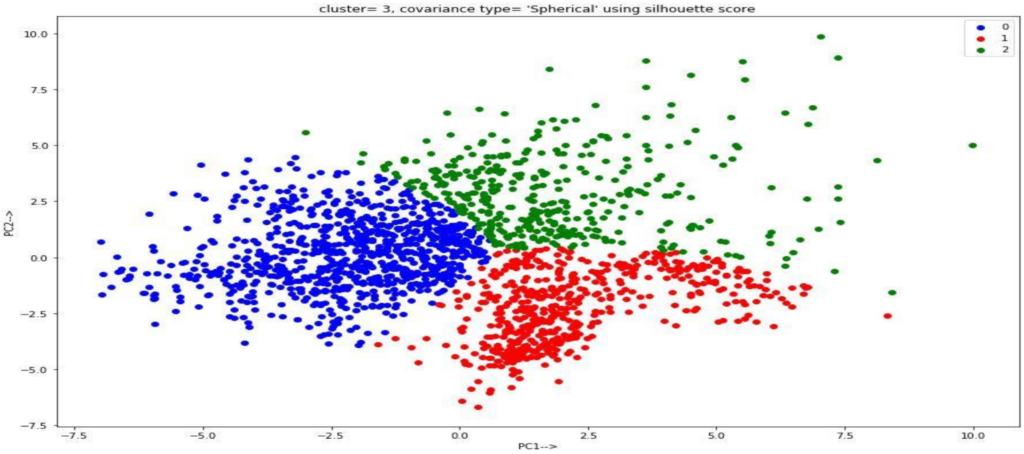


We can see that cluster K=3 i.e. 3 clusters and covariance matrix type: Spherical gives the highest Silhouette score.

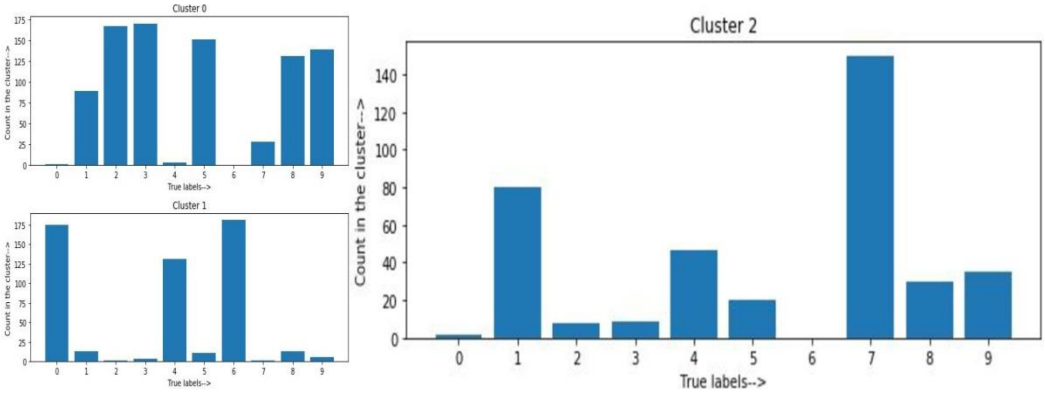
A covariance matrix-wise comparison:



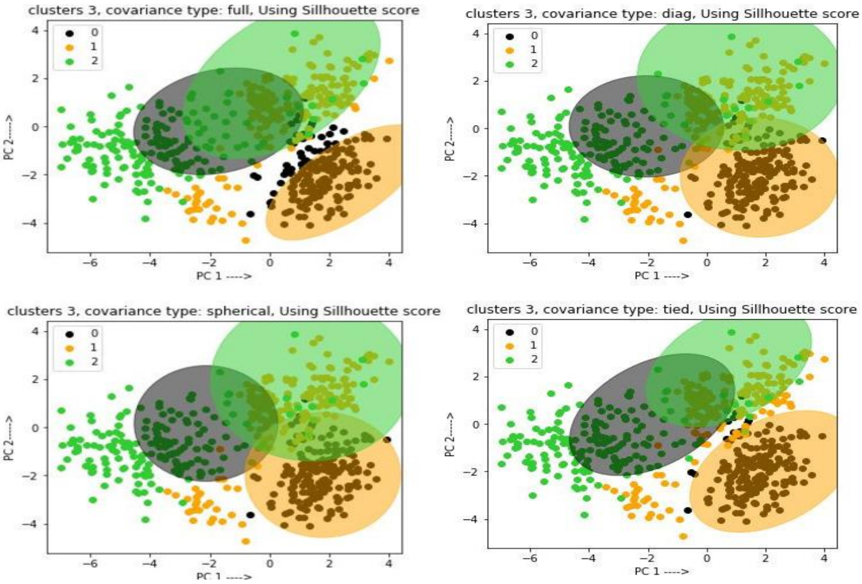
Outcome of choosing cluster K=3 and covariance matrix type: Spherical:



Checking which true label got split and grouped due to Silhouette model:

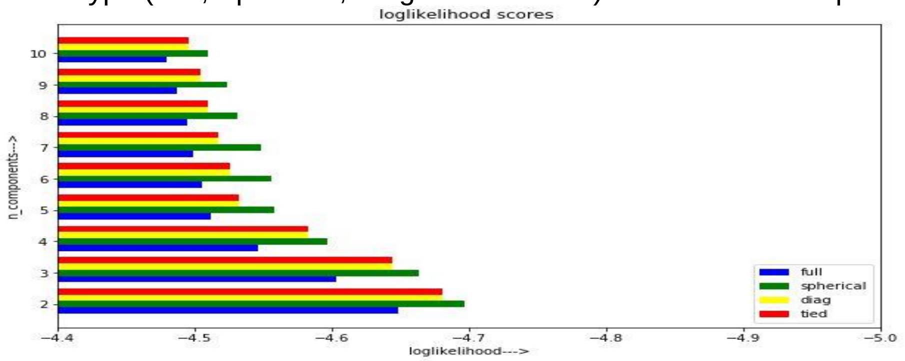


Covariance matrix wise visualization of 3 clusters:



4) **Cross-Validation and maximum likelihood on held out set**:

The maximum likelihood on test set value was taken for every cluster K and for every covariance matrix type (Full, Spherical, Diagonal and Tied). Below is the bar plot:



We can see that cluster K=10 i.e. 10 clusters and covariance matrix type: Full, when chosen, give the maximum likelihood value.

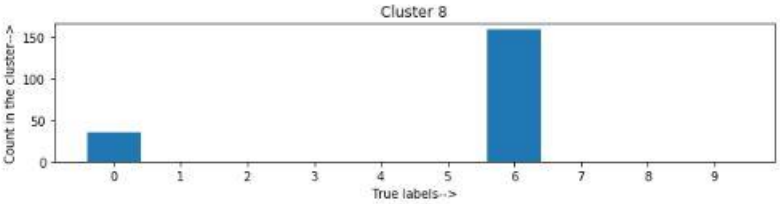
- **Outcome of choosing cluster K=10 and covariance matrix type: Full is same as in AIC.**
- **Checking which cluster contains which true labels is also same as in AIC.**
- **Covariance matrix-wise plot will be same as in AIC as well because both AIC and CV methods say that cluster 10 with covariance type Full is the optimum one.**

Discussion:

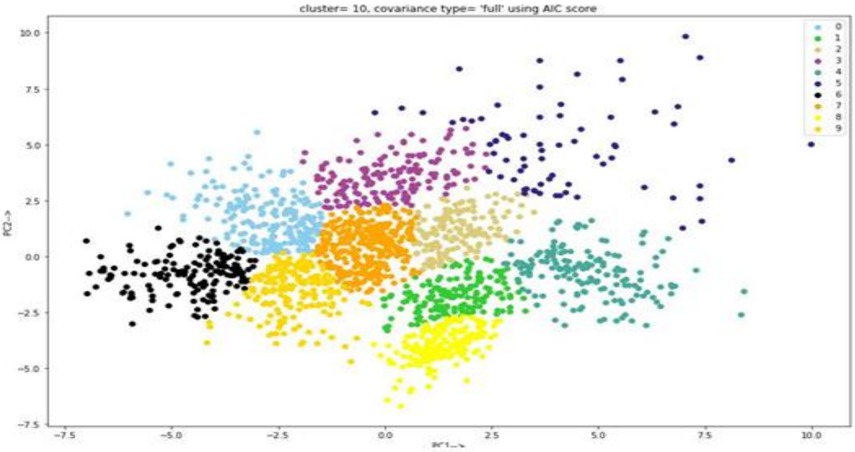
1. Observation from all the methods:
   - BIC optimum cluster: 4 and covariance type: Full
   - AIC optimum cluster: 10 and covariance type: Full
   - Silhouette score suggests: 3 and covariance type: Spherical
   - Cross-Validation suggests: 10 and covariance type: Full

Since there are 10 true labels, and majority of covariance type suggested is Full, I believe 10 clusters with Full covariance type should be a better choice for the digits dataset.

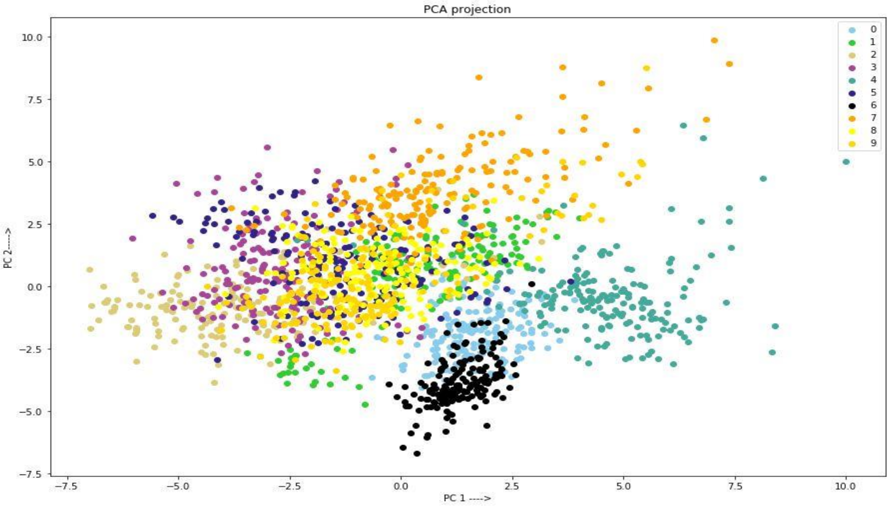
2. To further reinforce, checking how confident the clusters are about containing the correct label according to AIC and cross-validation i.e. clusters =10 and covariance ='Full' :



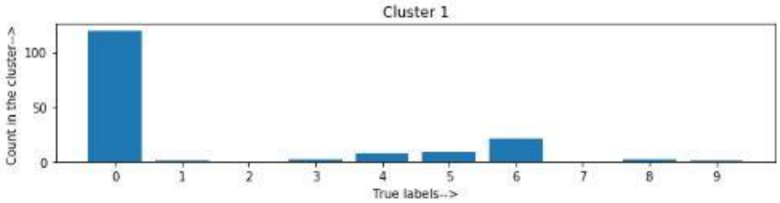
- Cluster=10 and covariance =Full suggest that cluster no. 8 is confident that almost all true label 6s are contained by it.
- Where is all 6 according to 10 clusters and Full covariance model?



- All 6s are in the bright Yellow region and if we compare with our PCA plot :



- All the 6s are spread between black and cyan region at the same location. Moreover, our cluster 10 model suggests that cluster number 1 also contains some 6s:



and geographically cluster 1 (which is the limegreen cluster) is the cyan cluster in the PCA projection. This proves that having 10 clusters help to differentiate between overlapping clusters because otherwise cyan and black regions would have been fused together, which is exactly what is happening with cluster 3 model and cluster 4 model as selected by BIC and Silhouette score.

- However, 4 cluster model (BIC model) is also accurately clustering the digit 4s into the pink region which is the same region on PCA projection for all 4s. This can be the scenario because this region is not overlapping.
- Conclusion, even though, the elbow is at 4(AIC line plot), I went for 10 clusters as per the bar plot because I know for sure that there are 10 distinct classes in my data. More clusters is desirable until it is lower or equal to maximum classes in the data (above the maximum classes will lead to overfitting). But, this holds true till we have the knowledge of the total distinct classes. Nevertheless, cluster 10 model can distinguish between overlapping digits.

References:

[1] Cluster analysis Available at: < https://en.wikipedia.org/wiki/Cluster_analysis>

[2] sklearn.datasets.load_digits Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html>

[3] sklearn.model_selection.StratifiedKFold Available at: < https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html >

[4] score(self, X, y=None) Available at: < https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html#sklearn.mixture.GaussianMixture.score

[5] sklearn.metrics.silhouette_score Available at : < https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html#sklearn.metrics.silhouette_score>