Title: Model selection in K-Means

Name: Xu Chentao

ID: 2438322X

Introduction

In this case study, we need to use a variety of methods for model selection on the basis of k-means clustering. K-means is one of the most commonly used methods in clustering, which is to calculate the best category attribution based on the similarity of distance between points. However, in the process of clustering, the excessive number of classes does not necessarily represent the number of real clusters of data obtained. Therefore, we need to obtain the real clustering number based on the data, which means the clustering number with the best effect for the sample data, because k-means algorithm is very sensitive to the initial value, the difference in the initial value will affect the clustering effect of the algorithm and the number of iterations.

The data in this case study comes from handwritten digital data set in Sklearn own digits, which contains 1,797 handwritten digits ranging from 0 to 9, each of them is in an 8 by 8 matrix. Since PCA has been used to reduce the dimensionality of these data in the experiment, I only need to cluster the data with the best effect after dimensionality reduction.

Methods

I will use BIC, AIC, silhouette score and cross validation to choose the best clustering numbers and covariance type in this case study.

AIC:

AIC (Akaike Information Criterion) is a standard to test the fitting effect of statistical model and a weighted function. The formula for AIC is as follows:

$$AIC=2k-2\ln(L)$$

k is the number of unknown parameters in the model, and L is the value likelihood function of the maximum likelihood function in the model. In model selection, we usually choose the model with the lowest AIC score.

BIC:

The principle of BIC (Bayesian Information Criterion) is similar to AIC and also applies to a method of model selection. In the process of training the model, the increase in the number of parameters will lead to the increase in the complexity of the model, which will increase the likelihood function and result in overfitting. Therefore, BIC and AIC introduced penalty terms related to model parameters for the phenomenon of overfitting. But BIC has a bigger penalty, the formula for BIC is as follows:

$$BIC=k\ln(n)-2\ln(L)$$

Silhouette score:

Silhouette score is a measure of a node and its subordinate degree of clustering compared to other similar. In this case, I use ‘sklearn.metrics.silhouette_score’ method, the method calculating the average of all the samples, its values range between 1 to 1, the greater value indicates that the cluster nodes is more matching the current, but if the score approaching ‘-1’ means that the sample should be classified into another cluster.

Cross validation:

Cross validation can make data partitioning more sporadically, reduce the accident caused by random division, improve the generalization ability of the model. In this case, the cross validation is mainly in

order to determine the k value which makes the square error of distance between center and point to decrease, so as to select one of the biggest cross-validation score.

My methods:

In the methods of BIC and AIC, I first trained the gaussian mixture model of four different types of covariance matrix including full, tied, diag and spherical, and used these models to fit the data after dimension reduction, then calculated the score of the model with AIC and BIC algorithms, and extracted the covariance matrix type with the lowest score and the number of clusters. Finally, I used the extracted covariance matrix type to create a specific gaussian mixture model, and then visualized the latest clustering to check the rationality.

In the method of silhouette score, the whole process is in a loop. I first instantiate the k-means classifier and then fit the dimensionally-reduced data. After fitting the model, I drew the results of each classification, printed the average contour coefficient, and finally calculated the number of points with positive contour coefficient.

In the method of cross validation, the method of comparison is similar with AIC and BIC, I compare the likelihood of four different covariance types and choose the largest likelihood.

Results

- **Comparison of AIC, BIC scores in four covariance types:**

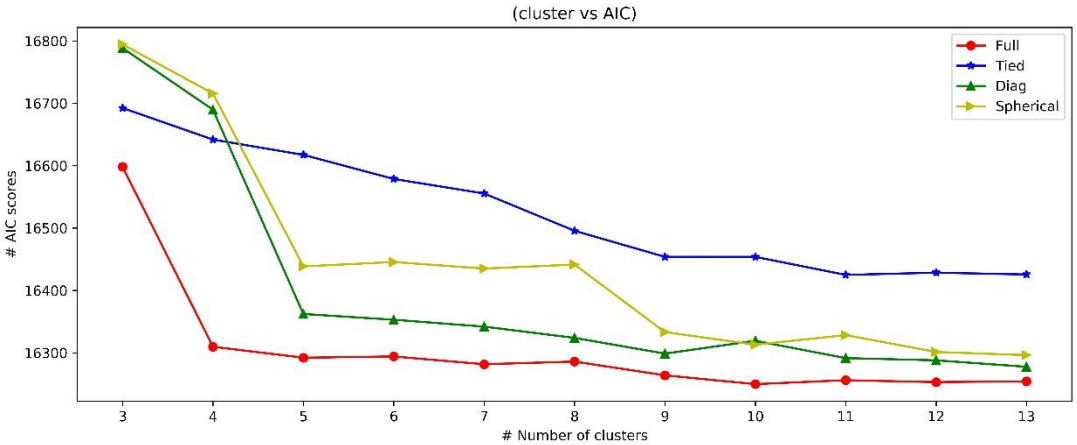


Fig 1: The scores from AIC

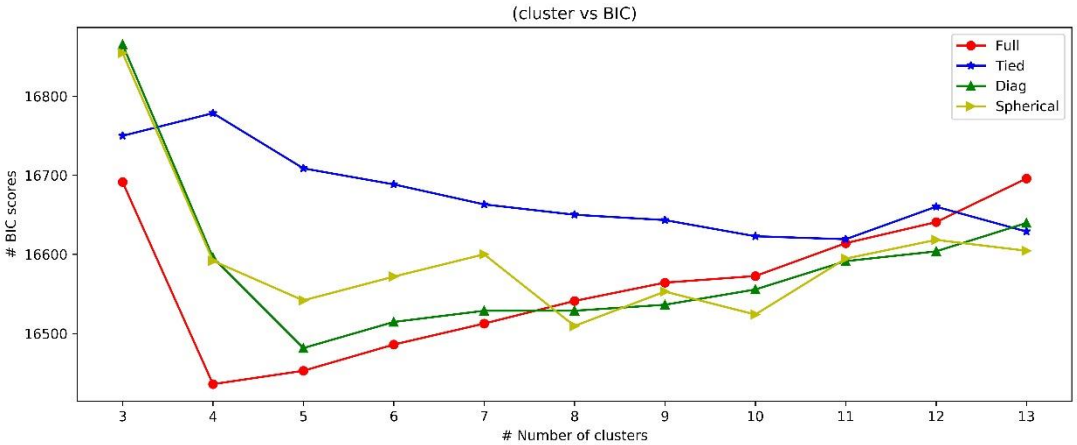


Fig 2: The scores from BIC

- **Silhouette analysis for K-means clustering:**
  The figure below shows the average silhouette score in four different covariance types of Gaussian mixture model.

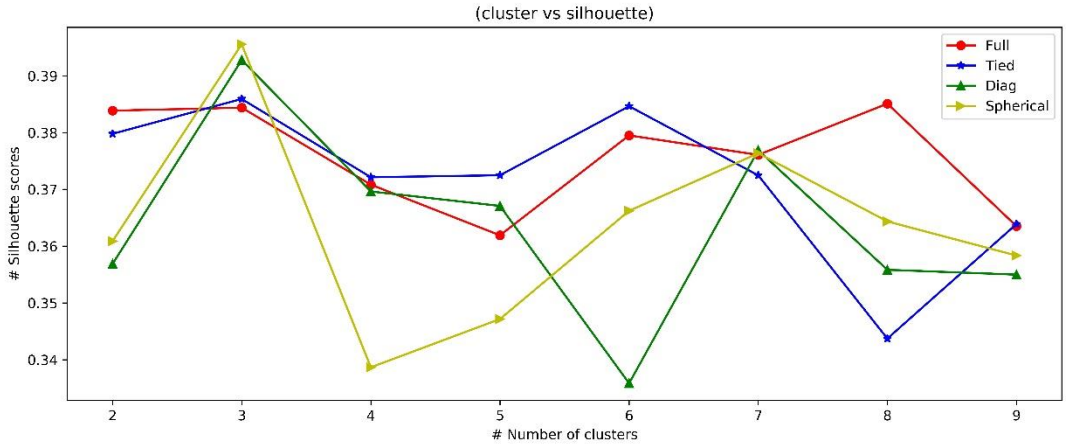
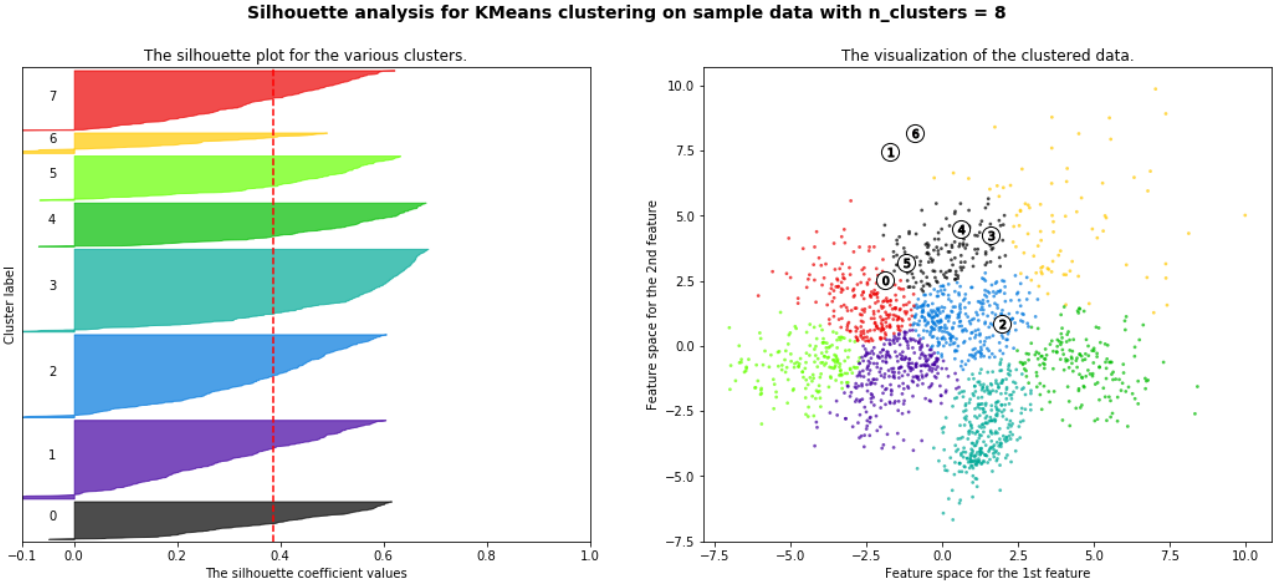
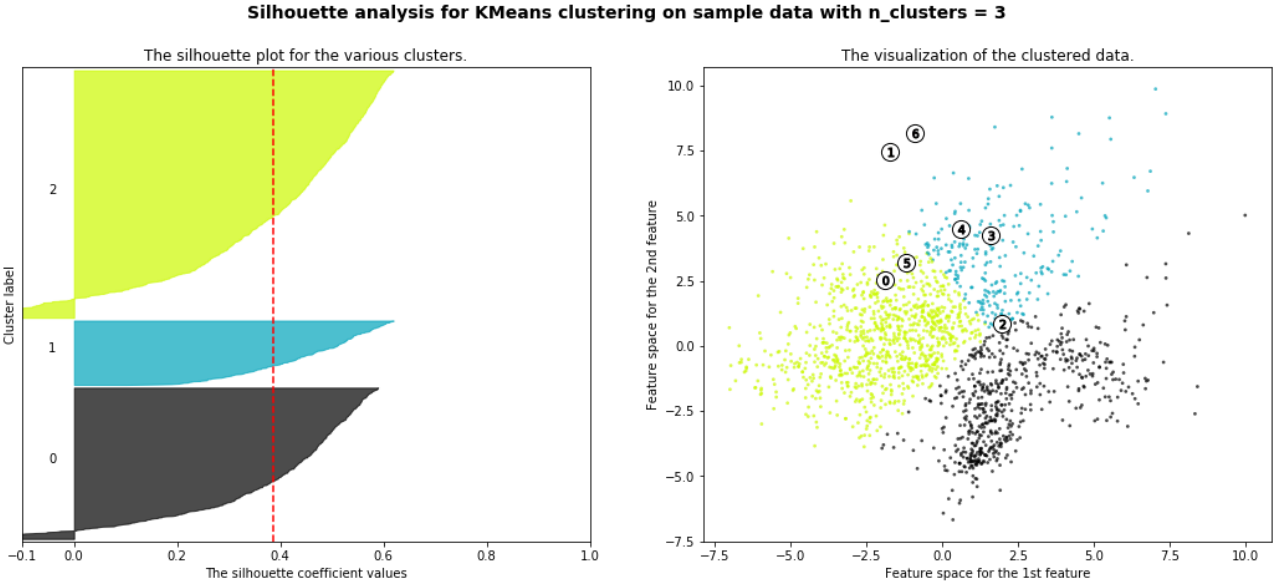


Fig 3: The comparison of silhouette scores in four covariance types

The four figures below show the highest average silhouette score in four types of covariance ('full' 'tied', 'diag', 'spherical') of Gaussian mixture model respectively.





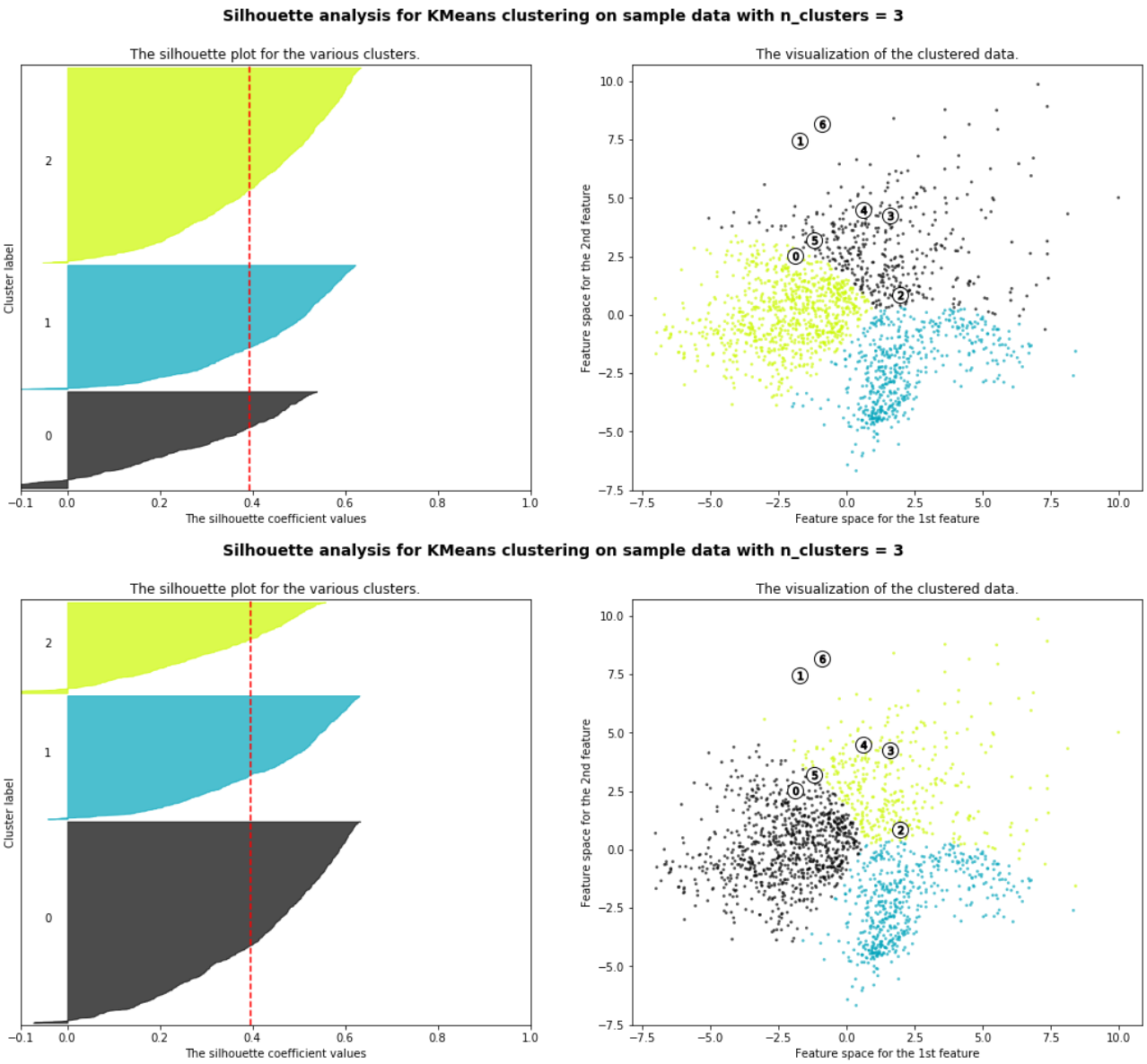


Fig 4: The best silhouette scores in four covariance types

- **Comparison of cross validation scores in four covariance types:**

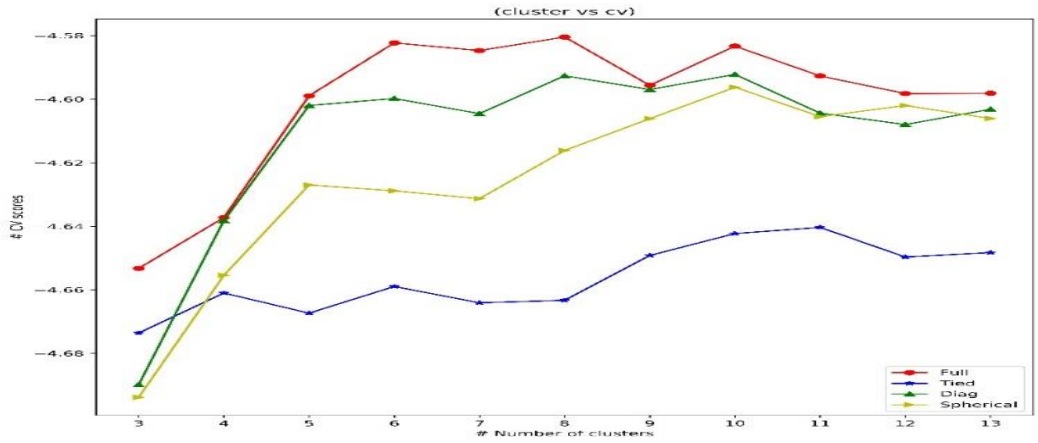


Fig 5: The scores from CV

- **Use the best cluster numbers and covariance type after the methods of BIC and AIC to create the GMM model:**

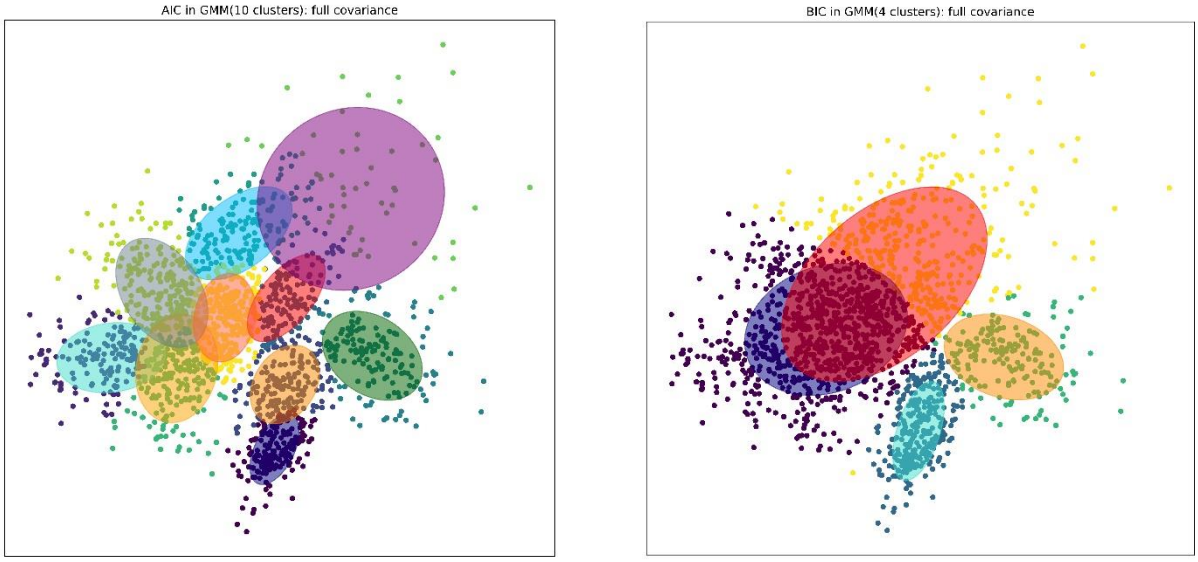


Fig 6: 10 clusters in AIC and 5 clusters in BIC (Full covariance)

- From the method of silhouette score, the situation with the highest score is when the cluster number is 3, so I set four GMM models, the cluster number is 8 in the 'n_components' parameters, and the covariance matrix type is spherical, respectively. Additionally, I used the model to fit the data after dimension reduction, and predicted the labels value with the model after training.

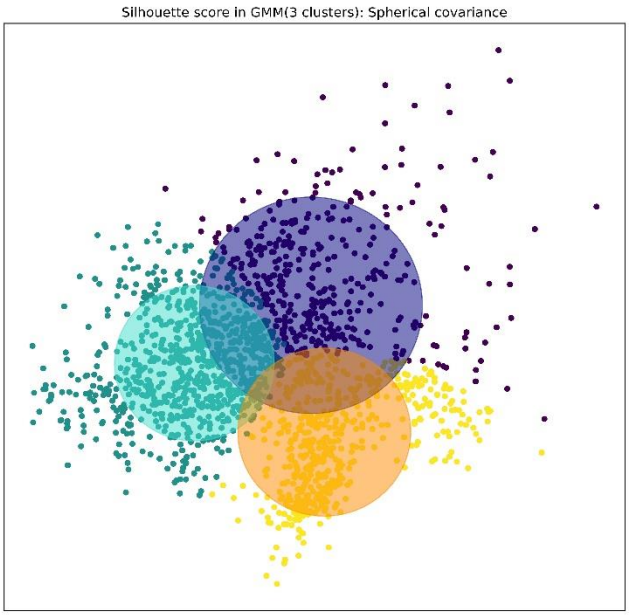


Fig 7:3 clusters from silhouette score (spherical covariance)

- **The comparison of lowest likelihood between these 4 methods:**
  The following figure shows the likelihood value comparison of the four model selection methods. The function of likelihood value here is to compare the fitting degree of different models obtained by different methods.

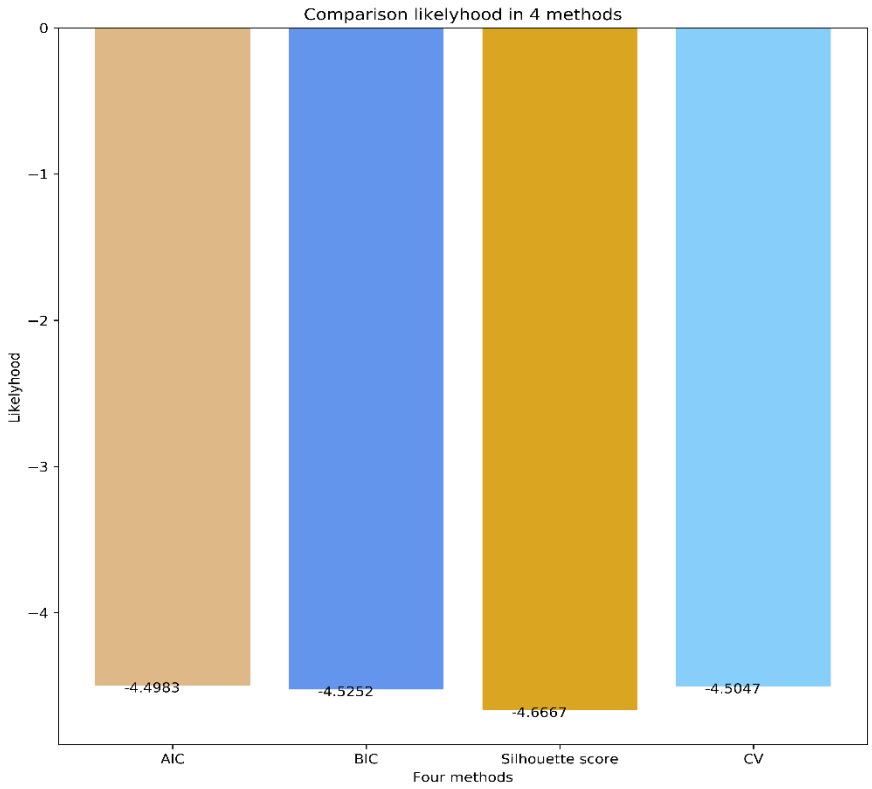


Fig 8: Comparison of lowest likelihood

- **Several covariance types of gaussian mixture clustering cross - validation models**

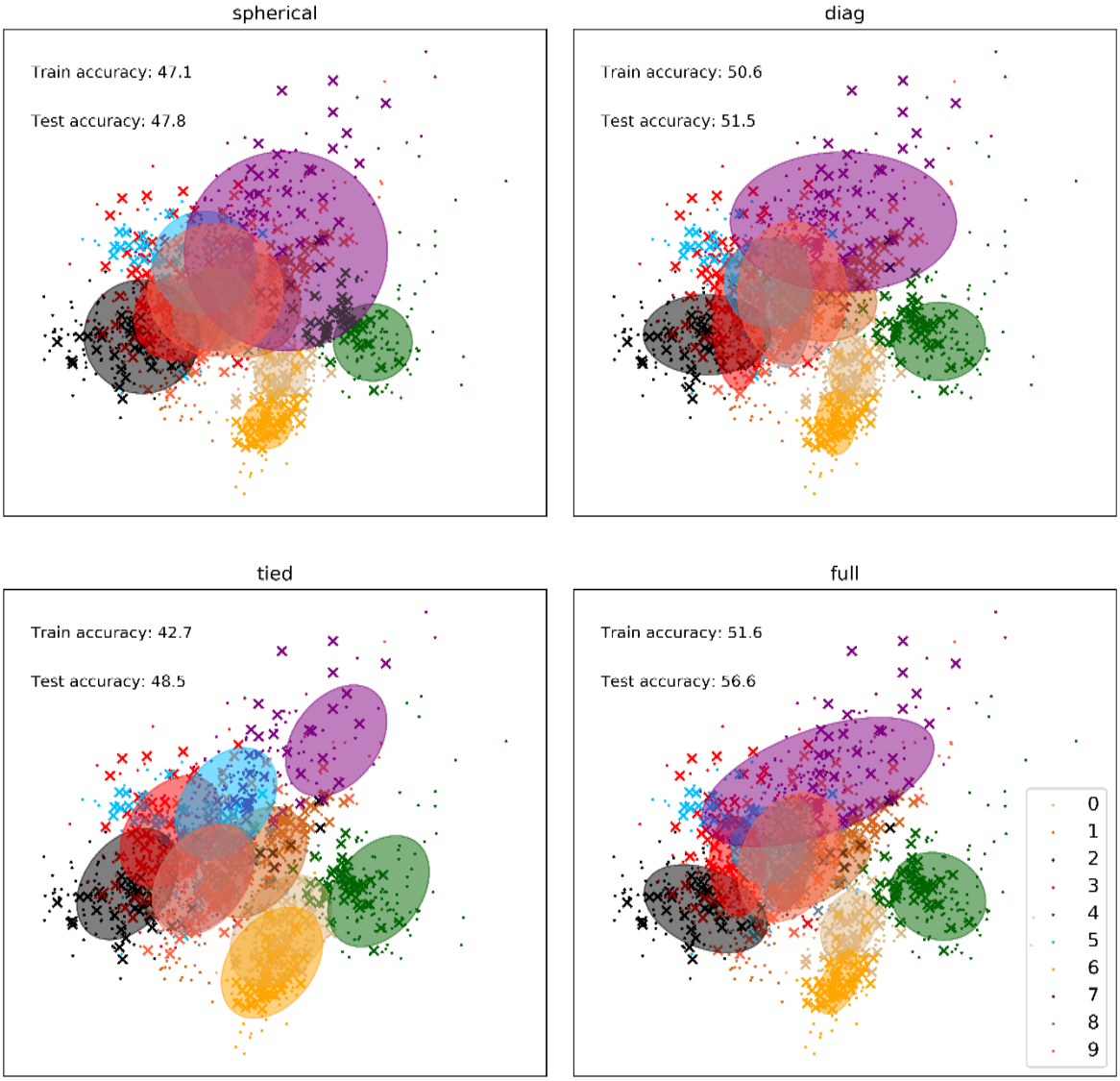


Fig 9: Several covariance types of gaussian mixture clustering cross - validation models

Discussion

According from the figure 1 to figure 5, I used four different methods to select the most appropriate K value and the most appropriate covariance matrix type in the k-means cluster. In AIC and BIC, I chose the situation with the lowest score because of the penalty. In the silhouette score method, I choose the corresponding number of clusters with the maximum coefficient. In the cross-validation method, the likelihood value of the gaussian model is obtained by this method, so I choose the clustering value and covariance matrix type corresponding to the maximum value.

In figure 6, I redefined the gaussian mixture model with the optimal cluster number and covariance matrix types selected from AIC and BIC methods, and visualized the optimal clustering. In figure 6, since the optimal clustering number can be found from the contour coefficient method to be 3, I visualized the clustering conditions obtained from the training of the gaussian mixture model with a clustering number of 3 and spherical covariance matrix type.

In figure 8, I compare the likelihood scores of the k and covariance matrix type defined gaussian mixture model, which is the most suitable of the four methods, with the reduced dimension data. It can be seen from the four values that AIC has the best effect.

Figure 9 shows the four covariance types of cross-validation models in gaussian mixture model clustering. I used the 5-fold cross-validation method and set the number of clusters to 10. In addition, I split the data into training data and test data and expressed them in two symbols. From accuracy, it can be seen that the 'full' type has the best performance among the four covariance types. In this figure, the intersecting clustering indicates that the labels are mixed and the clustering effect is poor, while the separated clustering indicates that the clustering effect is good.