Coursework 2: Case Study – Model Selection

Name: YIZE LI

ID: 2447939L

Introduction

- This case study has two parts. The first part mainly focus on some basic operations like choosing K value using elbow method in a K-means model. And define multiple measurements for scoring different K values like silhouette score and BIC scores. The second part is manly focus on using gaussian mixtures model trained by mnist dataset. By exploring multiple methods (BIC, AIC, Silhouette, and cross-validation) to determine the number of clusters required for a gaussian mixture model, and get a optimal number of clusters for the mnist dataset(which best number is 10).

  Mnist dataset contains 10 groups of handwritten numbers from 0 to 9. The aim is to divide the data into train, test and validation and apply them in the Gaussian Mixture model. Before training the model, it is essential to decide the number of clusters for the model. The ideal number is ten, but through different methods, it can be shown on plots what is the optimal number decided by different measurements.

Methods

- Part 1: synthetic data

First, using KMeans method to calculate the total sum of squared errors at different number of K and then draw a plot. As the Elbow method decided, the "elbow" on the arm is the value of k that is the best, in this case, like the plot shown below, is 4.

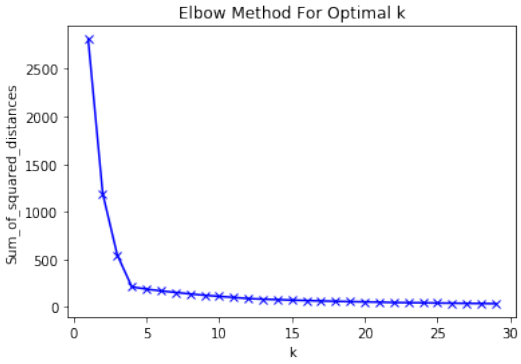
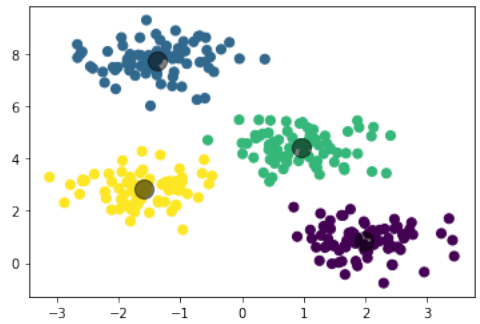


Figure 1.1 Elbow method plot

Then using the K I just chose to train the KMeans model and plot all the points in different color. As the picture shown below, 4 is a perfect number for dividing those numbers into groups. As a result, this method is quite effective for this dataset.



Fugure 1.2 Scatter plot for grouped data

Here is another measurement called silhouette score for K for KMeans method.

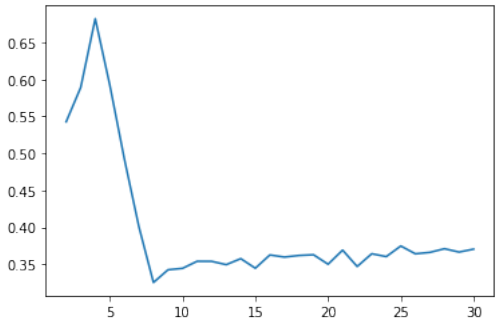


Figure 1.3 Silhouette score for different K

The plot also shows that the best number for K is 4.

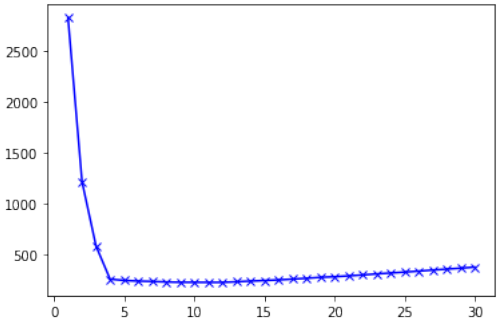


Figure 1.4 BIC values for different K

The plot shows that the smallest number is 11, which is not agree with the previous result.

- Part 2: Mixture models

After loading and projecting the data into 2D, I can start using different model selection strategies to measure what is the best cluster number for gaussian mixture model. The covariance_type parameter in mixture.GaussianMixture function is important. By changing this value and plot different scores, I can decide which covariance matrix structure should I choose.

Results

When choosing “full” covariance matrix structure, the BIC, AIC, Silhouette and cross-validation scores are listed below:

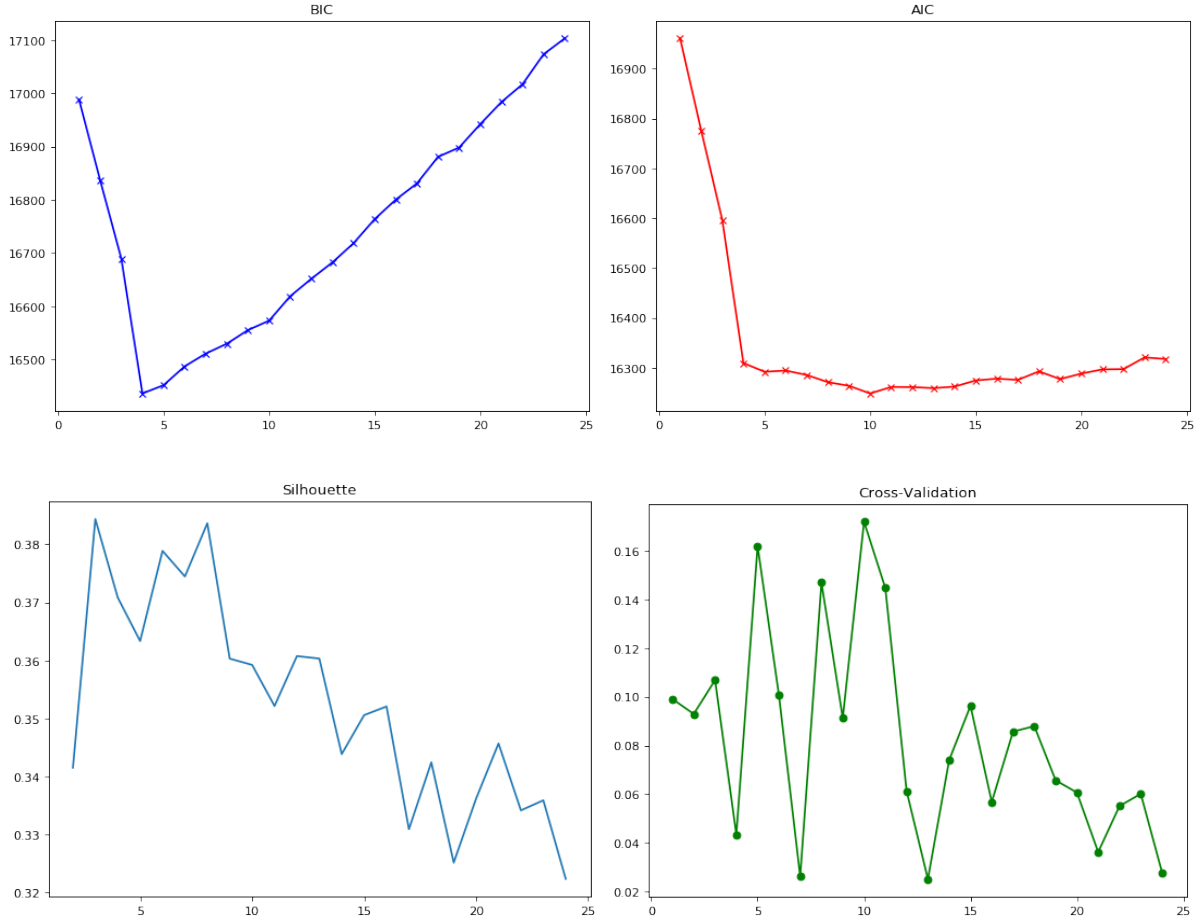
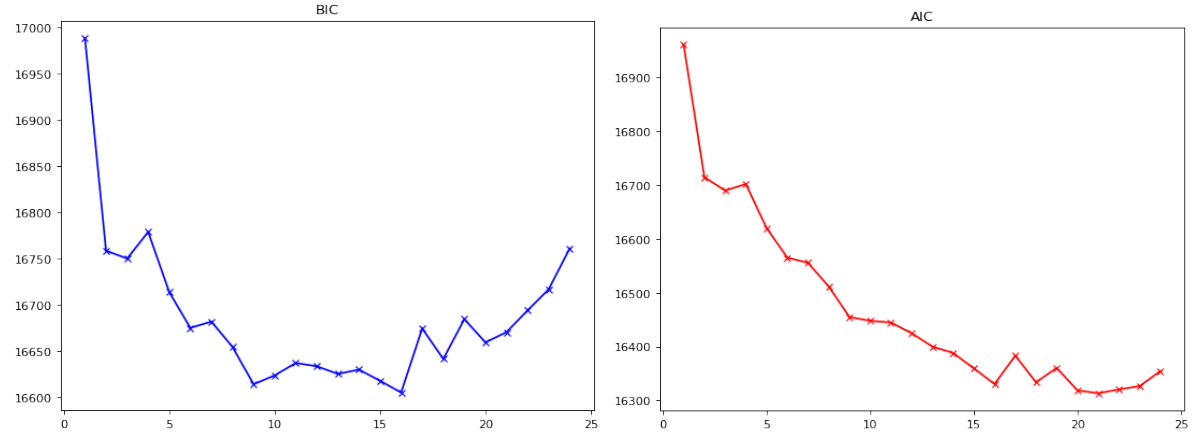


Figure 1.5 BIC, AIC Silhouette and Cross-Validation scores of “Full” structure

According to those plots, best cluster number is 4, 10, 3, 10 respectively.
When choosing “tied” covariance matrix structure, the BIC, AIC, Silhouette and cross-validation scores are listed below:



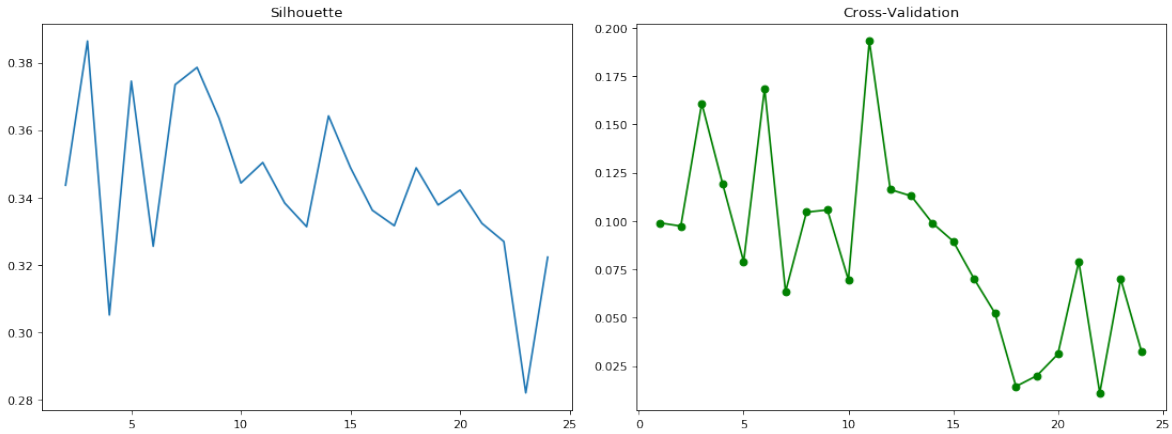


Figure 1.6 BIC, AIC Silhouette and Cross-Validation scores of "Tied" structure

According to those plots, best cluster number is 16, 21, 3, 11 respectively.
When choosing "diag" covariance matrix structure, the BIC, AIC, Silhouette and cross-validation scores are listed below:

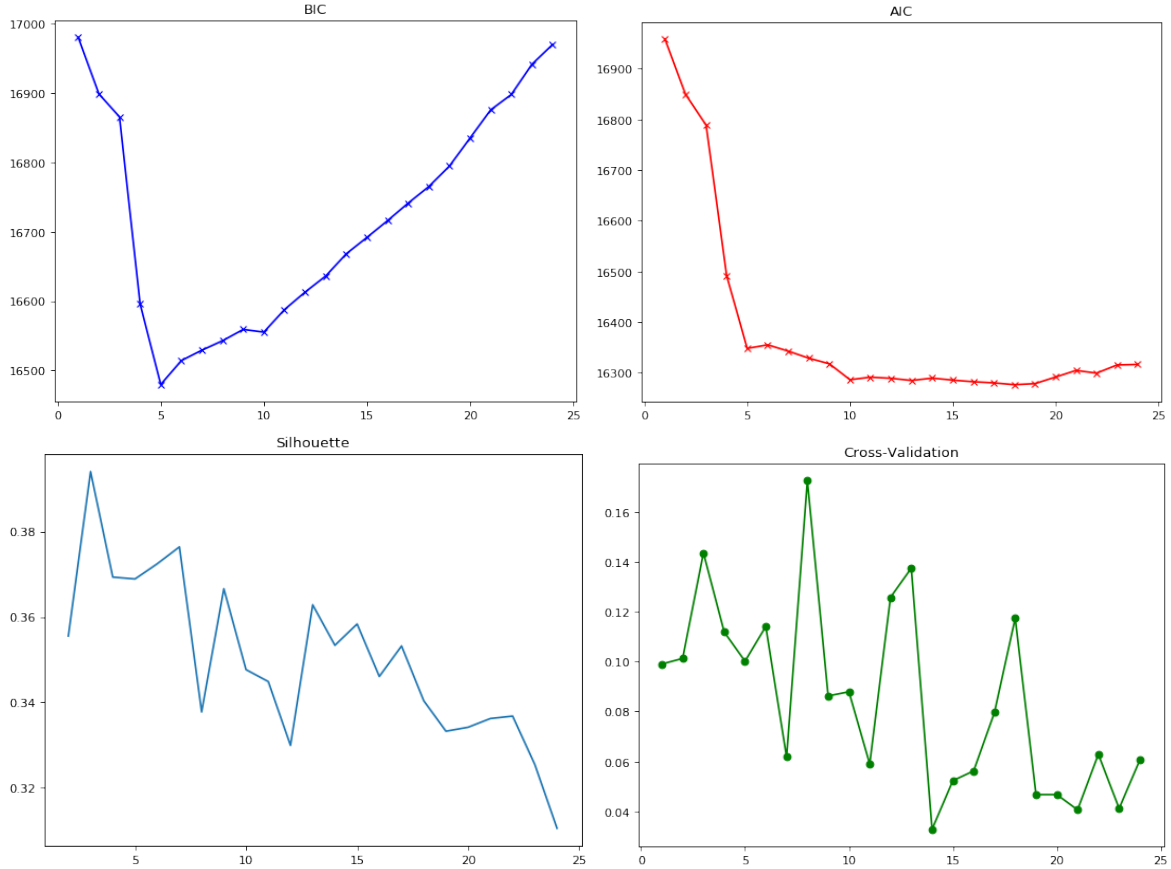


Figure 1.7 BIC, AIC Silhouette and Cross-Validation scores of "Diag" structure

According to those plots, best cluster number is 5, 18, 3, 8 respectively.
When choosing "spherical" covariance matrix structure, the BIC, AIC, Silhouette and cross-validation scores are listed below:

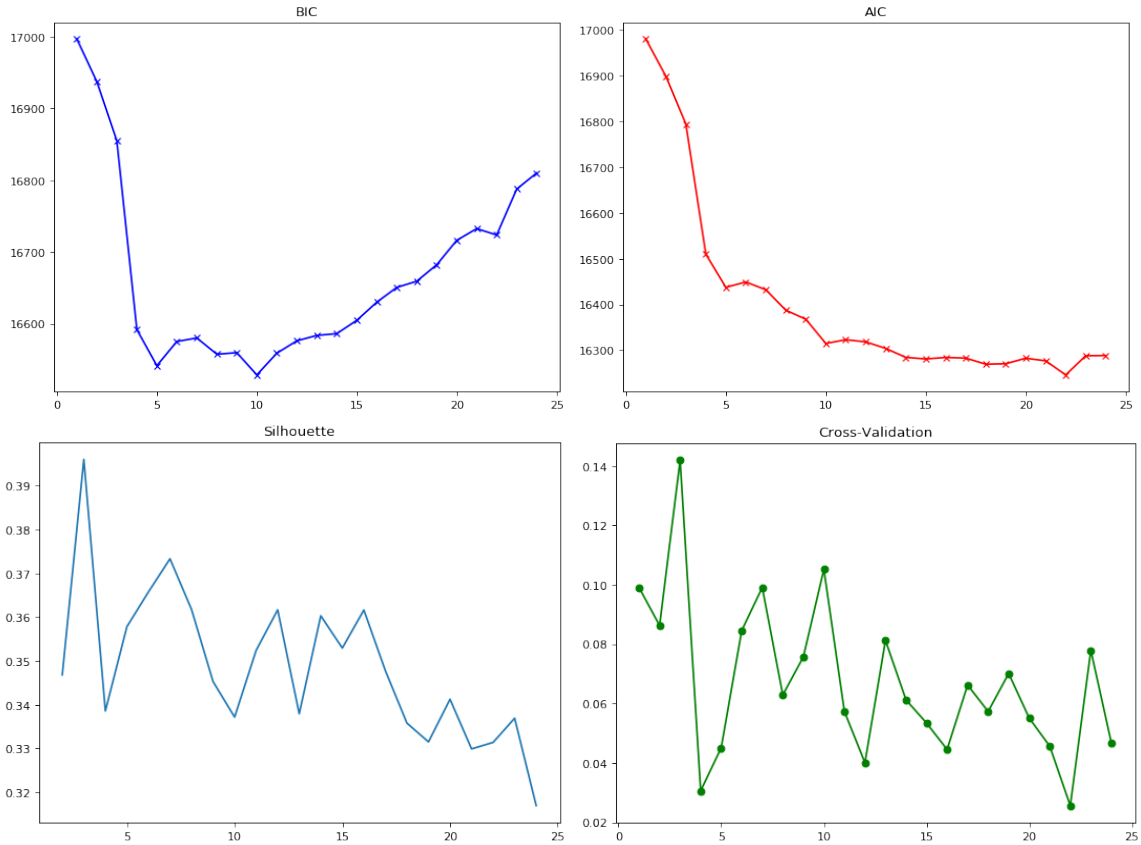


Figure 1.8 BIC, AIC Silhouette and Cross-Validation scores of “Spherical” structure

According to those plots, best cluster number is 10, 22, 3, 3 respectively.
To be concluded, the best covariance matrix structure is “full” cause all four measurements combined have the minimum variance with 10 in total.
The figure below shows the final result of clustering.

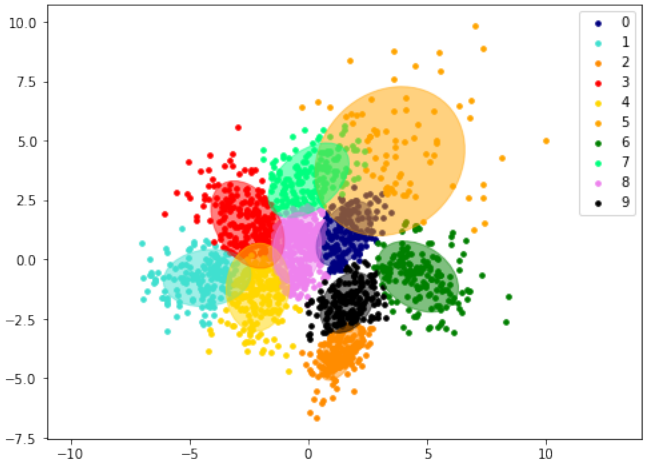


Figure 1.9 Final Test Result

Discussion

No matter which kind of covariance matrix structure, four measurements cannot reach an consensus about what is the best number. However, the “full” structure has the minimum variance with 10 in total. And AIC and CV plot of this structure perfectly determined 10 as the cluster number. Since we have already knew that MNIST dataset has 10 clusters, it is not hard to decide which structure is the best option.