Report

1. Results

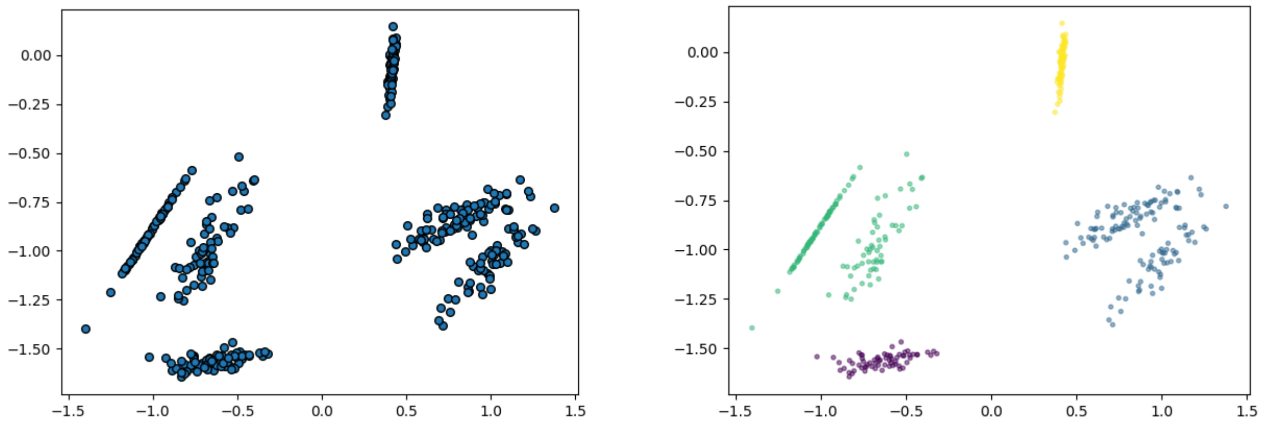

Fig.1 Cluster result for dataset A

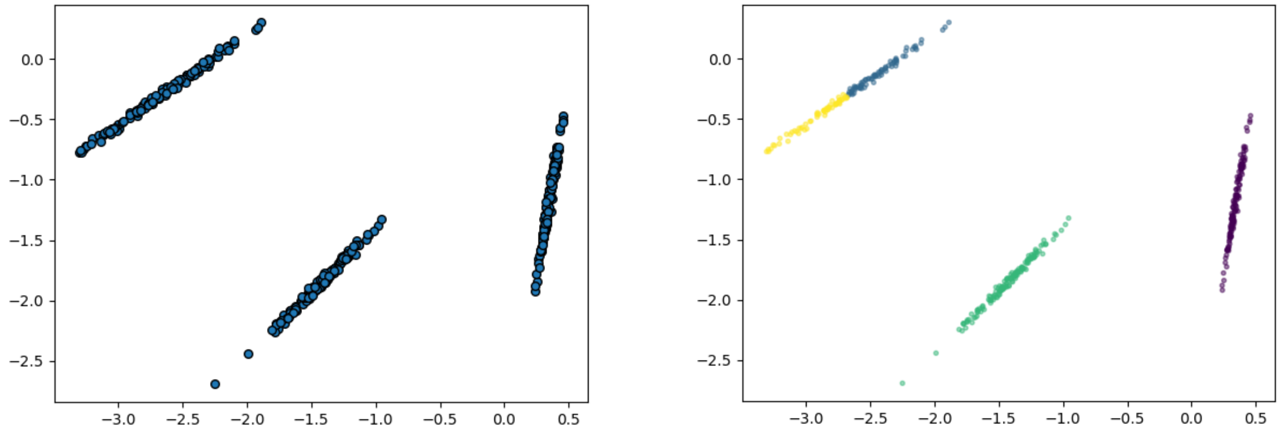

Fig.2 Cluster result for dataset B

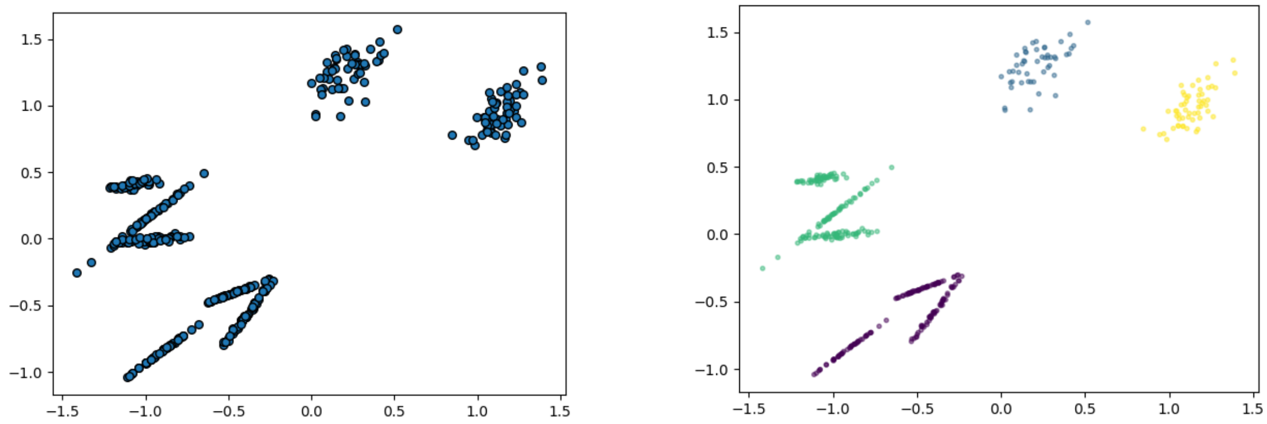

Fig.3 Cluster result for dataset C

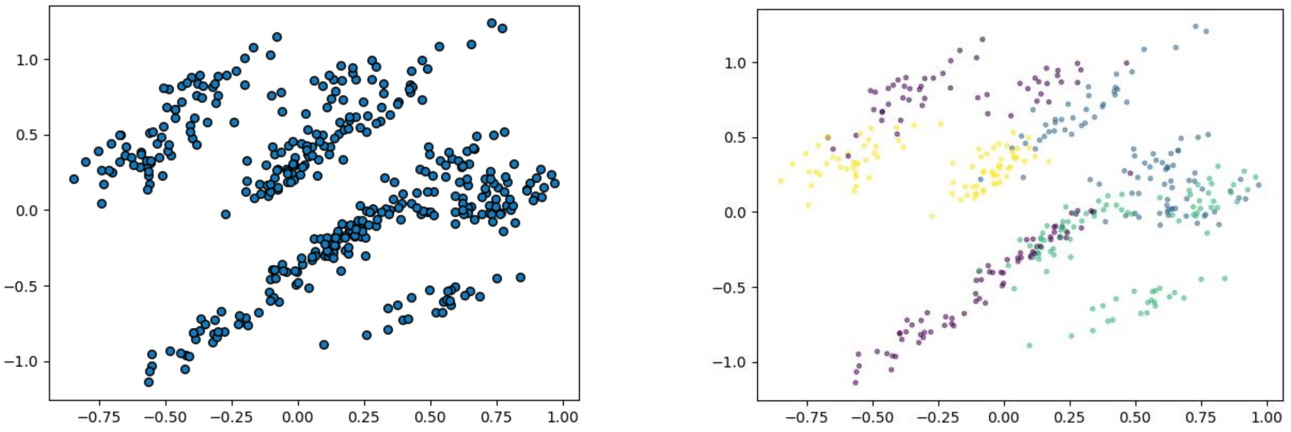

Fig.4 Cluster result for dataset Z

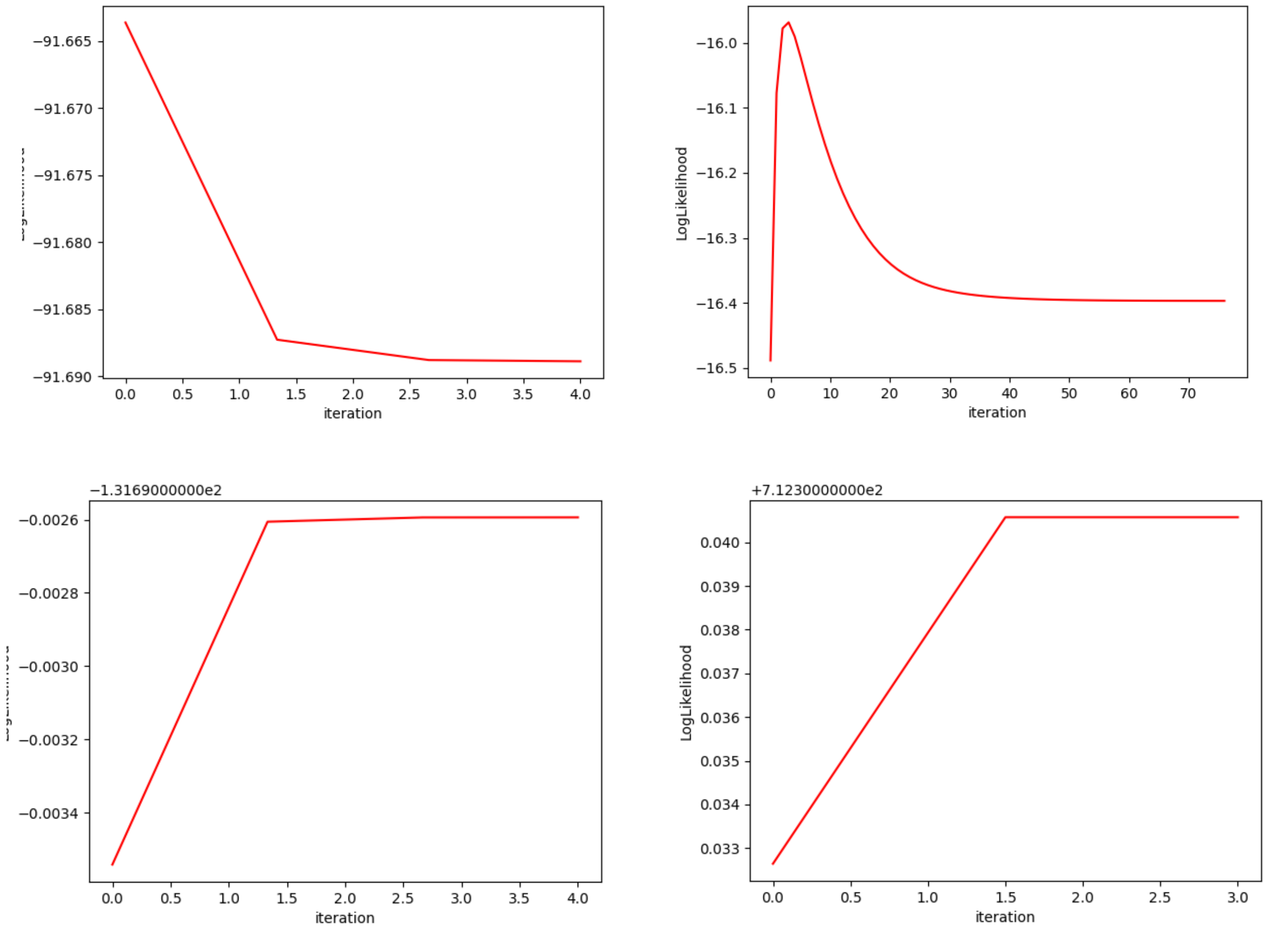

Fig.5 Log likelihood for A, B, C and Z

2. Analysis

I found that for dataset A and C, the number of clusters is 4. For dataset B, the number of clusters is 3. For dataset Z, since it is a multidimension data, I use exhaustive search of cluster number from 2 to 20 using Calinski-Harabaz score to evaluate the performance. I found that for Z, the number of clusters is 12, I only plot the result by first two dimensions.

To prevent overfitting, I set up a strategy that use a variable named ‘eps’ as a threshold of the likelihood. If the loglikelihood difference calculated by the previous and subsequent EM is less than ‘eps’, the iteration will be stopped.