Abstract

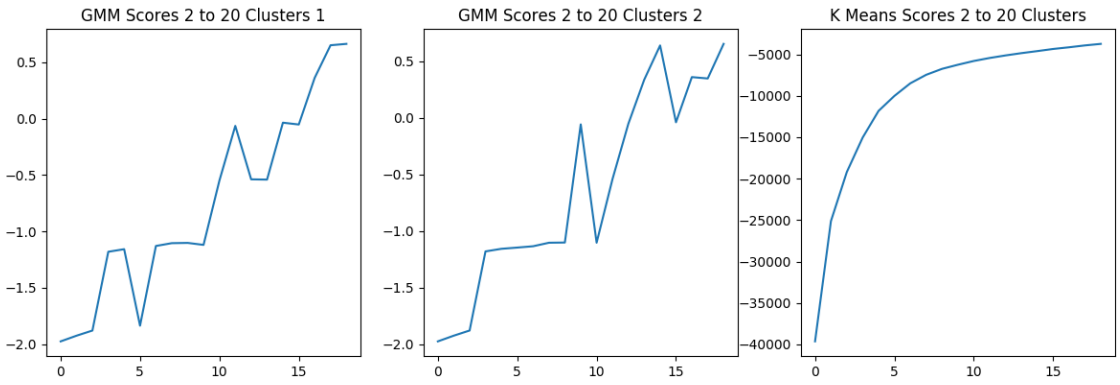
Its algorithm based on random number, GMM yields different result each trial. Our elbow test and multi-step clustering amplify such difference, so no single trial can represent the best clustering pattern. I send you the results of two trials with identical input. Besides the regression analysis you plan to do, you might also want to compare how similar these two trials are.

Scores

Browsing the internet, I do not see much information about elbow test in GMM. In K means, however, the built-in function "score" is a value that elbow test is commonly based on. Both K means and GMM have the built-in "score" function. In the next section I will compare them.

I am also including this score together with variance, so you can evaluate how the score represent the quality of clustering pattern. Note that the score is generated only when a clustering process happens, which means sub-clusters under the same parent cluster have the same score. Also, larger score should indicate better clustering pattern.

GMM score vs. K Means score



Higher score indicates better clustering pattern.

In this graph, we can see two surprising features of GMM:

- The result is not the same between each trial. In the first two graphs, with all inputs fixed, the graphs differ.
- The quality of clustering pattern jumps up and down when the number of cluster increases.

Actually, GMM is partly based on random number, so each time we cluster the same data set, the result differs. We did not notice it before because large data set hides the randomness. Here, the score of GMM amplifies the difference.

So now we face a dilemma. On one hand, GMM score differs in each trial, so we will get different clustering pattern in each trial, which is obviously not good.

On the other hand, K means score is also a bad choice. As we can see in the graph above, the graph of GMM and K means are really different.

After thinking thru it, I decide to use GMM score for elbow test. I am hoping that our multiple step clustering can offset the difference in each step. Whether this is actually the case waits for you to analyze.

Algorithm of Elbow Test and Further Clustering

The idea of elbow test is to find the point where increasing the number of clusters does not make the clustering pattern much better. To achieve this, I apply the following formula:

F(x) = score(x) * 2 – score (x + 1), x belongs to [3, 19]

The x with highest F(x) becomes the number of clusters.

To assess whether further clustering is required, first I need to ensure that I have at least 20 data points in each cluster, or there would be a problem in elbow test. Then, I calculate the data points' distance to origin, and then work out standard deviation. If standard deviation is more than 0.25, perform further clustering; otherwise, do not perform further clustering. In formula:

Return False if $\text{length}([x, y]) < 20$

$f(x,\ y) = \ std([\sqrt{x^2+y^2}\ ])$

Return True if $f(x, y) > .25$, False otherwise.