**Introduction**

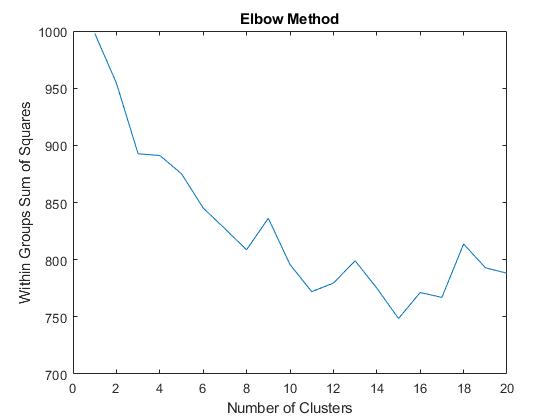
For this homework assignment I was tasked with using Expectation Maximization algorithm to fit the mixture of N Gaussian mixtures to the data. To accomplish this, I first ran the algorithm to estimate the covariance and mean for k clusters, where k was varied from one to twenty. Then I used the Mahalanobis distance to assign the points to the correct cluster. Finally, I used the elbow method to determine which number of clusters was optimal.

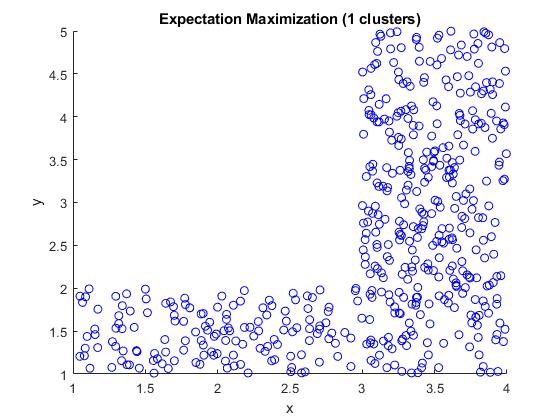
**Methods**

Below are algorithms that are critical to the implementation of the algorithm:

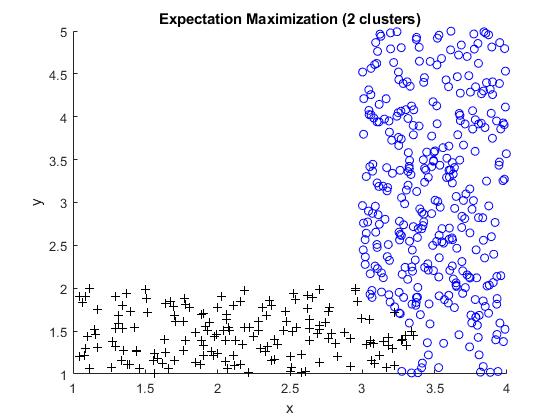
1. function [X,X1,X2] = generateData(N)
   1. Generates the three datasets which represent each true cluster.
   2. The purpose of this assignment is to use the fact that the data is normally distributed, to separate the data into these known clusters.
2. function [d] = md(p,mu,sigma)
   1. Calculate the Mahalanobin distance. Which is the measure of the distance between a point and a distribution.
3. function [sigma, mu] = gmme(X,X1,X2,k)
   1. Performs the Expectation-Maximization algorithm for Gaussian mixtures.
   2. Returns the estimated covariance and mean values for each of the k clusters.
4. function [sse] = calcSSE(X,ls,mu,sigma)
   1. Calculates the sum of squared error, which is the sum of Mahalanobis distance between each point in a cluster and that cluster’s mean, for each cluster.
   2. This is used to implement the elbow method, which idealy shows us the optimal number of clusters.
5. function plotElbow(k,sse)
   1. Displays the plot of the cluster count versus the sum of squared errors.
6. function displayMixtures(X,ls,k)
   1. Shows the clustered data.

**Results**

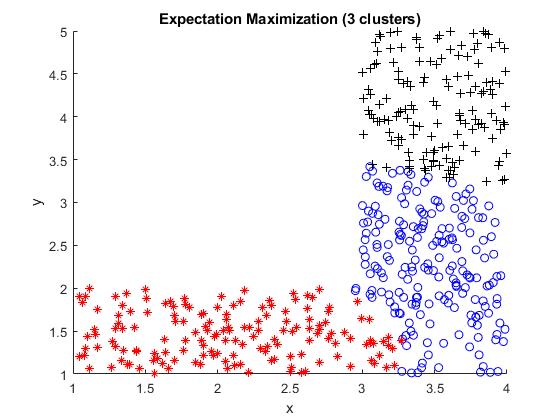
****

**Figure 1**. Elbow Method

**Figure 2.** Gaussian Mixture (one cluster)



**Figure 3.** Gaussian Mixture (two clusters)



**Figure 4.** Gaussian Mixture (three cluster)

**Discussion**

Figure 1 shows the plot of the cluster count versus the sum of squared error which is used to determine the optimum number of clusters for the dataset. This technique is known as the elbow method. The sum of squared error is calculated by measuring the Mahalanobis distance between each point in each cluster, and its mean. The elbow method is not a foolproof technique in finding the optimal cluster count. For example, according to the plot generated on the homework dataset, the optimal number of clusters is three, which is obviously incorrect.

Figure 2 displays the Gaussian mixture given only one cluster. Here, all the points are assigned to a single cluster. Meanwhile, figures 3 and four show the mixture given two and three clusters. The Expectation-Maximization algorithm is utilized here to estimate the covariance and mean of each (k) clusters. On the other hand, the Mahalanobis distance, which corresponds to the distance between a point and a distribution, is used to assign points to their corresponding cluster.

**Software listing and executable software**

Start the program by pressing “Run”. The program will display the confusion matrix and the results from the grid search.