CS342 Coursework Report

Kai Meller 1920229

Figures and Plots

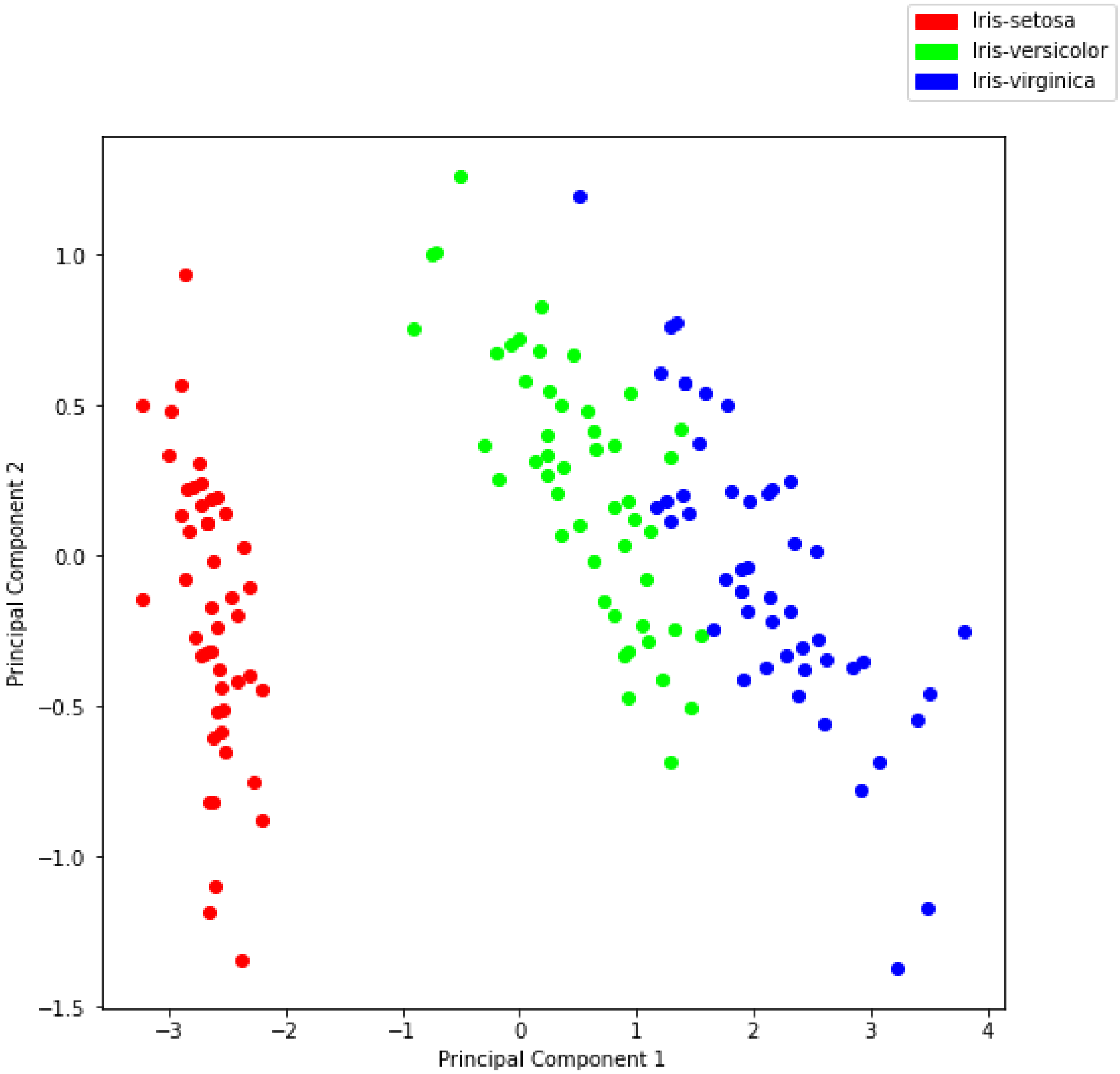


Fig 1. Iris Dataset after PCA conversion from 4D to 2D, sorted by class.

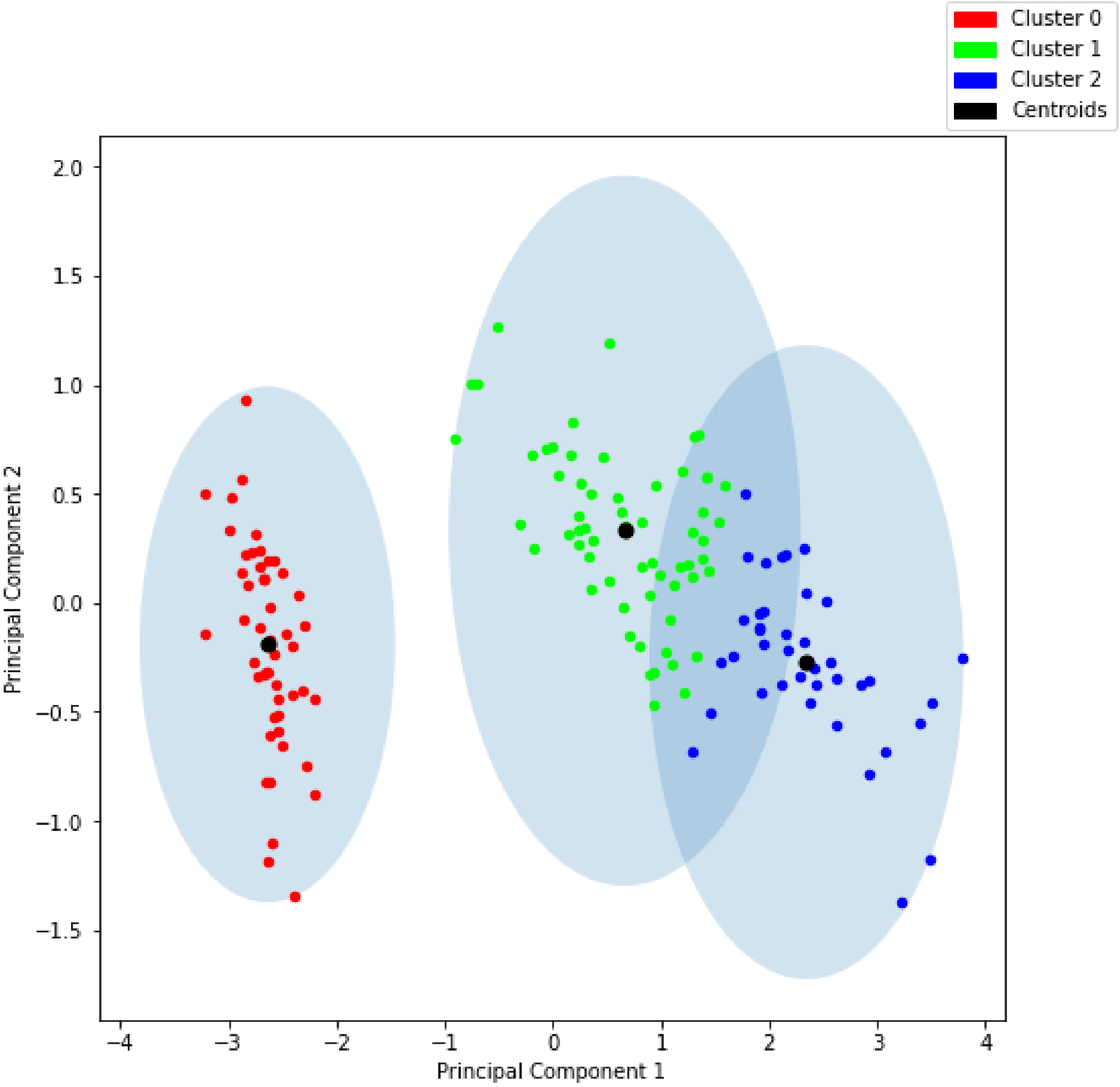


Fig. 2. Iris Dataset clustered into 3 classes using sklearn.cluster.KMeans, with a circle for each cluster. Circles are centered at the cluster centers, and have radius equal to the furthest away data point within their cluster.

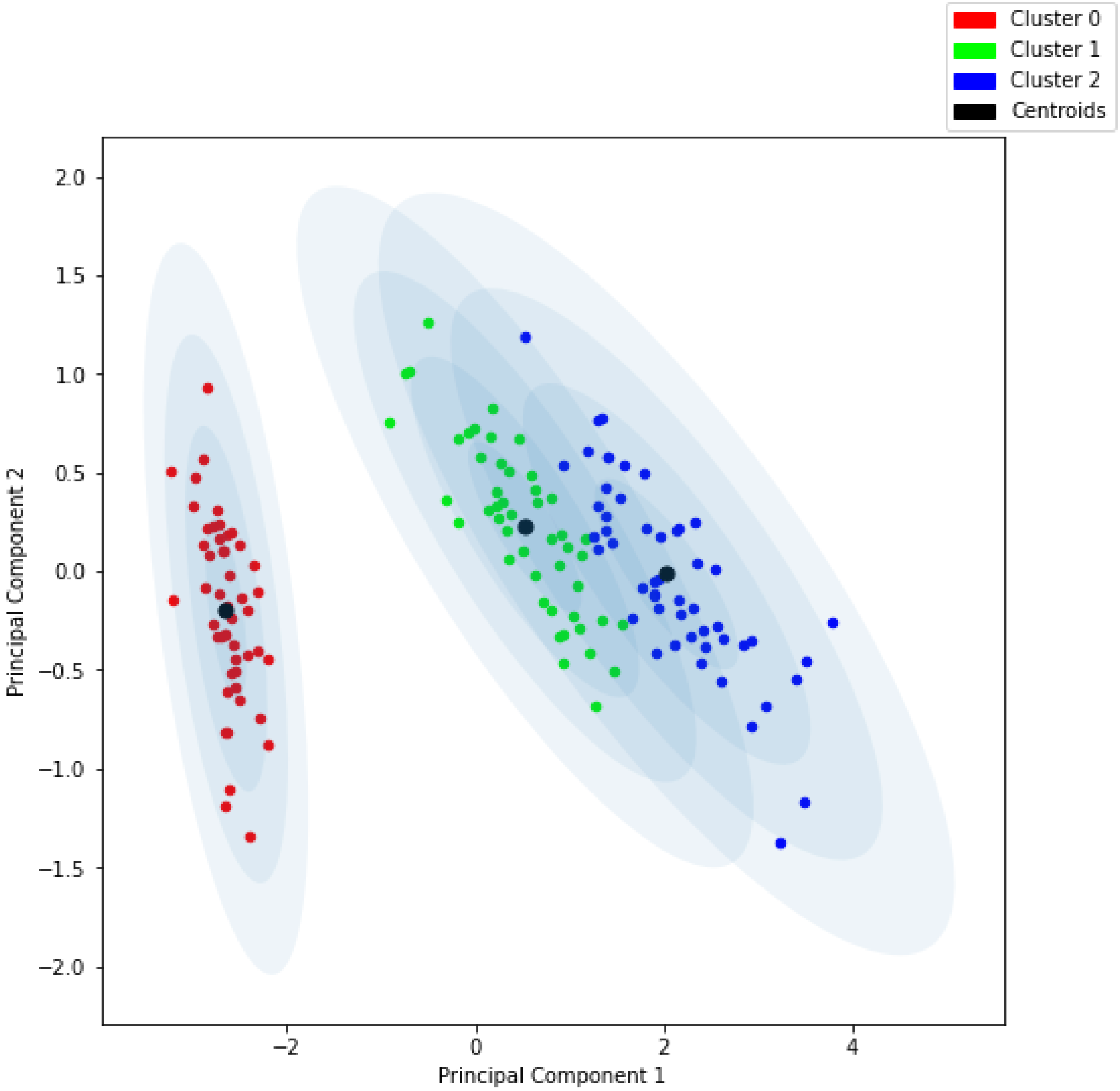
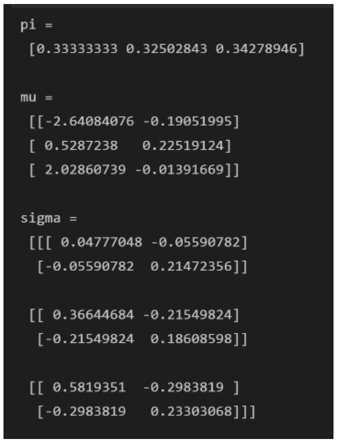


Fig. 3. Iris Dataset clustered into 3 classes using my implementation of the EM algorithm, after KMeans. Ellipses represent the Gaussian distributions for each cluster, with shade decreasing representing lower probability.



```
pi =
 [0.33333333 0.32502843 0.34278946]

mu =
 [[-2.64084076 -0.19051995]
 [ 0.5287238   0.22519124]
 [ 2.02860739 -0.01391669]]

sigma =
 [[[ 0.04777048 -0.05590782]
  [-0.05590782  0.21472356]]

 [[ 0.36644684 -0.21549824]
  [-0.21549824  0.18608598]]

 [[ 0.5819351  -0.2983819 ]
  [-0.2983819   0.23303068]]]
```

Fig. 4. Final set of parameters given by one run of my implementation of the EM algorithm. Each parameter has a value for each cluster, hence three values each.

Discussion of Number of Iterations

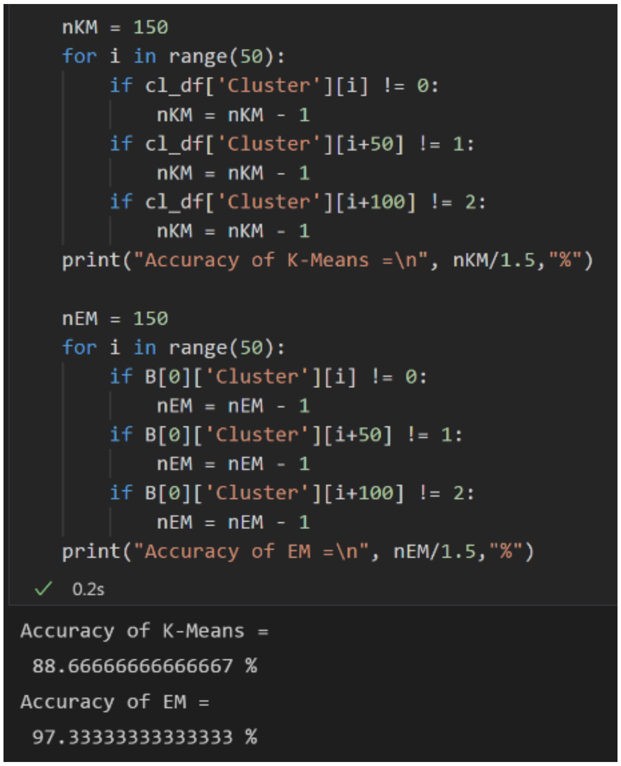
I chose mu as my parameter to measure for convergence as it is the most representative feature for each cluster. I decided that my EM algorithm converged sufficiently when the average change in each of the coordinates of each of the three means between the last two M-steps was less than 0.001. Although fairly arbitrary, changes of less than 0.006 by any one coordinate of any cluster's mean would be very unlikely to affect the assignments of data to clusters as the data in the set are measured to the nearest 0.1. Because of the convergence of EM, any further changes after the first mean change of less than 0.006 would be significantly less than 0.006, and so would sum to roughly 0.006. Experimentally, this led to the number of iterations of EM being 20 in my tests.

Discussion of Accuracy of K-Means against EM

For classification of flowers in the Iris dataset, it is apparent that the EM algorithm is superior. This is due to the fact that K-Means groups data based on distance from each other, which means it assigns circular clusters. However the Iris dataset is one which has non-circular clusters, and two of the classes (Iris-versicolor and Iris-virginica) have very close distributions. This meant that when K-Means was run on Iris, it split these two classes in the wrong direction as there are Iris-versicolor near the mean that are closer to many Iris-virginica than many Iris-versicolor, and vice versa.

EM with GMM is a much better machine learning model to classify the Iris dataset as in reality each datum has 4 values, each following a different Gaussian distribution depending on the species of flower - a multivariate Gaussian. This means the model takes into consideration the non-uniformity of the distribution and learns the shape. As shown in figure 5, the accuracy of K-Means was 88.7%, whereas the accuracy of EM was 97.3%.

However, the dataset is not a perfect representation of the real distributions with sample size only n=50 each. There are data that are identified as another class as they lie closer to another distribution than they do to their own. These are the outliers of the dataset - values that are so close to both versicolor and virginica that the algorithm calculates responsibilities that are out by such a small factor that they are comparable (ie. of the same order of magnitude).



```
nKM = 150
for i in range(50):
    if cl_df['Cluster'][i] != 0:
        nKM = nKM - 1
    if cl_df['Cluster'][i+50] != 1:
        nKM = nKM - 1
    if cl_df['Cluster'][i+100] != 2:
        nKM = nKM - 1
print("Accuracy of K-Means =\n", nKM/1.5,"%")

nEM = 150
for i in range(50):
    if B[0]['Cluster'][i] != 0:
        nEM = nEM - 1
    if B[0]['Cluster'][i+50] != 1:
        nEM = nEM - 1
    if B[0]['Cluster'][i+100] != 2:
        nEM = nEM - 1
print("Accuracy of EM =\n", nEM/1.5,"%")
```

```
Accuracy of K-Means =
 88.66666666666667 %
Accuracy of EM =
 97.33333333333333 %
```

Fig. 5. Program that calculates the accuracies of the two clustering algorithms by comparing to the ground truth labels, displaying the calculated accuracy of a run.