Nicholas Fang

Assignment 7

1) **EM falls under unsupervised learning. But what specific kind of unsupervised learning setting would you use it for? Put differently, you could use both k-means and EM for clustering, but when will you pick EM over k-means?**

Expectation maximization is a way to find the maximum likelihood estimates for model parameters. This method is useful when the data is incomplete, has missing data points, or has some hidden latent variables.

"Unlike K-means, in EM, the clusters are not limited to spherical shapes." (Naveed, 2018) In EM, you can constrain the algorithm to provide different covariance metrics. The different covariance metrics in return allow you to control the shape of the cluster. The benefit to this is that you would be able to identify sub-populations in the data with different characteristics. Not to say that one method is perfect, but it would be useful to use both methods in order to look at different clusters that each method produces, with the EM method offering greater control over characteristics of the cluster. With that being said, one would also have to keep in mind that there is no guarantee that the algorithm will produce a model/models that are the best fit for the dataset.

2) Use EM for clustering "similar" countries. Report how many groups you got and why you chose that number with the help of AIC and BIC.

Concluded that having 10 components/clusters produced the best model in terms of lowest average AIC and lowest BIC (for 1000 iterations). It seems that choosing an arbitrary higher number of clusters lowers the AIC and BIC. However, it is worth to note that the number of specified clusters between 3-7 seems to be fairly similar.

N_custers = 8

```
In [137]: print(model.aic(data))
1173.210825069109

In [138]: print(model.bic(data))
1400.7415627150585
```

N_clusters=10

```
In [143]: print(model.aic(data))
935.5826497368117

In [144]: print(model.bic(data))
1220.4740775456055
```