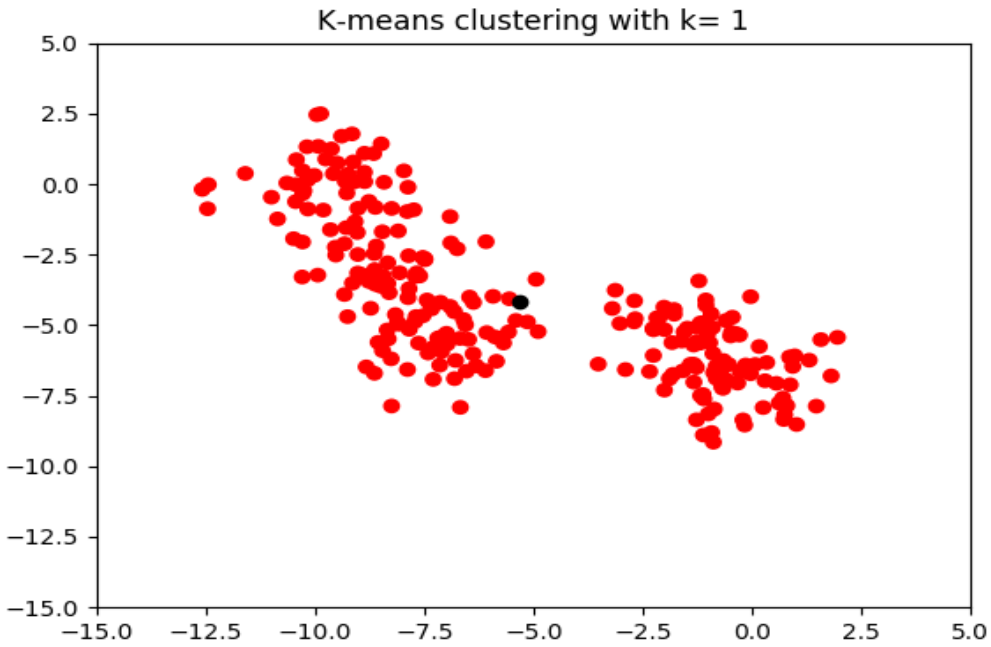
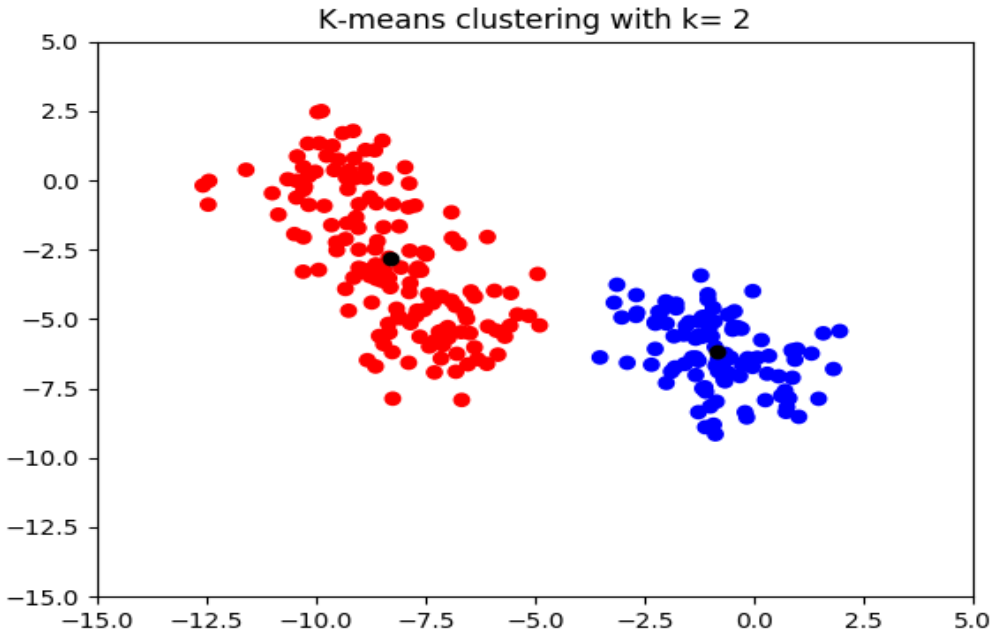
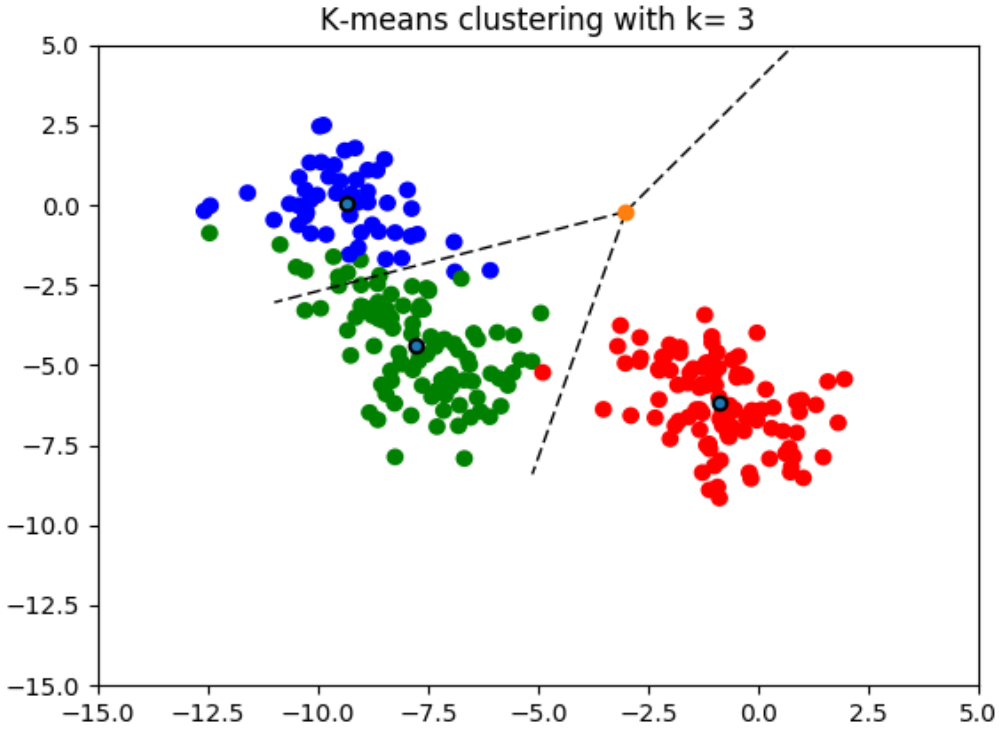
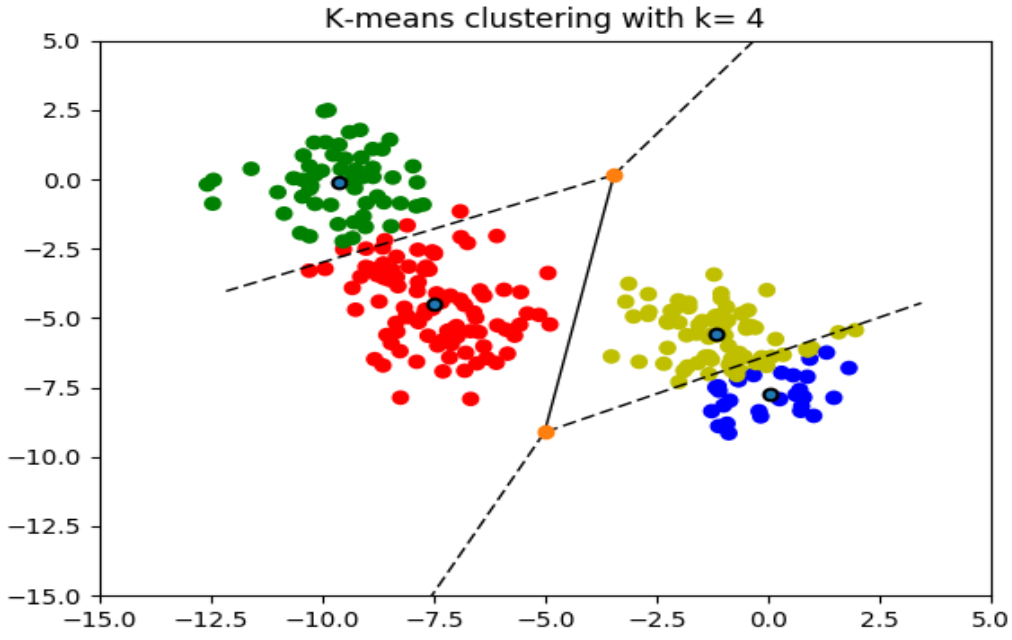
6.036 Machine Learning Project 3

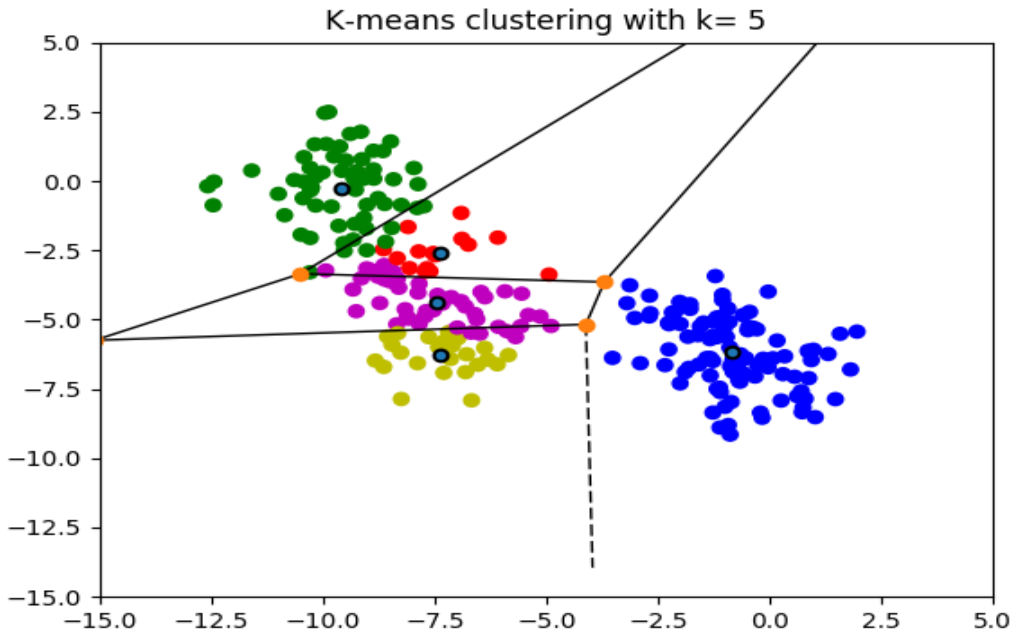
K-Means Clustering Graphs



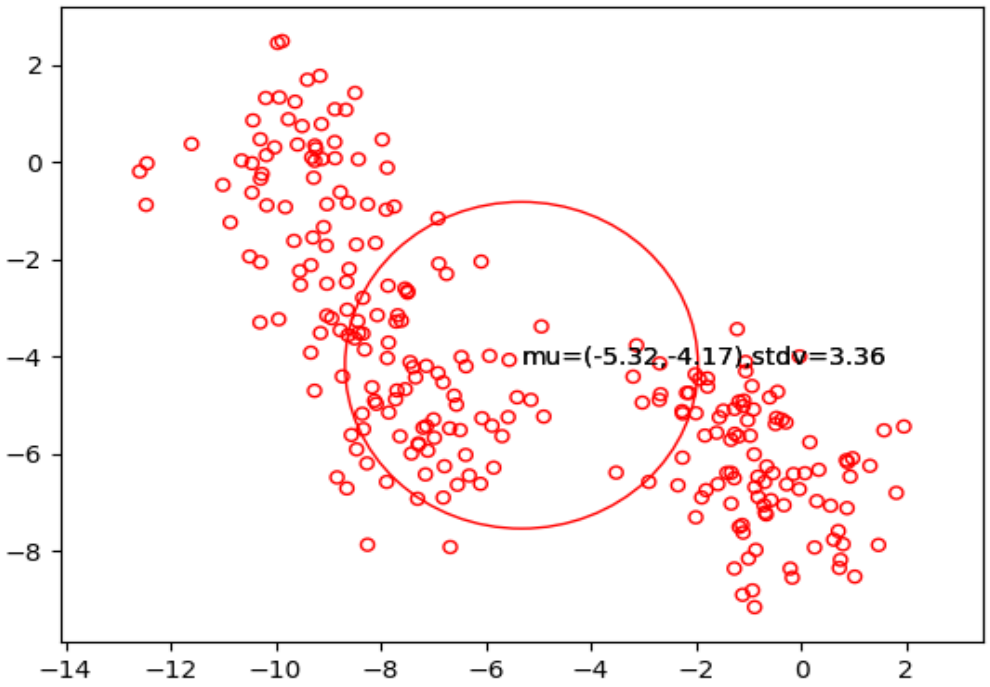


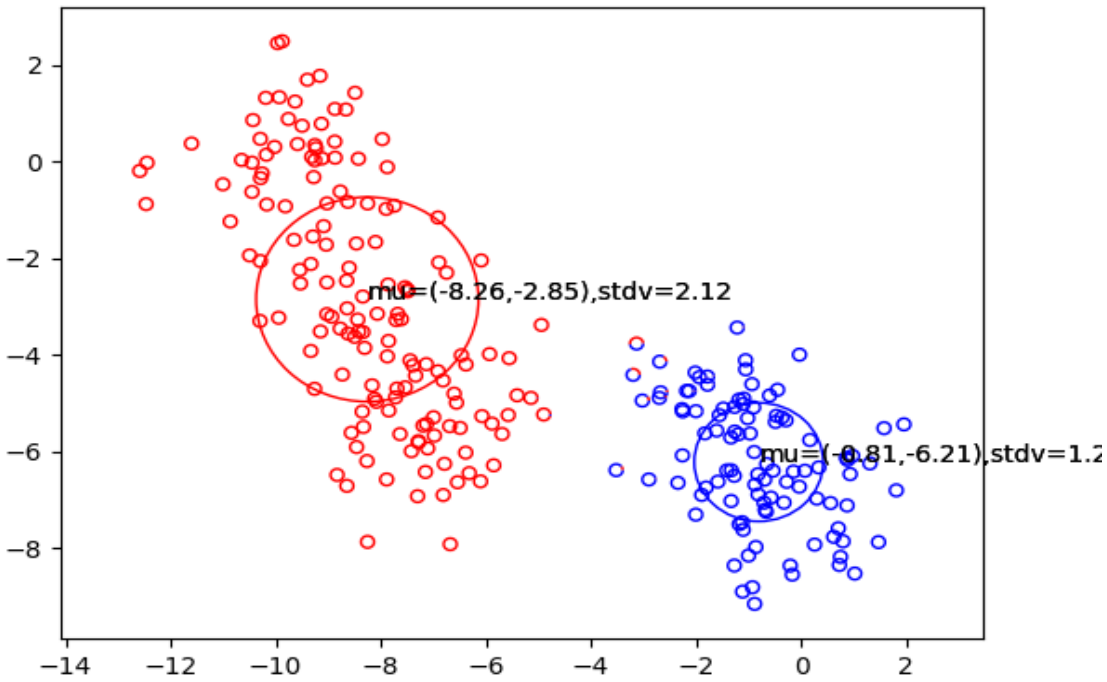

K-means clustering with k= 3
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
−15.0
−15.0
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0

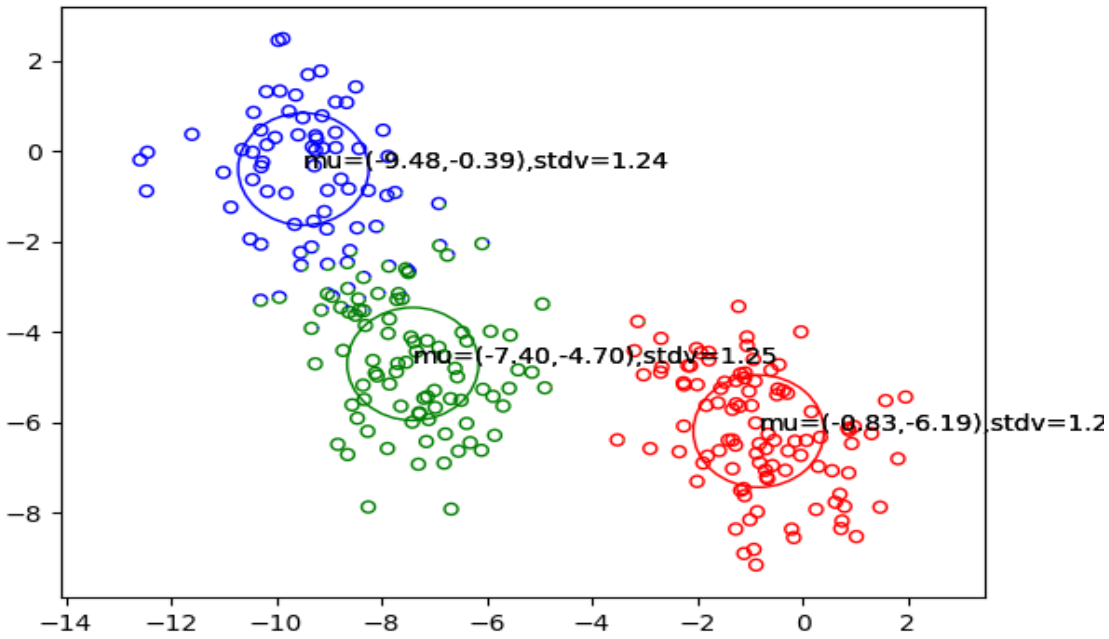

K-means clustering with k= 4
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
−15.0
−15.0
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0

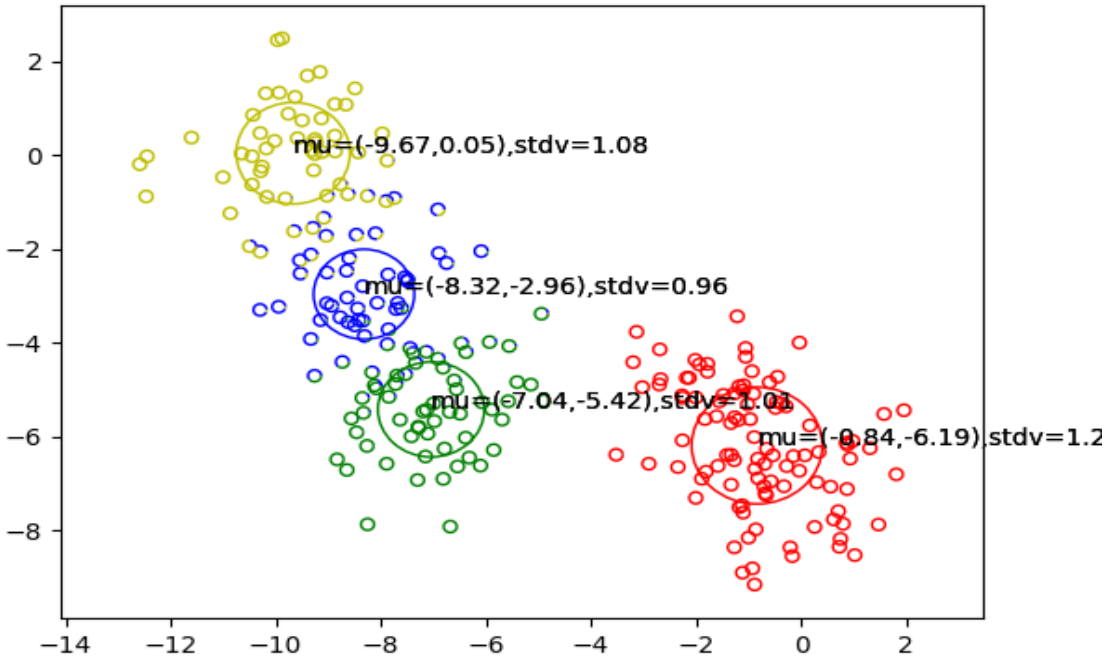


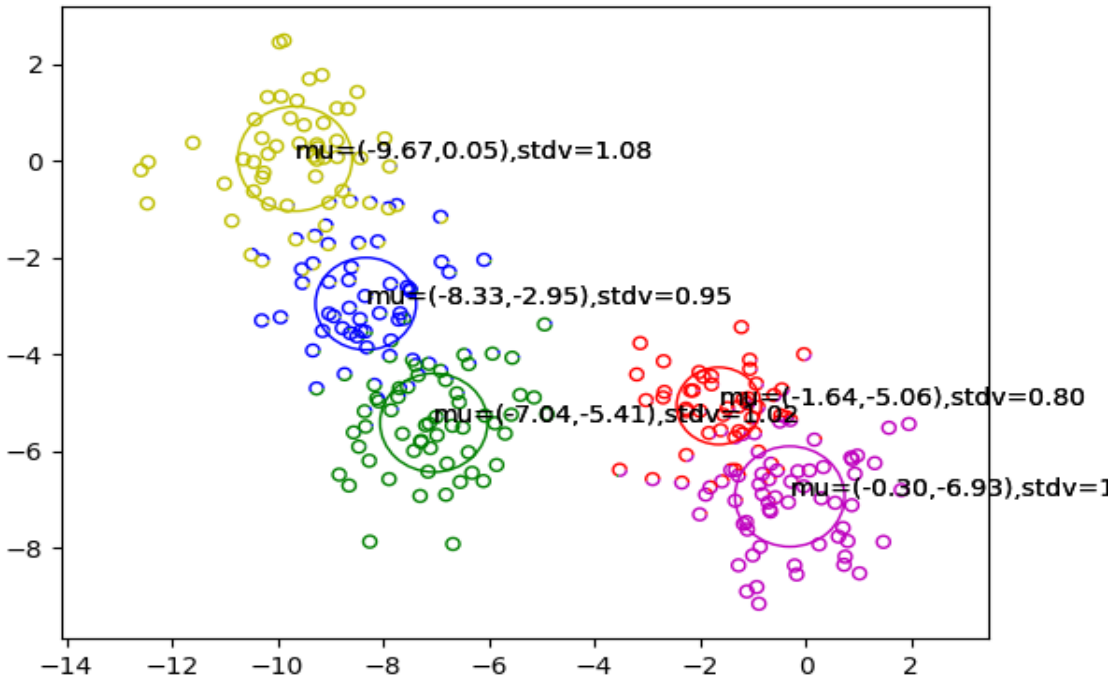
E and M Algorithm Plots for Gaussian Mixture


mu=(-5.32,-4.17),stdv=3.36
2
0
−2
−4
−6
−8
−14
−12
−10
−8
−6
−4
−2
0
2


mu=(-8.26,-2.85),stdv=2.12
mu=(-6.81,-6.21),stdv=1.2
2
0
−2
−4
−6
−8
−14
−12
−10
−8
−6
−4
−2
0
2


mu=(-9.48,-0.39),stdv=1.24
mu=(-7.40,-4.70),stdv=1.25
mu=(-0.83,-6.19),stdv=1.2


mu=(-9.67,0.05),stdv=1.08
mu=(-8.32,-2.96),stdv=0.96
mu=(-7.04,-5.42),stdv=1.01
mu=(-0.84,-6.19),stdv=1.2



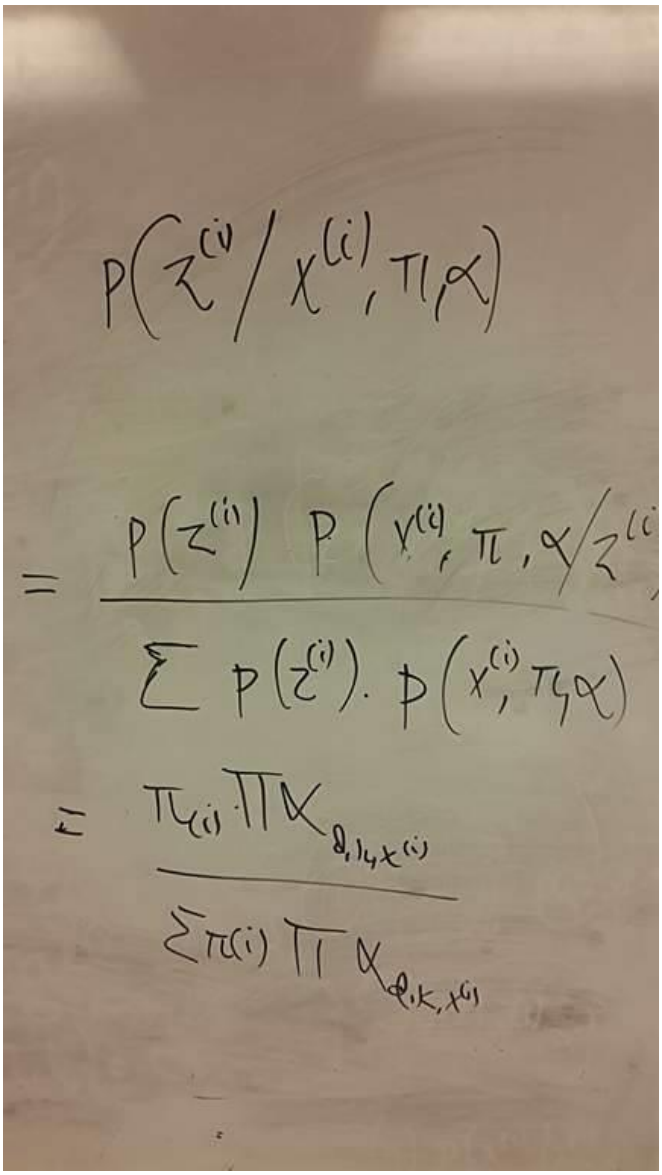
E and M algorithm gives the soft assignment of the data point to a particular cluster but K –means assigns a data point only into a particular cluster. K-means algorithm only considers the centroid or central tendency of the data whereas E and M also considers the variance from the centroids. I.e mu and sigma. In the above dataset, for k = 4 and k = 5, we got different clusters by k-means and E and M whereas for other k's we have pretty similar clusters.

2.1

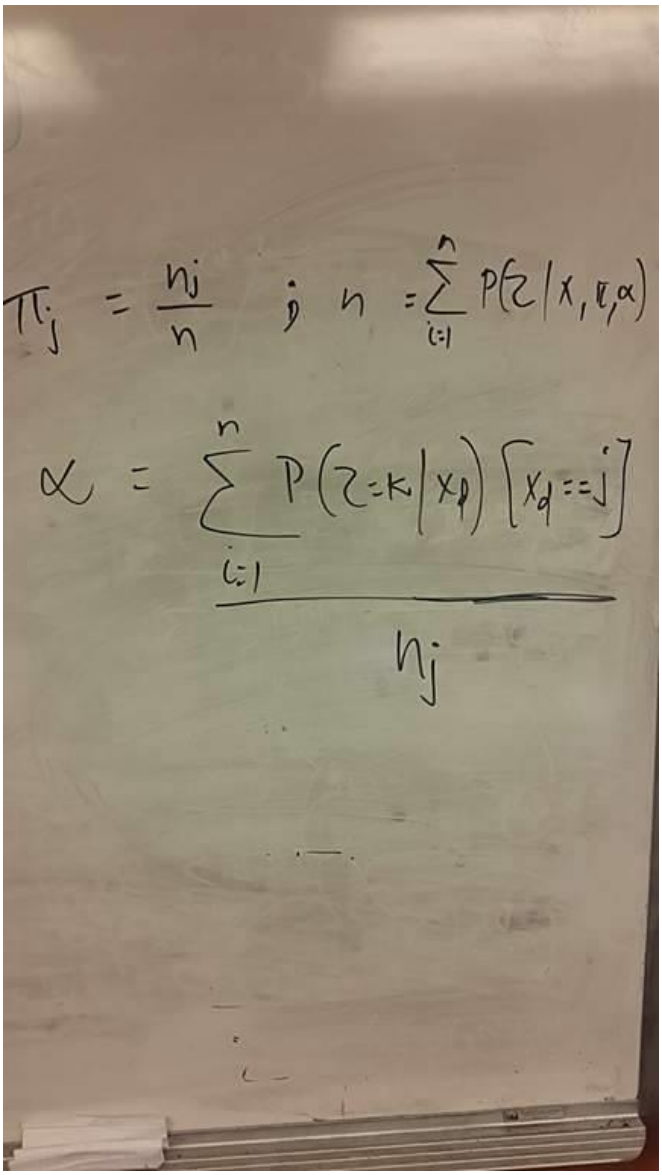
We are assuming the unrealistic feature independence and that will add more weight to the redundant correlated features which will lead to overfit those particular features while doing cluster assignments.

Using the indicator variable to remove the unknown values helps us to take in account the features that we know and discard the unknowns while taking the log.

2.2


$p(z^{(i)} / x^{(i)}, \pi, \alpha)$
$= \frac{p(z^{(i)})\ p(x^{(i)}, \pi, \alpha / z^{(i)}}{\sum p(z^{(i)}) \cdot p(x^{(i)}; \pi, \alpha)}$
$= \frac{\pi_{(i)} \cdot \prod \alpha_{\theta_{i,k}, x^{(i)}}}{\sum \pi_{(i)} \prod \alpha_{\theta_{i,k}, x^{(i)}}}$

2.3



2.5

Fitting k = 2: max ll = -2504522.64377 (0.11 min, 15 iters)

Fitting k = 3: max ll = -2415036.88500 (0.35 min, 40 iters)

Fitting k = 4: max ll = -2387972.10330 (0.17 min, 16 iters)

Fitting k = 5: max ll = -2414989.08775 (0.19 min, 17 iters)

Fitting k = 6: max ll = -2389499.95372 (0.25 min, 20 iters)

Fitting k = 7: max ll = -2369860.71326 (0.82 min, 52 iters)

Fitting k = 8: max ll = -2374467.41955 (0.24 min, 16 iters)

Fitting k = 9: max ll = -2375773.69350 (0.84 min, 47 iters)

Fitting k = 10: max ll = -2391576.38601 (1.01 min, 54 iters)

Fitting k = 11: max ll = -2394010.46214 (0.64 min, 32 iters)

Fitting k = 12: max ll = -2384466.64403 (1.11 min, 52 iters)

Fitting k = 13: max ll = -2419412.80252 (1.07 min, 47 iters)

Fitting k = 14: max ll = -2392110.81446 (1.55 min, 59 iters)

Fitting k = 15: max ll = -2422184.12663 (0.84 min, 29 iters)

Fitting k = 16: max ll = -2426362.58518 (2.55 min, 67 iters)

Fitting k = 17: max ll = -2406778.88568 (1.35 min, 35 iters)

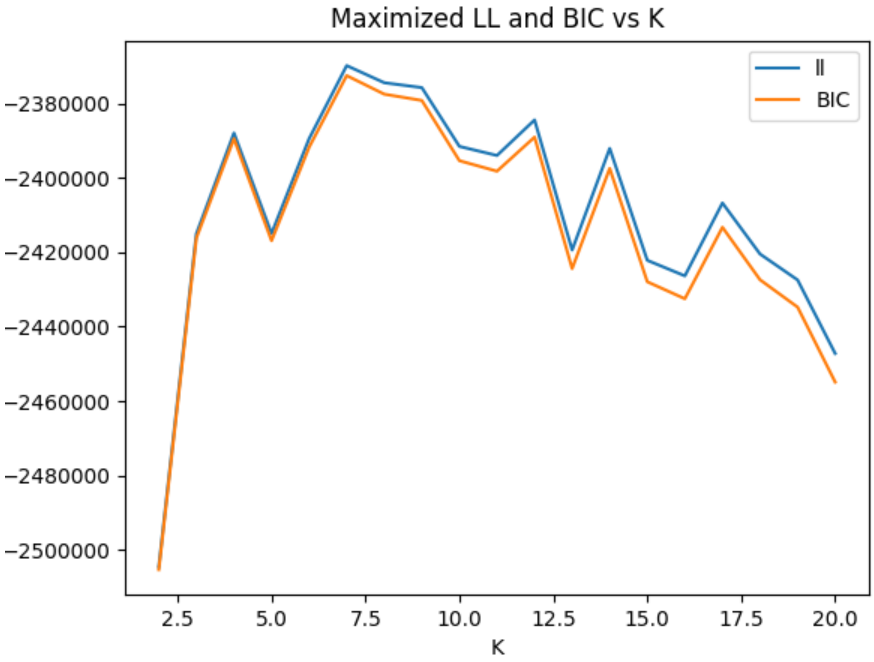
Fitting k = 18: max ll = -2420487.98730 (1.65 min, 41 iters)

Fitting k = 19: max ll = -2427460.23038 (2.97 min, 68 iters)

Fitting k = 20: max ll = -2447215.94378 (1.81 min, 43 iters)

The log likelihood is monotonically increasing when we iterate over the Iterations.

2.6



We should choose the value of 7.5.

Both results, max ll and BIC, agree as we can see from the graph.

2.7

(a).

Cluster 1:

age: 0 - 12

sex: male

birthplace: US

ancestry1: Western Europe (except Spain)

citizen: born in US

income: none

edlevel: 1st - 4th grade

employer: n/a, under 16

This cluster belongs to children born in US having ancestry from non-spanish western europe.

Cluster 2:

age: 30 - 39

sex: male

birthplace: US

ancestry1: Western Europe (except Spain)

citizen: born in US

income: $15k - $29999

edlevel: high school or ged

employer: private, for profit

This cluster belongs to middle aged, middle income people born in US having western European ancestry.

age: 65 and above

sex: female

birthplace: US

ancestry1: Western Europe (except Spain)

citizen: born in US

income: $1 - $14999

edlevel: high school or ged

employer: private, for profit

This is the cluster of old aged people with very less income who are born in US and ancestry from Western Europe except Spain.

Cluster 4:

age: 30 - 39

sex: female

birthplace: America (non US)

ancestry1: Hispanic (including Spain)

citizen: not a US citizen

income: $1 - $14999

edlevel: high school or ged

employer: private, for profit

This is the cluster of mid aged non American born foreign national who work for private or non-profit and earn under 15000 dollars.

Cluster 5:

age: 13 - 19

sex: female

birthplace: US

ancestry1: Western Europe (except Spain)

citizen: born in US

income: none

edlevel: 5th - 8th grade

employer: n/a, under 16

This is the cluster of female teen agers born in US and ancestry from western Europe except spain.

Cluster 6:

age: 20 - 29

sex: male

birthplace: America (non US)

ancestry1: Hispanic (including Spain)

citizen: not a US citizen

income: $1 - $14999

edlevel: 5th - 8th grade

employer: private, for profit

This cluster represents male population in their twenties who are born in Americas (not US) and having the hispanic origin.

Cluster 7:

age: 65 and above

sex: female

birthplace: US

ancestry1: Western Europe (except Spain)

citizen: born in US

income: $1 - $14999

edlevel: high school or ged

employer: n/a, under 16

This cluster represents female elderly populations born in US with European heritage.

All cluster points prints are pretty much obvious.

With the lower value of k, some clusters disappeared like the 20-29 male clusters but 0-12 female cluster, which I found prominent in all the runs stayed. I think some prominent clusters are stable and are more likely to be clustered while other disappear/ or appear depending on k or other parameters.

2.7(b)

After three runs I got pretty much the same clusters even with different initializations. Some cluster like female teenagers and old age clusters are quite frequent in all the iterations.

2.7(c)

We can check the dominant clusters from both states and do the qualitative analysis of the characteristics of such clusters. We can infer about the prominent demographics in individual states, their ancestry and others with respect to each other. For a given value of best k, the prominent clusters give the demographics characteristics of the states to compare.