Homework 4 Part 1 EM, K-Means
Qianying Lin

EM Algorithm

Question 1:

The two initiations I used
1 The first method: random data points Random data points. This method sets the means of the modes by sampling at random a corresponding number of data points, sets the covariance matrices of all the modes are to the covariance of the entire dataset, and sets the prior probabilities of the Gaussian modes to be uniform.
2 The second method: Kmeans initialization KMeans initialization This method uses KMeans to pre-cluster the points. It then sets the means and covariances of the Gaussian distributions the sample means and covariances of each KMeans cluster. It also sets the prior probabilities to be proportional to the mass of each cluster.

reference:
1 http://www.vlfeat.org/api/gmm-fundamentals.html
2 Simple Methods for Initializing the EM Algorithm for Gaussian Mixture Models Blomer and Bujna, http://arxiv.org/pdf/1312.5946.pdf

The results from initiation type 1:

```
iteration 0:
cluster 0: mean = 5.509279400176419, covariance =
1.0302576846296705
cluster 1: mean = 25.486654429329228, covariance =
0.9980966181791335
cluster 2: mean = 15.449160792031073, covariance =
0.9671159353647741
iteration 1:
cluster 0: mean = 25.48665442932891, covariance =
0.9980966181809378
cluster 1: mean = 15.449160792031293, covariance =
0.9671159353614397
cluster 2: mean = 5.509279400177353, covariance =
1.0302576846355043
iteration 2:
cluster 0: mean = 15.449160792031073, covariance =
0.9671159353647854
cluster 1: mean = 25.486654429329228, covariance =
0.9980966181791335
cluster 2: mean = 5.509279400176412, covariance =
1.030257684629634
iteration 3:
```

```
cluster 0: mean = 5.509279400176102, covariance =
1.030257684627704
cluster 1: mean = 15.449160792030884, covariance =
0.9671159353652593
cluster 2: mean = 25.486654429329203, covariance =
0.9980966181793028
iteration 4:
cluster 0: mean = 5.509279400176458, covariance =
1.0302576846299132
cluster 1: mean = 15.449160792031092, covariance =
0.9671159353647077
cluster 2: mean = 25.486654429329228, covariance =
0.998096618179134
iteration 5:
cluster 0: mean = 15.449160792031071, covariance =
0.9671159353647797
cluster 1: mean = 25.486654429329228, covariance =
0.9980966181791335
cluster 2: mean = 5.509279400176416, covariance =
1.0302576846296518
iteration 6:
cluster 0: mean = 15.449160792031117, covariance =
0.9671159353646348
cluster 1: mean = 25.486654429329228, covariance =
0.9980966181791341
cluster 2: mean = 5.5092794001764975, covariance =
1.030257684630169
iteration 7:
cluster 0: mean = 5.509279400176454, covariance =
1.0302576846298934
cluster 1: mean = 25.486654429329228, covariance =
0.998096618179134
cluster 2: mean = 15.449160792031094, covariance =
0.9671159353647132
iteration 8:
cluster 0: mean = 25.486654429329228, covariance =
0.9980966181791335
cluster 1: mean = 5.50927940017642, covariance =
1.0302576846296718
cluster 2: mean = 15.449160792031073, covariance =
0.9671159353647738
iteration 9:
cluster 0: mean = 25.486654429329192, covariance =
0.9980966181792538
```

```
cluster 1: mean = 15.449160792040155, covariance =
0.967115935334775
cluster 2: mean = 5.50927940019324, covariance =
1.0302576847353861
So the maximum log likelihood is: -2.5166286249411267
```

The results from initiation type 2:

```
iteration 0:
cluster 0: mean = 25.486654429329228, covariance =
0.9980966181791335
cluster 1: mean = 5.50927940017642, covariance =
1.0302576846296758
cluster 2: mean = 15.449160792031075, covariance =
0.9671159353647717
iteration 1:
cluster 0: mean = 25.48665442932923, covariance =
0.9980966181791533
cluster 1: mean = 15.449160792031078, covariance =
0.9671159353647346
cluster 2: mean = 5.509279400176431, covariance =
1.0302576846297442
iteration 2:
cluster 0: mean = 5.50927940017642, covariance =
1.0302576846296736
cluster 1: mean = 25.486654429329228, covariance =
0.9980966181791341
cluster 2: mean = 15.449160792031073, covariance =
0.9671159353647727
iteration 3:
cluster 0: mean = 25.486654429329228, covariance =
0.9980966181791338
cluster 1: mean = 15.449160792031085, covariance =
0.967115935364743
cluster 2: mean = 5.509279400176437, covariance =
1.030257684629784
iteration 4:
cluster 0: mean = 5.50927940017642, covariance =
1.0302576846296794
cluster 1: mean = 15.449160792031075, covariance =
0.9671159353647695
cluster 2: mean = 25.486654429329224, covariance =
0.9980966181791358
iteration 5:
```

```
cluster 0: mean = 25.486654429329228, covariance =
0.9980966181791346
cluster 1: mean = 15.449160792031073, covariance =
0.9671159353647716
cluster 2: mean = 5.50927940017642, covariance =
1.0302576846296756
iteration 6:
cluster 0: mean = 15.449160792031076, covariance =
0.9671159353647657
cluster 1: mean = 5.509279400176425, covariance =
1.0302576846296994
cluster 2: mean = 25.486654429329228, covariance =
0.9980966181791335
iteration 7:
cluster 0: mean = 25.486654429329228, covariance =
0.9980966181791341
cluster 1: mean = 15.449160792031106, covariance =
0.9671159353646743
cluster 2: mean = 5.509279400176476, covariance =
1.0302576846300269
iteration 8:
cluster 0: mean = 25.486654429329228, covariance =
0.9980966181791329
cluster 1: mean = 15.44916079203107, covariance =
0.9671159353648017
cluster 2: mean = 5.509279400176405, covariance =
1.0302576846295872
iteration 9:
cluster 0: mean = 5.509279400177054, covariance =
1.0302576846336637
cluster 1: mean = 25.48665442932923, covariance =
0.9980966181791391
cluster 2: mean = 15.44916079203141, covariance =
0.9671159353636427
So the maximum log likelihood is: -2.5166286249411227
```

Question 2: variance =1

The results from initiation type 1:

```
iteration 0:
cluster 0: mean = 5.509279400176421, covariance =
1.0302576846296811
cluster 1: mean = 15.449160792031075, covariance =
0.9671159353647703
cluster 2: mean = 25.486654429329228, covariance =
```

0.9980966181791335
iteration 1:
cluster 0: mean = 15.449160792031071, covariance =
0.9671159353648038
cluster 1: mean = 5.509279400176404, covariance =
1.03025768462958
cluster 2: mean = 25.486654429329228, covariance =
0.9980966181791324
iteration 2:
cluster 0: mean = 15.449160792031776, covariance =
0.9671159353576232
cluster 1: mean = 5.50927940017866, covariance =
1.0302576846436757
cluster 2: mean = 25.486654429328667, covariance =
0.9980966181823033
iteration 3:
cluster 0: mean = 25.486654429329228, covariance =
0.9980966181791335
cluster 1: mean = 15.449160792031073, covariance =
0.9671159353647738
cluster 2: mean = 5.50927940017642, covariance =
1.0302576846296718
iteration 4:
cluster 0: mean = 5.50927940017642, covariance =
1.0302576846296723
cluster 1: mean = 15.449160792031073, covariance =
0.9671159353647737
cluster 2: mean = 25.486654429329228, covariance =
0.9980966181791335
iteration 5:
cluster 0: mean = 25.74113493305759, covariance =
1.0111656572268515
cluster 1: mean = 25.30348063965848, covariance =
0.7423444394670239
cluster 2: mean = 10.646229338121298, covariance =
27.673368243825664
iteration 6:
cluster 0: mean = 20.306693235148316, covariance =
28.089908774475848
cluster 1: mean = 6.097955104270551, covariance =
0.5008835186162726
cluster 2: mean = 4.822099593352498, covariance =
0.6374875296804788
iteration 7:

```
cluster 0: mean = 5.5092794001765295, covariance =
1.0302576846303695
cluster 1: mean = 15.449160792031131, covariance =
0.9671159353645771
cluster 2: mean = 25.48665442932923, covariance =
0.9980966181791355
iteration 8:
cluster 0: mean = 5.5092794001764185, covariance =
1.030257684629660
cluster 1: mean = 15.449160792031073, covariance =
0.9671159353647768
cluster 2: mean = 25.486654429329228, covariance =
0.9980966181791336
iteration 9:
cluster 0: mean = 5.5092794001764, covariance =
1.030257684629557
cluster 1: mean = 25.486654429329228, covariance =
0.998096618179133
cluster 2: mean = 15.449160792031064, covariance =
0.9671159353648099
So the maximum log likelihood is: -2.5166286249411245
```

The results from initiation type 2:

```
iteration 0:
cluster 0: mean = 25.48665442932922, covariance =
0.9980966181791667
cluster 1: mean = 15.449160792031078, covariance =
0.9671159353647095
cluster 2: mean = 5.509279400176437, covariance =
1.0302576846297813
iteration 1:
cluster 0: mean = 5.509279400176415, covariance =
1.0302576846296427
cluster 1: mean = 25.486654429329228, covariance =
0.9980966181791335
cluster 2: mean = 15.449160792031071, covariance =
0.9671159353647825
iteration 2:
cluster 0: mean = 25.486654429329228, covariance =
0.9980966181791335
cluster 1: mean = 15.449160792031071, covariance =
0.9671159353647825
cluster 2: mean = 5.509279400176415, covariance =
```

1.0302576846296427
iteration 3:
cluster 0: mean = 5.509279400176415, covariance =
1.0302576846296456
cluster 1: mean = 25.486654429329228, covariance =
0.9980966181791335
cluster 2: mean = 15.449160792031073, covariance =
0.9671159353647814
iteration 4:
cluster 0: mean = 25.486654429329228, covariance =
0.9980966181791335
cluster 1: mean = 5.509279400176415, covariance =
1.0302576846296447
cluster 2: mean = 15.449160792031071, covariance =
0.9671159353647817
iteration 5:
cluster 0: mean = 25.486654429329228, covariance =
0.9980966181791335
cluster 1: mean = 15.449160792031071, covariance =
0.9671159353647825
cluster 2: mean = 5.509279400176415, covariance =
1.0302576846296427
iteration 6:
cluster 0: mean = 5.509279400176415, covariance =
1.0302576846296427
cluster 1: mean = 25.486654429329228, covariance =
0.9980966181791335
cluster 2: mean = 15.449160792031071, covariance =
0.9671159353647825
iteration 7:
cluster 0: mean = 25.486654429329228, covariance =
0.9980966181791335
cluster 1: mean = 5.509279400176414, covariance =
1.0302576846296396
cluster 2: mean = 15.449160792031073, covariance =
0.9671159353647835
iteration 8:
cluster 0: mean = 25.486654429329228, covariance =
0.9980966181791335
cluster 1: mean = 5.509279400176415, covariance =
1.0302576846296425
cluster 2: mean = 15.449160792031071, covariance =
0.9671159353647825
iteration 9:

```
cluster 0: mean = 25.486654429329228, covariance =
0.9980966181791335
cluster 1: mean = 5.509279400176415, covariance =
1.0302576846296463
cluster 2: mean = 15.449160792031073, covariance =
0.9671159353647812
So the maximum log likelihood is: -2.516628624941133
```

Conclusions:
1 Random initialization is more sensitive to the initialization settings and parameters than the K-means initialization. For random initialization, sometimes we can see means like {20, 6, 4}, (iteration 6 for variance =1) which means that it will sometimes get trapped in local minima. But the K-means methods do not show these exceptions.
2 With a low thresh hold value, the maximum log likehood of different methods seem to be similar. So initialization types do not influence the best result.
3 Changing covariance to 1 seems to have no significant impact on results.

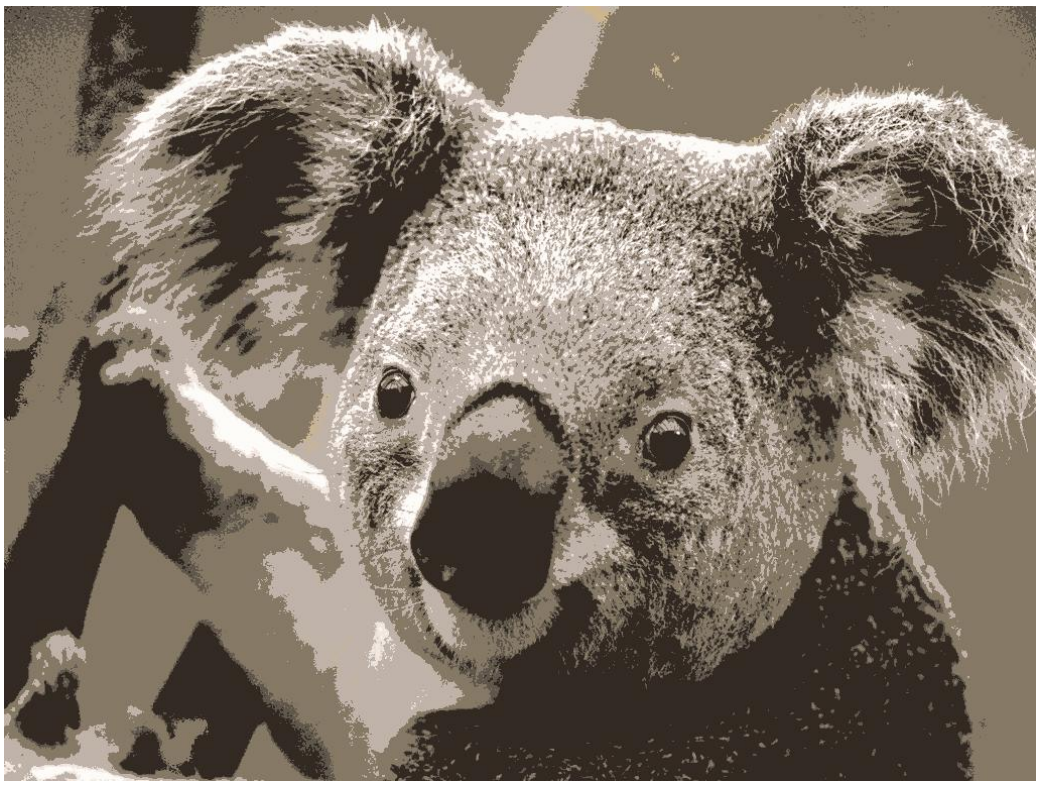
K-means clustering on images

Question 1:

1 Koala.jpg
K=2

K=5

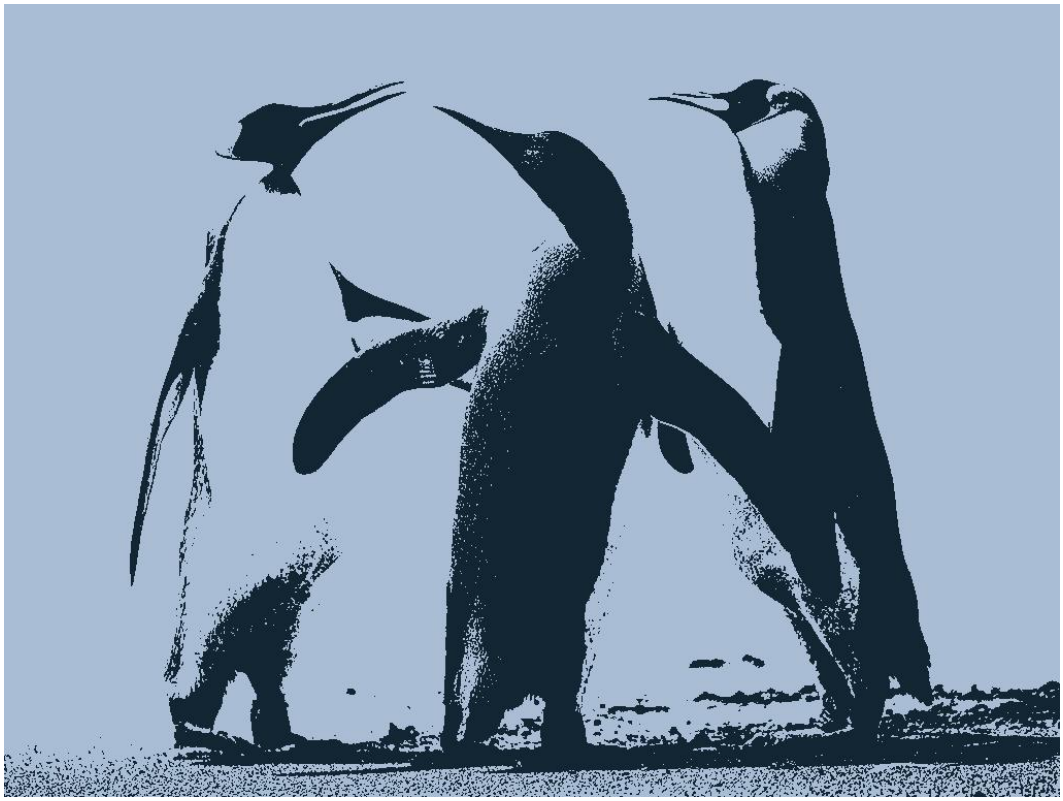
K=10



K=15

K=20

2 Penguin.jpg
K=2



K=5

K=10

K=15

K=20

Question 2:

Data compression ratio is defined as the ratio between the uncompressed size and compressed size:

$$\text{Compression Ratio} = \frac{\text{Uncompressed Size}}{\text{Compressed Size}}$$

1 Koala.jpg

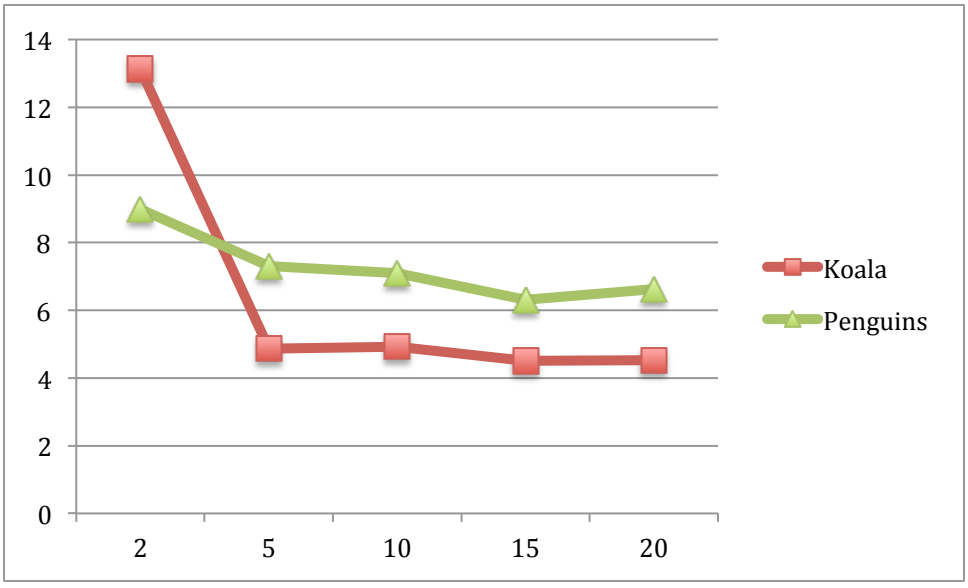
```
For K = 2
Compression rate mean is: 13.14043774586132, variance is:
256.69430051616257
For K = 5
Compression rate mean is: 4.867607195783325, variance is:
0.22295288372483496
For K = 10
Compression rate mean is: 4.922549948253031, variance is:
0.3230046888800588
For K = 15
Compression rate mean is: 4.504286506348433, variance is:
0.028488236692057016
For K = 20
```

```
Compression rate mean is: 4.536089016906287, variance is:
0.03605304173610866
```

2 Penguin.jpg

```
For K = 2
Compression rate mean is: 8.991735625412726, variance is:
1.5085747120884152
For K = 5
Compression rate mean is: 7.311607595738292, variance is:
0.7556654089492489
For K = 10
Compression rate mean is: 7.106388555733014, variance is:
0.24126482107265224
For K = 15
Compression rate mean is: 6.313097697483055, variance is:
0.07536079469821759
For K = 20
Compression rate mean is: 6.632673819804009, variance is:
0.07435799046773724
```

Question 3:



Yes, there is a tradeoff between image quality and degree of compression. The higher image quality, the lower the compression rate.

I plot the mean of compression rate v.s. K value. The mean compression rate changes with pictures. For each picture we can still find the point K=10 that achieve a compression rate that is close to the best one and with the smallest K value. So I think K =10 is the best K value.