Pranith Mullapudi pxm220087
CS6375.001

Project 4 Report

Data tables and plots are below, plots are also in the folders alongside the code.

Part 1 Discussion:

There wasn't a huge difference between the initialization types in this project, maybe because of the relatively low number of samples, how they were created, or just not enough models run. But between the dozen or so times I ran the models, both random and Kmeans++ both either converged to the same result or occasionally had slight differences but still very similar. There were only a handful of occasions where there was a big difference and those were when the model had more clusters than the "true" model (5 vs 3 for example). In those cases there was more room for slight differences to make an impact since the model had to split clusters into parts.

The Blob Dataset is by far better for kmeans as expected with much better silhouette scores across the board when compared to the moon dataset. This makes sense since kmeans has the limitation of only being able to have equal sized, circular clusters which doesn't work well for a shape like moons.

The choice of K affects the quality since too few or too many clusters both lead to problems of either one center having to span multiple clusters or multiple centers having to split one cluster. As a result, the best case is when the number of clusters the model assumes matches the true number of clusters as closely as possible.

Part 2 Discussion:

GMMs handle non circular clusters much better since they can actually form clusters that aren't circular through the covariance, this is shown off by the difference between GMM and Kmeans on the moon dataset where you can see in the plots how the GMM clusters are able to much better follow the curves.

EM outperforms Kmeans when the data is non convex/the clusters aren't circular. In addition, it also better models cases where it makes more sense for a given datapoint to belong to multiple clusters at the same time since multiple clusters can have the same point in their distribution.

The covariance type determines the shape by making it so the foci of the ellipses have to be aligned with either of the axis if you choose diag this makes it worse at fitting data.This can be seen in the plots for the moon set where the full covariance right side does a much better job at representing the moon shape since the ellipses can angle themselves along the curve of the moons.

Result Tables:

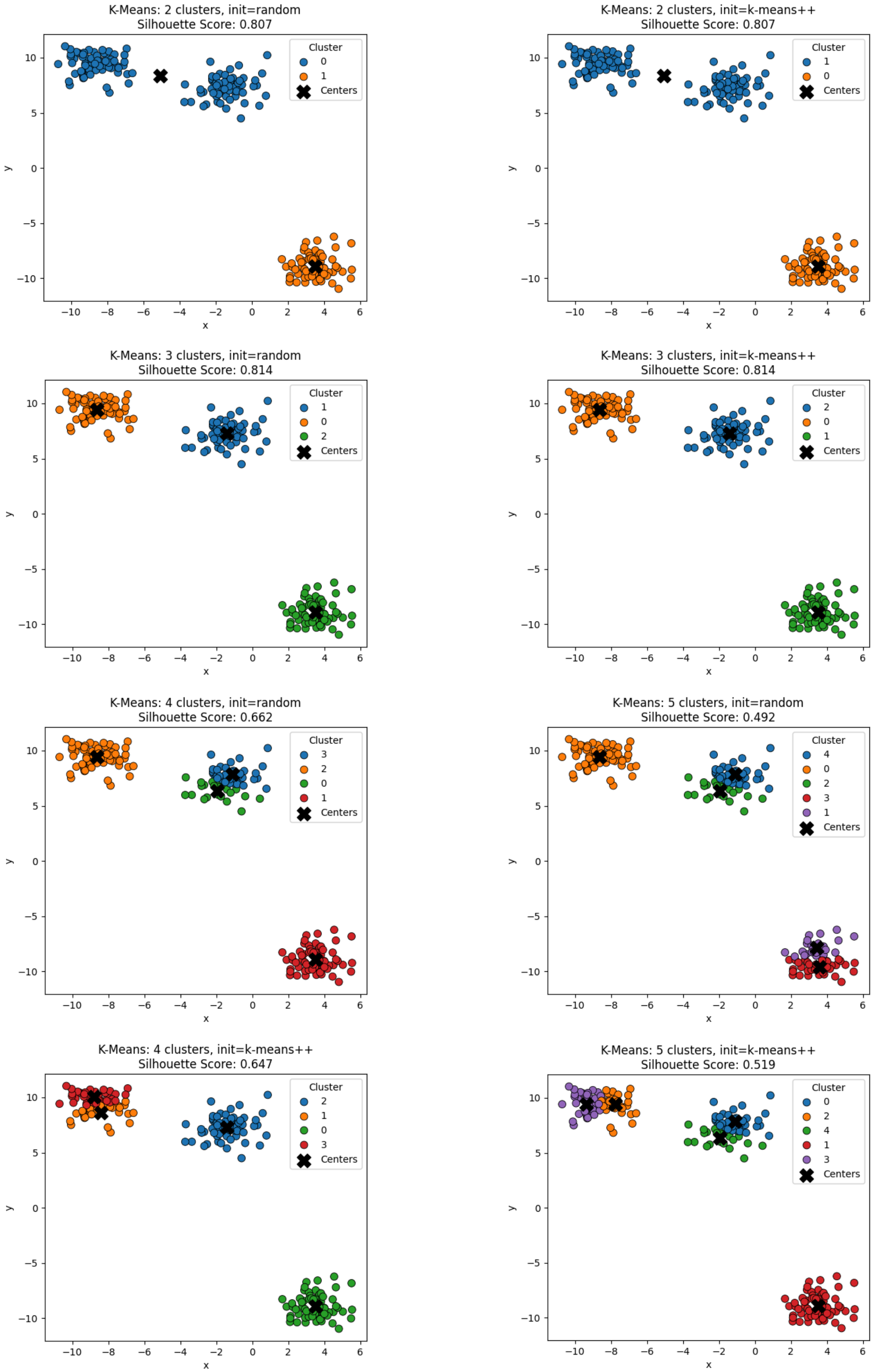
KMeans on Blob

Num Clusters	Initialization type	Silhouette Score
2	kmeans++	0.807
2	random	0.807
3	kmeans++	0.814
3	random	0.814
4	kmeans++	0.647
4	random	0.662
5	kmeans++	0.519
5	random	0.492

Kmeans on Moons

Num Clusters	Initialization type	Silhouette Score
2	random	0.489
2	kmeans++	0.489
3	random	0.422
3	kmeans++	0.423
4	random	0.458
4	kmeans++	0.457
5	random	0.486
5	kmeans++	0.487

Kmeans on Blobs, rows are cluster counts, left is rand, right is kmean++



K-Means: 2 clusters, init=random
Silhouette Score: 0.807

K-Means: 2 clusters, init=k-means++
Silhouette Score: 0.807

K-Means: 3 clusters, init=random
Silhouette Score: 0.814

K-Means: 3 clusters, init=k-means++
Silhouette Score: 0.814

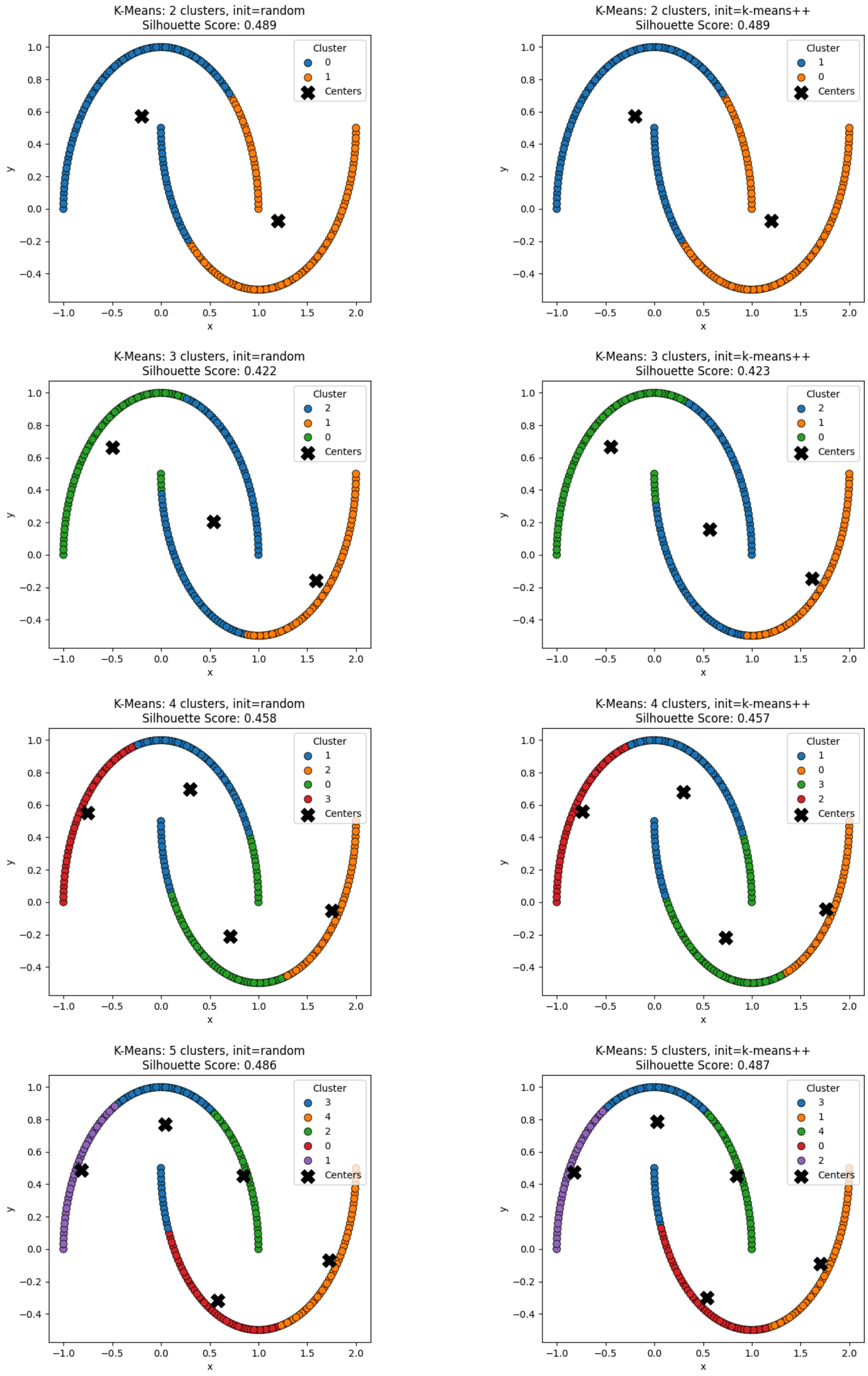
K-Means: 4 clusters, init=random
Silhouette Score: 0.662

K-Means: 5 clusters, init=random
Silhouette Score: 0.492

K-Means: 4 clusters, init=k-means++
Silhouette Score: 0.647

K-Means: 5 clusters, init=k-means++
Silhouette Score: 0.519

Kmeans on Moon plots, rows are cluster count, left is rand, right is kmeans++



K-Means: 2 clusters, init=random
Silhouette Score: 0.489

K-Means: 2 clusters, init=k-means++
Silhouette Score: 0.489

K-Means: 3 clusters, init=random
Silhouette Score: 0.422

K-Means: 3 clusters, init=k-means++
Silhouette Score: 0.423

K-Means: 4 clusters, init=random
Silhouette Score: 0.458

K-Means: 4 clusters, init=k-means++
Silhouette Score: 0.457

K-Means: 5 clusters, init=random
Silhouette Score: 0.486

K-Means: 5 clusters, init=k-means++
Silhouette Score: 0.487

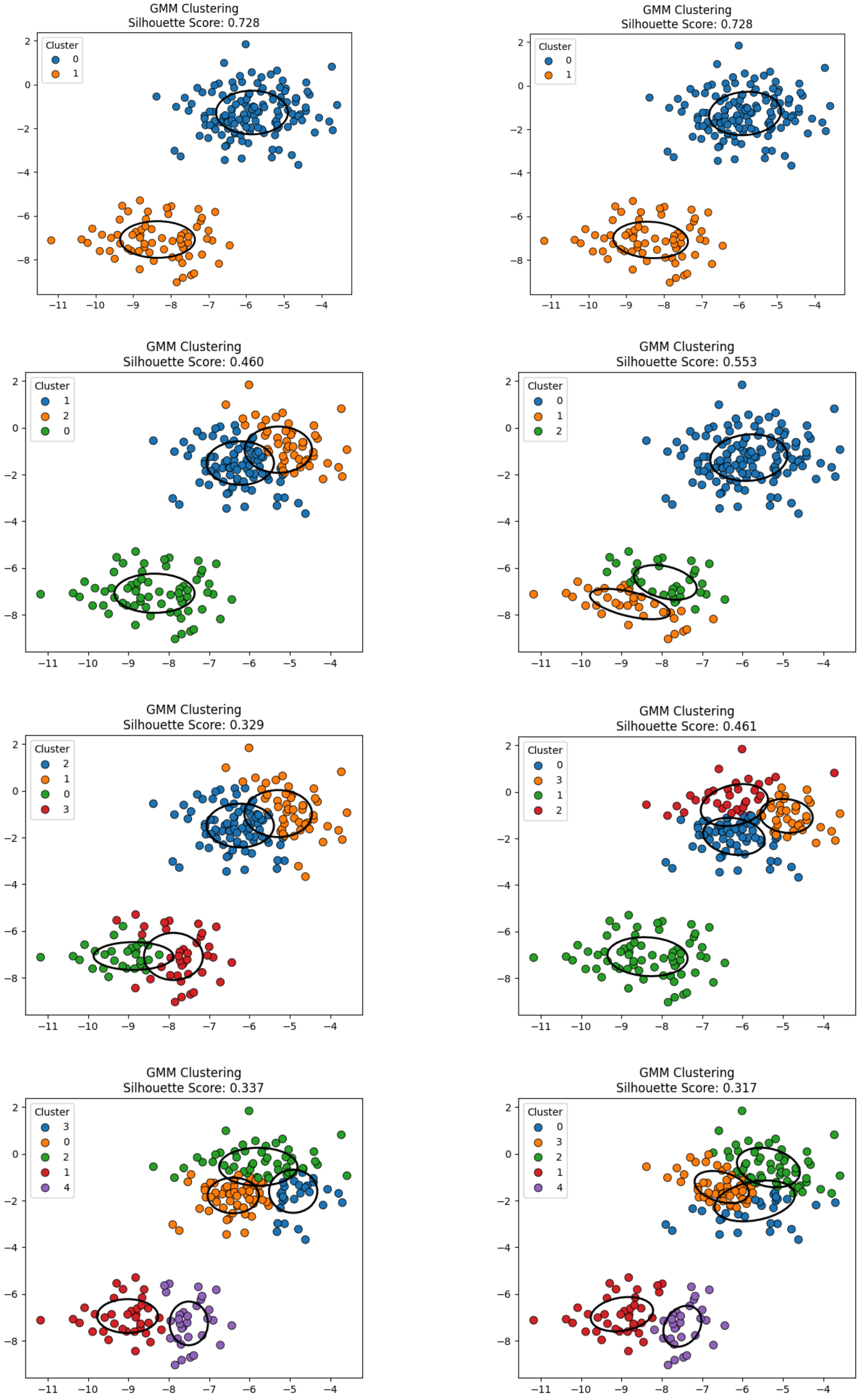
Gaussian EM on Blobs

num clusters	covariance type	silhouette score	avg log likelihood	bic
2	full	0.728	-3.374	1407.932
2	diag	0.728	-3.378	1398.943
3	full	0.553	-3.360	1434.134
3	diag	0.460	-3.375	1424.015
4	full	0.461	-3.370	1469.880
4	diag	0.329	-3.363	1445.781
5	full	0.219	-3.343	1490.925
5	diag	0.337	-3.351	1467.669

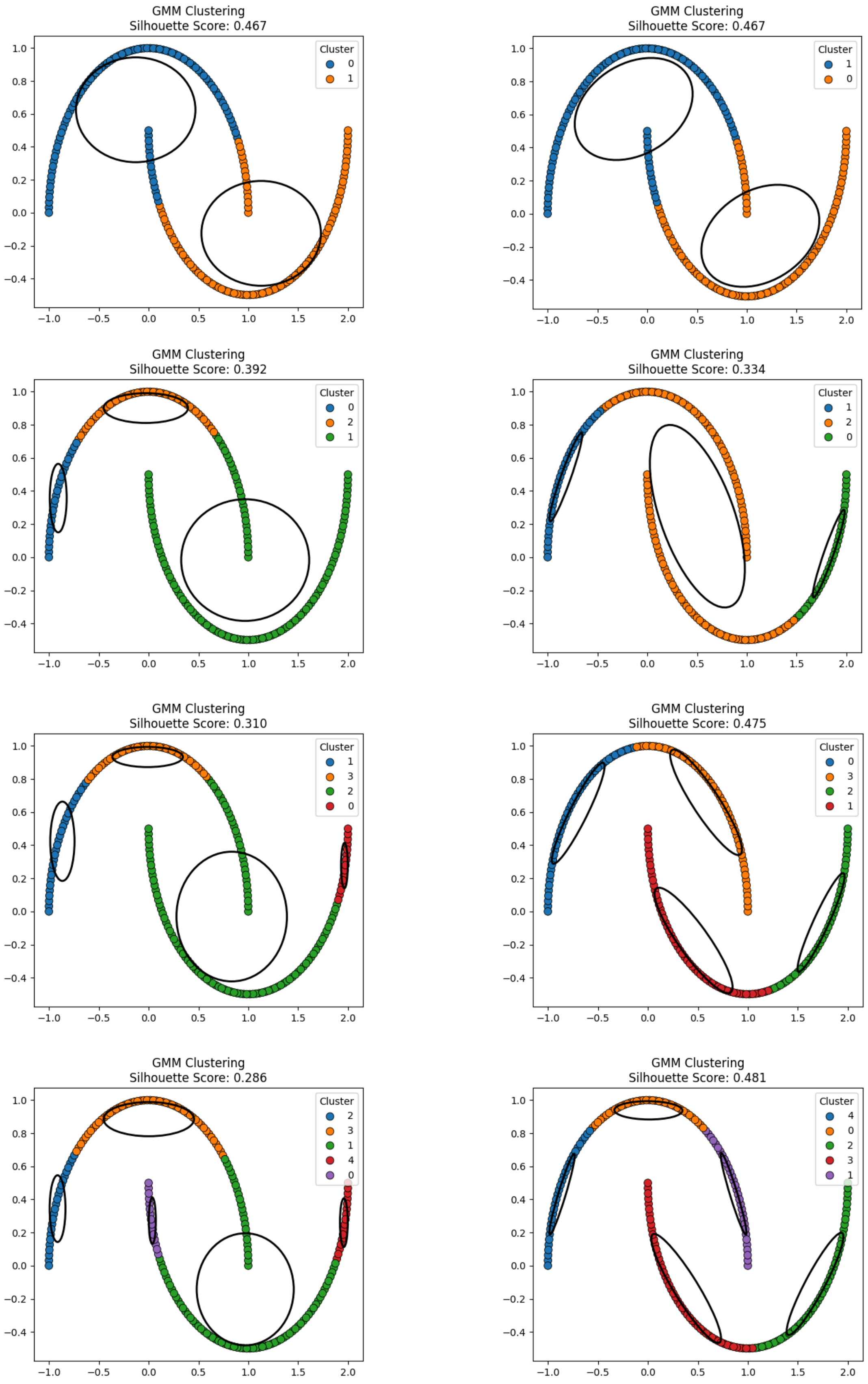
Gaussian EM on Moons

num clusters	covariance type	silhouette score	avg log likelihood	bic
2	full	0.467	-1.699	737.882
2	diag	0.467	-1.738	742.928
3	full	0.334	-1.201	570.312
3	diag	0.392	-1.470	662.368
4	full	0.475	-0.964	507.577
4	diag	0.310	-1.321	628.925
5	full	0.481	-0.582	386.315
5	diag	0.286	-1.176	597.543

Gaussian Mixture Model on Blobs, each row is cluster count, each column is covar type



Gaussian Mixture Model on Moons, each row is cluster count, each column is covar type


GMM Clustering
Silhouette Score: 0.467
Cluster
GMM Clustering
Silhouette Score: 0.467
Cluster
GMM Clustering
Silhouette Score: 0.392
Cluster
GMM Clustering
Silhouette Score: 0.334
Cluster
GMM Clustering
Silhouette Score: 0.310
Cluster
GMM Clustering
Silhouette Score: 0.475
Cluster
GMM Clustering
Silhouette Score: 0.286
Cluster
GMM Clustering
Silhouette Score: 0.481
Cluster