REPORT

Project 4: K-Means Clustering and Expectation-Maximization for Gaussian Mixture Models

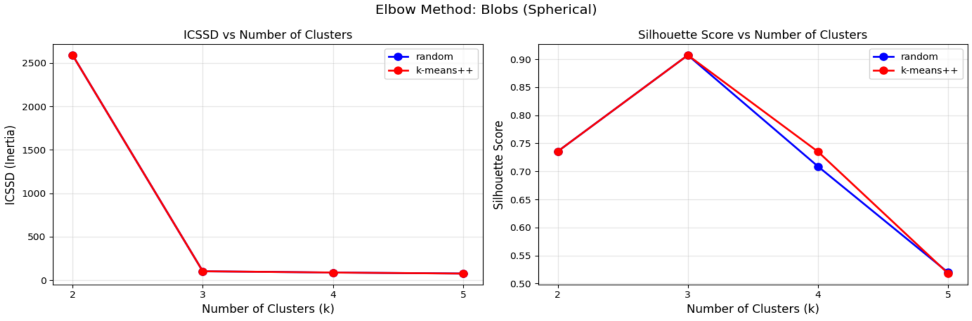
PART 1 (K-means Clustering)

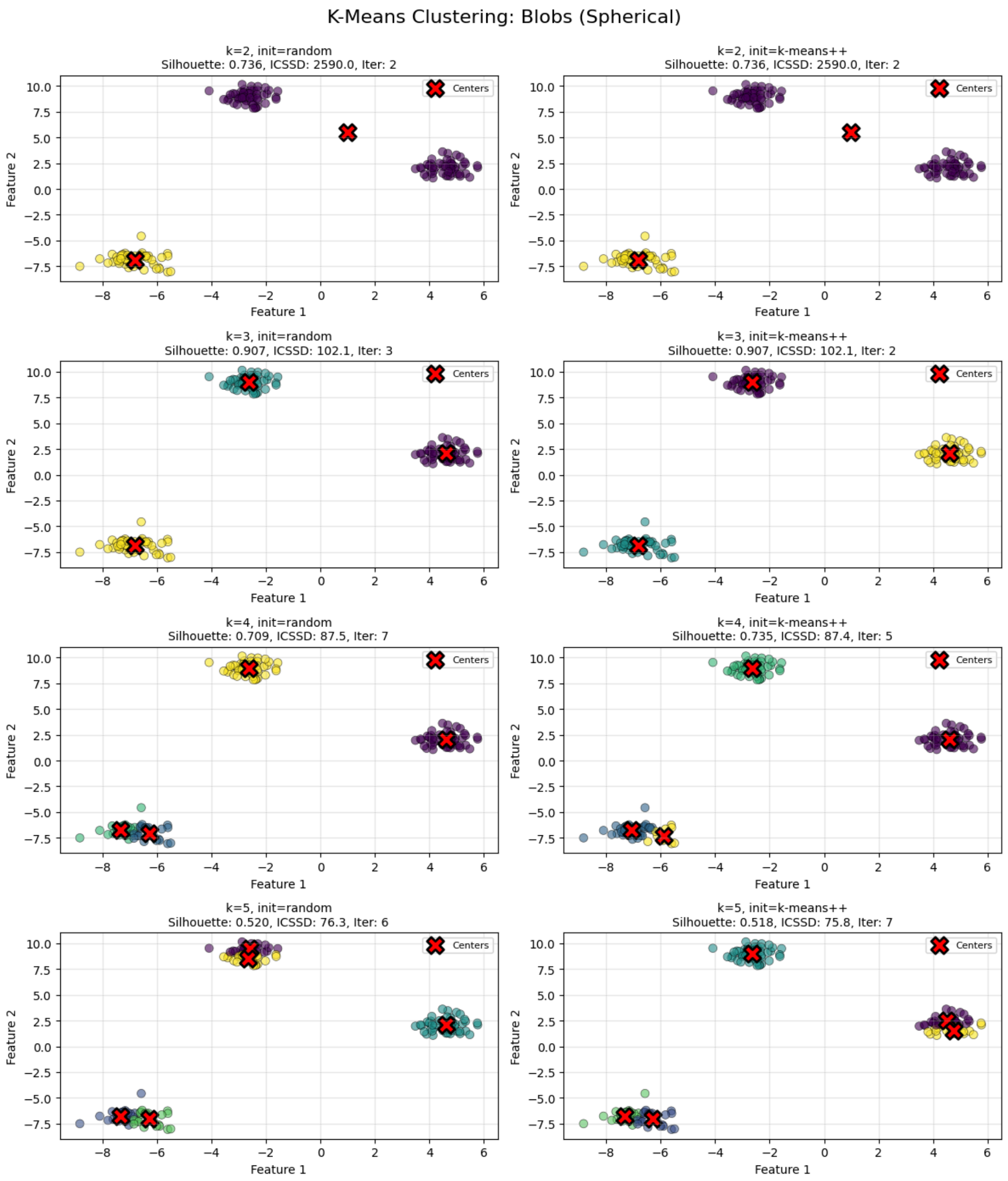
Table with Silhouette scores and ICSSD values using K-Means

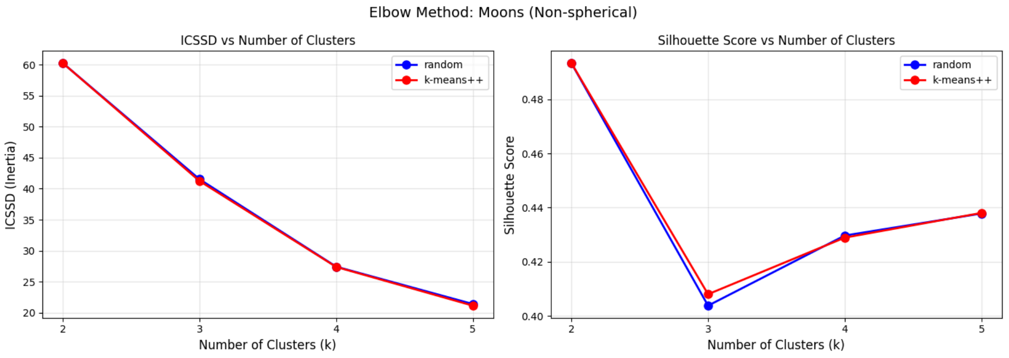
Dataset	k	Silhouette (Random)	Silhouette (k-means++)	ICSSD (Random)	ICSSD (k-means++)
Blobs	2	0.7355	0.7355	2589.97	2589.97
	3	0.9072	0.9072	102.10	102.10
	4	0.7088	0.7351	87.48	87.43
	5	0.5197	0.5178	76.34	75.82
Moons	2	0.4934	0.4934	60.27	60.27
	3	0.4038	0.4080	41.53	41.24
	4	0.4297	0.4289	27.42	27.38
	5	0.4378	0.4380	21.39	21.13

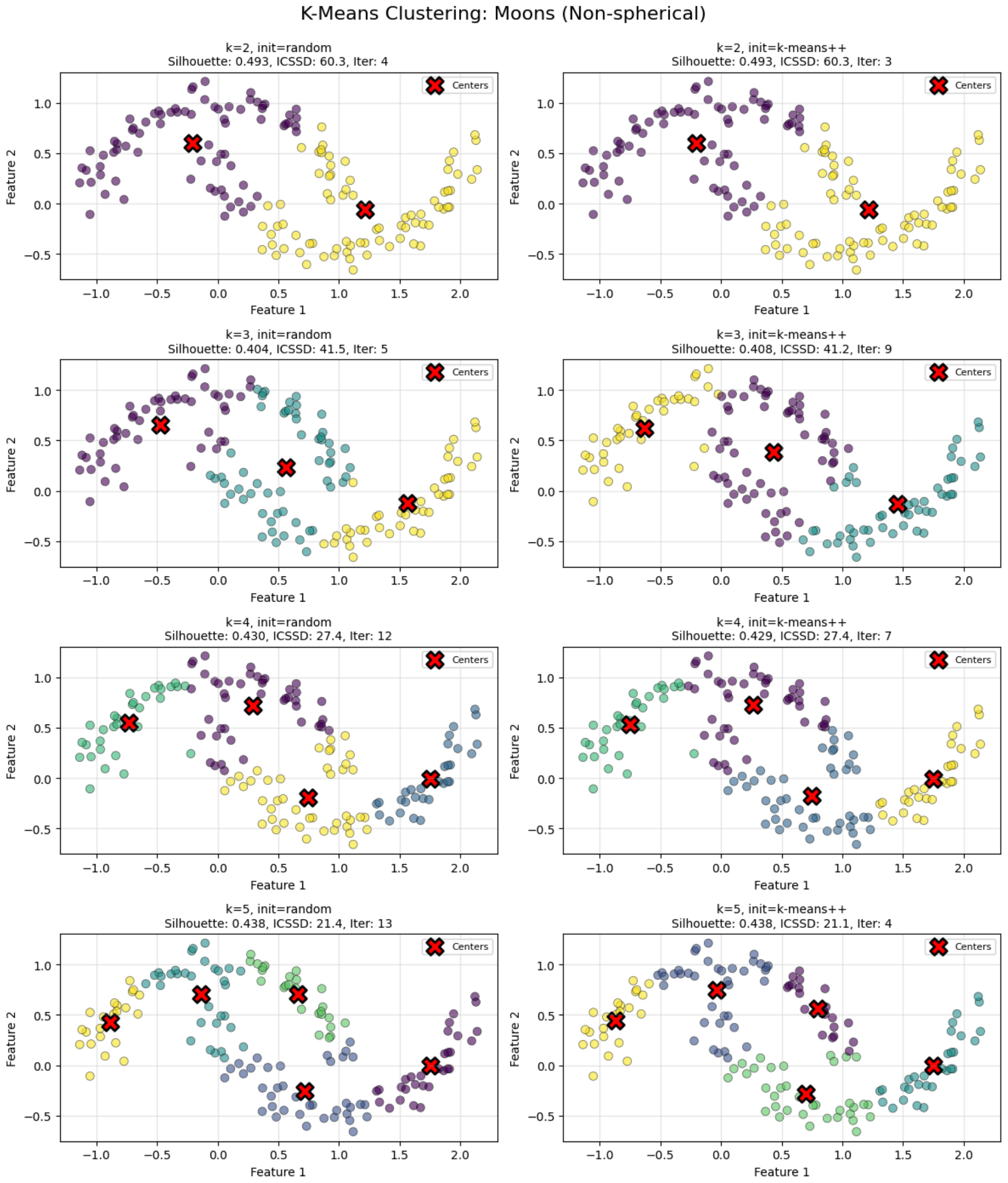
- The Blobs dataset: Clear optimal clustering at **k = 3**, where the silhouette score peaks **(0.9072)** and ICSSD drops sharply, confirming well-defined spherical clusters.
- The Moons dataset: The Silhouette scores remain low across all k values and ICSSD decreases smoothly.
- Initialization effects are minimal, with nearly identical scores between random and k-means++.

KEY PLOTS: Scatterplots of clusters for K-Means




K-Means Clustering: Blobs (Spherical)
k=2, init=random
Silhouette: 0.736, ICSSD: 2590.0, Iter: 2
k=2, init=k-means++
Silhouette: 0.736, ICSSD: 2590.0, Iter: 2
k=3, init=random
Silhouette: 0.907, ICSSD: 102.1, Iter: 3
k=3, init=k-means++
Silhouette: 0.907, ICSSD: 102.1, Iter: 2
k=4, init=random
Silhouette: 0.709, ICSSD: 87.5, Iter: 7
k=4, init=k-means++
Silhouette: 0.735, ICSSD: 87.4, Iter: 5
k=5, init=random
Silhouette: 0.520, ICSSD: 76.3, Iter: 6
k=5, init=k-means++
Silhouette: 0.518, ICSSD: 75.8, Iter: 7
Centers
Feature 1
Feature 2
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−8
−6
−4
−2
0
2
4
6


Elbow Method: Moons (Non-spherical)
ICSSD vs Number of Clusters
ICSSD (Inertia)
60
55
50
45
40
35
30
25
20
random
k-means++
Number of Clusters (k)
2
3
4
5
Silhouette Score vs Number of Clusters
Silhouette Score
0.48
0.46
0.44
0.42
0.40
random
k-means++
Number of Clusters (k)
2
3
4
5



DISCUSSIONS

1. Compare the effect of different initializations

Across both datasets, initialization had minimal practical impact on clustering quality, though k-means++ consistently improved convergence stability.

Blobs dataset:

In both initializations, the Silhouette scores for k = 2 and k = 3 were the same, 0.7355 and 0.9072, respectively, demonstrating that both techniques converged to virtually the same final clusters. The values of ICSSD were also nearly identical, with similar local minima. k-means++ typically required fewer iterations, reflecting better starting centroids.

Moons dataset:

The Silhouette score differences were extremely small across all k (differences < 0.005), meaning initialization did not affect the clustering outcome. k-means++ sometimes converged faster, but overall K-Means remained unsuitable for this dataset due to its non-convex structure.

Comparison:
k-means++ gives slightly faster, more stable convergence, but cluster quality stayed nearly the same for both initialization methods. For well-structured data (Blobs), both perform equally well; for non-convex data (Moons), neither initialization overcomes K-Means' geometric limitations.

2. **Which dataset is better suited for K-Means, and why?**

   The Blobs dataset is a much better fit for K-Means because its clusters are naturally spherical and well-separated. This is reflected in the consistently high Silhouette scores (peaking at 0.907) and the clear elbow at k = 3, showing that K-Means can recover the true structure easily.
   In contrast, the Moons dataset is non-convex and crescent-shaped, which K-Means cannot model using simple Euclidean distance. As a result, the algorithm splits the curved clusters incorrectly, leading to noticeably lower Silhouette scores, around 0.40–0.49, and no meaningful elbow. This demonstrates that Moons dataset is poorly suited for K-Means.
   Thus, Blobs dataset is far better suited for K-Means clustering.

3. **How does the choice of k affect clustering quality?**

   The effect of k is clearly visible in both datasets. For the Blobs data, clustering quality improves as k increases from 2 to 3, where the Silhouette score peaks at 0.9072, matching the true number of clusters. The ICSSD value also drops sharply from 2589.97 to 102.10, and the elbow in the curve becomes evident at k=3, which indicates that this is the most meaningful segmentation. Increasing k values greater than 3 only splits natural clusters, which lowers the Silhouette score to 0.7088 and 0.5197, showing signs of over-segmentation.

   For the Moons dataset, changing k does not meaningfully improve cluster quality because K-Means cannot represent the curved cluster shapes. Silhouette scores remain low across all k (ranging only from 0.4038 to 0.4934), and although ICSSD decreases with higher k, there is no clear elbow, confirming that increasing k does not recover the true structure.

   Overall, the optimal k reflects the underlying geometry: k=3 works best for Blobs, while no value of k produces good separation for Moons due to its non-convex shape.

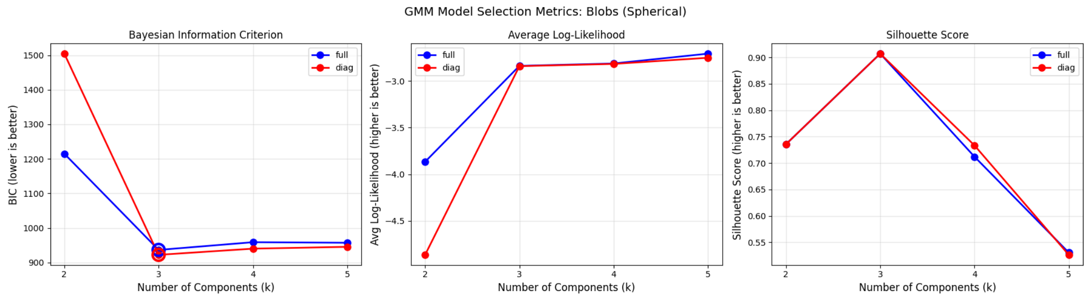
PART 2 (GAUSSIAN MIXTURE MODELS (EM))

Table with Silhouette Scores, Average Log-Likelihood, and BIC for GMM (Both Datasets, Both Covariance Types)

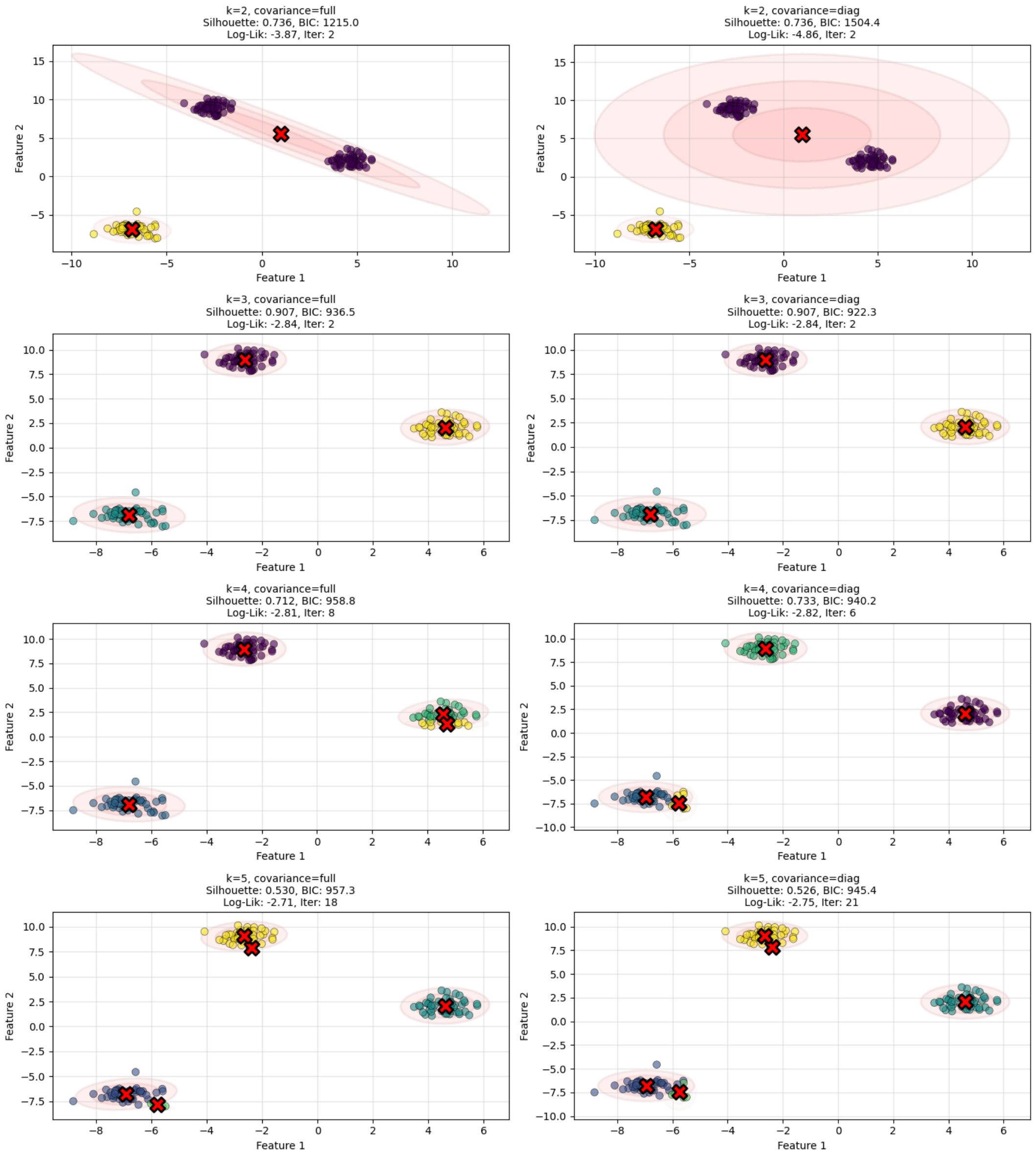
Dataset	**k**	**Covariance**	**Silhouette**	**Avg Log-Likelihood**	**BIC**
Blobs	2	full	0.7355	-3.8662	1214.97
		diag	0.7355	-4.8645	1504.45
	3	full	0.9072	-2.8378	936.52
		diag	0.9072	-2.8405	922.30
	4	full	0.7120	-2.8120	958.84
		diag	0.7334	-2.8166	940.17
	5	full	0.5305	-2.7065	957.25
		diag	0.5262	-2.7506	945.44
Moons	2	full	0.4673	-1.7238	572.27
		diag	0.4673	-1.7665	575.05
	3	full	0.3858	-1.4618	523.71
		diag	0.3795	-1.6110	553.46
	4	full	0.4472	-1.2795	499.09
		diag	0.3535	-1.5330	555.11
	5	full	0.4483	-1.1573	492.49
		diag	0.3767	-1.4172	545.42

- The Blobs dataset: Best clustering at k = 3, where both covariance types achieve the highest silhouette score (0.9072) and BIC reaches its minimum, confirming it as the optimal model.
- The Moons dataset: Full covariance consistently outperforms diag, achieving higher silhouette scores and better log-likelihood, with BIC lowest at k = 5, indicating that multiple elliptical components are needed to approximate the curved structure.
- Overall, GMM provides more flexibility than K-Means, especially for non-spherical data.

KEY PLOTS: Scatterplots of clusters for EM



Gaussian Mixture Models (EM): Blobs (Spherical)



k=2, covariance=full
Silhouette: 0.736, BIC: 1215.0
Log-Lik: -3.87, Iter: 2

k=2, covariance=diag
Silhouette: 0.736, BIC: 1504.4
Log-Lik: -4.86, Iter: 2

k=3, covariance=full
Silhouette: 0.907, BIC: 936.5
Log-Lik: -2.84, Iter: 2

k=3, covariance=diag
Silhouette: 0.907, BIC: 922.3
Log-Lik: -2.84, Iter: 2

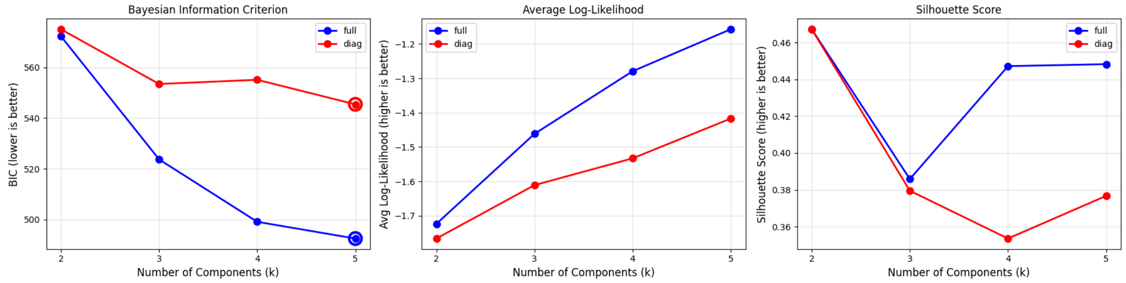
k=4, covariance=full
Silhouette: 0.712, BIC: 958.8
Log-Lik: -2.81, Iter: 8

k=4, covariance=diag
Silhouette: 0.733, BIC: 940.2
Log-Lik: -2.82, Iter: 6

k=5, covariance=full
Silhouette: 0.530, BIC: 957.3
Log-Lik: -2.71, Iter: 18

k=5, covariance=diag
Silhouette: 0.526, BIC: 945.4
Log-Lik: -2.75, Iter: 21

GMM Model Selection Metrics: Moons (Non-spherical)

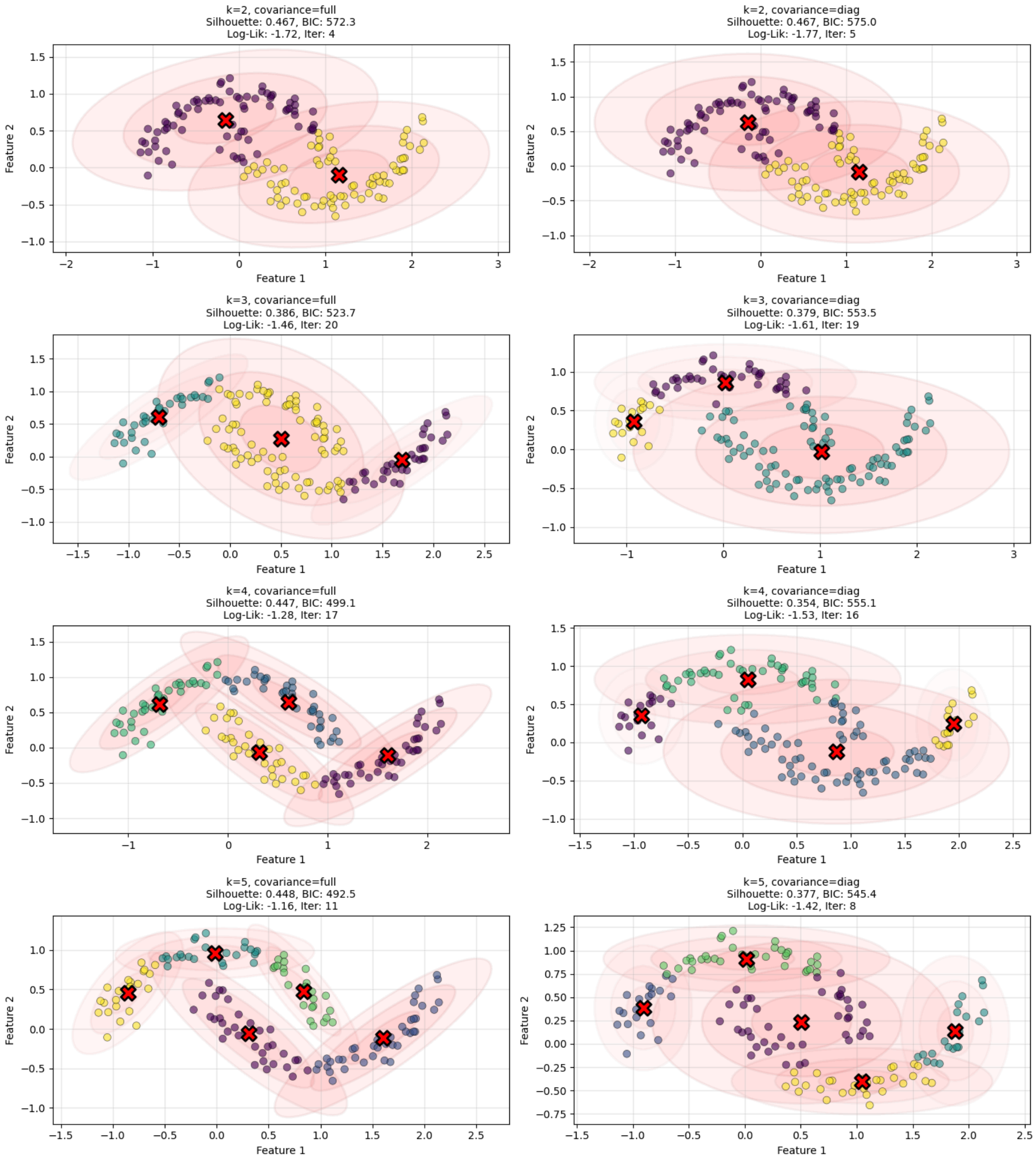


Bayesian Information Criterion

Average Log-Likelihood

Silhouette Score

Gaussian Mixture Models (EM): Moons (Non-spherical)



k=2, covariance=full
Silhouette: 0.467, BIC: 572.3
Log-Lik: -1.72, Iter: 4

k=2, covariance=diag
Silhouette: 0.467, BIC: 575.0
Log-Lik: -1.77, Iter: 5

k=3, covariance=full
Silhouette: 0.386, BIC: 523.7
Log-Lik: -1.46, Iter: 20

k=3, covariance=diag
Silhouette: 0.379, BIC: 553.5
Log-Lik: -1.61, Iter: 19

k=4, covariance=full
Silhouette: 0.447, BIC: 499.1
Log-Lik: -1.28, Iter: 17

k=4, covariance=diag
Silhouette: 0.354, BIC: 555.1
Log-Lik: -1.53, Iter: 16

k=5, covariance=full
Silhouette: 0.448, BIC: 492.5
Log-Lik: -1.16, Iter: 11

k=5, covariance=diag
Silhouette: 0.377, BIC: 545.4
Log-Lik: -1.42, Iter: 8

DISCUSSIONS

1. Analyze how GMMs handle elliptical clusters better than K-Means, especially on non-convex datasets.

   GMMs outperform K-Means on datasets where clusters are not perfectly spherical because they model each cluster using a full covariance matrix, allowing ellipses of different shapes, sizes, and orientations. This flexibility is evident in the Moons dataset: while K-Means produces low Silhouette scores, with maximum around 0.49, and incorrectly splits the curved structures, GMMs achieve higher scores up to 0.4483 with full covariance and can approximate the crescent shapes by combining multiple Gaussian components. The ability of the full covariance GMM to rotate and stretch ellipses enables it to better follow the intrinsic geometry of the moons, something K-Means cannot do with its rigid circular boundaries. On the Blobs dataset, both methods perform very well because the clusters are already spherical. However, GMM still offers slightly more flexibility by adapting to mild shape variations and producing soft assignments rather than hard labels.

Overall, GMMs handle elliptical or uneven clusters more effectively, and although they still struggle with truly non-convex shapes, they consistently provide a better fit than K-Means on such datasets.

2. In what scenarios does EM outperform K-Means?

   EM (GMM) outperforms K-Means when clusters are elliptical, uneven in size, or overlapping, because it can model full covariance and provide soft, probabilistic assignments. This is the reason why it performs better on the Moons dataset, achieving higher log-likelihoods and better Silhouette scores than K-Means, forcing circular boundaries and splitting the curved structure incorrectly. EM is also superior when model selection is needed, since BIC provides a principled way to choose the number of components, which K-Means is unable to do.

3. How does the covariance type influence cluster shape and separation?

   The covariance type directly controls the flexibility of each Gaussian component and thus determines what cluster shapes the GMM can represent. With full covariance, the model can learn ellipses of any orientation; clusters can thus stretch and rotate, which also captures dependencies between features. This again translates into better separation of clusters in complex datasets, as reflected in higher log-likelihoods, e.g., Moons at k = 5: -1.1573 with full versus -1.4172 with diag, along with better Silhouette scores. Full covariance is a particularly potent method for non-convex or stretched structures, because it can approximate curved clusters by combining several oriented ellipses.

   In contrast, diag covariance restricts each component to axis-aligned ellipses with no feature correlation, making it less flexible. Although this simpler structure often yields lower BIC values due to fewer parameters, it fits more poorly, particularly on datasets with rotated or otherwise oddly-shaped clusters. This is clearly seen in some cases, for instance Moons at k=4, where the Silhouette score drops from 0.4472 (full) to 0.3535 (diag).

   Overall, full covariance allows for better modeling of cluster shapes and their separation, while diag covariance trades some flexibility for simplicity and better generalization when data is closer to axis-aligned structure.

OBSERVATIONS AND INTERPRETATIONS

- K-Means works well for the Blobs dataset because its assumption of spherical clusters is met in the data.

- Silhouette scores are highest for k = 3 (0.9072) and ICSSD has a clear elbow at that point, so the algorithm catches the true cluster structure.
- Scores plunge for k > 3, suggesting over-segmentation.
- Initialization makes little difference to the quality of the clusterings - silhouette values are almost identical, but k-means++ does converge faster and yields slightly more stable ICSSD values.
- By contrast, for the Moons dataset, Silhouette scores remain low regardless of k, and ICSSD decreases monotonically with no elbow; this shows that K-Means is incapable of capturing the curved, non-convex structure.
- For GMMs, the optimal k value is 3 for the Blobs dataset, with identical silhouette scores compared to K-Means, while having slightly better flexibility given that full covariance provides higher log-likelihood.
- BIC also chooses k = 3 (diag) as the best model, confirming that for spherical clusters, simple covariance structure already suffices.
- On the Moons dataset, GMM outperforms K-Means since its elliptical components can roughly approximate curved shapes: full covariance consistently yields higher silhouette scores (best 0.4483) and better log-likelihood with increasing k. Its BIC reaches its minimum at k = 5 (full), indicating that fitting this kind of crescent structure requires multiple oriented Gaussians.
- Overall, GMM copes with non-spherical geometry way better than K-Means; even so, EM can't capture highly non-convex shapes.