Clustering Analysis of Yeast Protein Localization

Group 6

April 3, 2025

1 Introduction

This report compares K-Means and Gaussian Mixture Models (GMM) for clustering yeast protein localization data. We evaluate performance through three metrics: Silhouette Score (SS), Davies-Bouldin Index (DBI), and Calinski-Harabasz Score (CHS). Our analysis reveals that K-Means outperforms GMM in both clustering quality and computational efficiency.

2 Objective

Our objective is to cluster Proteins into groups based on their attributes to identify localization patterns in within cells. To achieve our objectives we aim to find the most optimal number of clusters and the best performing distance metric. We find optimal number of clusters using the elbow method and confirm it using the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score.

3 Model Architecture: K-Means Algorithm

1. **Initialization**: Initialization is done by using greedy kmeans++ algorithm provided as a part of KMeans() function in sklearn.
2. **Assignment**: Cluster points using Euclidean distance.
3. **Update**: Recalculate centroids as the means of assigned clusters.
4. **Iteration**: Repeat until convergence (maximum 300 iterations).

4 Evaluation Metrics

Metric	**Range**	**Optimal**	**Description**
Silhouette Score (SS)	[-1, 1]	High	Measures cluster cohesion and separation.
Davies-Bouldin Index (DBI)	$[0, \infty)$	Low	Ratio of within-cluster to between-cluster distances.
Calinski-Harabasz Score (CHS)	$[0, \infty)$	High	Ratio of between-cluster to within-cluster dispersion.

5 Results

5.1 Algorithm Comparison

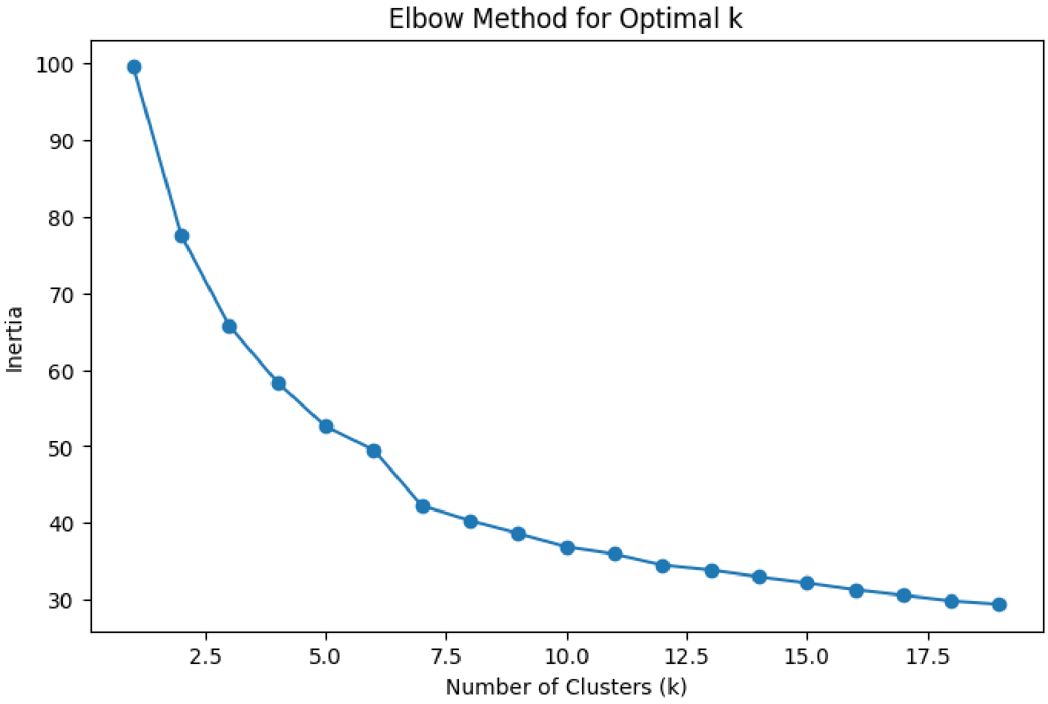
Method	SS	DBI	CHS
K-Means	**0.21**	**1.27**	**265**
GMM	0.05	2.45	75

5.2 Key Observations

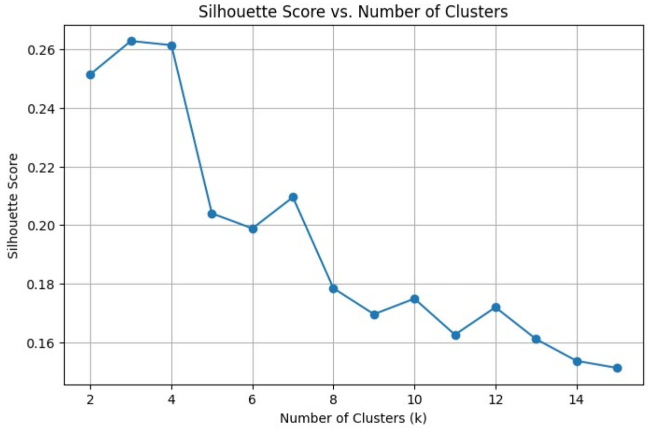
- **Optimal Clusters**: All score reach their optimal value for k = 7, thus, strongly suggesting presence of 7 clusters in the data.
- **Optimal Distance Metric**: Euclidean distance metric performs the best having one of the highest Silhouette Score, lowest Davies-Bouldin Index, and highest Calinski-Harabasz Score.

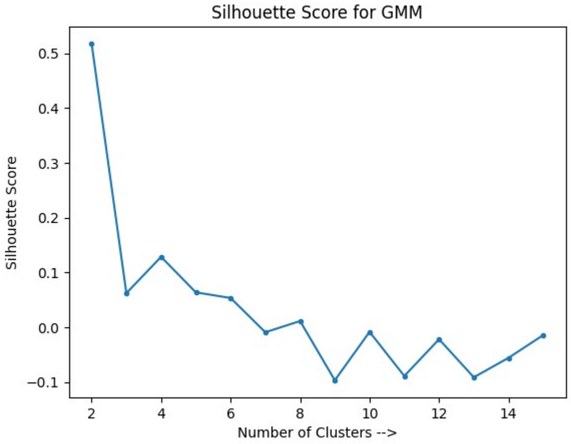
5.3 Plots

5.3.1 Elbow Method for Optimal K

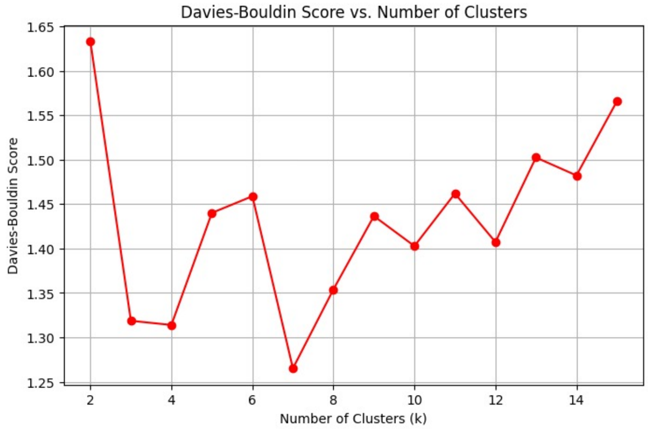


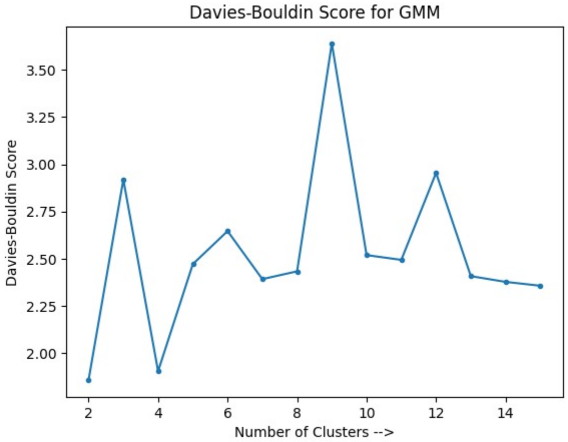
5.3.2 Silhouette Score for KMeans V/S GMM



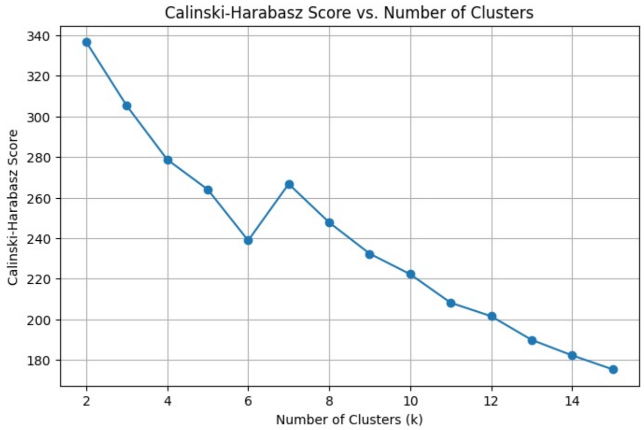


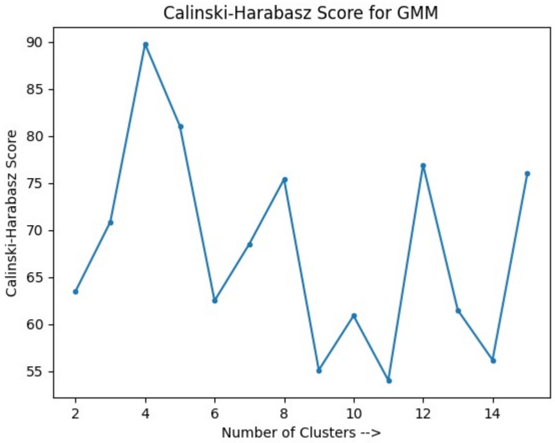
5.3.3 Davies-Bouldin Index for KMeans V/S GMM



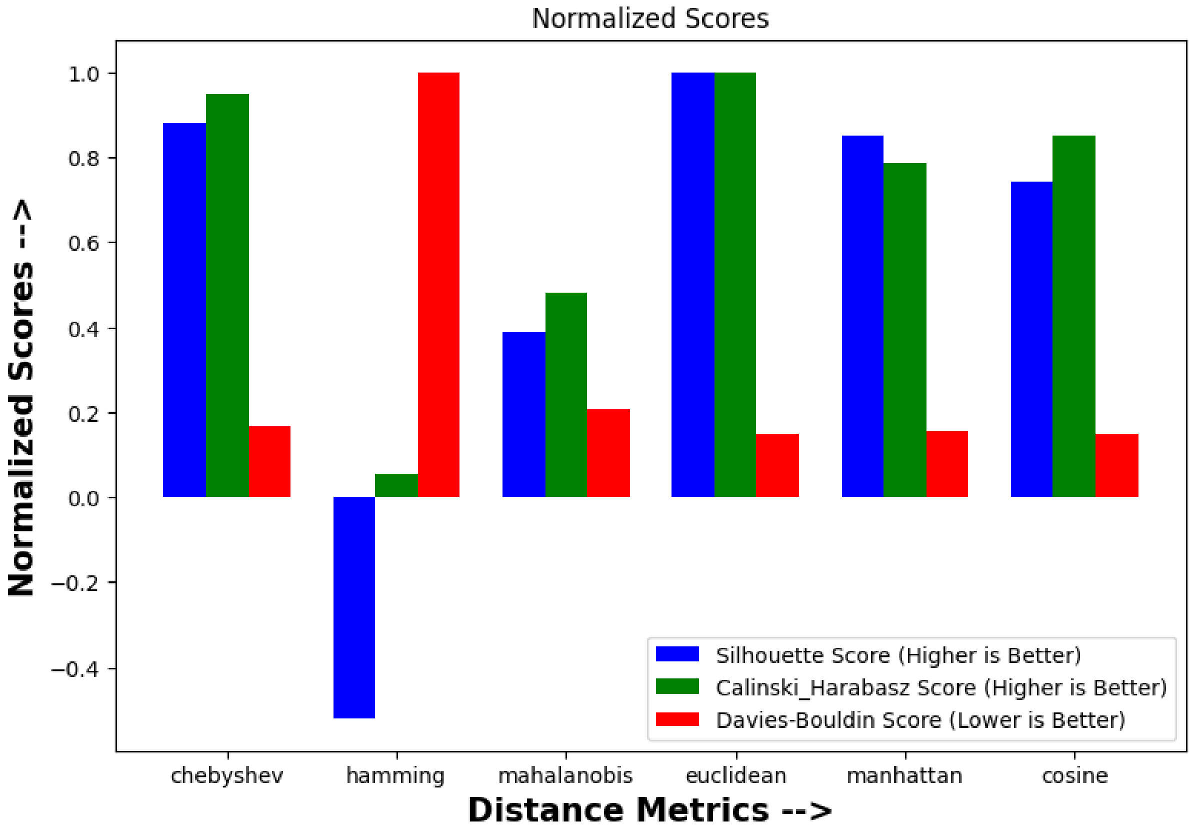


5.3.4 Calinski-Harabasz Score for KMeans V/S GMM





5.3.5 Comparing Different Distances Metrics



6 Conclusion

K-Means demonstrated superior performance for yeast protein localization analysis by achieving the best scores across all evaluation metrics for number of clusters = 7 using the euclidean distance metric.