Assignment 6

Xuchen Zhu, Jingxue Yan

December 18, 2025

1 Exercise 1

1.1 Q1:

The principal components is 10101-dimensional vectors.

1.2 Q2:

The scatter plot of the first two principal components of X:

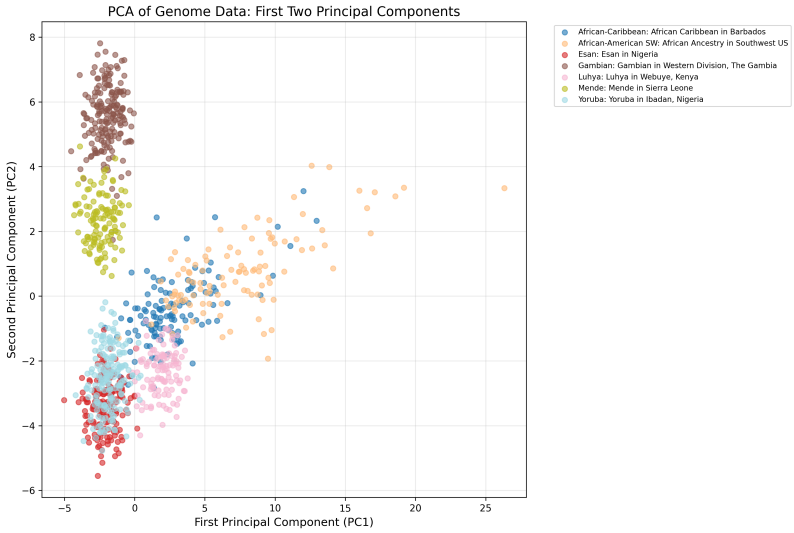

Figure 1: Q2: The first two principal components of X

1.3 Q3:

The scatter plot reveals that the first principal component captures genetic ancestry and admixture, separating West African populations from African-American and Caribbean individuals who show drift due to historical mixing with non-African groups. The second principal component captures intra-continental geography, distinguishing East African populations (Luhya) from the tightly clustered West African groups.

1.4 Q4:

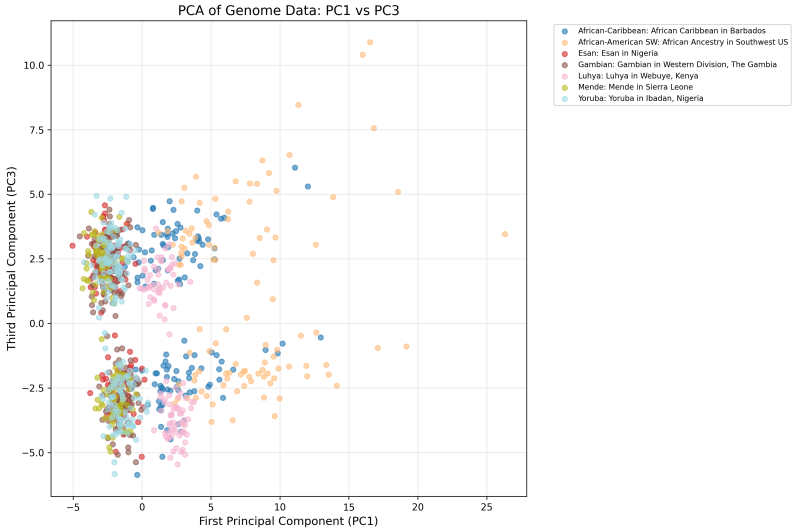

Figure 2: Q4: First and third principal components

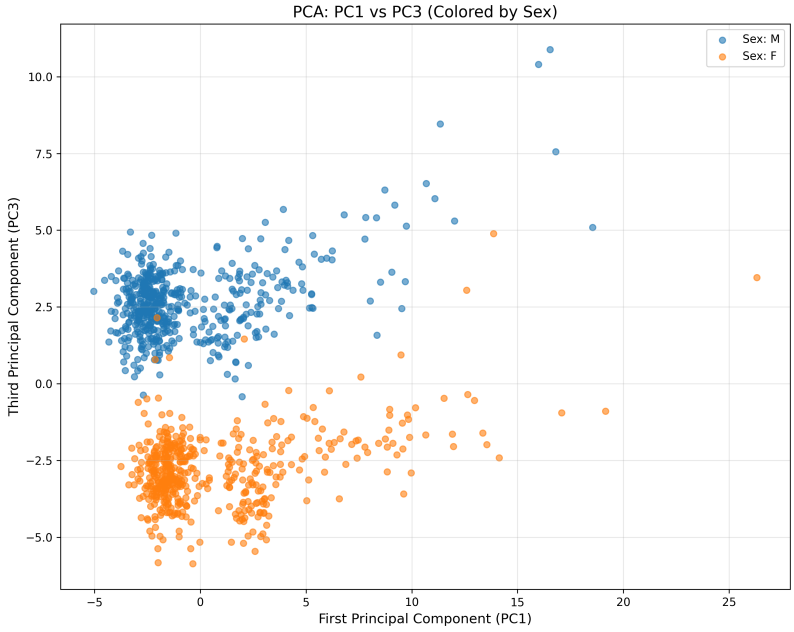

Figure 3: Q4: Cluster by sex

1.5 Q5:

The third principal component captures the biological sex of the individuals, separating the data into two distinct clusters corresponding to males and females.

1.6 Q6:

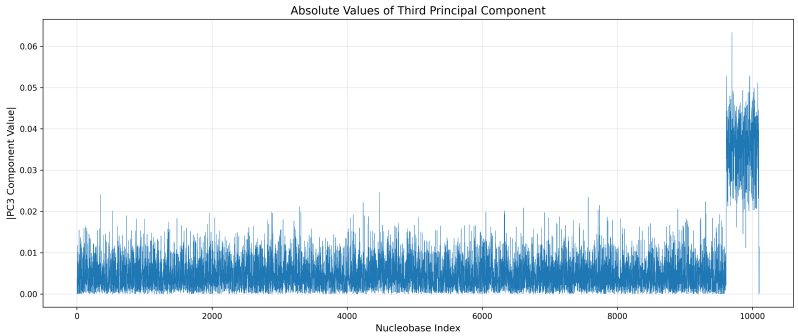

Figure 4: Q6: PC3 Value Analysis

The plot of the third principal component's loadings displays a massive spike in value at the very end of the nucleobase indices, while the rest of the genome remains near zero. This indicates that the variation driving PC3 is localized to a specific region of the genome, specifically the sex chromosomes (X or Y). Since males (XY) and females (XX) possess fundamentally different genetic markers in this region, PCA identifies this chromosomal difference as a major source of variance in the dataset.

2 Exercise 2

2.1 Q1:

As seen in the Fig, the centroid is around [-1, -1], and [1, 1], the variance is also under the variance. The plot is verified.

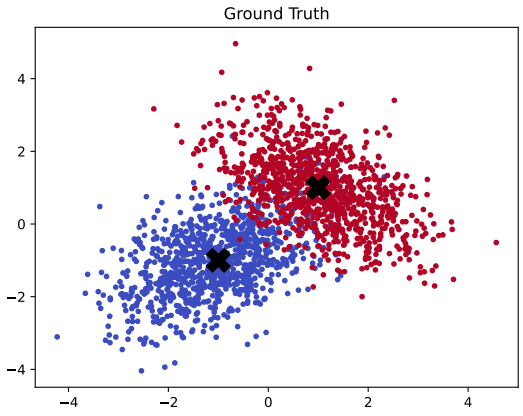

Figure 5: Verify the Ground Truth.

2.2 Q2:

As seen in the Fig, the K-means is below

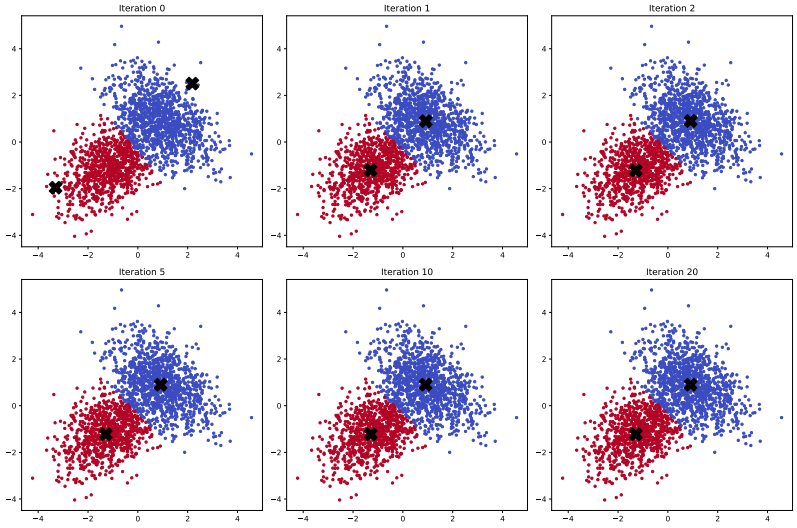

Figure 6: K-means iteration.

2.3 Q3:

As seen in the Fig, it's the EM algorithm.

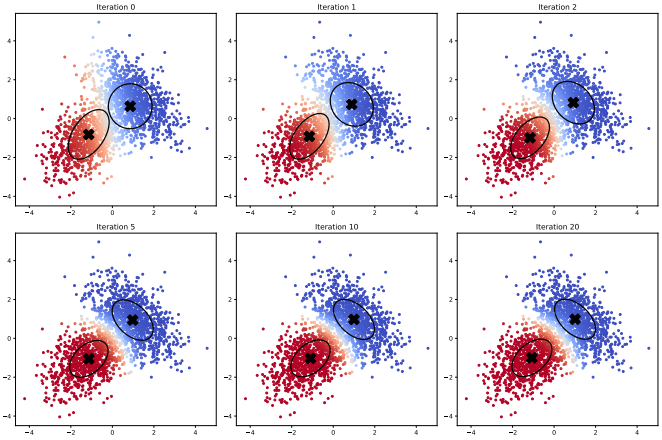

Figure 7: EM algorithm.

2.4 Q4:

1)EM ARI: 0.771 vs k-means ARI: 0.704. ARI is the metric to measure the performance of unsupervised learning. The larger it is, the better the performance is. Since EM ARI is larger, EM algorithm is better.

2)K-means is hard-assigned to the nearest centroid as opposed to EM which computes probability dense to determine the latent feature by γ. EM can partially allow overlapped data to clusters.

If the data distribution is away from Gasussian distributuion, the performance will decrease. If the assumption holds true, then the prior belief with observed data affects responsibility γ, responsibility γ affects mean, covariance, and prior belief.

3) During the intermediate process, we can see the boundary in the EM is a curve line as opposed to K-means which is straight line.

2.5 Q5:

When the number of clusters is overestimated, K-means degrades rapidly because its hard assignments force data points into rigid, linear partitions, leading to over-fragmentation. In contrast, EM (GMM) degrades more gracefully since its soft responsibilities allow multiple Gaussian components to represent a single true cluster, resulting in smoother, curved decision boundaries.

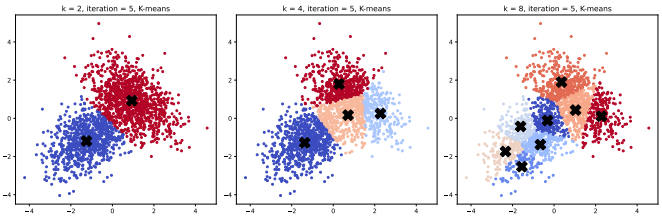

Figure 8: K-means overesitmated k.

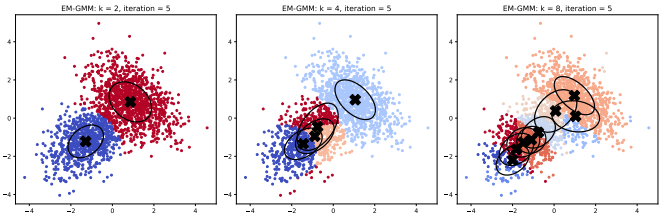

Figure 9: EM overesitmated k.

2.6 Q6:

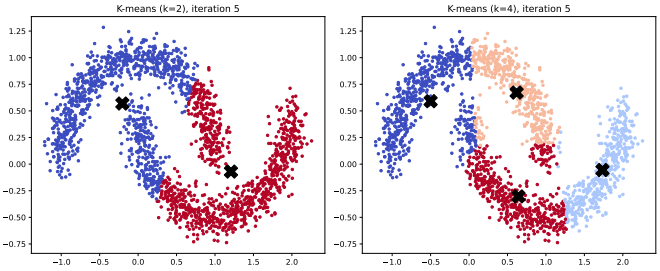

Figure 10: non-circular K-means.

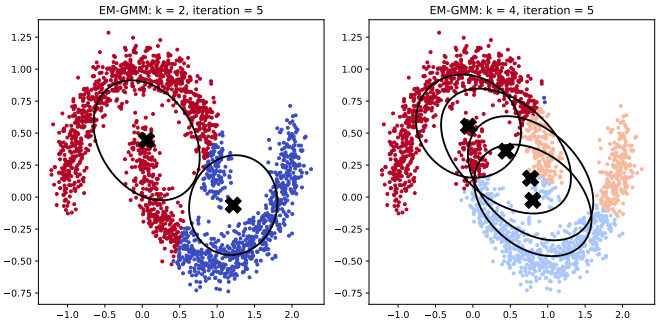

Figure 11: non-circular EM.

EM-GMM performs better than K-means on non-circular clusters because it relaxes the spherical-cluster assumption and uses probabilistic, covariance-aware modeling, whereas K-means is fundamentally limited by its distance-based, hard-assignment framework.

3 Exercise 3

3.1 Q1:

Figure 12: The original one.

3.2 Q2:

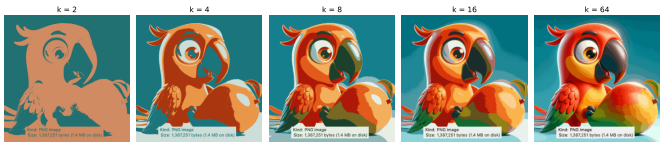

Figure 13: The segmented one.

3.3 Q3:

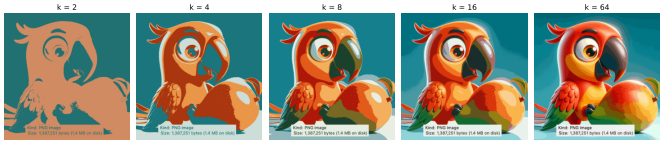

Figure 14: The segmented one.

As we increased K, the segmented image is more similar to the original one.

The small k will compute faster, less overlapped data points, but it will not look alike the original one. The large k will compute slower, more overlapped data points, but it resembles more the original one.