Salma Chaaban – 301216551

COMP257 - Assignment 4 – Analysis Report

1. The Olivetti dataset loaded like the previous assignment.
2. The data was split into training, testing, and validation, like the previous assignment.
3. PCA was applied on the training data with n\_components set to 0.99 to preserve 99% of the variance, reducing dimensionality from 4096 to 199 components.
4. A list of covariance types was tested across cluster counts (2–32). AIC and BIC scores were printed and plotted for each type and cluster count. The 'full' covariance type had the lowest AIC (-215197.50), while 'diag' had the lowest BIC (24865.95). Lower values indicate a better fit, with AIC generally outperforming BIC here.
5. The minimum number of clusters that best represents the dataset was 6 for AIC and 2 for BIC. Again, AIC seems better than BIC for this example, as 2 clusters is a very low number, even 6 clusters is on the lower side; however, it is better than 2.
6. The results for steps 3 and 4 were plotted. The explained variance after PCA continued to increase with component count, indicating information retention. The AIC scores were also plotted, showing that 6 clusters had the least score (provides the best balance of fit and complexity) and after that it kept increasing.
7. A final Gaussian Mixture model with 6 components (clusters) and ‘full’ covariance was fitted to the PCA transformed training data. Hard clustering assignments for each image were printed to show which cluster an image belongs to, and scatter plot was generated to show how well-separated and compact the clusters are, most of the clusters were compact and well-separated. With a 'diag' covariance and 2 clusters, clusters were poorly compacted.

Images within clusters were visualized. While clusters contained different faces due to the limited number of clusters, similarities were noticeable.

1. Using the same final GMM model, soft clustering probabilities showed 0s and 1s for each instance, indicating the model assigned images to clusters with high confidence. After visualizing the images in each cluster, it can be seen that different faces were grouped together but was unavoidable given the low cluster value of 6.
2. Using sample(), 10 new faces were generated and transformed back to the original space via inverse\_transform(). The new faces resemble the original data but are synthetic. Although the images are not as clear as the original and did not seem to capture the finer details of the faces; this approach still shows the GMM’s ability to capture and recreate subtle variations between different faces in the dataset.
3. The images were flipped sideways 90 degrees and darkened, then plotted.
4. Log-likelihood scores were used to differentiate normal and modified images. The positive log-likelihood scores (696 to 919) for normal images indicate a high probability that these samples align well with the GMM. The higher the scores, the closer the images fit within the model’s learned clusters. On the other hand, the very large negative log-likelihood scores for anomalies (-11 million to -14 million) show a dramatic deviation from the patterns learned by the GMM. The more negative the score, the less likely the image fits within the normal distribution. This confirmed the model's capacity to identify anomalies, as shown in the plotted distributions.