**Assignment 4: Gaussian Mixture Model**

**Analysis Report**

In this assignment, we applied unsupervised learning techniques, specifically PCA and Gaussian Mixture Models (GMM), to the Olivetti faces dataset. First, we retrieved and loaded the dataset, which contains grayscale images of faces across 40 classes, and split it into training, validation, and test sets using stratified sampling to ensure balanced representation. We then applied Principal Component Analysis (PCA) to the training data, preserving 99% of the variance, which reduced the dataset’s dimensionality to about 100 components (Figure 1). Next, we evaluated different covariance types for the GMM and found that the ‘spherical’ covariance type had the best score according to the Bayesian Information Criterion (BIC).

Using the BIC measure, we determined that the optimal number of clusters was 1, which was unexpected given the dataset’s 40 classes. This suggested that the data in the PCA-reduced space did not exhibit distinct clusters, likely due to the high similarity of faces in this space, making it difficult for the GMM to distinguish between different clusters. Additionally, the PCA approach might have caused some loss of information that could have helped in distinguishing between clusters. In step 6, we plotted the cumulative variance explained by PCA components (Figure 1) and the AIC scores (Figure 2), which showed a nearly linear increase, indicating that adding more clusters did not significantly improve the model fit. For example, at 1 cluster, there was a small change in the scores, but as the number of clusters increased, the BIC scores increased almost linearly. This suggests that adding more clusters increased model complexity without significantly improving fit, leading to higher penalty terms in BIC. This issue could be due to the data not having well-separated clusters in the PCA-reduced space, and both criteria are designed to avoid overfitting. If the data does not strongly support more clusters, AIC and BIC will favor simpler models with fewer clusters, potentially leading to underfitting. When we assigned clusters to the test data in step 7, all instances were assigned to a single cluster (0), confirming the lack of distinct groupings and the underfitting issue of the GMM model. The soft clustering probabilities also indicated high confidence in this single cluster assignment. In step 9, PCA was used to encode the data into a lower-dimensional space and then decode it back to the original space. This method was effective in capturing the most important information from the faces, as evidenced by the recognizable generated faces in Figure 3. Finally, we compared the log-likelihood scores of normal and modified images, where we added noise (rotation, flip, darken) to the original images to consider them as anomalies. We found that anomalies had lower scores, demonstrating the GMM model’s ability to detect anomalies.

In conclusion, the GMM model demonstrated efficiency in detecting anomalies by identifying data points that do not conform to the learned distribution. This was achieved using the same trained model that was used for clustering. However, the model struggled with finding the optimal number of clusters, likely due to the high similarity of faces in the PCA-reduced space and the influence of the BIC measure, which penalized model complexity and led to underfitting. Additionally, PCA effectively extracted and retained the most important features of the Olivetti faces dataset.

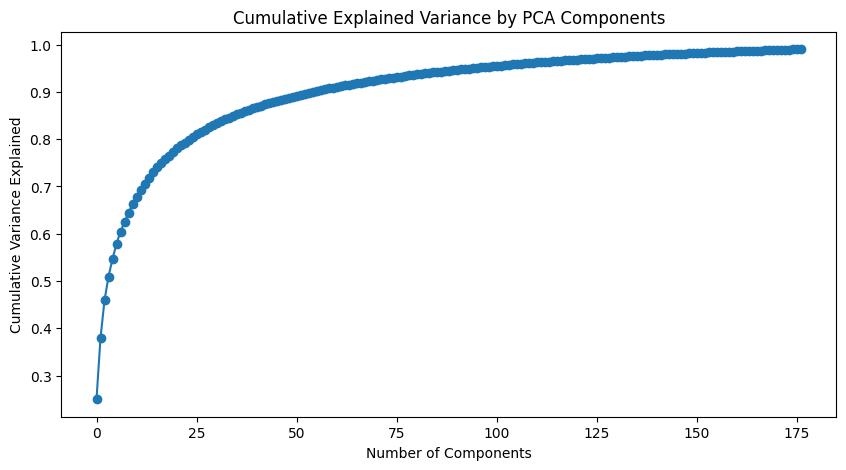


Figure 1. Cumulative Explained Variance by PCA Components

A graph of a number of clusters

Description automatically generated

Figure 2. BIC Scores for Different Cluster Counts

A collage of a person's face

Description automatically generated

Figure 3. Using PCA to generate new images