# Unsupervised Learning: PCA and EM Algorithm Report

Asif Azad
Student ID: 1905004
Department of CSE, Bangladesh University of Engineering and Technology
Email: asifazad0178@gmail.com

## I. Introduction

This report covers the implementation and analysis of two widely used unsupervised learning methods: Principal Component Analysis (PCA) and the Expectation-Maximization (EM) algorithm. The tasks include:

1) PCA for dimensionality reduction and visualization.
2) UMAP and t-SNE plots for high-dimensional data visualization.
3) EM algorithm to estimate parameters of a Poisson mixture model.

## II. Instructions to Run the Code

- Place the provided datasets `pca_data.txt` and `em_data.txt` in the same directory as the notebook.
- Run the code cells sequentially in the Jupyter Notebook. Ensure that the required Python libraries (e.g., NumPy, pandas, matplotlib, seaborn, sklearn, umap-learn, and scipy) are installed.
- The output includes PCA, UMAP, t-SNE visualizations, EM algorithm results, and a log-likelihood plot.

## III. Principal Component Analysis (PCA)

### A. Implementation

PCA was implemented from scratch, leveraging linear algebra for eigen decomposition. The dataset `pca_data.txt` (1000 samples, 500 features) was reduced to two dimensions for visualization.

### B. Results

- Figure 1 shows the 2D scatter plot of the PCA-transformed data.
- The first two principal components capture the majority of the variance in the dataset.

## IV. UMAP and t-SNE Visualizations

The UMAP and t-SNE visualizations of the original dataset provide alternative perspectives on the data structure.

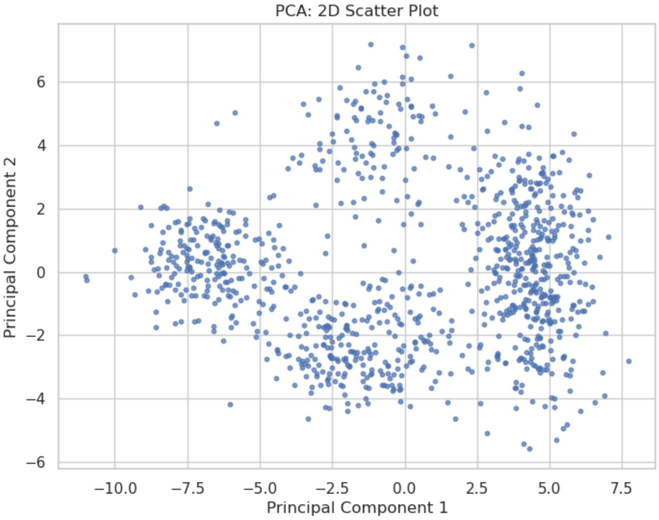

Fig. 1. 2D Scatter Plot of PCA Transformed Data

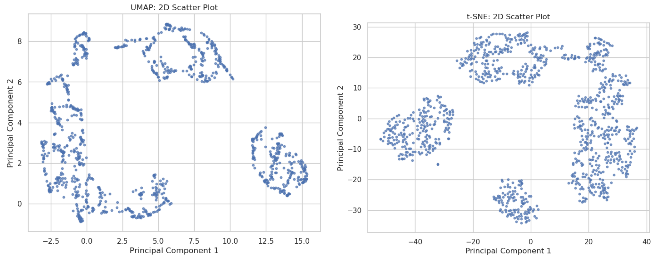

Fig. 2. (Left) UMAP Plot, (Right) t-SNE Plot

## V. Expectation-Maximization (EM) Algorithm

### A. Implementation

The EM algorithm was implemented to estimate:

- The mean number of children in families with and without family planning.
- The proportions of families with and without family planning.

The dataset `em_data.txt` was modeled as a mixture of two Poisson distributions.

### B. Results

The EM algorithm converged in $X$ iterations. The estimated parameters are:

- Mean (with family planning): $\mu_1 = 1.78$
- Mean (without family planning): $\mu_2 = 4.91$
- Proportion (with family planning): $\pi_1 = 0.36$
- Proportion (without family planning): $\pi_2 = 0.64$

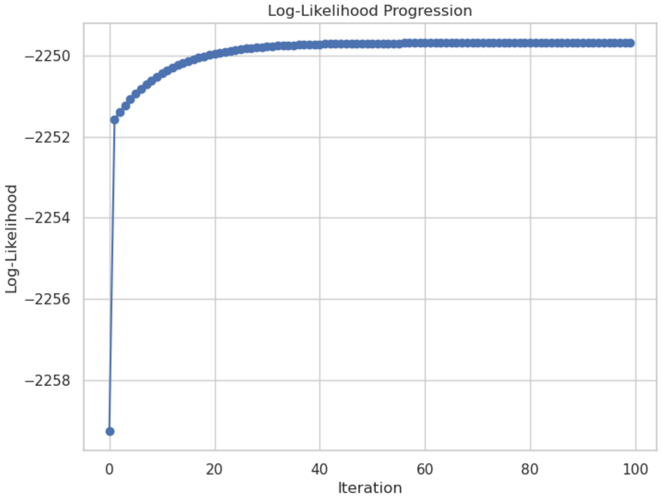

Fig. 3. Log-Likelihood Progression

## VI. Conclusion

This report demonstrates the implementation of PCA and EM algorithms for unsupervised learning tasks. The results highlight the utility of dimensionality reduction and parameter estimation in analyzing real-world datasets.