Report for Assignment 4: PCA and EM Algorithm

Course: CSE472 (Machine Learning Sessional)

Assignment Title: Principal Component Analysis (PCA) and Expectation-Maximization (EM) Algorithm

Student ID: 1905028

1. How to Run the Code

1. Ensure the datasets `pca_data.txt` and `em_data.txt` are in the same directory as 1905028.ipynb file.
2. Install the required libraries by running the following command in your terminal:

pip install numpy matplotlib sklearn umap-learn

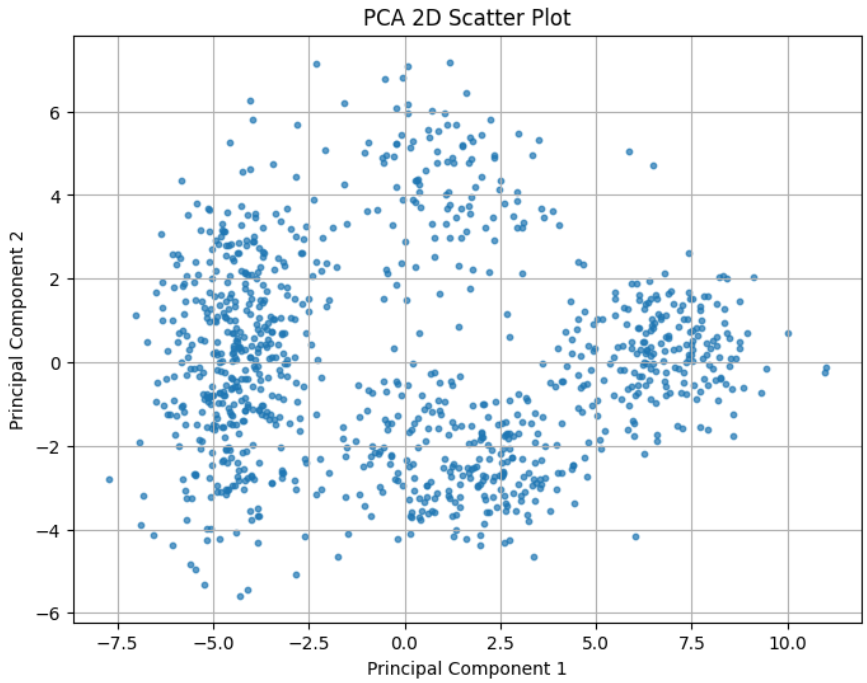
3. Open the provided Python code in Jupyter Notebook or any Python IDE that supports `.ipynb` files.
4. Run the code cell by cell. The code will:
   - Generate PCA, UMAP, and t-SNE plots.
   - Output the estimated parameters from the EM algorithm.

The datasets provided were:
1. pca_data.txt for PCA tasks (1000 samples, 500 features).
2. em_data.txt for EM tasks (1000 rows representing the number of children in hypothetical families).

2. Results

2.1 Principal Component Analysis (PCA)

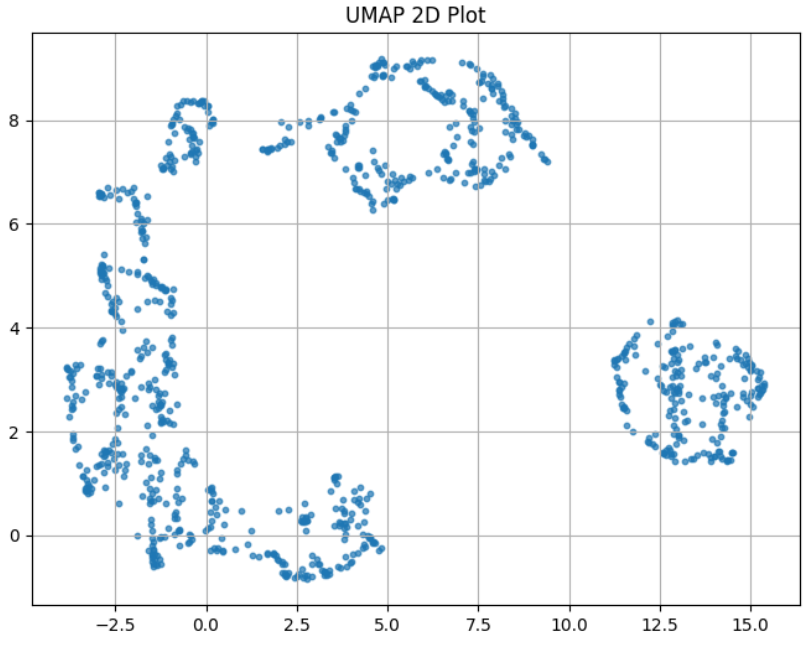
The scatter plot of the data projected onto the first two principal components is shown below.



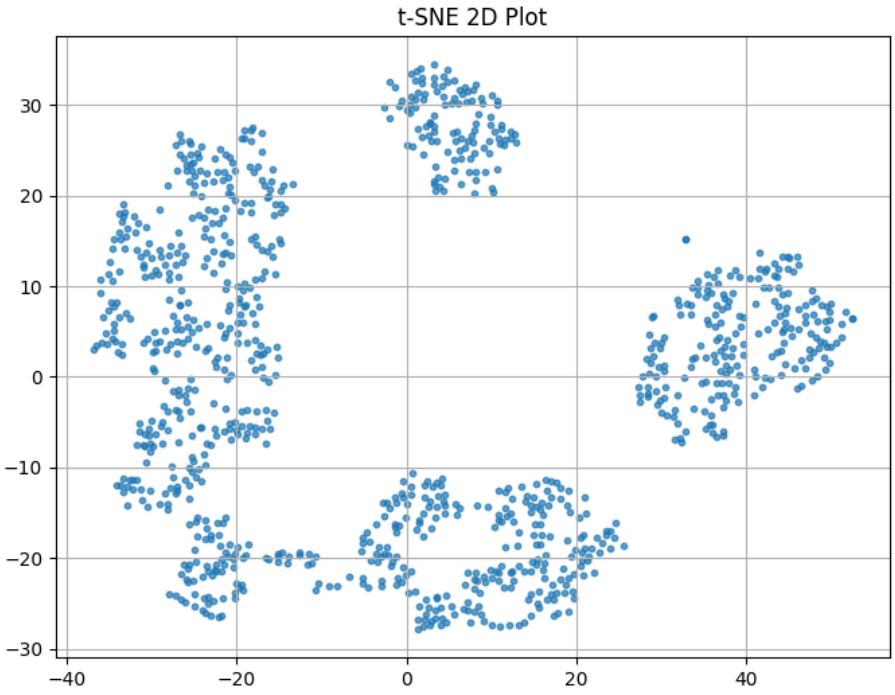
2.2 UMAP and t-SNE

Using library implementations, the original high-dimensional data was also visualized using UMAP and t-SNE.

- UMAP Plot:



- t-SNE Plot:



3. Expectation-Maximization (EM) Algorithm

After running the EM algorithm, the following parameters were estimated:
- Proportion of families with family planning (π): 0.357
- Proportion of families without family planning (1-π): 0.643
- Mean number of children in families with family planning (λ1): 1.785
- Mean number of children without family planning (λ2): 4.913