Principal Component Analysis and Expectation-Maximization Algorithm

Mahfuzzaman Sizan
Student ID: 1905054

November 22, 2024

Introduction

This report presents the implementation and results of two widely used unsupervised learning methods in machine learning:

- **Principal Component Analysis (PCA):** For dimensionality reduction.
- **Expectation-Maximization (EM):** To estimate the mean number of children and the proportion of families with and without family planning.

1 Running the Codes

To execute the provided Python scripts, follow these steps:

1. Ensure Python (version 3.10 or later) is installed on your system.
2. Install the required libraries using the following command:

```
pip install numpy matplotlib scikit-learn umap-learn
```

3. For the EM algorithm, ensure the file `em_data.txt` is in the same directory as the script.

4. Execute the scripts:

   - **PCA, UMAP, and t-SNE:**

     ```
     python pca_umap_tsne.py
     ```

   - **EM Algorithm:**

     ```
     python em_algorithm.py
     ```

5. Outputs such as plots and estimated parameters will be saved to the current directory.

2 Results

2.1 PCA, UMAP, and t-SNE Plots

The dimensionality of the dataset was reduced from 500 to 2 dimensions using PCA, UMAP, and t-SNE. The following plots were generated:

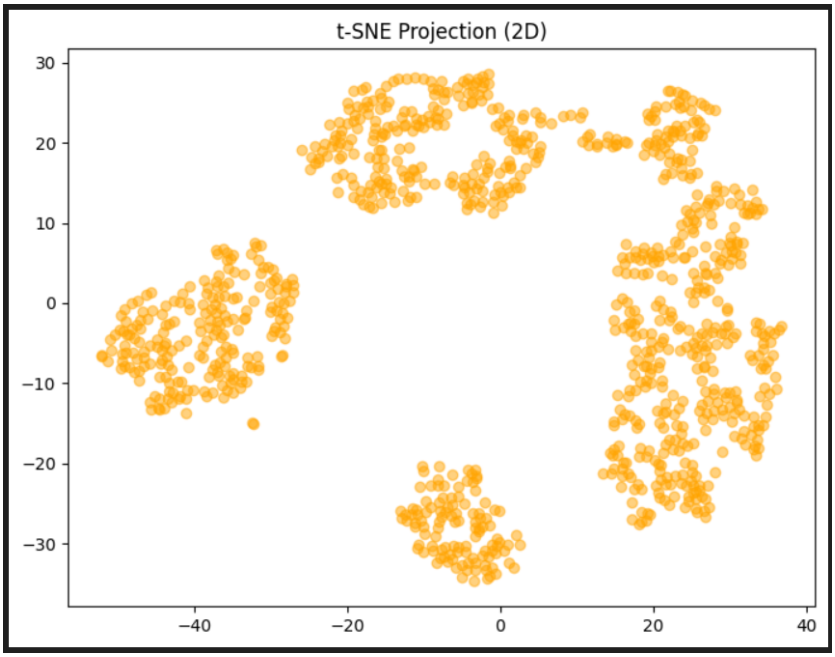

Figure 1: t-SNE Projection (2D)

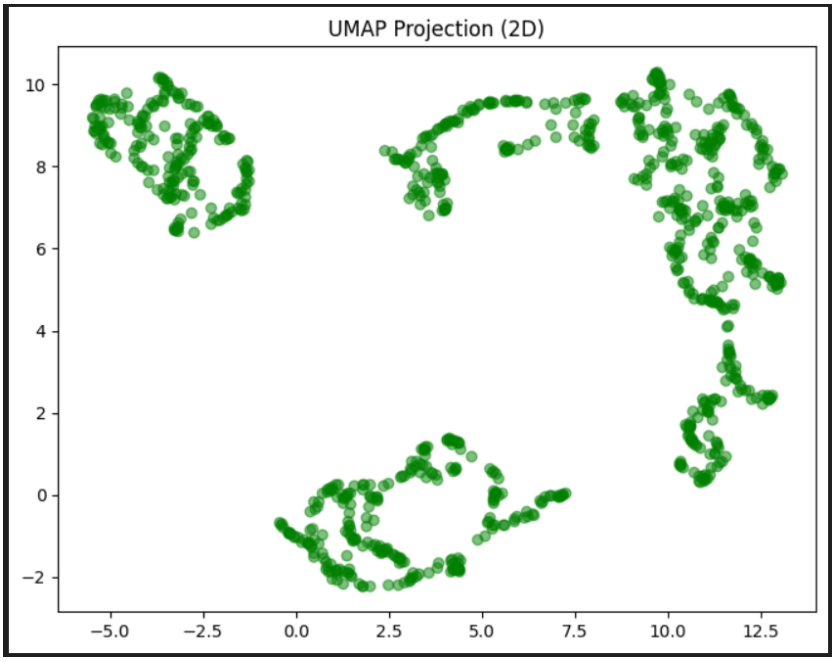

Figure 2: UMAP Projection (2D)

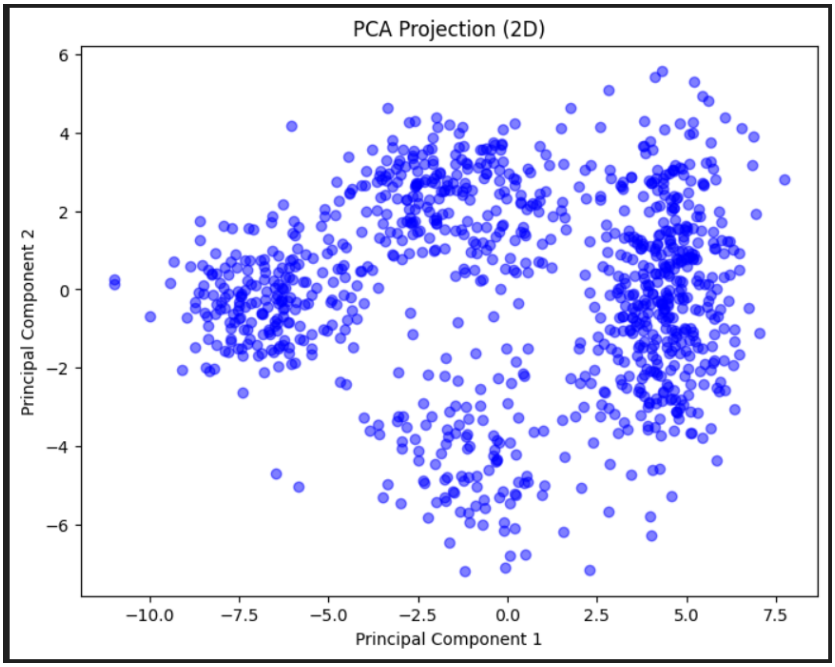

Figure 3: PCA Projection (2D)

2.2 Expectation-Maximization Results

The EM algorithm estimated the following parameters for families with and without family planning, assuming a Poisson distribution:

Parameter	**With Family Planning**	**Without Family Planning**
Mean (λ)	1.7924	4.9192
Proportion (π)	0.3589	0.6411

Table 1: Estimated Parameters from EM Algorithm

3 Conclusion

This report demonstrates the implementation of PCA, UMAP, and t-SNE for dimensionality reduction and the EM algorithm for parameter estimation. The visualizations provide insights into the structure of the dataset, while the EM algorithm effectively estimates family planning statistics.