Assignment 4: Principal component analysis (PCA) and expectation-maximization (EM) algorithm

1905113 - Anamul Hoque Emtiaj

November 22, 2024

1 Introduction

This report presents the implementation and results of two machine learning tasks:

1. Principal Component Analysis (PCA) for dimensionality reduction, along with UMAP and t-SNE visualizations

2. Expectation-Maximization (EM) algorithm for analyzing family planning data using a Poisson mixture model

2 Environment Setup and Requirements

The implementation was done in Python using Jupyter Notebook. Here are the required packages and their versions:

```
numpy==2.0.0
scipy
matplotlib
sklearn
umap-learn
```

To install the required packages, run:

```
pip install numpy==2.0.0 scipy matplotlib scikit-learn umap-learn
```

3 Code Execution Instructions

The code is contained in the Jupyter notebook file `1905113.ipynb`. To run the analysis:

1. Ensure all required packages are installed

2. Place the data files (`pca_data.txt` and `em_data.txt`) in the same directory as the notebook

3. Open Jupyter Notebook:

   ```
   jupyter notebook

   ```

4. Navigate to and open `1905113.ipynb`

5. Run all cells in the notebook (Cell → Run All)

4 Results

4.1 Dimensionality Reduction Results

The high-dimensional data was visualized using three different techniques:

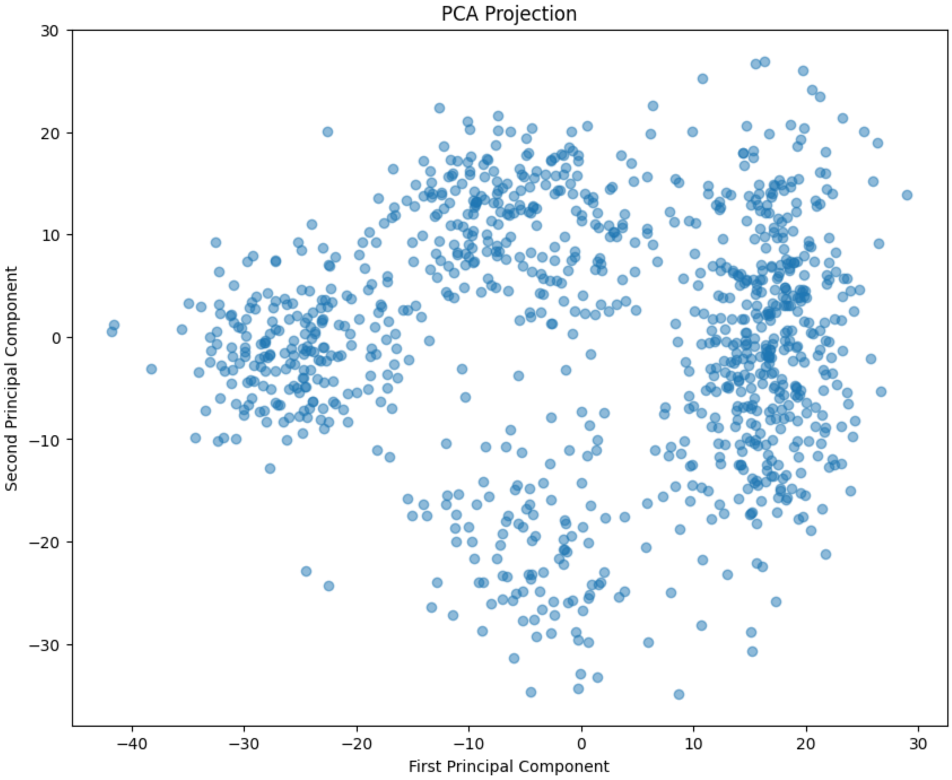

Figure 1: PCA projection of the data

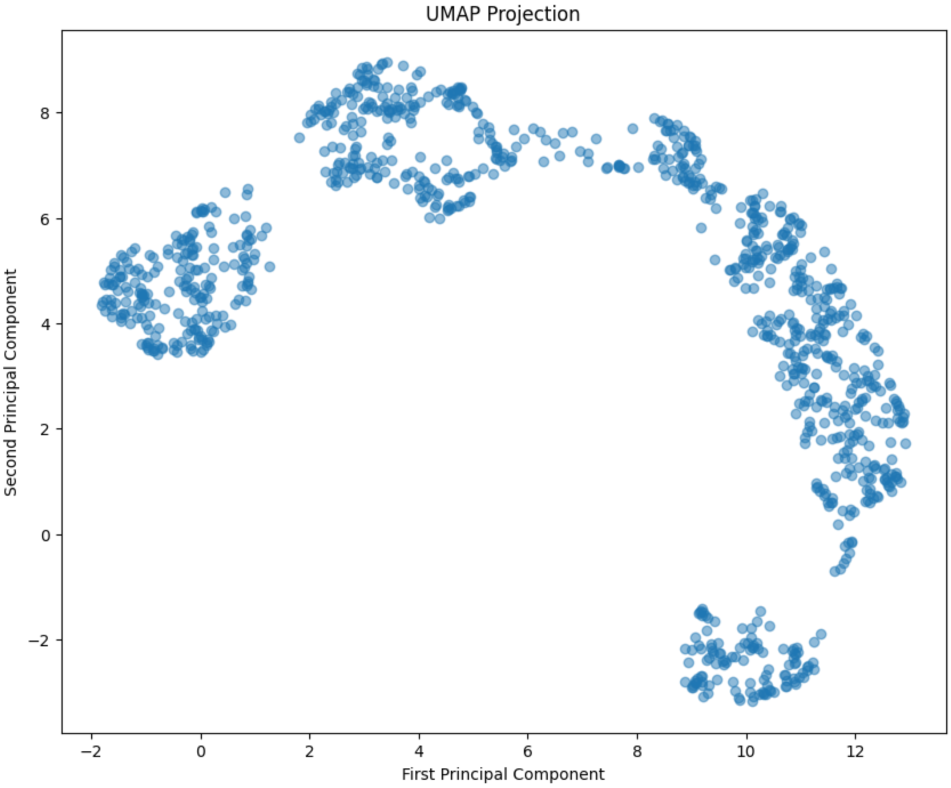

Figure 2: UMAP projection of the data

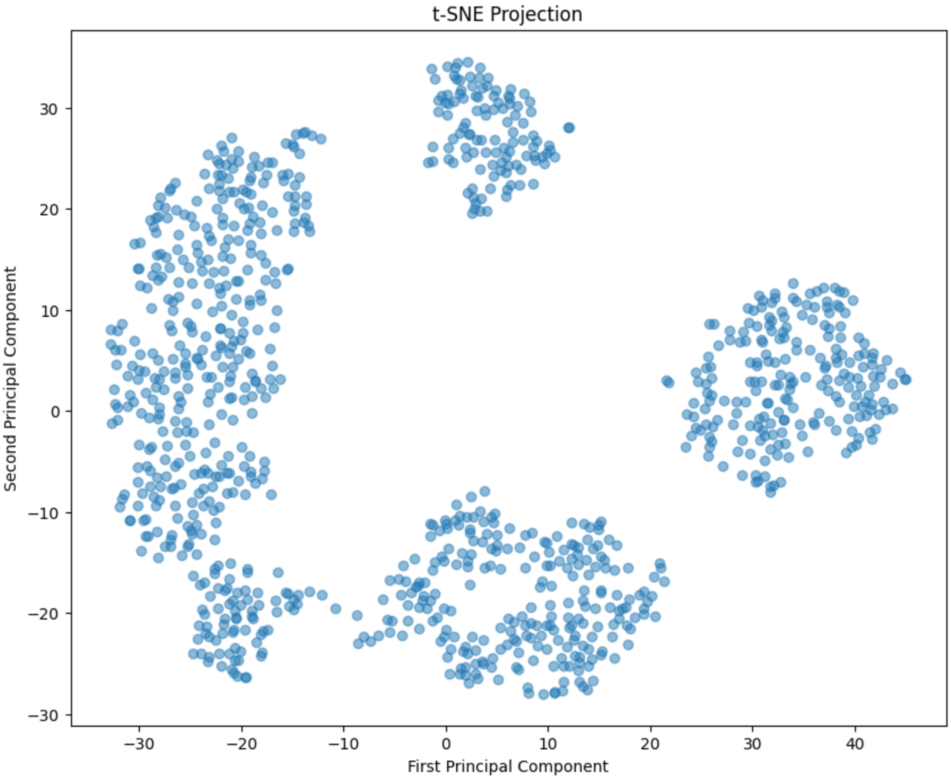

Figure 3: t-SNE projection of the data

4.2 Family Planning Analysis Results

The EM algorithm was applied to analyze the distribution of children in families with and without family planning advice. The results are visualized below:

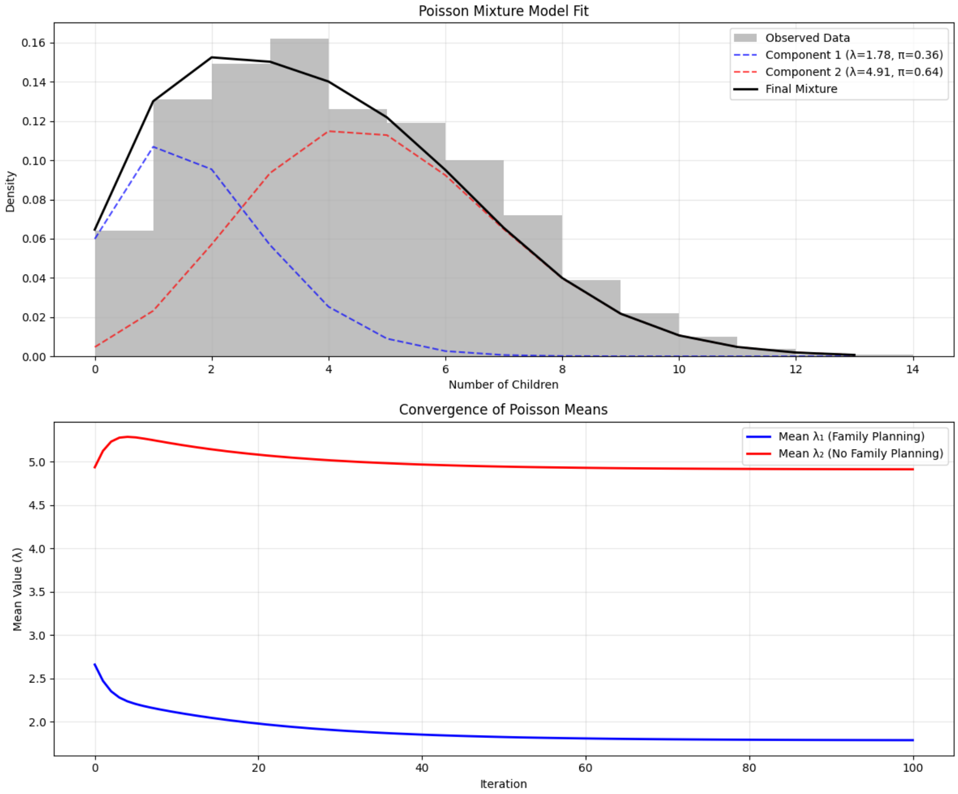

Figure 4: Poisson Mixture Model Fit

4.3 Family Planning Statistics

The EM algorithm revealed two distinct groups in the population:

1. **Families with Family Planning:**
   - Mean number of children: 1.78
   - Proportion of families: 35.67%
2. **Families without Family Planning:**
   - Mean number of children: 4.91
   - Proportion of families: 64.33%

5 Conclusion

The analysis successfully demonstrated both dimensionality reduction techniques and the application of the EM algorithm. The PCA implementation provided comparable results to established techniques like UMAP and t-SNE. The EM algorithm effectively separated the population into two groups, revealing distinct patterns in family planning outcomes.