Bangladesh University of Engineering and Technology

CSE 472: Machine Learning Sessional

Offline-4: PCA and EM

Student ID: 1905012

Section: A1

Level: 4

Term: II

PCA (Principal Component Analysis):

PCA is a **linear** dimensionality reduction technique with application in visualization, data processing. There are other two dimensionality reduction techniques **UMAP** (Uniform Manifold Approximation and Projection) and **t-SNE** (t-distributed Stochastic Neighbor Embedding) which are **non-linear and non-deterministic**.

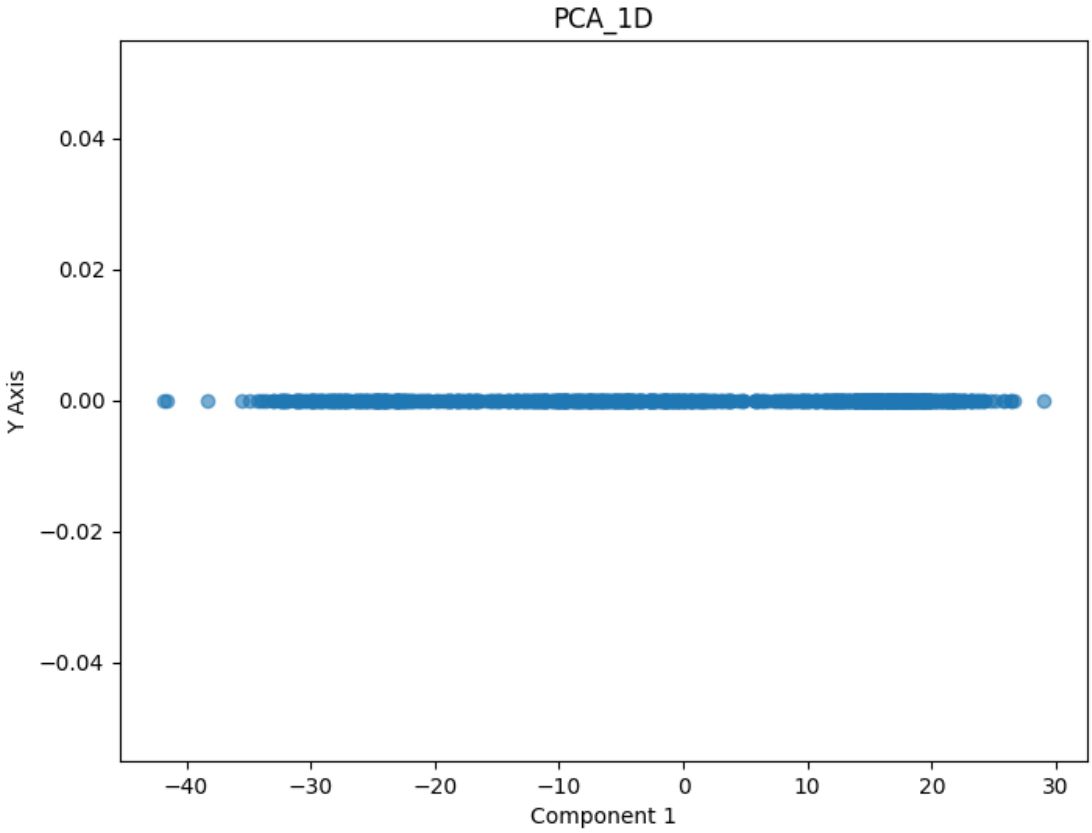
Dataset:

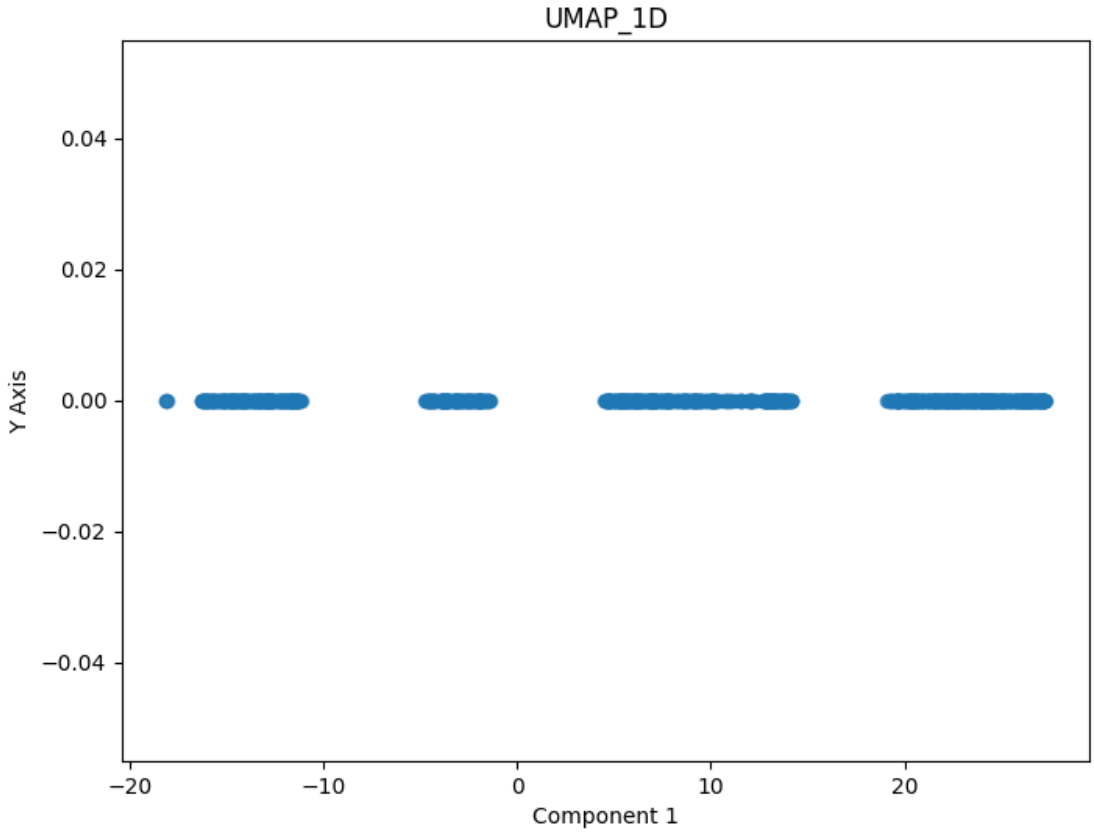
The dataset consists of 1000 sample points and each sample point is 500 dimensional. We need to perform PCA on this dataset to visualize the data. So we can reduce it to 1, 2 and 3 dimensions and plot the figures.

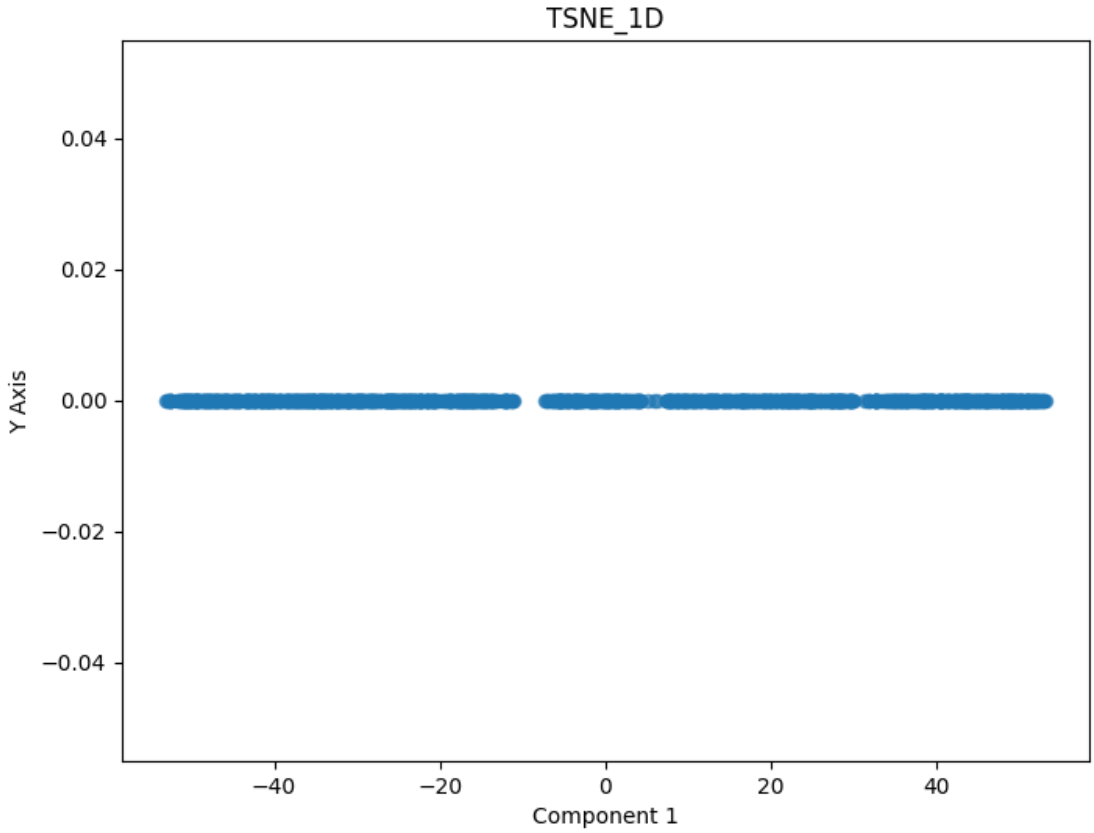
We also need to perform UMAP and t-SNE on the dataset.

Comparison

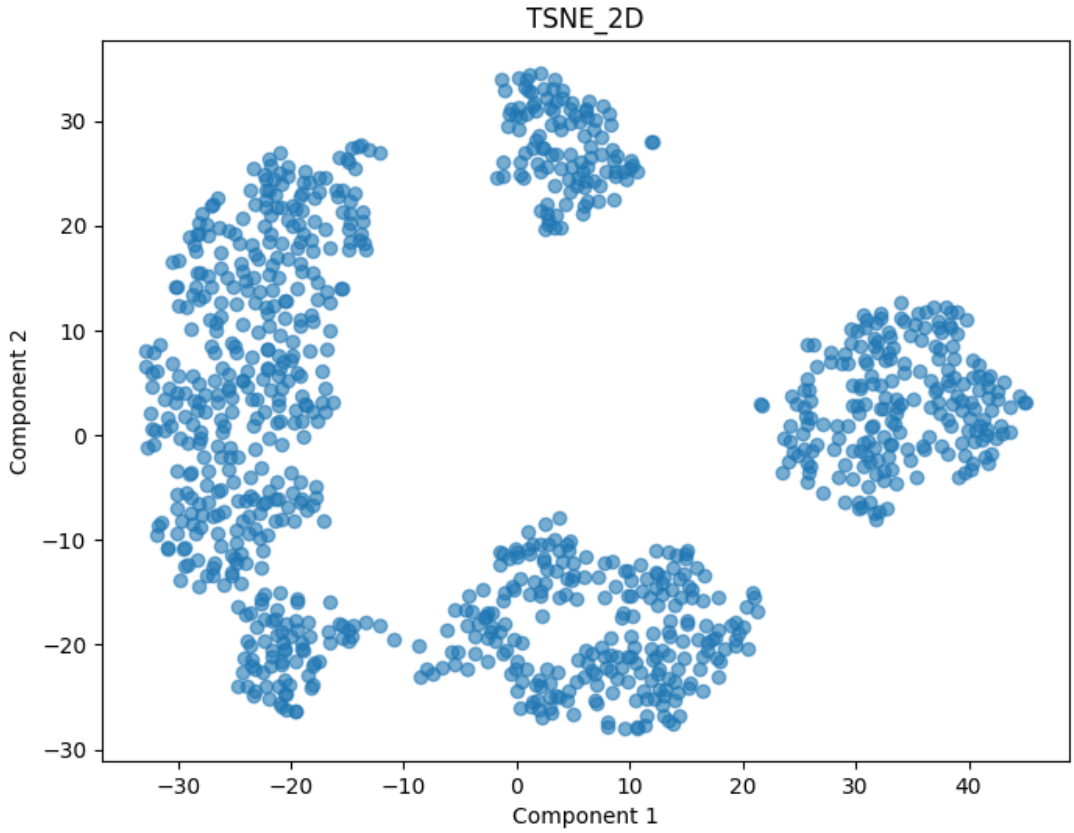
1-Dimensional Plot:

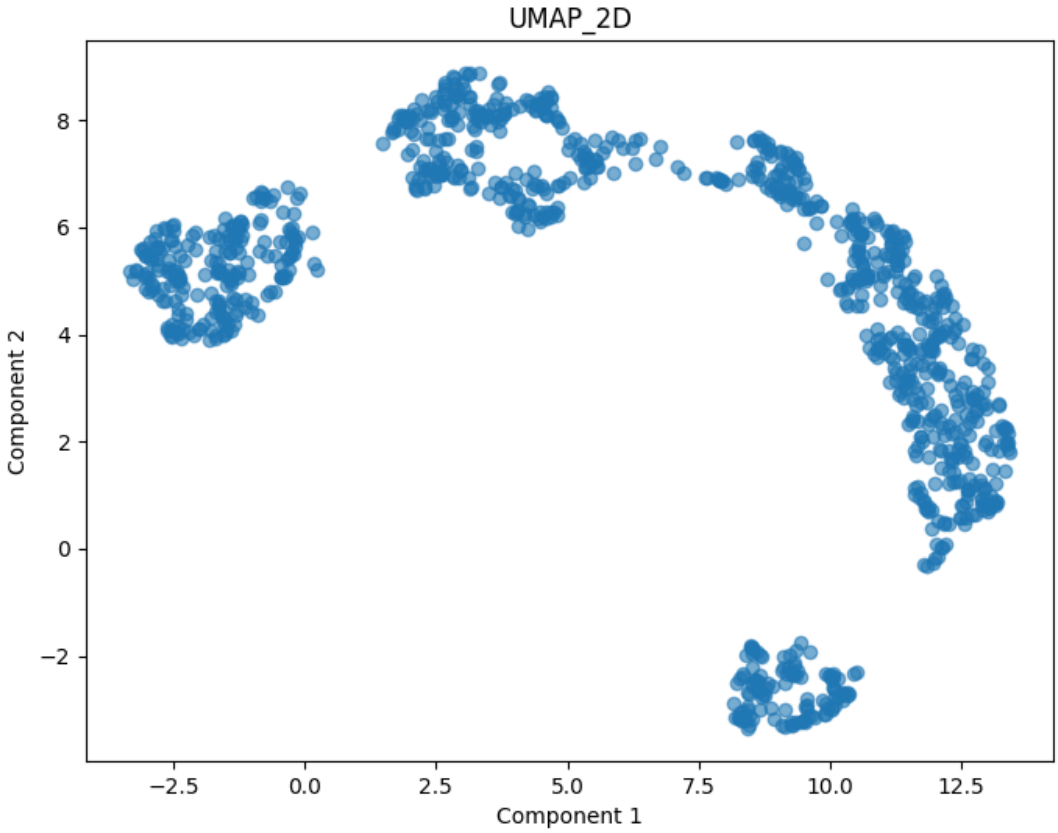

PCA_1D
0.04
0.02
0.00
−0.02
−0.04
Y Axis
−40
−30
−20
−10
0
10
20
30
Component 1

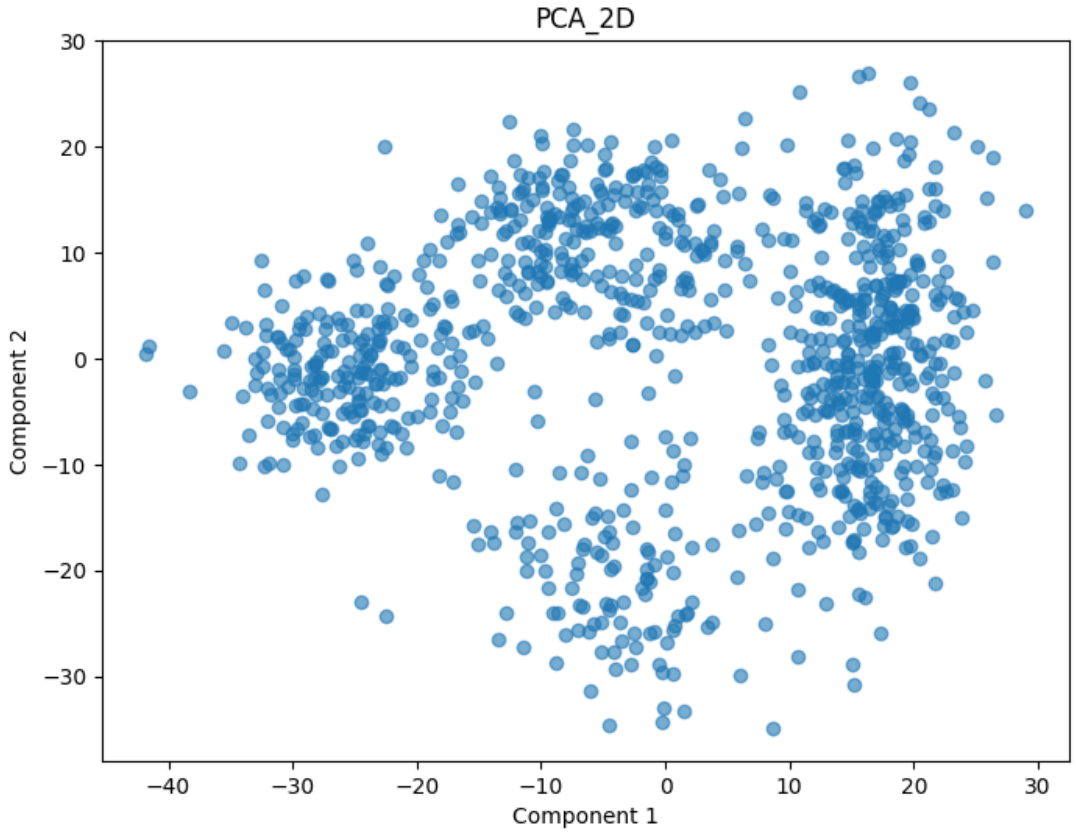

UMAP_1D
0.04
0.02
0.00
−0.02
−0.04
Y Axis
−20
−10
0
10
20
Component 1



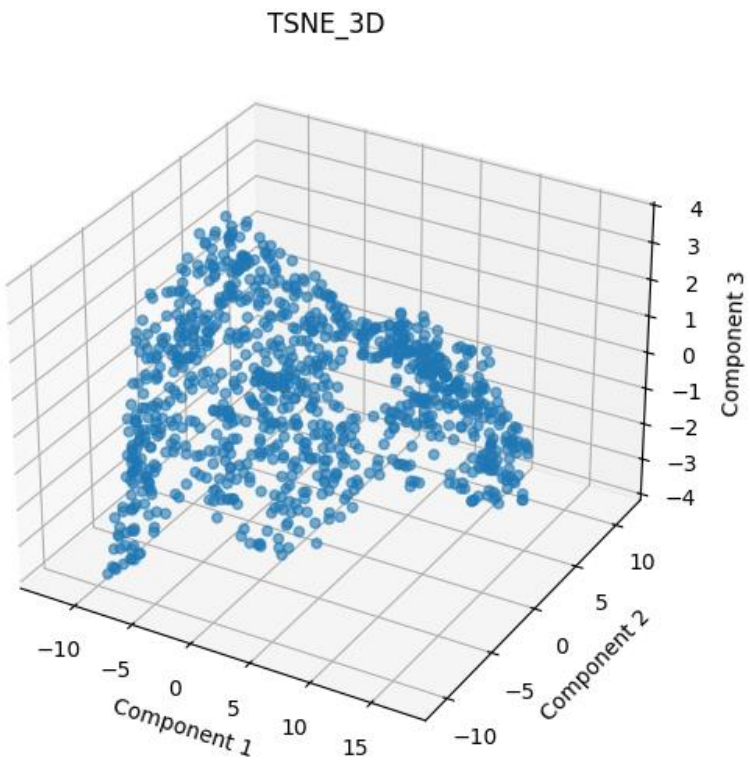
2-Dimensional Plot:

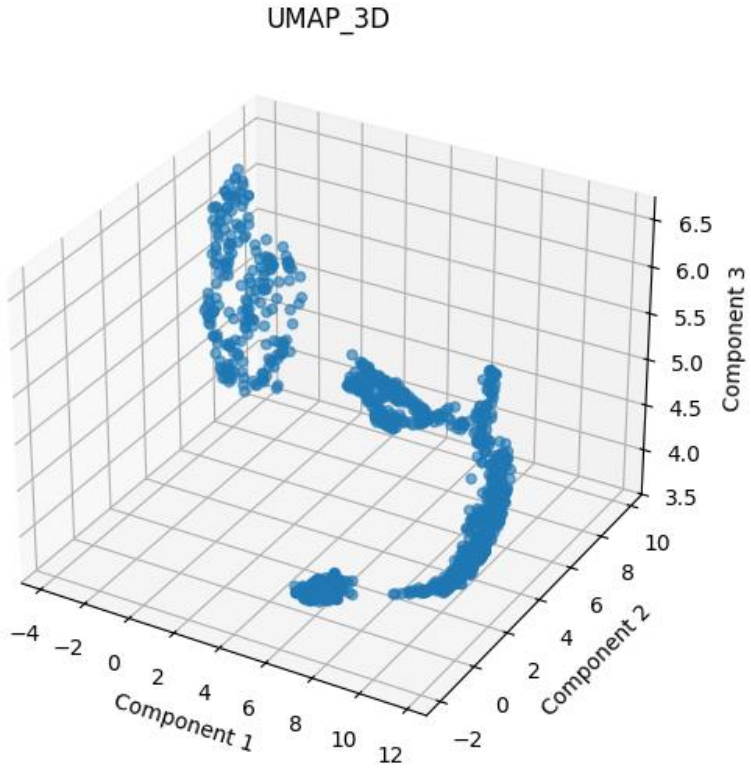

TSNE_2D
Component 2
Component 1
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30
40

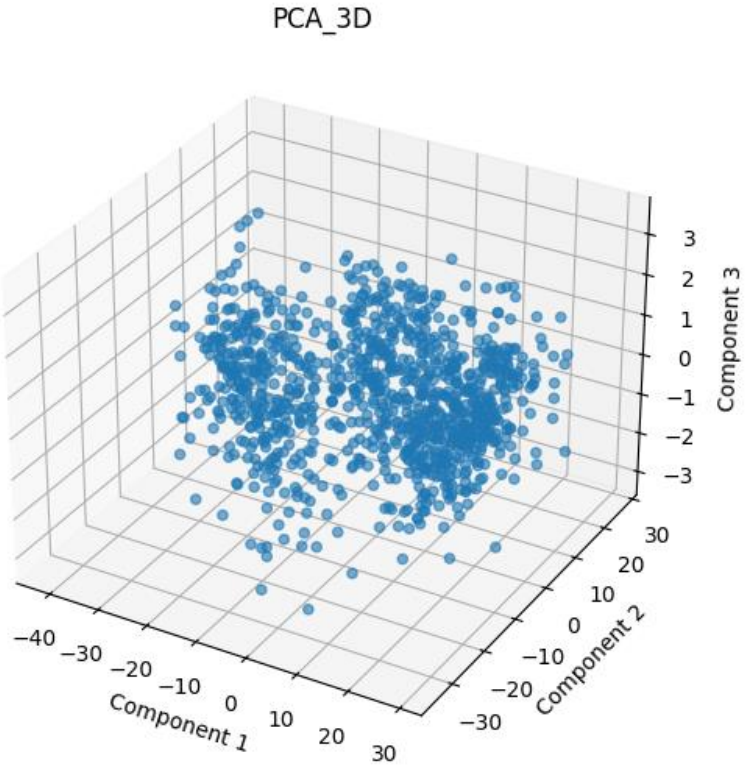

UMAP_2D
Component 2
Component 1
8
6
4
2
0
−2
−2.5
0.0
2.5
5.0
7.5
10.0
12.5



3-Dimensional Plot:


TSNE_3D
Component 1
Component 2
Component 3
−10
−5
0
5
10
15
−10
−5
0
5
10
4
3
2
1
0
−1
−2
−3
−4


UMAP_3D
Component 1
Component 2
Component 3
−4
−2
0
2
4
6
8
10
12
−2
0
2
4
6
8
10
3.5
4.0
4.5
5.0
5.5
6.0
6.5



Summary:

PCA can’t capture non-linearity relationship among the features like UMAP and t-SNE. The variance captured is reported as 57.8%, 86.46% and 86.79% for 1, 2 and 3 reduced dimensions respectively.

EM (Expectation Maximization) Algorithm:

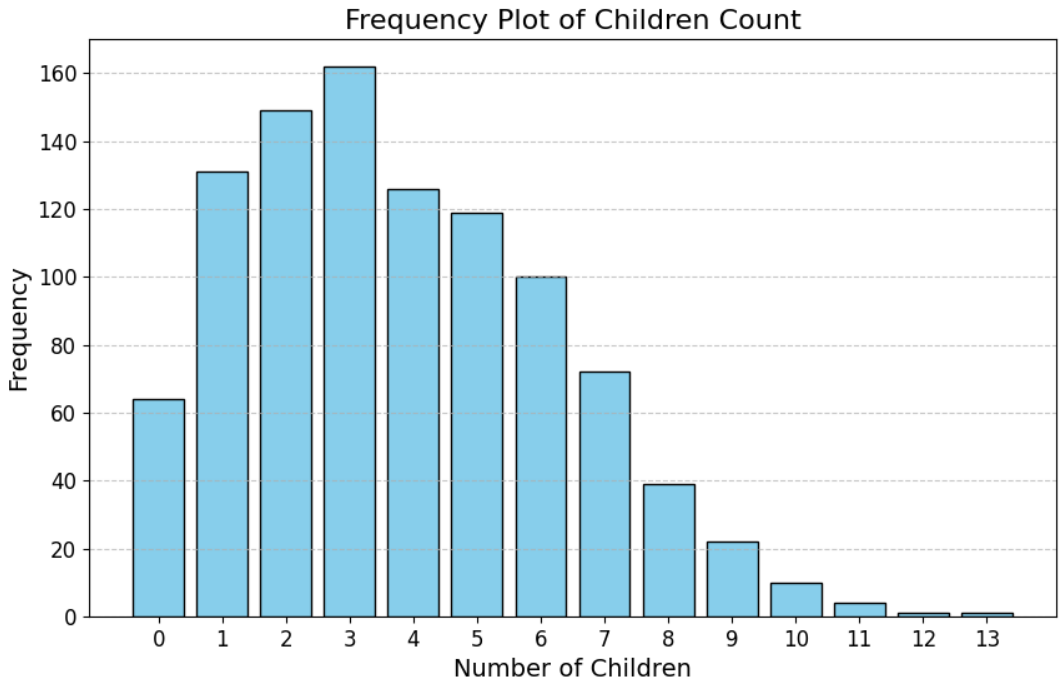
Expectation Maximization algorithm is an iterative method to find local maximum likelihood or maximum a posteriori (MAP) estimates of parameters where the model depends on unobserved or latent variables.

The E-M iterations alternate between performing an Expectation step which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the *E* step.

Dataset:

The dataset comprises of 1000 rows which denotes the number of hypothetical families with or without family planning advice. Each row value denotes the number of children of that family.

Mean of the dataset: 3.797



Derivation of the formulae for Poisson Mixture Model:

We assume the number of children X in each family follows a mixture of two Poisson distributions.

Let, z_i be the random variable indicating whether a family received family planning or not.

1. Families that are provided with family planning (*z_i = 1*)

Therefore, x_i equals to Poisson(*lambda_1*).

2. Families that are not provided with family planning (*z_i = 0*)

Therefore, x_i equals to Poisson(*lambda_2*).

The complete likelihood involves the joint probability of the observed data x_i and latent variable z_i:

The complete likelihood :

$$p(x_i, z_i) = \left[\pi_1 \frac{\lambda_1^{x_i} e^{-\lambda_1}}{x_i!}\right]^{z_i} \cdot \left[\pi_2 \frac{\lambda_2^{x_i} e^{-\lambda_2}}{x_i!}\right]^{(1-z_i)}$$

where, π_1, π_2 = prior probabilities of a family being in group 1 and 2 respectively.

Here, $\pi_1 + \pi_2 = 1$

So, the complete log-likelihood given by —

$$\log p(x, z \mid \lambda_1, \lambda_2, \pi_1)$$

$$= \sum_{i=1}^{N} \left[z_i \log\left(\pi_1 \frac{\lambda_1^{x_i} e^{-\lambda_1}}{x_i!}\right) + (1-z_i) \log\left(\pi_2 \frac{\lambda_2^{x_i} e^{-\lambda_2}}{x_i!}\right) \right]$$

$$= z_i \{\log \pi_1 + x_i \log \lambda_1 - \lambda_1 - \log(x_i!)\}$$

$$+ (1-z_i) \{\log \pi_2 + x_i \log \lambda_2 - \lambda_2 - \log(x_i!)\}$$

$$= z_i \{ \log \pi_1 + x_i \log \lambda_1 - \lambda_1 \} + (1 - z_i) \{ \log \pi_2 + x_i \log \lambda_2 - \lambda_2 \} - \log (x_i!)$$

E-Step:

The posterior probability of a family's receiving family planning:

$$\gamma_i = P(z_i = 1 \mid x_i)$$

$$= \frac{P(z_i = 1) \, P(x_i \mid z_i = 1)}{P(z_i = 1) \, P(x_i \mid z_i = 1) + P(z_i = 0) \, P(x_i \mid z_i = 0)}$$

$$= \frac{\pi_1 \, P(x_i \mid \lambda_1)}{\pi_1 P(x_i \mid \lambda_1) + \pi_2 P(x_i \mid \lambda_2)}$$

∴ The posterior probability that a family has no family planning $= 1 - \gamma_i$

M-Step:

After the E-step, we plug-in $z_i \leftarrow \gamma_i$ and get the equation:

$$L = \log P(x, z | \lambda_1, \lambda_2, \Pi_1)$$

$$= \gamma_i \{\log \Pi_1 + x_i \log \lambda_1 - \lambda_1\}$$

$$+ (1-\gamma_i)\{\log \Pi_2 + x_i \log \lambda_2 - \lambda_2\}$$

$$- \log(x_i!) \quad \ldots\ldots (1)$$

Taking the first order derivative of L wrt λ_1,

$$\frac{\partial L}{\partial \lambda_1} = \sum_{i=1}^{N} \gamma_i \left\{ \frac{x_i}{\lambda_1} - 1 \right\} = 0$$

$$\Rightarrow \sum_{i=1}^{N} \frac{\gamma_i x_i}{\lambda_1} = \sum_{i=1}^{N} \gamma_i$$

$$\therefore \lambda_1 = \frac{\sum_{i=1}^{N} \gamma_i x_i}{\sum_{i=1}^{N} \gamma_i}$$

Taking the first order derivative of L wrt. λ_2,

$$\frac{\partial L}{\partial \lambda_2} = \sum_{i=1}^{N} (1-\gamma_i)\left(\frac{x_i}{\lambda_2} - 1\right) = 0$$

$$\Rightarrow \sum_{i=1}^{N} \frac{(1-\gamma_i)x_i}{\lambda_2} = \sum_{i=1}^{N} (1-\gamma_i)$$

$$\therefore x_2 = \frac{\sum_{i=1}^{N} (1-\gamma_i)x_i}{\sum_{i=1}^{N} (1-\gamma_i)}$$

Taking the first order derivative wrt. π_1,

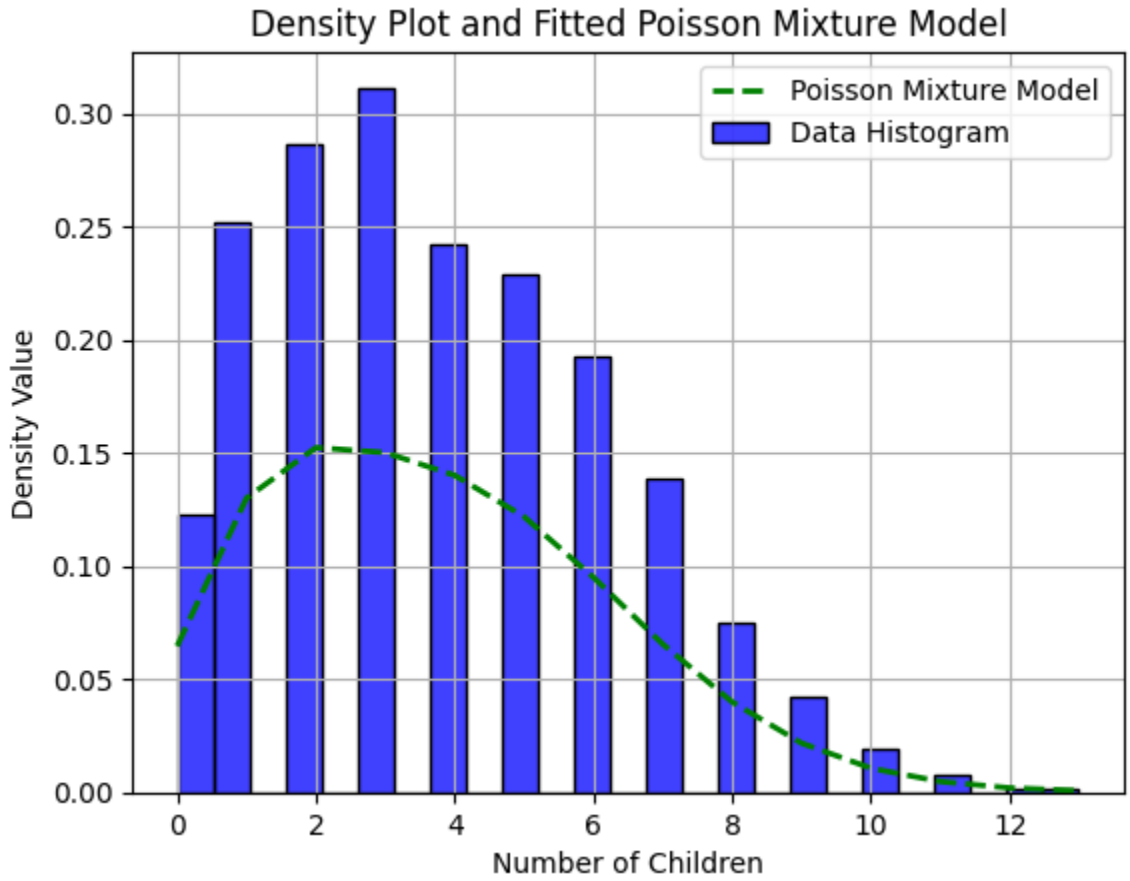
$$\frac{\partial L}{\partial \pi_1} = \sum_{i=1}^{N} \left\{ \frac{\gamma_i}{\pi_1} - \frac{1-\gamma_i}{1-\pi_1} \right\} = 0$$

$$\Rightarrow \sum_{i=1}^{N} \frac{\gamma_i}{\pi_1} = \sum_{i=1}^{N} \frac{1-\gamma_i}{1-\pi_1}$$

$$\Rightarrow \pi_1 = \frac{\sum_{i=1}^{N} \gamma_i}{N}$$

And hence, $\pi_2 = 1 - \pi_1$

Density Plot and Fitted Mixture Models:



Parameters Output:

Estimated lambda_1 (mean children with family planning): 1.7830212418443654

Estimated lambda_2 (mean children without family planning): 4.911238969730369

Estimated pi_1 (proportion with family planning): 0.3561897114122400 4

Estimated pi_2 (proportion without family planning): 0.643810288587759 9

Instructions on how to run the code:

PCA:

The class customPCA is implemented which takes number of components to be reduced as an input for creating an object instance of the class. The get_eigenvectors() functions requires the data to be expected in the format that rows represent the samples and columns represent the features.

After getting the eigenVectors, we can get the reduced coordinates for the reduced dimensions and plot them. The plot function takes a parameter config which controls which dimensional plot to be generated.

The data feeded into the PCA model must be normalized.

EM:

The function em_algorithm takes the first argument as the dataset, second argument is the maximum number of iterations and the third argument is the tolerance parameter which controls when the iterations can be stopped and taken as converged. It returns 4 parameters for the mixture model. Finally, a density plot is created along with the fitted Poisson mixture model.