PCA (Principal Component Analyze)

1. Standardize the Data (This ensures all features have a mean of 0 and a variance of 1)

2. Compute the Covariance Matrix.

$$\text{Cov}(X_j, X_k) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

$$C = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_p) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \dots & \text{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_p, X_1) & \text{Cov}(X_p, X_2) & \dots & \text{Var}(X_p) \end{bmatrix}$$

3. Compute the Eigenvalues and Eigenvectors.

$$(C - \lambda_i I)v_i = 0$$

$$\det(C - \lambda I) = 0$$

4. Select Principal Components.

- Arrange the eigenvalues in descending order.
- Select the top k eigenvalues (based on explained variance criteria).
- Corresponding eigenvectors form the principal components.

5. Finding Best K (explained variance ratio)

The explained variance ratio tells us how much variance each principal component captures.

$$\text{Explained Variance Ratio}_i = \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j}$$

where λ_i is the eigenvalue of the i-th principal component.

The **cumulative explained variance** up to K components is:

$$\text{Cumulative Explained Variance}(K) = \sum_{i=1}^{K} \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j}$$

6. Transform the Data.

$$Z = XW$$

where:

- X is the standardized dataset (size $n \times p$).
- W is the matrix of top k eigenvectors (size $p \times k$).
- Z is the transformed dataset (size $n \times k$).

7. Reconstructed Matrix and MSE.

$$A_{\text{reconstructed}} = A_{\text{new}} \cdot V_k^T + \mu$$

- A_{new} → The transformed data in the reduced-dimensional space (scores), obtained by projecting the original data onto the top k principal components.
- V_k^T → The transpose of the matrix of the top k eigenvectors (principal components).
- μ → The mean of the original dataset before standardization (added back to reverse mean-centering).
- $A_{\text{reconstructed}}$ → The approximate reconstruction of the original data using the reduced components.

$$MSE = \frac{1}{n \times p} \sum_{i=1}^{n} \sum_{j=1}^{p} (X_{ij} - \hat{X}_{ij})^2$$

Result

Adult Dataset

Total time PCA code without library

1.7990135960008047

Total time PCA code with library

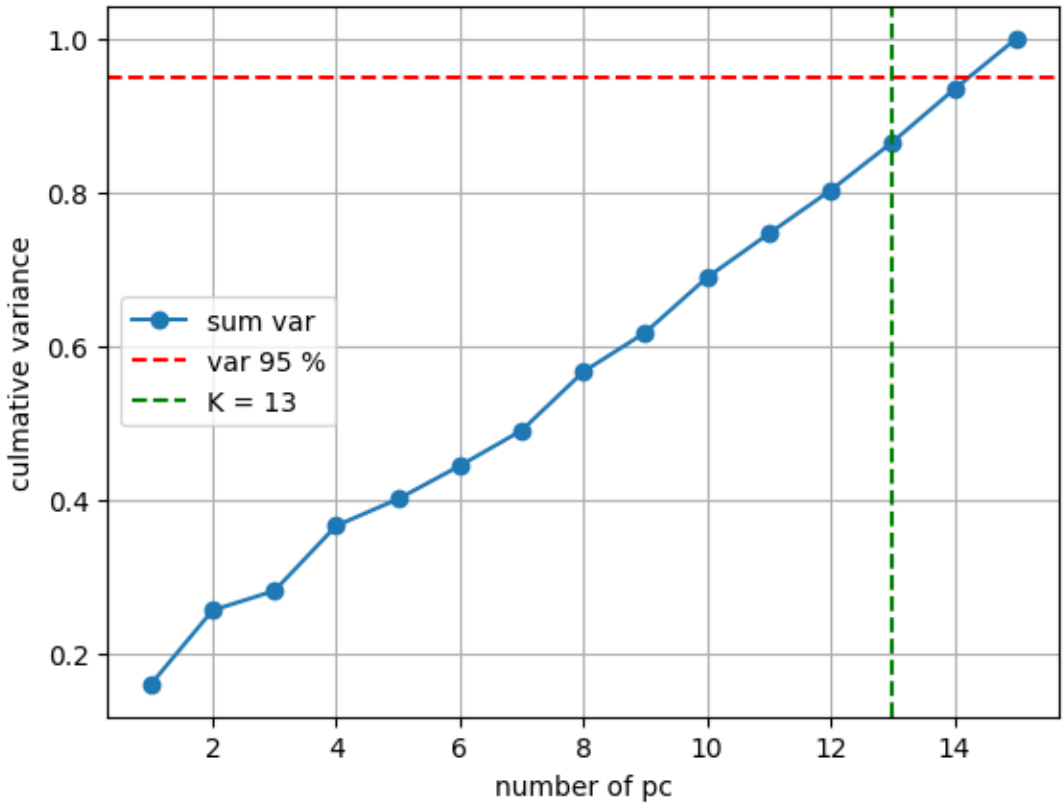
0.007373170999926515

Reconstructed matrix MSE Error for library

(MSE): 0.7434989915360593

Reconstructed matrix MSE Error for "without library"

0.13724514846386968



Result

Flower Dataset

Total time PCA code without library

0.4593327889997454

Total time PCA code with library

0.0019394470000406727

Reconstructed matrix Error for library

(MSE): 0.04199024638518026

Reconstructed matrix Error for "without library"

0.04199024638518034