Principal Component Analysis (PCA): Step-by-Step Implementation

Step 1: Define the Dataset

The dataset consists of two features:

$$\text{Feature 1: } [2.5, 0.5, 2.2, 1.9, 3.1, 2.3, 2.0, 1.0, 1.5, 1.1]$$

$$\text{Feature 2: } [2.4, 0.7, 2.9, 2.2, 3.0, 2.7, 1.6, 1.1, 1.6, 0.9]$$

We organize this into a data frame:

Feature 1	Feature 2
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2.0	1.6
1.0	1.1
1.5	1.6
1.1	0.9

Step 2: Standardize the Data

PCA is sensitive to the scale of data, so we standardize each feature to have a mean of 0 and a standard deviation of 1. The standardization formula is:

$$z = \frac{x - \mu}{\sigma}$$

where x is the feature value, μ is the mean, and σ is the standard deviation.

The standardized data is:

$$\text{Standardized Data: } \begin{bmatrix} 0.69 & 0.60 \\ -1.49 & -1.44 \\ 0.39 & 0.94 \\ 0.05 & 0.26 \\ 1.29 & 1.36 \\ 0.49 & 0.99 \\ -0.12 & -0.86 \\ -1.10 & -1.34 \\ -0.70 & -0.86 \\ -0.91 & -1.06 \end{bmatrix}$$

Step 3: Apply PCA

PCA transforms the standardized data into new features (principal components) that are linear combinations of the original features. These components maximize the variance captured.

Using two principal components, the transformed data is:

$$\text{Principal Components:} \begin{bmatrix} 0.827 & 0.176 \\ -2.279 & -0.364 \\ 0.992 & 0.176 \\ 0.274 & -0.112 \\ 1.676 & 0.299 \\ 1.099 & 0.190 \\ -0.372 & -0.516 \\ -1.962 & -0.304 \\ -1.058 & -0.089 \\ -1.202 & 0.253 \end{bmatrix}$$

Step 4: Analyze Variance Explained

PCA explains the variance in the dataset through its components. The explained variance ratio for each component is:

$$\text{Explained Variance Ratio: } [0.977, 0.023]$$

The first component captures 97.7% of the variance, while the second captures 2.3%.

Step 5: Plot Cumulative Variance Explained

The cumulative variance explained by the components is plotted. This helps determine the number of components required to capture most of the variance.

The Python code:

```
plt.plot(np.cumsum(explained_variance), marker='o', linestyle='--')
plt.title('Cumulative Explained Variance')
plt.xlabel('Number of Components')
plt.ylabel('Cumulative Variance Explained')
```

The resulting plot shows that one component is sufficient to capture most of the variance.

Step 6: Visualize Principal Components

The data in the principal component space is visualized using a scatter plot. Each point represents an observation in terms of the first two principal components.

The Python code:

```
plt.scatter(pca_df['PC1'], pca_df['PC2'], edgecolor='k', s=100)
plt.title('Data in Principal Component Space')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
```

The scatter plot shows the distribution of the data in the reduced space.

Results and Interpretation

- The first principal component ($PC1$) captures most of the variance (97.7%).
- The scatter plot indicates how the data is distributed in the new reduced-dimensional space.
- Dimensionality reduction using PCA is effective here, as one component captures almost all the variance.

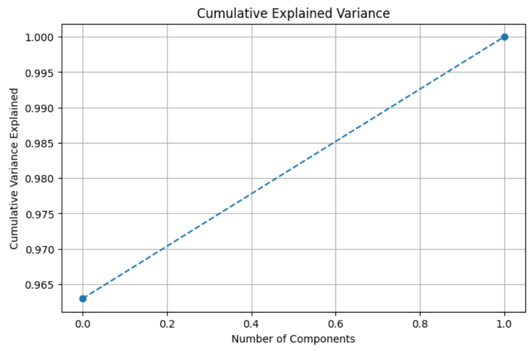
Compile Solution

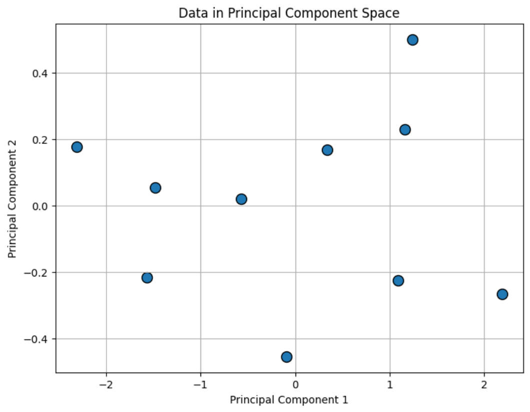
```
# Compile Solution
Original Data:
   Feature1  Feature2
0       2.5       2.4
1       0.5       0.7
2       2.2       2.9
3       1.9       2.2
4       3.1       3.0
5       2.3       2.7
6       2.0       1.6
7       1.0       1.1
8       1.5       1.6
9       1.1       0.9

Standardized Data:
[[ 0.92627881  0.61016865]
 [-1.7585873  -1.506743  ]
 [ 0.52354889  1.23278973]
 [ 0.12081898  0.36112022]
 [ 1.73173864  1.35731394]
 [ 0.6577922   0.9837413 ]
 [ 0.25506228 -0.38602507]
 [-1.08737078 -1.00864614]
 [-0.41615425 -0.38602507]
 [-0.95312747 -1.25769457]]

Principal Components:
        PC1       PC2
0  1.086432 -0.223524
1 -2.308937  0.178081
2  1.241919  0.501509
3  0.340782  0.169919
4  2.184290 -0.264758
5  1.160739  0.230481
6 -0.092605 -0.453317
7 -1.482108  0.055667
8 -0.567226  0.021305
9 -1.563287 -0.215361

Explained Variance Ratio:
[0.96296464 0.03703536]
```


Cumulative Explained Variance
Cumulative Variance Explained
Number of Components


Data in Principal Component Space
Principal Component 2
Principal Component 1