As a data science expert, **Principal Component Analysis (PCA)** is the preferred method for reducing the dimensionality of data, especially when handling highly complex datasets where computational cost scales cubically with the number of dimensions (D). The objective is to transform the data to a lower-dimensional subspace that maximizes the variance of the projected data.

Per your request, I will detail the process of dimensionality reduction from 2 dimensions to 1 using the PCA algorithm via **manual calculation**, applying the key steps defined in the sources.

PCA Application: Dimensionality Reduction (2D to 1D)

We will use the following dataset, which consists of $N = 4$ examples and $D = 2$ features (X1, X2):

Feature	Example 1	Example 2	Example 3	Example 4
X1	4	8	13	7
X2	11	4	5	14

The raw data matrix, $\mathbf{X}$, is:

$$\mathbf{X} = \begin{pmatrix} 4 & 11 \\ 8 & 4 \\ 13 & 5 \\ 7 & 14 \end{pmatrix}$$

Step 1: Mean Subtraction (Centering the Data)

The first step is to center the data by computing the mean (μ) of the entire dataset and subtracting it from every single data point. This ensures the resulting dataset has a mean of 0.

1. **Calculate the Mean (μ):**
   - Mean of Feature X1 (μ_1):

$$\mu_1 = \frac{4 + 8 + 13 + 7}{4} = \frac{32}{4} = 8$$

   - Mean of Feature X2 (μ_2):

$$\mu_2 = \frac{11 + 4 + 5 + 14}{4} = \frac{34}{4} = 8.5$$

$$\mu = \begin{pmatrix} 8 \\ 8.5 \end{pmatrix}$$

2. **Center the Data ($\mathbf{X}_{centered}$):** Subtract the corresponding mean from each row vector in $\mathbf{X}$.

$$\mathbf{X}_{centered} = \mathbf{X} - \mu = \begin{pmatrix} 4-8 & 11-8.5 \\ 8-8 & 4-8.5 \\ 13-8 & 5-8.5 \\ 7-8 & 14-8.5 \end{pmatrix} = \begin{pmatrix} -4 & 2.5 \\ 0 & -4.5 \\ 5 & -3.5 \\ -1 & 5.5 \end{pmatrix}$$

Step 2: Eigendecomposition of the Covariance Matrix ($\mathbf{S}$)

We now compute the data covariance matrix $\mathbf{S}$ and calculate its eigenvalues and corresponding eigenvectors. The eigenvectors associated with the largest eigenvalues form the **principal subspace**.

1. **Calculate the Covariance Matrix ($\mathbf{S}$):**
   We first calculate the unnormalized covariance matrix $\mathbf{C} = \mathbf{X}_{centered}^T \mathbf{X}_{centered}$. (For simplicity in manual calculation, we use N instead of $N-1$ for normalization). $\mathbf{S} = \frac{1}{N}\mathbf{C}$.

$$\mathbf{C} = \begin{pmatrix} -4 & 0 & 5 & -1 \\ 2.5 & -4.5 & -3.5 & 5.5 \end{pmatrix} \begin{pmatrix} -4 & 2.5 \\ 0 & -4.5 \\ 5 & -3.5 \\ -1 & 5.5 \end{pmatrix}$$

   - S_{11} (Variance of X1): $(-4)^2 + 0^2 + 5^2 + (-1)^2 = 16 + 0 + 25 + 1 = 42$
   - S_{22} (Variance of X2): $(2.5)^2 + (-4.5)^2 + (-3.5)^2 + (5.5)^2 = 6.25 + 20.25 + 12.25 + 30.25 = 69.0$
   - S_{12} (Covariance): $(-4)(2.5) + (0)(-4.5) + (5)(-3.5) + (-1)(5.5) = -10 + 0 - 17.5 - 5.5 = -33.0$

   The normalized Covariance Matrix ($\mathbf{S}$):

$$\mathbf{S} = \frac{1}{4}\begin{pmatrix} 42 & -33 \\ -33 & 69 \end{pmatrix} = \begin{pmatrix} 10.5 & -8.25 \\ -8.25 & 17.25 \end{pmatrix}$$

2. **Calculate Eigenvalues (λ):** Solve the characteristic equation $|\mathbf{S} - \lambda\mathbf{I}| = 0$:

$$\det\begin{pmatrix} 10.5-\lambda & -8.25 \\ -8.25 & 17.25-\lambda \end{pmatrix} = 0$$

$$(10.5-\lambda)(17.25-\lambda) - (-8.25)^2 = 0$$

$$181.125 - 10.5\lambda - 17.25\lambda + \lambda^2 - 68.0625 = 0$$

$$\lambda^2 - 27.75\lambda + 113.0625 = 0$$

Using the quadratic formula $\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$:

$$\lambda = \frac{27.75 \pm \sqrt{(-27.75)^2 - 4(1)(113.0625)}}{2}$$

$$\lambda = \frac{27.75 \pm \sqrt{770.0625 - 452.25}}{2} = \frac{27.75 \pm \sqrt{317.8125}}{2} \approx \frac{27.75 \pm 17.827}{2}$$

- **Largest Eigenvalue (λ_1):** This corresponds to the variance captured by the first principal component.

$$\lambda_1 \approx \frac{27.75 + 17.827}{2} \approx 22.79$$

- **Second Eigenvalue (λ_2):**

$$\lambda_2 \approx \frac{27.75 - 17.827}{2} \approx 4.96$$

3. **Calculate the Eigenvector ($\mathbf{v}_1$) for λ_1 (First Principal Component):**
   We seek the eigenvector $\mathbf{v}_1$ corresponding to the largest eigenvalue, $\lambda_1 \approx 22.79$.
   Solve $(\mathbf{S} - \lambda_1 \mathbf{I})\mathbf{v}_1 = \mathbf{0}$:

$$\begin{pmatrix} 10.5 - 22.79 & -8.25 \\ -8.25 & 17.25 - 22.79 \end{pmatrix} \begin{pmatrix} v_{11} \\ v_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} -12.29 & -8.25 \\ -8.25 & -5.54 \end{pmatrix} \begin{pmatrix} v_{11} \\ v_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

   Using the first row equation:

$$-12.29v_{11} - 8.25v_{12} = 0$$

$$v_{12} = -\frac{12.29}{8.25}v_{11} \approx -1.49v_{11}$$

   Setting $v_{11} = 1$, we get $v_{12} \approx -1.49$.

The unit eigenvector $\mathbf{B}$ (the projection matrix):

$$\mathbf{B} \approx \begin{pmatrix} 0.556 \\ -0.830 \end{pmatrix}$$

Note: This eigenvector, associated with the largest eigenvalue, forms the basis of the 1-dimensional principal subspace.

Step 3: Projection onto the Principal Subspace

To reduce the dimension from 2 to 1, we project the centered data ($\mathbf{X}_{centered}$) onto the first principal component, $\mathbf{B}$. The output of PCA is the set of low-dimensional coordinates (z^*), not the full projection.

The low-dimensional coordinate vector $\mathbf{Z}$ is calculated as:

$$\mathbf{Z} = \mathbf{X}_{centered}\mathbf{B}$$

$$\mathbf{Z} \approx \begin{pmatrix} -4 & 2.5 \\ 0 & -4.5 \\ 5 & -3.5 \\ -1 & 5.5 \end{pmatrix} \begin{pmatrix} 0.556 \\ -0.830 \end{pmatrix}$$

1. **Example 1 (z_1):** $(-4)(0.556) + (2.5)(-0.830) = -2.224 - 2.075 = -4.299$
2. **Example 2 (z_2):** $(0)(0.556) + (-4.5)(-0.830) = 0 + 3.735 = 3.735$
3. **Example 3 (z_3):** $(5)(0.556) + (-3.5)(-0.830) = 2.78 + 2.905 = 5.685$
4. **Example 4 (z_4):** $(-1)(0.556) + (5.5)(-0.830) = -0.556 - 4.565 = -5.121$

Final Reduced 1D Data ($\mathbf{Z}$):
The original 4 examples, previously defined by two features (X1, X2), are now represented by a single principal component coordinate:

$$\mathbf{Z} \approx \begin{pmatrix} -4.30 \\ 3.74 \\ 5.69 \\ -5.12 \end{pmatrix}$$

As a data science expert, **Principal Component Analysis (PCA)** is a powerful technique to reduce the dimension of the data while preserving maximal variance. PCA finds orthogonal linear combinations (principal components) that project the data onto a lower-dimensional principal subspace.

Below is the detailed manual calculation to reduce the dimension of the given 2D dataset to 1D using the PCA algorithm.

PCA Application: Dimensionality Reduction (Problem Set 2)

We are given a dataset with $N = 6$ examples and $D = 2$ features (X1, X2):

Feature	Ex 1	Ex 2	Ex 3	Ex 4	Ex 5	Ex 6
X1	2	3	4	5	6	7
X2	1	5	3	6	7	8

The raw data matrix, $\mathbf{X}$, is:

$$\mathbf{X} = \begin{pmatrix} 2 & 1 \\ 3 & 5 \\ 4 & 3 \\ 5 & 6 \\ 6 & 7 \\ 7 & 8 \end{pmatrix}$$

Step 1: Mean Subtraction (Centering the Data)

We must center the data by computing the mean (μ) for each feature and subtracting it from all corresponding data points. This ensures the dataset has a mean of 0.

1. **Calculate the Mean (μ):**
    - Mean of Feature X1 (μ_1):

$$\mu_1 = \frac{2+3+4+5+6+7}{6} = \frac{27}{6} = 4.5$$

    - Mean of Feature X2 (μ_2):

$$\mu_2 = \frac{1+5+3+6+7+8}{6} = \frac{30}{6} = 5.0$$

$$\mu = \begin{pmatrix} 4.5 \\ 5.0 \end{pmatrix}$$

2. **Center the Data ($\mathbf{X}_{centered}$):**

$$\mathbf{X}_{centered} = \mathbf{X} - \mu = \begin{pmatrix} 2-4.5 & 1-5 \\ 3-4.5 & 5-5 \\ 4-4.5 & 3-5 \\ 5-4.5 & 6-5 \\ 6-4.5 & 7-5 \\ 7-4.5 & 8-5 \end{pmatrix} = \begin{pmatrix} -2.5 & -4 \\ -1.5 & 0 \\ -0.5 & -2 \\ 0.5 & 1 \\ 1.5 & 2 \\ 2.5 & 3 \end{pmatrix}$$

Step 2: Eigendecomposition of the Covariance Matrix ($\mathbf{S}$)

The data covariance matrix $\mathbf{S}$ must be computed, followed by calculating its eigenvalues (λ) and eigenvectors ($\mathbf{v}$).

1. **Calculate the Covariance Matrix ($\mathbf{S}$):**
   We calculate the unnormalized matrix $\mathbf{C} = \mathbf{X}_{centered}^T \mathbf{X}_{centered}$.
   - S_{11}: $(-2.5)^2 + (-1.5)^2 + (-0.5)^2 + (0.5)^2 + (1.5)^2 + (2.5)^2 = 17.5$
   - S_{22}: $(-4)^2 + 0^2 + (-2)^2 + 1^2 + 2^2 + 3^2 = 34.0$
   - S_{12} (Covariance): $(-2.5)(-4) + (-1.5)(0) + (-0.5)(-2) + (0.5)(1) + (1.5)(2) + (2.5)(3) = 10 + 0 + 1 + 0.5 + 3 + 7.5 = 22.0$

   Unnormalized Covariance Matrix ($\mathbf{C}$):

$$\mathbf{C} = \begin{pmatrix} 17.5 & 22.0 \\ 22.0 & 34.0 \end{pmatrix}$$

2. **Calculate Eigenvalues (λ):** We solve the characteristic equation $|\mathbf{C} - \lambda \mathbf{I}| = 0$:

$$(17.5 - \lambda)(34.0 - \lambda) - (22.0)^2 = 0$$

$$595 - 17.5\lambda - 34.0\lambda + \lambda^2 - 484 = 0$$

$$\lambda^2 - 51.5\lambda + 111 = 0$$

   Using the quadratic formula, we find the eigenvalues:

$$\lambda = \frac{51.5 \pm \sqrt{(51.5)^2 - 4(1)(111)}}{2} \approx \frac{51.5 \pm \sqrt{2652.25 - 444}}{2} \approx \frac{51.5 \pm 46.99}{2}$$

   - **Largest Eigenvalue (λ_1):**

$$\lambda_1 \approx \frac{51.5 + 46.99}{2} \approx 49.25$$

- **Second Eigenvalue (λ_2):**

$$\lambda_2 \approx \frac{51.5 - 46.99}{2} \approx 2.25$$

3. **Calculate the Eigenvector ($\mathbf{v}_1$) for λ_1 (First Principal Component):**
   The projection matrix $\mathbf{B}$ uses the eigenvector corresponding to the largest eigenvalue ($\lambda_1 \approx 49.25$). We solve $(\mathbf{C} - \lambda_1 \mathbf{I})\mathbf{v}_1 = \mathbf{0}$:

$$\begin{pmatrix} 17.5 - 49.25 & 22.0 \\ 22.0 & 34.0 - 49.25 \end{pmatrix} \begin{pmatrix} v_{11} \\ v_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} -31.75 & 22.0 \\ 22.0 & -15.25 \end{pmatrix} \begin{pmatrix} v_{11} \\ v_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

   Using the first row equation:

$$-31.75 v_{11} + 22.0 v_{12} = 0$$

$$v_{12} = \frac{31.75}{22.0} v_{11} \approx 1.443 v_{11}$$

   We normalize this vector to find the projection basis $\mathbf{B}$.
   Unit Vector calculation: $v_{11}^2 + (1.443 v_{11})^2 = 1 \implies 3.085 v_{11}^2 \approx 1$.
   $v_{11} \approx 0.568$. Thus, $v_{12} \approx 1.443 \times 0.568 \approx 0.820$.

$$\mathbf{B} \approx \begin{pmatrix} 0.568 \\ 0.820 \end{pmatrix}$$

Step 4: Projection onto the Principal Subspace

We project the centered data ($\mathbf{X}_{centered}$) onto the first principal component, $\mathbf{B}$, to obtain the final 1D coordinates ($\mathbf{Z}$).

The low-dimensional coordinate vector $\mathbf{Z}$ is calculated as $\mathbf{Z} = \mathbf{X}_{centered}\mathbf{B}$:

$$\mathbf{Z} \approx \begin{pmatrix} -2.5 & -4 \\ -1.5 & 0 \\ -0.5 & -2 \\ 0.5 & 1 \\ 1.5 & 2 \\ 2.5 & 3 \end{pmatrix} \begin{pmatrix} 0.568 \\ 0.820 \end{pmatrix}$$

Example	**Calculation (z_n)**	**Reduced Coordinate (z_n)**
1	$(-2.5)(0.568) + (-4)(0.820) = -1.42 - 3.28$	**-4.70**
2	$(-1.5)(0.568) + (0)(0.820) = -0.852 + 0$	**-0.85**
3	$(-0.5)(0.568) + (-2)(0.820) = -0.284 - 1.64$	**-1.92**
4	$(0.5)(0.568) + (1)(0.820) = 0.284 + 0.82$	**1.10**
5	$(1.5)(0.568) + (2)(0.820) = 0.852 + 1.64$	**2.49**
6	$(2.5)(0.568) + (3)(0.820) = 1.42 + 2.46$	**3.88**

The final dataset, reduced from 2 dimensions to 1, is approximately:

$$\mathbf{Z} \approx \begin{pmatrix} -4.70 \\ -0.85 \\ -1.92 \\ 1.10 \\ 2.49 \\ 3.88 \end{pmatrix}$$