

Detailed PCA Math Calculations - Hotel Booking Behavior

This document provides a detailed step-by-step mathematical calculation of Principal Component Analysis (PCA) using a hotel booking dataset with two features: Nights Stayed (X1) and Room Service Orders (X2).

Step 1: Sample Dataset

Nights Stayed (X1)	Room Service Orders (X2)
2	3
3	5
4	6
5	7
6	9
7	11

Step 2: Mean and Standard Deviation

Calculate the mean of each feature:

$$\text{Mean}_{X1} = (2 + 3 + 4 + 5 + 6 + 7) / 6 = 4.5$$

$$\text{Mean}_{X2} = (3 + 5 + 6 + 7 + 9 + 11) / 6 = 6.83 \text{ (approx)}$$

Calculate the standard deviation of each feature:

► Standard Deviation (σ):

$$\sigma_{X1} = \sqrt{\frac{(2 - 4.5)^2 + (3 - 4.5)^2 + (4 - 4.5)^2 + (5 - 4.5)^2 + (6 - 4.5)^2 + (7 - 4.5)^2}{6 - 1}} = \sqrt{\frac{17.5}{5}} = \sqrt{3.5} \approx 1.87$$

$$\sigma_{X2} = \sqrt{\frac{(3 - 6.83)^2 + (5 - 6.83)^2 + (6 - 6.83)^2 + (7 - 6.83)^2 + (9 - 6.83)^2 + (11 - 6.83)^2}{5}} \approx 2.86$$

Step 3: Standardization (Z-Score)

$$Z = (X - \text{Mean}) / \text{Standard Deviation}$$

Data Point	Z(X1)	Z(X2)
1	-1.34	-1.34
2	-0.80	-0.64
3	-0.27	-0.29
4	0.27	0.06
5	0.80	0.76
6	1.34	1.46

Step 4: Covariance Matrix

The Covariance matrix shows how much each feature varies from the mean with respect to each other.

$$Cov(X, Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{N - 1}$$

Where:

- X_i and Y_i represent the observed values of X and Y.
- $\bar{X}$ and $\bar{Y}$ denote their respective means.
- N is the number of observations.

Once the data is standardized using Z-scores, each value of X and Y is transformed into:

$$Z_X = \frac{X_i - \bar{X}}{SD_X} \quad \text{and} \quad Z_Y = \frac{Y_i - \bar{Y}}{SD_Y}$$

This means that each value now has a mean of 0 and a standard deviation of 1.

As a result, the covariance formula simplifies when applied to standardized data.

▶ New Covariance Formula (Standardized Data):

$$Cov(Z_X, Z_Y) = \frac{\sum Z_{X_i} \cdot Z_{Y_i}}{N - 1}$$

Since the means of standardized variables are zero, we no longer need to subtract the mean again inside the formula — we directly multiply the Z-scores.

Products of Z-scores:

$Z1 \cdot Z2 = [1.80, 0.51, 0.08, 0.02, 0.61, 1.96]$

Sum = 4.98

$Cov(Z1, Z2) = 4.98 / 5 = 0.996$

Covariance Matrix:

Covariance Matrix Formula (2D case):

$$\text{Covariance Matrix (C)} = \begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) \\ Cov(X_2, X_1) & Cov(X_2, X_2) \end{bmatrix}$$

Here

Final Covariance Matrix (Standardized Data):

$$C = \begin{bmatrix} 1 & 0.996 \\ 0.996 & 1 \end{bmatrix}$$

Step 5: Eigenvalues and Eigenvectors

Solve $|C - \lambda I| = 0$

$|1 - \lambda \quad 0.996|$

$|0.996 \quad 1 - \lambda|$

Determinant: $(1-\lambda)^2 - (0.996)^2 = 0$

$$(1-\lambda)^2 = 0.992$$

$$1 - 2\lambda + \lambda^2 = 0.992$$

$$\lambda^2 - 2\lambda + 0.008 = 0$$

$$\lambda_1 \approx 1.996, \lambda_2 \approx 0.004$$

Variance Explained:

$$PC1 = (1.996 / 2.0) * 100 \approx 99.8\%$$

$$PC2 = (0.004 / 2.0) * 100 \approx 0.2\%$$

Using the equation:

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

For $\lambda = 1.996$:

$$\begin{bmatrix} 1 - 1.996 & 0.996 \\ 0.996 & 1 - 1.996 \end{bmatrix} \cdot \begin{bmatrix} a \\ b \end{bmatrix} = 0 \Rightarrow -0.996a + 0.996b = 0 \Rightarrow a = b$$

Eigenvector (not normalized):

$$v = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \text{Normalize: } \frac{1}{\sqrt{2}} = 0.71 \Rightarrow \text{PC1 Vector} = \begin{bmatrix} 0.71 \\ 0.71 \end{bmatrix}$$

PCA gives us an eigenvector direction (here $[1, 1]$), which gets normalized to $[0.71, 0.71]$ for unit length. This means both features contribute equally to PC1 — a new composite dimension that simplifies customer behavior interpretation.

Step 6: Principal Component Equations

Eigenvectors indicate directions of maximum variance.

$$PC1 = 0.71 * Z1 + 0.71 * Z2$$

$$PC2 = -0.71 * Z1 + 0.71 * Z2$$

Step 7: PC1 Projection - Guest Activity Index

Guest	PC1 Score
1	-1.89
2	-1.02
3	-0.40

4	0.23
5	1.10
6	1.98

Final Conclusion

Through PCA, we've reduced a 2D dataset (two correlated features) into a single new component (PC1) while retaining most of the meaningful variance. The Guest Activity Index (PC1) provides a simplified yet powerful way to represent customer behavior patterns for further analysis, clustering, or segmentation.