Các bước thực hiện PCA:

1. **Tính vector kỳ vọng (mean vector):**

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \text{với } \bar{x} \text{ là vector trung bình của toàn bộ dữ liệu.}$$

2. **Chuẩn hóa dữ liệu:** Trừ mỗi điểm dữ liệu đi vector kỳ vọng:

$$\hat{x}_n = x_n - \bar{x}, \quad \text{tạo thành dữ liệu đã chuẩn hóa } \hat{X}.$$

3. **Tính ma trận hiệp phương sai:**

$$S = \frac{1}{N}\hat{X}\hat{X}^T, \quad \text{với } S \text{ là ma trận hiệp phương sai kích thước } m \times m.$$

4. **Tính trị riêng và vector riêng:** - Giải bài toán trị riêng cho ma trận S:

$$Sv_i = \lambda_i v_i, \quad \text{với } v_i \text{ là vector riêng và } \lambda_i \text{ là trị riêng.}$$

- Sắp xếp các trị riêng λ_i theo thứ tự giảm dần:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m.$$

5. **Chọn số lượng thành phần chính:** - Chọn K vector riêng đầu tiên tương ứng với K trị riêng lớn nhất, tạo thành ma trận $U_K \in \mathbb{R}^{m \times K}$ với các cột là các vector riêng. - K vector riêng này (còn gọi là các thành phần chính) tạo thành không gian con gần nhất với phân bố dữ liệu ban đầu.

6. **Chiếu dữ liệu lên không gian con mới:**

$$Z = U_K^T \hat{X}, \quad \text{với } Z \text{ là dữ liệu mới trong không gian } K \text{ chiều.}$$

7. **Xấp xỉ dữ liệu ban đầu từ dữ liệu mới:**

$$x \approx U_K Z + \bar{x}, \quad \text{với } U_K Z \text{ là dữ liệu xấp xỉ trong không gian gốc.}$$

Hàm loss của PCA:

PCA sử dụng hàm loss để tối thiểu hóa sai số tái tạo giữa dữ liệu ban đầu và dữ liệu xấp xỉ. Hàm loss được định nghĩa như sau:

$$\mathcal{L} = \sum_{i=1}^{N} \left\| x_i - \left(U_K U_K^T (x_i - \bar{x}) + \bar{x} \right) \right\|^2,$$

hoặc dạng đơn giản hơn (khi dữ liệu đã chuẩn hóa):

$$\mathcal{L} = \|\hat{X} - U_K U_K^T \hat{X}\|_F^2,$$

trong đó: - $\|\cdot\|_F$: Chuẩn Frobenius (tổng bình phương tất cả các phần tử của ma trận). - $U_K U_K^T$: Phép chiếu dữ liệu lên không gian con K chiều.

—

- Số lượng thành phần chính K được chọn sao cho hàm loss $\mathcal{L}$ nhỏ nhất, đồng thời cân bằng giữa độ chính xác và số chiều của dữ liệu. - Tổng phương sai giải thích bởi K thành phần chính được đo bằng:

$$\text{Explained Variance Ratio} = \frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{m} \lambda_i}.$$

- Chọn K sao cho tỷ lệ phương sai giải thích đạt ngưỡng mong muốn (ví dụ: 95% hoặc 99%).