Optimization Techniques for Machine Learning

AMLZC326 · #11 PCA I

Anshid Aboobacker

Introduction

- We will look at principal components analysis and dimension reduction in this lecture.
- High-dimensional data is hard to visualize and interpret. Can we project this data into lower dimensions while preserving the semantics of the data?
- Higher dimensional data is often overcomplete, meaning there are redundant dimensions which can be explained by combinations of other dimensions.
- Dimensions in high-dimensional data might be correlated, so the actual data may have an intrinsic lower-dimensional structure.

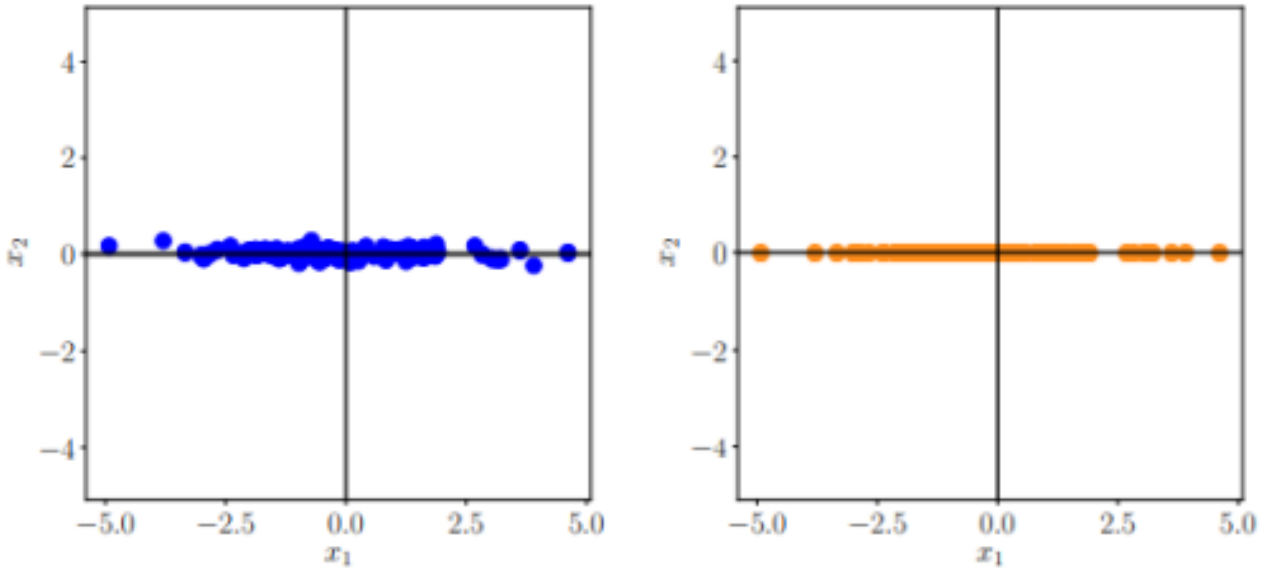
Learning Objectives

By the end of this lecture you should be able to:

- Motivate PCA as the optimal linear dimensionality reduction technique
- Derive the maximum-variance objective and show it leads to an eigenvalue problem for the covariance matrix
- Define principal components as eigenvectors of S sorted by decreasing eigenvalue
- Project data onto the principal subspace and compute the explained variance ratio

Principal Components Analysis

- PCA is a technique for linear dimensionality reduction.
- It was first proposed by Pearson in 1900 and independently rediscovered by Hotelling in 1933.



(a) Dataset with x_1 and x_2 coordinates.

(b) Compressed dataset where only the x_1 coordinate is relevant.

Problem Setting

- Given: $\{x_1, x_2, \ldots, x_N\}$ where $x_n \in \mathbb{R}^D$ is an i.i.d dataset with mean 0.
- The data covariance matrix is: $S = \frac{1}{N} \sum_{n=1}^{N} x_n x_n^T$
- Aim: Find projections $\tilde{x}_n \in U \subseteq \mathbb{R}^D$ such that $\dim(U) = M << D$, while preserving similarity to original data.
- We seek a compressed representation: $z_n = B^T x_n$ where $B = [b_1, \ldots, b_M] \in \mathbb{R}^{D \times M}$.
- The columns of B are orthonormal:

$$b_i^T b_j = 0 \quad (i \neq j), \quad b_i^T b_i = 1$$

Problem Setting

- There exists a linear relationship: $z = B^T x, \quad \tilde{x} = Bz$
- The variable z acts as a lower-dimensional bottleneck representation controlling information flow between x and $\tilde{x}$.

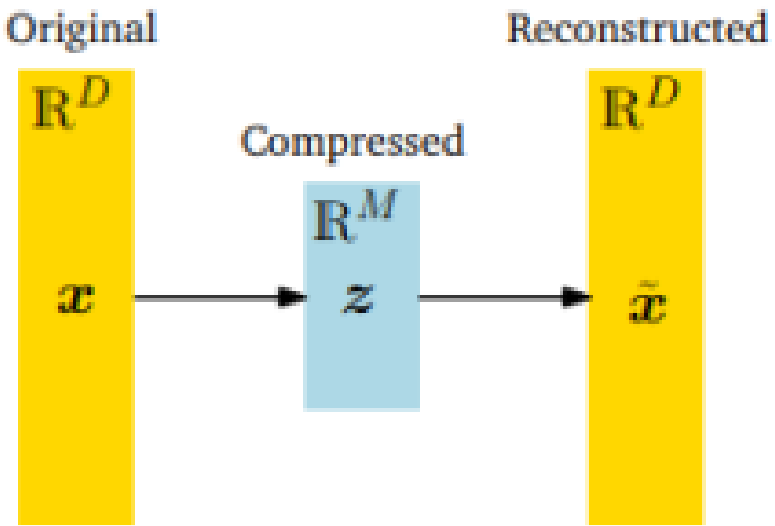


Table of Contents

Maximum Variance Perspective

- We interpret information content in data as how "space-filling" it is.
- The spread of the data captures this information, which we measure using variance.
- PCA can be viewed as a dimensionality reduction algorithm that maximizes the variance in the lower-dimensional representation.
- Our goal is to find a matrix B such that projections onto $b_1, b_2, \ldots, b_M$ retain maximum information.

Centred Data

- Let μ be the mean of the data. Centered data means working with $x - \mu$ instead of x.
- Note that:

$$V_z(z) = V_x(B^T(x - \mu)) = V_x(B^T x - B^T \mu) = V_x(B^T x)$$

- Thus, centering does not change the variance.
- Hence, we assume:

$$\mathbb{E}_x(x) = 0 \quad \Rightarrow \quad \mathbb{E}_z(z) = B^T \mathbb{E}_x(x) = 0$$

- Covariance matrix: $S = \frac{1}{N} \sum_{n=1}^{N} x_n x_n^T$

Direction with Maximal Variance

- We maximize the variance of the low-dimensional code by following a sequential approach.
- **Aim 1:** To maximize the variance V_1 of the first coordinate z_{1n} of $z \in \mathbb{R}^M$.
- i.e. to maximize

$$V_1 = V(z_1) = \frac{1}{N} \sum_{n=1}^{N} z_{1n}^2$$

since the data x is independent.

- Now $z_{1n} = b_1^T x_n$, and can be viewed as the orthogonal projection of x_n onto the one-dimensional subspace spanned by b_1.

Direction with Maximal Variance

$$
\begin{aligned}
V_1 &= \frac{1}{N}\sum_{n=1}^{N}(b_1^T x_n)^2 \\
&= \frac{1}{N}\sum_{n=1}^{N} b_1^T x_n x_n^T b_1 \\
&= b_1^T \left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n^T\right) b_1 \\
&= b_1^T S b_1
\end{aligned}
$$

Arbitrarily increasing the magnitude of the vector b_1 will increase the variance. Hence, we maximize variance subject to: $\|b_1\| = 1$

Direction with Maximal Variance

- To find the direction b_1 that maximizes variance, we solve:

$$\max \; b_1^T S b_1 \quad \text{subject to} \quad \|b_1\| = 1$$

- Form the Lagrangian:

$$\mathcal{L}(b_1, \lambda) = b_1^T S b_1 + \lambda(1 - b_1^T b_1)$$

Solving the Lagrangian

- To solve the Lagrangian, let $\frac{\partial \mathcal{L}}{\partial \lambda} = 0$ and $\frac{\partial \mathcal{L}}{\partial b_1} = 0$.
- So, we get:

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 1 - b_1^T b_1 = 0$$

$$\frac{\partial \mathcal{L}}{\partial b_1} = 2Sb_1 - 2\lambda b_1 = 0$$

- On simplification:

$$Sb_1 = \lambda b_1, \quad b_1^T b_1 = 1$$

- Thus, b_1 is an eigenvector of the covariance matrix S, and λ is the corresponding eigenvalue.

First Principal Component

- Substituting into the objective:

$$b_1^T S b_1 = b_1^T (\lambda b_1) = \lambda$$

- Maximizing the objective is equivalent to maximizing λ.
- Hence, we choose the eigenvector of S corresponding to the largest eigenvalue.
- This is called the **first principal component**.
- We now examine the inner workings of the Lagrangian method.

M-dimensional Subspace with Maximum Variance

- Assume we have found the first $m-1$ principal components as eigenvectors of S corresponding to the largest $m-1$ eigenvalues.
- Since S is symmetric, by the spectral theorem, these eigenvectors form an orthonormal basis for an $(m-1)$-dimensional subspace of $\mathbb{R}^D$.
- The m-th principal component is found by removing the contribution of the first $m-1$ components:

$$b_1, b_2, \ldots, b_{m-1}$$

- This ensures that each new component captures the remaining variance (information).

Table of Contents

M-DIMENSIONAL SUBSPACE WITH MAXIMUM VARIANCE

- $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{D \times N}$ contains the data points x_k as column vectors.
- Then, new data matrix $\hat{X}$ is given by

$$\hat{X} = X - \sum_{i=1}^{m-1} b_i b_i^T X$$

$$= X - B_{m-1} X$$

- Here $B_{m-1} = \sum_{i=1}^{m-1} b_i b_i^T$ is a projection matrix that projects X onto the subspace spanned by $b_1, b_2, \ldots, b_{m-1}$.

M-dimensional subspace with maximum variance

- To find the mth principal component we maximize

$$V_m = \mathbb{V}[z_m] = \frac{1}{N}\sum_{n=1}^{N} z_{mn}^2$$

$$= \frac{1}{N}\sum_{n=1}^{N}(b_m^T \hat{x}_n)^2$$

$$= b_m^T \hat{S} b_m$$

- Here $\hat{S}$ is the data covariance matrix of the transformed data set represented by $\hat{X} = [\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N]$.

M-dimensional subspace with maximum variance

- As before we set up a constrained optimization problem to find the first principal component, and establish that the optimal solution b_m is the eigenvector of $\hat{S}$ that corresponds to the largest eigenvalue.
- We now establish that b_m is an eigenvector of the original data matrix X.
- More generally the sets of eigenvectors for $\hat{S}$ and S are the same.

Eigenvectors of S and $\hat{S}$

- We now show that the eigenvectors of S and $\hat{S}$ are the same.
- Let b_i be an eigenvector of S, i.e. $Sb_i = \lambda b_i$.
- Now we can write

$$\hat{S}b_i = \frac{1}{N}(X - B_{m-1}X)(X - B_{m-1}X)^T b_i$$

$$= (S - SB_{m-1}^T - B_{m-1}S + B_{m-1}SB_{m-1}^T)b_i$$

$$= (S - SB_{m-1} - B_{m-1}S + B_{m-1}SB_{m-1})b_i$$

- Note that in the last line we have used the fact that B_{m-1} is a projection matrix and is therefore symmetric.

Eigenvectors of S and $\hat{S}$

- **Case 1:** $i \geq m$.
- b_i is an eigenvector not among the first $m-1$ components.
- Since $B_{m-1} = \sum_{j=1}^{m-1} b_j b_j^T$ and b_m is orthogonal to the b_i, $1 \leq i \leq m-1$, we have $B_{m-1} b_i = 0$.
- Plugging this into the previous equation,

$$\hat{S} b_i = (S - B_{m-1} S) b_i = S b_i = \lambda_i b_i.$$

- Thus $S b_m = \lambda_m b_m$. λ_m is the mth largest eigenvalue of S and is also the largest eigenvalue of $\hat{S}$.

Eigenvectors of S and $\hat{S}$

- **Case 2:** $i \leq m-1$.
- We have

$$B_{m-1} b_i = \sum_{j=1}^{m-1} b_j b_j^T b_i = b_i.$$

- Plugging this into

$$\hat{S} b_i = (S - S B_{m-1} - B_{m-1} S + B_{m-1} S B_{m-1}) b_i,$$

we get

$$\hat{S} b_i = 0.$$

- Thus the vectors $b_1, b_2, \ldots, b_{m-1}$ are eigenvectors of $\hat{S}$ with eigenvalue 0.

Eigenvectors of S and $\hat{S}$

- Since $V_m = b_m^T S b_m = \lambda_m$, we see that the variance of the data projected onto the mth principal component is λ_m.
- To find an M-dimensional subspace that retains as much information as possible, PCA tells us to choose the columns of matrix B as the M eigenvectors of the data covariance matrix S that have the largest eigenvalues.

Table of Contents

Projection perspective

- We derived the PCA as an algorithm that maximizes the variance in the projected space to retain as much information as possible.
- Now we can also derive the PCA using a projection perspective to minimize the average reconstruction error. The original data is modeled as x_n and the reconstruction is modeled as $\tilde{x}_n$. We seek to minimize the distance between x_n and $\tilde{x}_n$.

Table of Contents

An Example

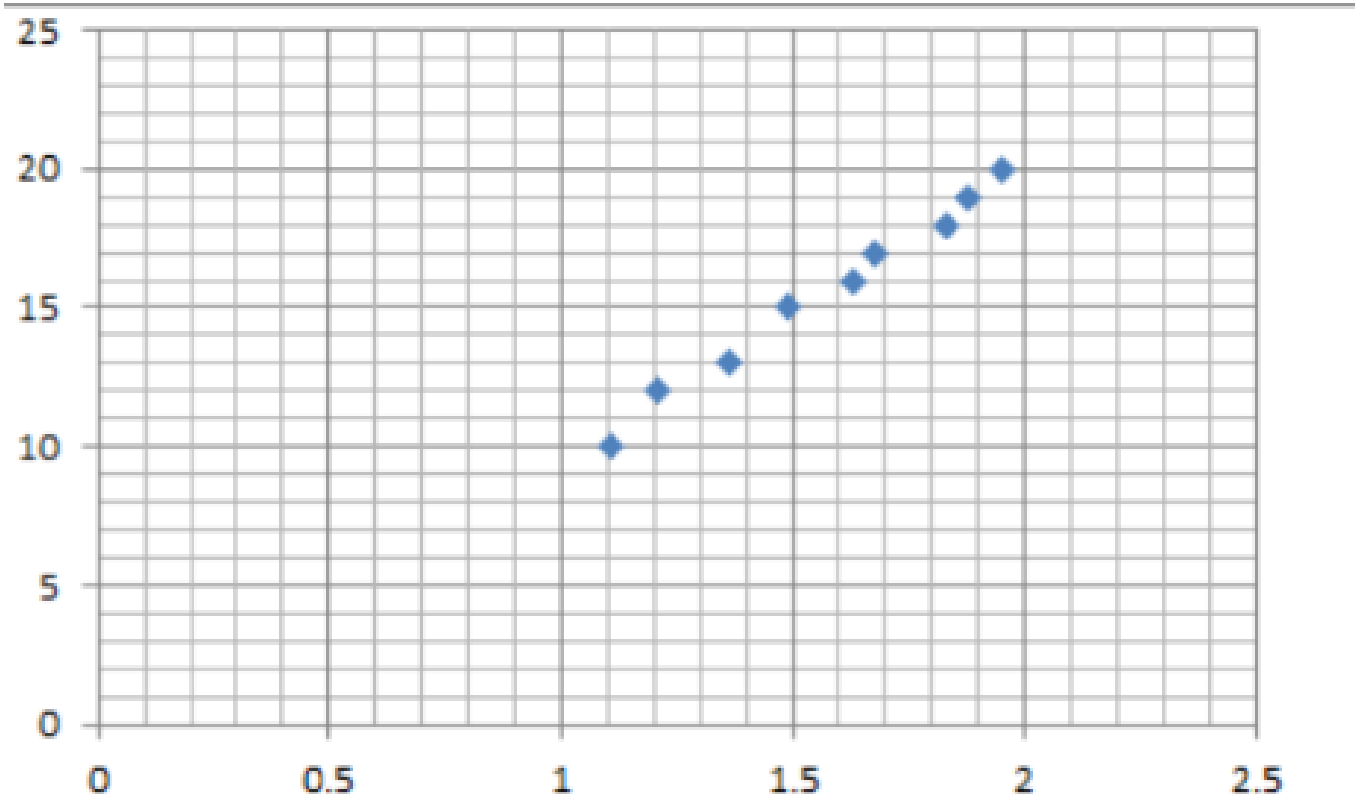
Data matrix (each row is a data point):

$$X = \begin{bmatrix} 1.11 & 10 \\ 1.21 & 12 \\ 1.36 & 13 \\ 1.49 & 15 \\ 1.63 & 16 \\ 1.68 & 17 \\ 1.83 & 18 \\ 1.88 & 19 \\ 1.95 & 20 \end{bmatrix}$$

Mean vector:

$$\mu = \begin{bmatrix} 1.5711 & 15.5556 \end{bmatrix}$$

Data Visualization



Centering the Data

Centered data:

$$\tilde{X} = X - \mathbf{1}\mu$$

$$\tilde{X} = \begin{bmatrix} -0.4611 & -5.5556 \\ -0.3611 & -3.5556 \\ -0.2111 & -2.5556 \\ -0.0811 & -0.5556 \\ 0.0589 & 0.4444 \\ 0.1089 & 1.4444 \\ 0.2589 & 2.4444 \\ 0.3089 & 3.4444 \\ 0.3789 & 4.4444 \end{bmatrix}$$

Covariance Matrix

$$S = \frac{1}{N}\tilde{X}^T\tilde{X}$$

$$S = \begin{bmatrix} 0.0795 & 0.8883 \\ 0.8883 & 10.0247 \end{bmatrix}$$

Largest eigenvalue: $\lambda_1 \approx 10.103$

Principal eigenvector:

$$b_1 = \begin{bmatrix} 0.0883 \\ 0.9961 \end{bmatrix}$$

Projection onto Principal Component

$$z = \tilde{X} b_1$$

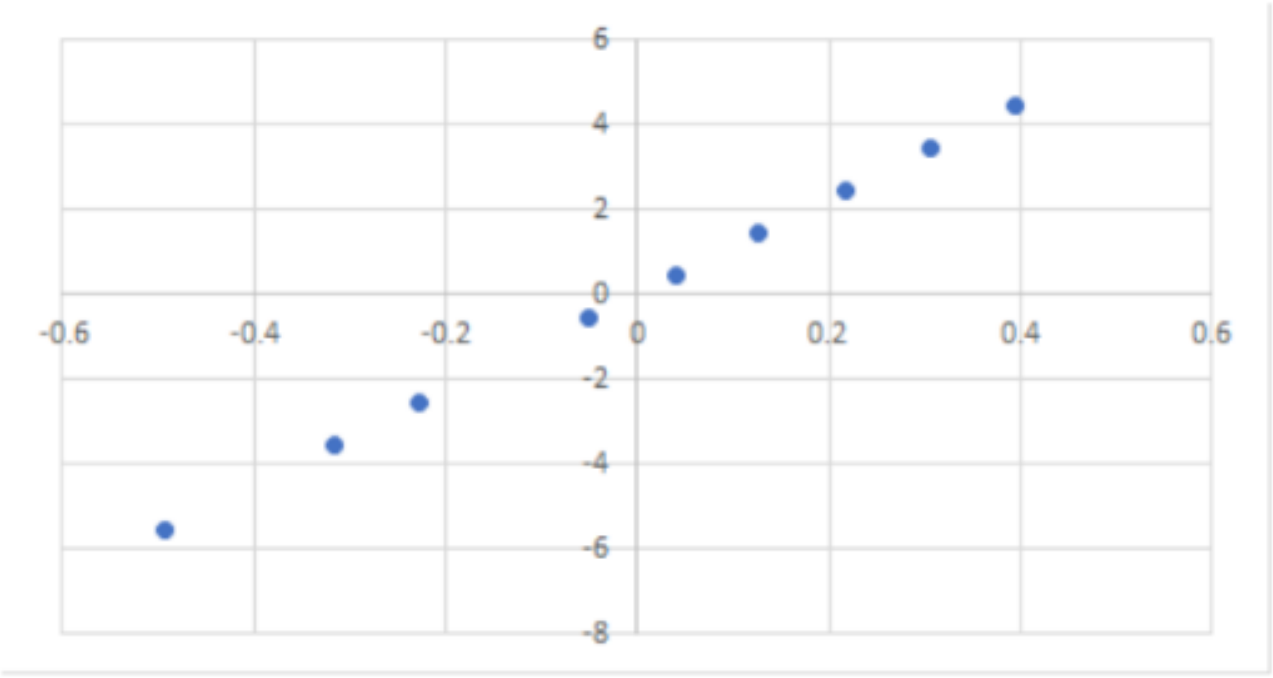
$$z = \begin{bmatrix} -5.57 \\ -3.57 \\ -2.56 \\ -0.56 \\ 0.45 \\ 1.45 \\ 2.46 \\ 3.46 \\ 4.46 \end{bmatrix}$$

Reconstruction

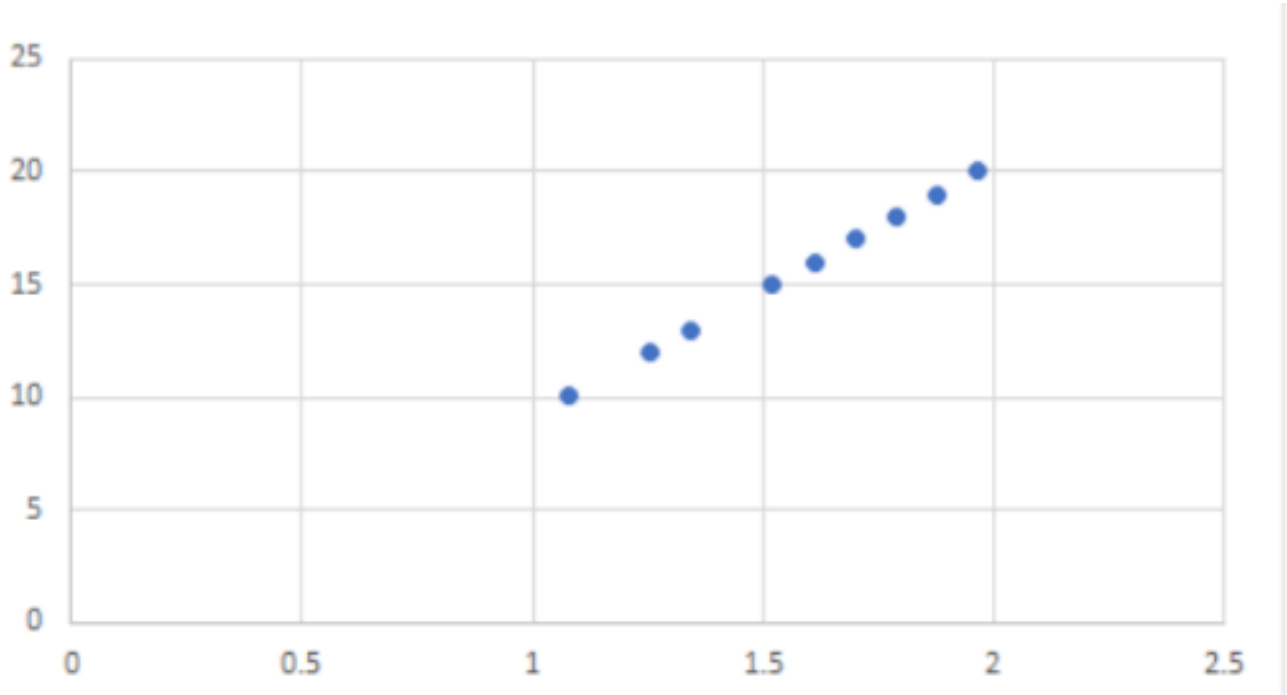
$$\hat{X} = z b_1^T$$

$$X_{\text{reconstructed}} = \hat{X} + \mathbf{1}\mu$$

Projected Data in Principal Subspace



Reconstructed Data in Original Space



Key Takeaways

- PCA finds M orthogonal directions of maximum variance — these are the top M eigenvectors of the data covariance matrix S
- Covariance: $S = \frac{1}{N} \sum_n \boldsymbol{x}_n \boldsymbol{x}_n^\top$ (computed on centred data); eigenvectors of S are the principal components
- Projected data: $\boldsymbol{y}_n = B^\top \boldsymbol{x}_n$ where $B = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M]$ are the top M eigenvectors
- Explained variance ratio: $\lambda_k \big/ \sum_j \lambda_j$ — choose M where the cumulative ratio exceeds 95%

Thank you :)