Principal Component Analysis

Machine Learning 2023-2024 - UMONS
Souhaib Ben Taieb

Exercise 1

Consider the following design matrix X:

$$X = \begin{array}{cc} X_1 & X_2 \\ \hline 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{array}$$

We want to represent the data in only one dimension using principal components analysis (PCA). To this end :

- Center the data.
- Compute the sample covariance matrix C.
- Compute the eigenvalues and eigenvectors of the covariance matrix C.
- Plot the dataset, and draw the first principal component direction (as a line) and the projections of all four sample points onto the principal direction.
- Label each data point with its principal component score.
- Compute the proportion of variance explained by the first principal component.
- Add the projections of the data points onto the second principal component, compute the second principal component scores, and show that the sum of the variance explained by each component is equal to the total variance of the data.

Recall that the eigenvalues of a square matrix A are obtained by solving the characteristic equation $\det(A - \lambda I) = 0$. If $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then $\det(A) = ad - bc$. Finally, an eigenvector v of A must satisfy $Av = \lambda v$. If X is centered, then its covariance matrix is equal to $C = \frac{1}{n} X^\intercal X$.

Solution:

The sample means are $\bar{X}_1 = 3$ and $\bar{X}_2 = 2$. Centering X yields:

$$\tilde{X} = \begin{array}{cc} X_1 & X_2 \\ \hline 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{array}$$

The sample covariance matrix is obtained as :

$$C = \frac{1}{4} \tilde{X}^T \tilde{X} = \frac{1}{4} \begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix}.$$

The eigenvalues of C are obtained by solving the characteristic equation :

$$\begin{aligned}
&\det(C-\lambda I)=0\\
\iff &\begin{vmatrix} \frac{5}{2}-\lambda & \frac{3}{2} \\ \frac{3}{2} & \frac{5}{2}-\lambda \end{vmatrix} = 0\\
\iff &(\frac{5}{2}-\lambda)^2-\frac{9}{4}=0\\
\iff &\lambda_1 = 4 \quad \lambda_2 = 1
\end{aligned}$$

The eigenvector v_1 associated to the first eigenvalue $\lambda_1 = 4$ must verify :

$$\begin{aligned}
&(C-\lambda_1 I)v_1 = 0\\
\iff &\begin{pmatrix} -3/2 & 3/2 \\ 3/2 & -3/2 \end{pmatrix}\begin{pmatrix} v_{11} \\ v_{12} \end{pmatrix} = 0\\
\iff &\begin{cases} -\frac{3}{2}v_{11}+\frac{3}{2}v_{12}=0 \\ \frac{3}{2}v_{11}-\frac{3}{2}v_{12}=0 \end{cases}
\end{aligned}$$

An arbitrary choice can be $v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, which once normalized, gives $\phi_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$

The eigenvector v_2 associated to the eigenvalue $\lambda_2 = 1$ must verify :

$$\begin{aligned}
&(C-\lambda_2 I)v_2 = 0\\
\iff &\begin{pmatrix} 3/2 & 3/2 \\ 3/2 & 3/2 \end{pmatrix}\begin{pmatrix} v_{21} \\ v_{22} \end{pmatrix} = 0\\
\iff &\begin{cases} \frac{3}{2}v_{21}+\frac{3}{2}v_{22}=0 \\ \frac{3}{2}v_{21}+\frac{3}{2}v_{22}=0 \end{cases}
\end{aligned}$$

An arbitrary choice can be $v_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, which once normalized, gives $\phi_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$

Figure 1 displays the plot of the centered dataset (left), and their projection onto the first principal component direction (right). The first component scores are given as:

$$Z_1 = \tilde{X}\phi_1 = \begin{pmatrix} 0 \\ 0 \\ \frac{4}{\sqrt{2}} \\ -\frac{4}{\sqrt{2}} \end{pmatrix}$$

The total variance in the centered data is obtained as :

$$\begin{aligned}
\text{TV} &= \sum_j^p \frac{1}{n}\sum_i^n \tilde{x}_{ij}^2\\
&= \sum_j^p \text{Var}(X_j)\\
&= 5
\end{aligned}$$

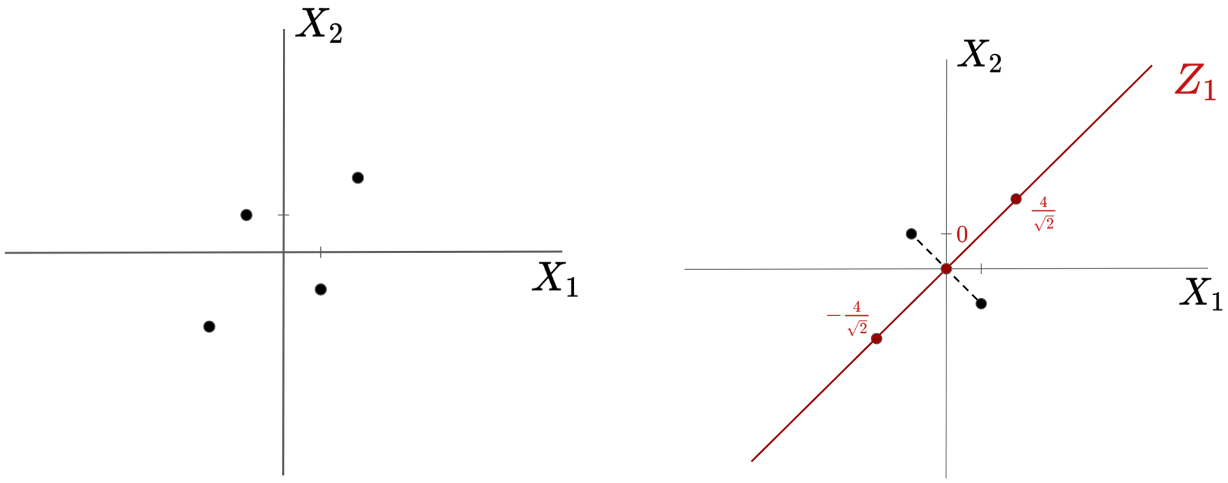

Figure 1: Plots of the centered data points (Left) and of their projections onto the direction of the first principal component.

The variance explained by the first principal component is given by :

$$\begin{aligned}\mathrm{Var}(Z_1) &= \frac{1}{n}\sum_{i}^{n} z_{1i}^2 \\ &= \frac{1}{4}(0+0+8+8) = 4\end{aligned}$$

And finally, the proportion of the variance explained by the first principal component is obtained as $\mathrm{TV}_1 = \frac{\mathrm{Var}(Z_1)}{TV} = \frac{4}{5}$.

The second principal components scores are obtained as:

$$Z_2 = \tilde{X}\phi_2 = \begin{pmatrix} \frac{2}{\sqrt{2}} \\ -\frac{2}{\sqrt{2}} \\ 0 \\ 0 \end{pmatrix}.$$

And thus, the coordinates of the data points in the principal components space are

$$Z = \begin{pmatrix} 0 & \frac{2}{\sqrt{2}} \\ 0 & -\frac{2}{\sqrt{2}} \\ \frac{4}{\sqrt{2}} & 0 \\ \frac{4}{\sqrt{2}} & 0 \end{pmatrix},$$

leading to Figure [2]. The variance explained by the second principal component is obtained as

$$\mathrm{Var}(Z_2) = \frac{1}{n}\sum_{i}^{n} z_{2i}^2 = 1.$$

From which we have

$$\mathrm{Var}(Z_1) + \mathrm{Var}(Z_2) = 5 = \mathrm{TV}.$$

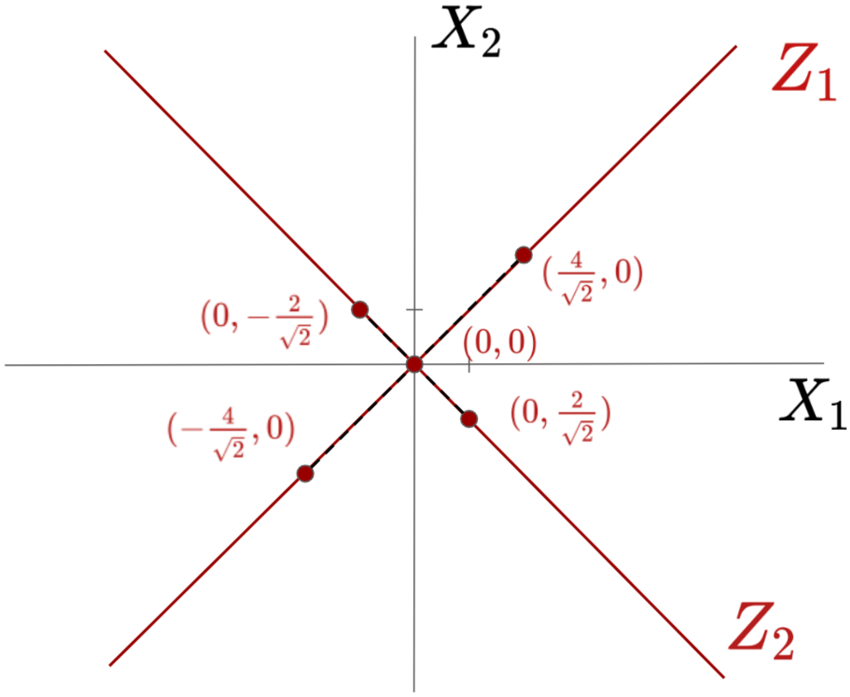

Figure 2: Projection of the centered data points onto the principal components.

Exercise 2

Suppose that the columns of $X \in \mathbb{R}^{n\times p}$ have been centered (i.e. they have sample mean zero). The total sample variance of X is defined as $\text{TV} = \sum_{j=1}^{p} \text{Var}(X_j) = \frac{1}{n}\text{Trace}(X^\intercal X)$.

Let $X = UDV^\intercal$ be the singular value decomposition of X where the columns of $U \in \mathbb{R}^{n\times p}$ and $V \in \mathbb{R}^{p\times p}$ are orthonormal and the matrix $D \in \mathbb{R}^{p\times p}$ is diagonal with positive real entries, $D = \text{diag}(d_1, ..., d_n)$, with $d_1 \geq d_2 \geq ... \geq d_n \geq 0$.

Prove that the total variance of X is given by $\frac{1}{n}\sum_j^n d_j^2$. How do you link this result to the eigenvalues of covariance matrix C ?

Hints :

- $\text{Trace}(AB) = \text{Trace}(BA)$
- $(AB)^\intercal = B^\intercal A^\intercal$
- If V is orthonormal, then $V^\intercal V = I$
- If A is diagonal, then $A^\intercal A = A^2$

Solution:

$$
\begin{aligned}
\text{TV} &= \sum_{j=1}^{p} \text{Var}(X_j) \\
&= \frac{1}{n}\text{Trace}(X^\intercal X) \\
&= \frac{1}{n}\text{Trace}\Big((UDV^\intercal)^\intercal UDV^\intercal\Big) \\
&= \frac{1}{n}\text{Trace}(VD^\intercal U^\intercal UDV^\intercal) \qquad (AB)^\intercal = B^\intercal A^\intercal \\
&= \frac{1}{n}\text{Trace}(VD^\intercal DV^\intercal) \qquad U^\intercal U = I \text{ as the columns of } U \text{ are orthonormal} \\
&= \frac{1}{n}\text{Trace}(VD^2V^\intercal) \qquad D^\intercal D = D^2 \text{ as } D \text{ is diagonal} \\
&= \frac{1}{n}\text{Trace}(D^2V^\intercal V) \qquad \text{Trace}(AB) = \text{Trace}(BA), \text{where } B = D^2V^\intercal \\
&= \frac{1}{n}\text{Trace}(D^2) \qquad V^\intercal V = I \text{ as the columns of } V \text{ are orthonormal} \\
&= \frac{1}{n}\sum_{j=1}^{p} d_j^2
\end{aligned}
$$

We have that $C = \frac{1}{n}X^\intercal X = \frac{1}{n}VDU^\intercal UDV^\intercal = \frac{1}{n}VD^2V^\intercal$, where $\frac{1}{n}D^2$ is the matrix of eigenvalues of C. This means that the total variance of X is equal to the sum of the eigenvalues of the covariance matrix C.