3.1 Principal Component Analysis

Principal component analysis (PCA) is the simplest and most widely used method for dimensionality reduction. It transforms the original variables or features (initial data dimension) into uncorrelated ones.
In the original dataset, some variables might be correlated (to be in a connection with each other), for example Height and Weight might be positively correlated (taller people tend to weigh more); correlated features provide redundant information, which PCA aims to reduce.
PCA creates new variables that are linearly uncorrelated:

- The **correlation** between **any** two principal **components is zero**
- This makes each principal component represent unique information from the data

We consider a data matrix $X \in \mathbb{R}^{N \times D}$ which contains $N \in \mathbb{N}$ feature vectors, each with $D \in \mathbb{N}$ features (N: number of data points or feature vectors (row vectors), D number of features).
Please see example below
That is, the i-th row for $i = 1, \dots, N$ contains the feature vector $x_i \in \mathbb{R}^D$. We assume that X has centered features, i.e. $\sum_{i=1}^{N} X_{ij} = 0$ for all $j = 1, \dots, D$.
This can always be achieved by replacing the data $(x_i)_{i=1}^{N}$ with $x_i - \overline{x}$, where $\overline{x} := \frac{1}{N} \sum_{i=1}^{N} x_i$

Example

Consider a data matrix $X \in \mathbb{R}^{3 \times 2}$ ($N = 3$ data points / samples (3 row vectors) and $D = 2$ features for each data point (2 columns))

$$X = \begin{bmatrix} 170 & 60 \\ 180 & 75 \\ 160 & 50 \end{bmatrix}$$

Here the first row $[170, 60]$ is the feature vector (D not really but only symbolically) for the first data point, the second row $[180, 75]$ is the feature vector for the second data point and $[160, 50]$ represents the third data point.^

We have two features:

- Feature 1: $[170, 180, 160]$ being for example Heights
- Feature 2: $[60, 75, 50]$ being for example Weights

So each data point is described by 2 features (one from each column)

Centering the features in X

Reconsider: Replacing the data $(x_i)_{i=1}^N$ with $x_i - \overline{x}$, where $\overline{x} := \frac{1}{N}\sum_{i=1}^N x_i$
Lets compute the mean for column 1:

$$\frac{170 + 180 + 160}{3} = 170$$

and the mean for column 2:

$$\frac{60 + 75 + 50}{3} = 61,66$$

So we need to do:

$$X - \overline{X} = \begin{bmatrix} 170 & 60 \\ 180 & 75 \\ 160 & 50 \end{bmatrix} - \begin{bmatrix} 170 & 61.66 \\ 170 & 61.66 \\ 170 & 61.66 \end{bmatrix} = \begin{bmatrix} 0 & -1.67 \\ 10 & 13.33 \\ -10 & -11.67 \end{bmatrix}$$

Now the features are centered because if you add each column they are 0:
$(0 + 10 + (-10)) = 0$ and $(-1.67 + 13.33 + (-11.67)) = 0$

3.1.1

The first goal is to transform the original features into uncorrelated ones. To do this, we consider the empirical covariance matrix

$$C := X^T X = \sum_{i=1}^N x_i x_i^T \in \mathbb{R}^{D \times D}$$

Since the features are assumed to be centered (i.e., $\sum_{i=1}^N x_i = 0$) the entry C_{ij} measures the empirical correlation between the i-th and j-th feature. We are now looking for a system of uncorrelated features. Note that C is a symmetric real-valued matrix ($C = C^T$ and all entries in the matrix are real numbers).
This means that the entry at row i, column j is equal to the entry at row j at column i.
$C_{ij} = C_{ji}$
An example for such a matrix would be

$$C = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}, \quad C^T = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}$$

Because of the matrix being symmetric real-valued, it is orthogonally diagonalizable, i.e., there exists an orthogonal matrix $P \in \mathbb{R}^{D \times D}$ such that $P^T X P$ is diagonal (all non-diagonal entries are 0).
Why is that matrix orthogonally diagonalizable?

1. All eigenvalues $\lambda_1, \ldots, \lambda_D$ are **real**
2. Eigenvectors corresponding to different eigenvalues are **orthogonal**

3. Using these eigenvectors, we can construct an orthogonal matrix P, whose columns are the eigenvectors of C
   An orthogonal matrix P satisfies $P^TP = I$, where I is the identity matrix. This means the columns of P are **orthonormal vectors**, i.e.:

- Each column has a length of 1 (normalization)
- Columns are perpendicular (senkrecht) to each other (orthogonality)
  Example

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad P^TP = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

The columns of $P := (v_1, \dots, v_D)$ are also eigenvectors of C, i.e., there are numbers $(\lambda_j)_{j=1}^D \subset \mathbb{R}_+$ such that $Cv_j = \lambda_j v_j$ (remember: $Cv = \lambda v$.
For example:

$$C = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}, \quad v = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \lambda = 2$$

then:

$$Cv = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} = \lambda v = 2 \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

We can assume without loss of generality that the columns of P are ordered such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_D$.

Example: Compute the covariance matrix

We have the centered X matrix already calculated above with

$$X - \overline{X} = \begin{bmatrix} 170 & 60 \\ 180 & 75 \\ 160 & 50 \end{bmatrix} - \begin{bmatrix} 170 & 61.66 \\ 170 & 61.66 \\ 170 & 61.66 \end{bmatrix} = \begin{bmatrix} 0 & -1.67 \\ 10 & 13.33 \\ -10 & -11.67 \end{bmatrix} \$\$Nowwecalculatetheempiricalcovariancem$$

3.1.2

We now want to use the eigenvectors v_j as new features. To do this, we consider the transformed data

$$Z := XP \in \mathbb{R}^{N \times D}$$

(XP gives the transformed data Z which now has the features in terms of the Eigenvectors of the covariance matrix C; P contains the eigenvectors of C)
Expanding the above matrix product, we obtain for $i = 1, \dots, N$ and $j = 1, \dots, D$

3.1.3

$$Z_{ij} = \sum_{k=1}^{D} X_{ik} P_{kj} = \sum_{k=1}^{D} (x_i)_k (v_j)_k = \langle x_i, v_j \rangle$$

(X_{ik} represents the k-th feature of the i-th data point, P_{kj} represents the k-th entry of the eigenvector v_j). The dot product means we're projecting the data point x_i onto the direction defined by the eigenvector v_j
$(x_i)_k$ refers to the k-th element of the row vector x_i, which is the i-th row of X:

$$x_1 = (X_{i1}, X_{i2}, \ldots, X_{iD})$$

Thus, the j-th new feature of the i-th new data point is given by the dot product of the old data point x_i with the j-th eigenvector of C. Another interpretation is that the new features are weighted sums of the old features, where the weights are given by the entries of the respective eigenvector.
With these insights, we can now achieve dimensionality reduction by considering only $j = 1, \ldots, d$ with $d < D$ in 3.1.3 (considering only the first d eigenvectors that correspond to the largest eigenvalues).
These eigenvectors capture the most variance in the data, and by using only the top d eigenvectors, we can reduce the number of features while preserving most of the information.
Equivalently, we can define the matrix $P^{(d)}$ as $P^{(d)} := (v_1, \ldots, v_d) \in \mathbb{R}^{D \times d}$ and instead of 3.1.2, consider the data transformation

3.1.4

$$Z^{(d)} := XP^{(d)} \in \mathbb{R}^{N \times d}$$

(this is the reduced dimensionality data)
Note that for $d = D, P^{(d)} = P$ and $Z^{(d)} = Z$. However, for $d < D, Z^d$ is a dimension-reduced representation of the data. Note that the components of $Z^{(d)}$ are exactly given by 3.1.3, i.e. for $i = 1, \ldots, N$ and $j = 1, \ldots, d, Z_{ij}^{(d)} = Z_{ij}$.

We first prove that the transformed data are indeed uncorrelated

Proposition 3.1.1

For all $j = 1, \ldots, d$ we have $\sum_{i=1}^{N} Z_{ij}^{(d)} = 0$. Moreover, the empirical covariance matrix $(Z^{(d)})^T Z^{(d)}$ is diagonal.

Proof: According to 3.1.3, we have

$$\sum_{i=1}^{N} Z_{ij}^{(d)} = \sum_{i=1}^{N} \langle x_i, v_j \rangle = \left\langle \sum_{i=1}^{N} x_i, v_j \right\rangle = 0$$

for all $j = 1, \ldots, d$. (the data is assumed to be centered!).
For the covariance matrix, we have

$$(Z^{(d)})^T Z^{(d)} = (XP^{(d)})^T XP^{(d)} = (P^{(d)})^T X^T XP^{(d)} = (P^{(d)})^T CP^{(d)}$$

Recall that $X^T X = C$, the empirical covariance matrix of the original data.
In the case where $d = D$, $P^T CP$ is diagonal, as P is chosen as the diagonalizing matrix for C.
In the general case, we note that we can write $P^{(d)} = PS$, where the matrix $S \in \mathbb{R}^{(D \times d)}$ is given by

$$S = \begin{pmatrix} \mathbb{I}_{d \times d} \\ 0_{D-d \times d} \end{pmatrix}$$

(S is a block matrix where the top block is the identity matrix and the bottom block is a zero matrix which is a $D - d \times d$ filled matrix with zeros)
Let's assume $D = 5$ and $d = 3$. Then S would look like

$$S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Thus, we obtain (substitute $P^{(d)}$ with PS)

$$(Z^{(d)})^T Z^{(d)} = S^T P^T CPS = S^T \Lambda S$$

where $\Lambda \in \mathbb{R}^{D \times D}$ is a diagonal matrix. It is now easy to see that $S^T \Lambda S \in \mathbb{R}^{d \times d}$ is also a diagonal matrix, specifically the upper-left corner of Λ

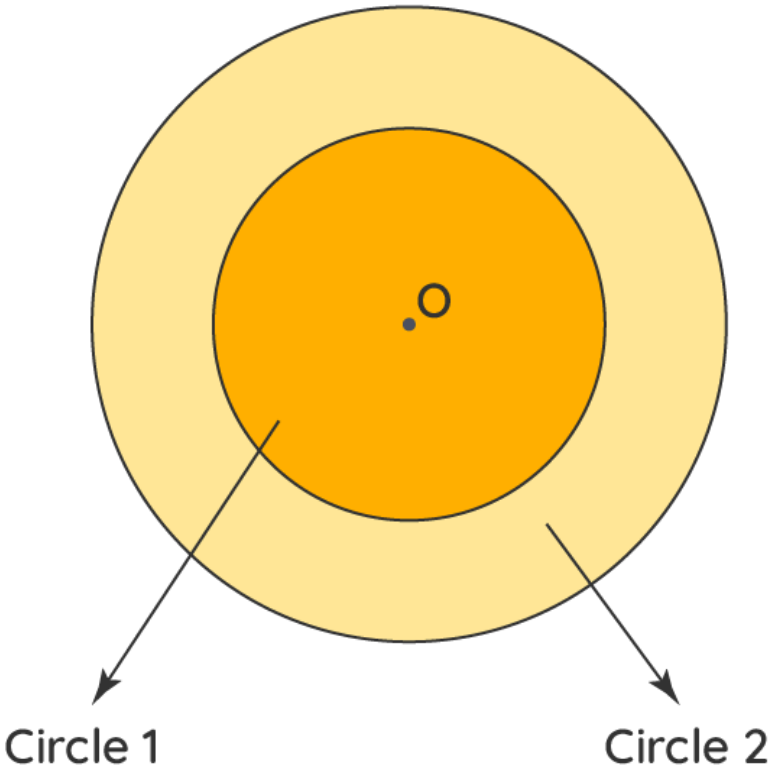
Summary

We summarize the principal component algorithm for dimensionality reduction as follows:

1. Create a data matrix $X \in \mathbb{R}^{N \times D}$ and normalize it so that all columns sum to zero
2. Compute the eigenvectors and eigenvalues of the covariance matrix $C = X^T X$
3. Retain the eigenvectors corresponding to the d largest eigenvalues and combine them into a matrix $P^{(d)} \in \mathbb{R}^{D \times d}$
4. Compute the transformed data $Z^{(d)} := XP^{(d)}$

Note that Principal Component Analysis assumes that the data can be well described by the empirical covariance matrix C. This is not always the case, for example, in the case of data points distributed on concentric circles

Concentric Circles



In this case, with appropriate scaling, C is proportional to the identity matrix, and it's eigenvectors are simply the coordinate directions. However, a more meaningful feature would be the distance of a data point from the origin. To extract this, it is useful as in Section 2.2.1, to first embed the data into a higher dimensional space and then perform PCA in this space. For instance, this can be done using the mapping $\phi : \mathbb{R}^2 \to \mathbb{R}^3$ with $\phi(x) := (x_1, x_2, x_1^2 + x_2^2)$. For the data described above, the dominant component would be the third coordinate, which corresponds precisely to the distance from the origin. Such situations can be handled much more systematically using the so-called kernel-based PCA, which goes beyond the scope of this lecture.