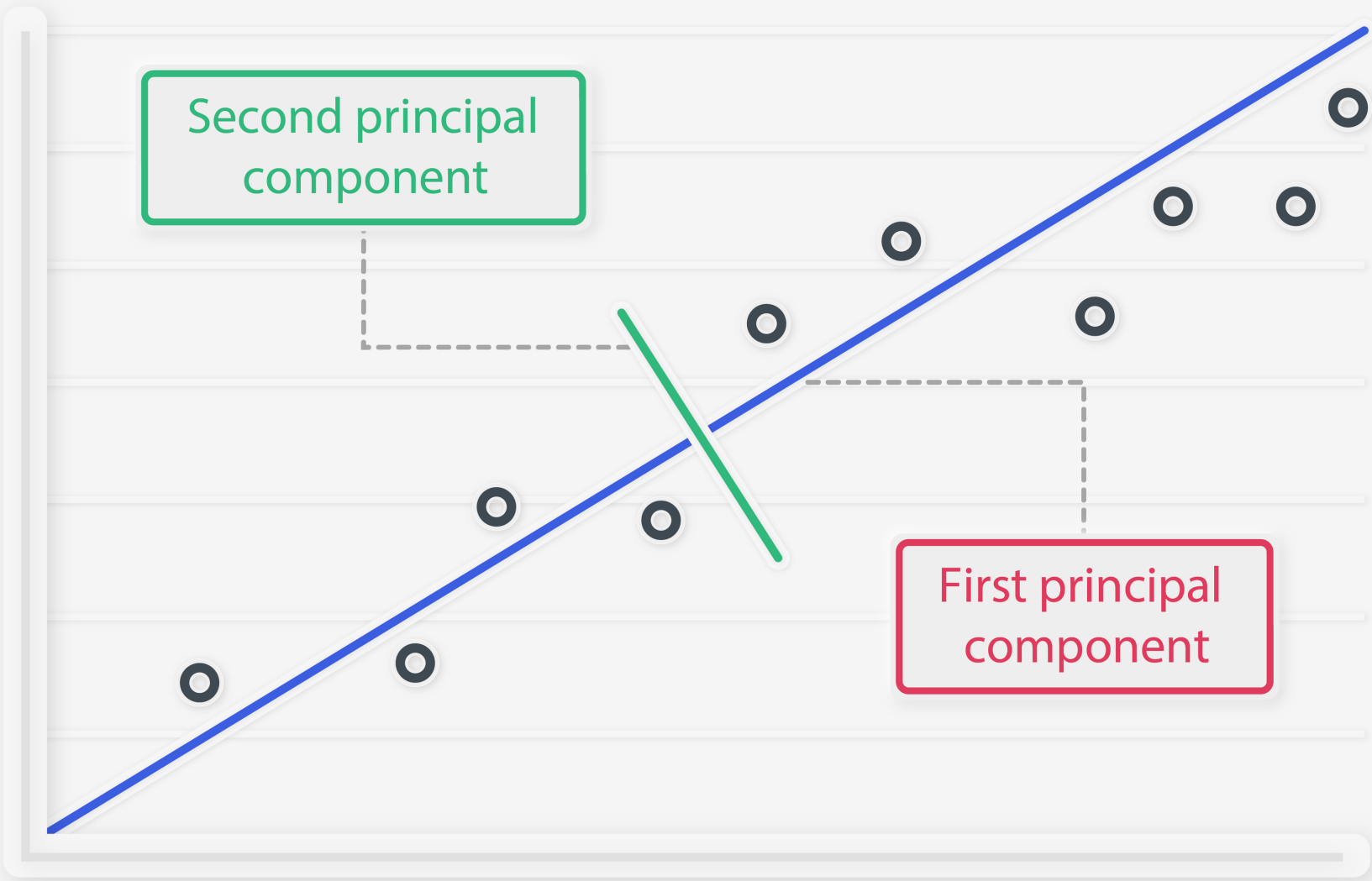
PRINCIPAL COMPONENT ANALYSIS

PCA projects the features onto the principal components.
The motivation is to reduce the features dimensionality while only losing a small amount of information.



DATA-DRIVEN
SCIENCE

Principal Component Analysis

STEP 1: STANDARDIZATION

The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.

$$z = \frac{value - mean}{standard\ deviation}$$

DATA-DRIVEN
SCIENCE

Principal Component Analysis

STEP 2: COVARIANCE MATRIX COMPUTATION

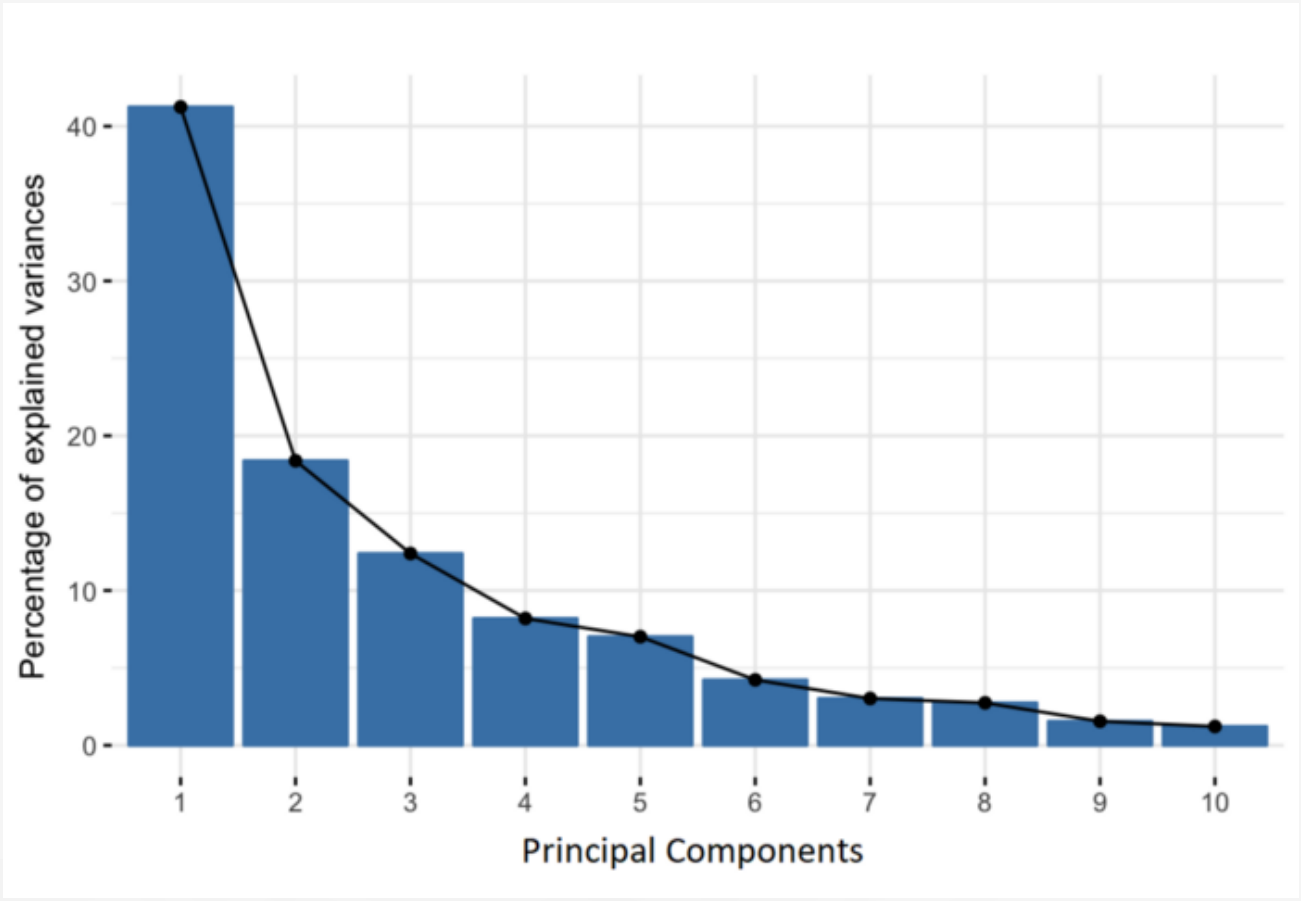
The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them.

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

DATA-DRIVEN
SCIENCE

Principal Component Analysis

STEP 3: COMPUTE THE EIGENVECTORS AND EIGENVALUES OF THE COVARIANCE MATRIX TO IDENTIFY THE PRINCIPAL COMPONENTS



DATA-DRIVEN
SCIENCE

Principal Component Analysis

STEP 4: FEATURE VECTOR

In this step, we choose whether to keep all these components or discard those of lesser significance (of low eigenvalues), and form with the remaining ones a matrix of vectors that we call Feature vector.

we can either form a feature vector with both of the eigenvectors v1 and v

$$\begin{bmatrix} 0.6778736 & -0.7351785 \\ 0.7351785 & 0.6778736 \end{bmatrix}$$

Or discard the eigenvector v2, which is the one of lesser significance, and form a feature vector with v1 only:

$$\begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix}$$

DATA-DRIVEN
SCIENCE

Principal Component Analysis

LAST STEP: RECAST THE DATA ALONG THE PRINCIPAL COMPONENTS AXES

In the previous steps, apart from standardization, you do not make any changes on the data, you just select the principal components and form the feature vector, but the input data set remains always in terms of the original axes (i.e, in terms of the initial variables).

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

DATA-DRIVEN SCIENCE

Get daily updates on
Data Science concepts

FOLLOW NOW

DATA-DRIVEN
SCIENCE