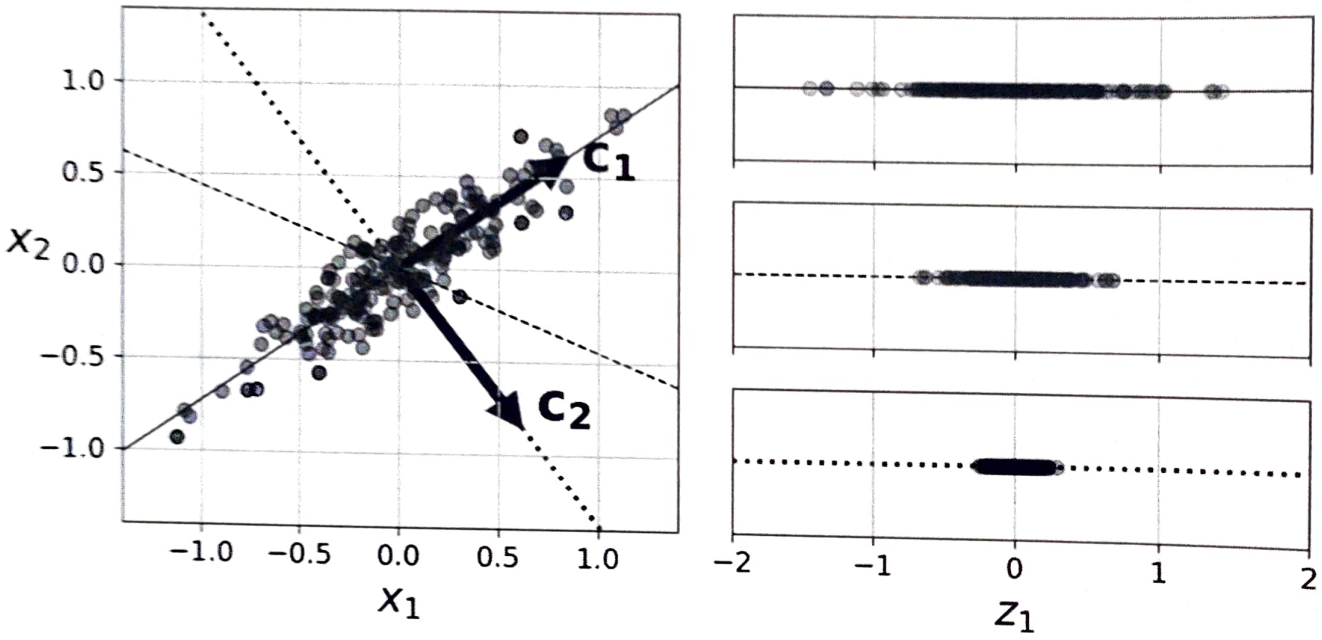
Principal Component Analysis

Principal Component Analysis (PCA) is by far the most popular dimensionality reduction algorithm. First it identifies the hyperplane that lies closest to the data, and then it projects the data onto it.

Before you can project the training set onto a lower-dimensional hyperplane, you first need to choose the right hyperplane.



As you can see, the projection onto the solid line preserves the maximum variance, while the projection onto the dotted line preserves very little variance, and the projection onto the dashed line preserves an intermediate amount of variance.

It seems reasonable to select the axis that preserves the maximum amount of variance, as it will most likely lose less information than the other projections. Another way to justify this choice is that it is the axis that minimizes the mean squared distance between the original dataset and its projection onto that axis. This is the rather simple idea behind PCA. There are as many axes as the number of dimensions in the dataset.

- Principal Component Analysis is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning.
- It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation.
- These new transformed features are called the Principal Components.
- PCA works by considering the variance of each attribute because the high variance shows the good split between the classes, and hence it reduces the dimensionality.

Some common terms used in PCA algorithm:

Dimensionality: It is the number of features or variables present in the given dataset. More easily, it is the number of columns present in the dataset.

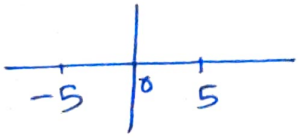
Correlation: It signifies that how strongly two variables are related to each other. Such as if one changes, the other variable also gets changed. The correlation value ranges from -1 to +1. Here, -1 occurs if variables are inversely proportional to each other, and +1 indicates that variables are directly proportional to each other.

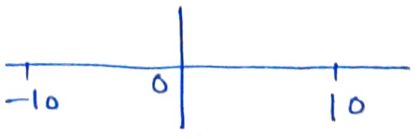
Orthogonal: It defines that variables are not correlated to each other, and hence the correlation between the pair of variables is zero.

Variance:

Variance is proportional to spread of data.

$$\text{Variance} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$





$V = 25 + 0 + 25/3$

$= 50/3$

$V = 100 + 0 + 100/3$

$= 200/3$

Variance is more for red graph.

Covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma(x,x) & \sigma(x,y) \\ \sigma(y,x) & \sigma(y,y) \end{bmatrix}$$

If x is positively correlated with y, y is also positively correlated with x. In other words, we can state that

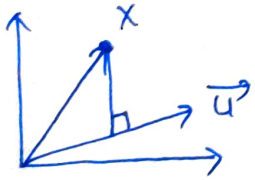
$\sigma(x,y) = \sigma(y,x)$

Therefore, the covariance matrix is always a symmetric matrix with the variances on its diagonal and the covariances off-diagonal.

Eigen decomposition of a covariance matrix:

The largest eigenvector of the covariance matrix always points into the direction of the largest variance of the data, and the magnitude of this vector equals the corresponding eigenvalue. The second largest eigenvector is always orthogonal to the largest eigenvector, and points into the direction of the second largest spread of the data.

Projection: If we are converting 2D to 1D, find single axis wherein you can project the data points



$$\text{projection} = \frac{\vec{u}\cdot\vec{x}}{\|u\|} = \vec{u}\cdot\vec{x} = u^T x$$

$\|u\| = 1$ as it is unit vector.

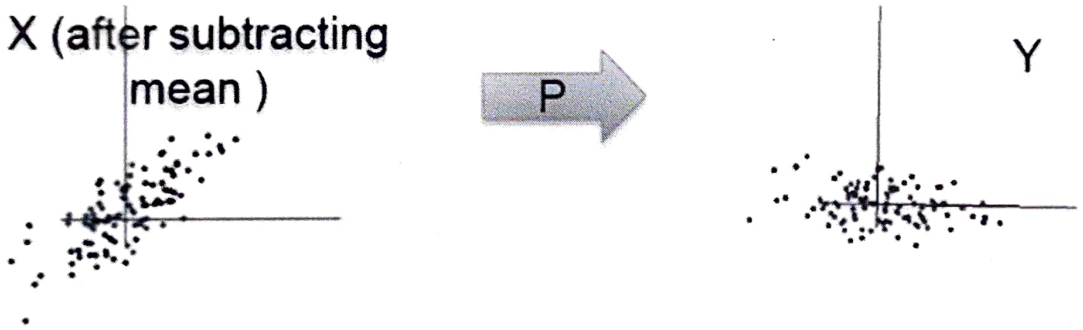
Each point in 2D gets projected on unit vector. Unit vector can be anywhere. Choose the one having maximum variance.

$$\therefore \text{variance} = \sum_{i=1}^{n} \left(\frac{u^T x_i - u^T \bar{x}}{n} \right)^2$$

Steps for PCA algorithm:

Step1: Standardizing the independent variables:

Data When we standardize the data points then what happens is that the central values become the dimensions and the data is scattered around it. on all the dimensions are subtracted from their means to shift the data points to the origin.



Step 2: Generating the covariance matrix for all dimensions.

We find the covariance between x1 and x2 and represent it in the form of a matrix. This matrix is the numerical representation of how much information is contained between the two-dimensional space of X1 and X2. In the matrix, the elements on the diagonals are the variance or spread of x1 with itself and of x2 with itself implying how much information is contained within the variable itself.

Step 3: Eigen Decomposition:

During this process, we get two outputs as below:

- Eigen Vectors: These are the new dimensions of the new mathematical space, and
- Eigenvalues: This is the information content of each one of these eigenvectors. It is the spread or the variance of the data on each of the eigenvectors.

Step 4: Sort the Eigenvectors corresponding to their respective eigenvalues.

Step5: Compute principal components:

Multiply the original standardized data by the selected principal components to obtain the new, lower-dimensional representation of the data.

Principal Component Analysis (PCA) Examples:

Image Compression: PCA reduces image dimensionality for efficient storage without losing critical information.

Genomic Data Analysis: PCA identifies patterns in gene expression data, aiding in disease research.

Financial Data Analysis: PCA analyzes covariance in asset returns for portfolio optimization.

Spectral Analysis: PCA helps in signal processing to identify dominant spectral features.

Customer Segmentation: PCA clusters customers based on behavior for targeted marketing.