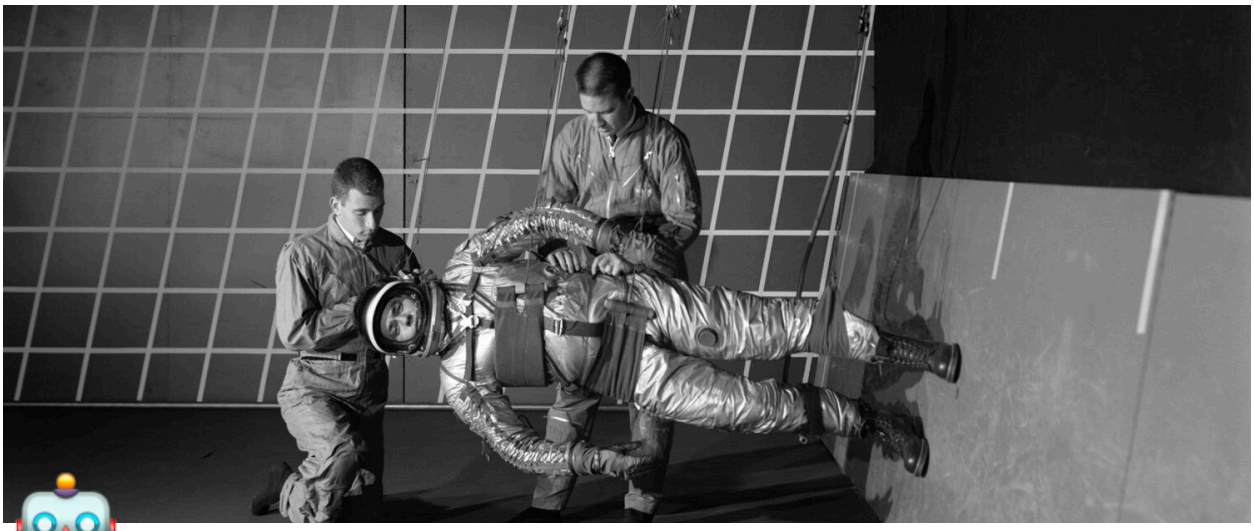


Module 2 Lab2 - Principal Components Analysis (PCA)

Created	@July 6, 2025 8:34 PM
Class	IIITH
Event	

refs:https://towardsdatascience.com/scikit-learn-vs-sklearn-6944b9dc1736/

Step By Step Computation Of PCA

The below steps need to be followed to perform dimensionality reduction using PCA:

1. Standardization of the data
2. Computing the covariance matrix
3. Calculating the eigenvectors and eigenvalues
4. Computing the Principal Components
5. Reducing the dimensions of the data set

Standardization of the data:

While applying StandardScaler, each feature of your data should be normally distributed such that it will scale the distribution to a mean of zero and a standard deviation of one.

Example:

Let's say you have a feature with the following values: `[10, 20, 30, 40, 50]`

1. **Calculate the mean (μ):** (10 + 20 + 30 + 40 + 50) / 5 = 30
2. **Calculate the standard deviation (σ):**
   - Find the variance first: `((10-30)^2 + (20-30)^2 + (30-30)^2 + (40-30)^2 + (50-30)^2) / 5 = (400 + 100 + 0 + 100 + 400) / 5 = 1000 / 5 = 200`
   - Standard deviation is the square root of the variance: `√200 ≈ 14.14`
3. **Standardize each value:**
   - (10 - 30) / 14.14 ≈ -1.414
   - (20 - 30) / 14.14 ≈ -0.707
   - (30 - 30) / 14.14 = 0
   - (40 - 30) / 14.14 ≈ 0.707
   - (50 - 30) / 14.14 ≈ 1.414

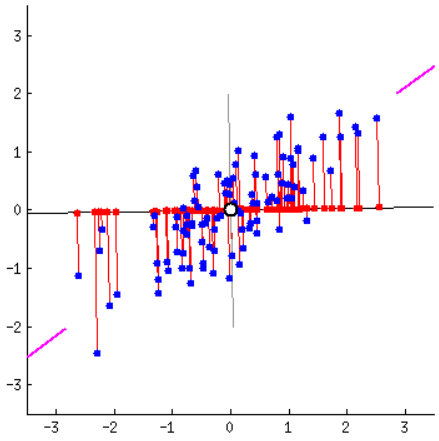
The standardized values are approximately `[-1.414, -0.707, 0, 0.707, 1.414]`. You can see that the mean of these values is 0, and the standard deviation is 1.

An important thing to realize here is that the principal components are less interpretable and don't have any real meaning since they are constructed as linear combinations of the initial variables.

Geometrically speaking, principal components represent the directions of the data that explain a maximal amount of variance, that is to say, the lines that capture most information of the data. The relationship between variance and information here, is that, the larger the variance carried by a line, the larger the dispersion of

the data points along it, and the larger the dispersion along a line, the more information it has

let's assume that the scatter plot of our data set is as shown below, can we guess the first principal component ? Yes, it's approximately the line that matches the purple marks because it goes through the origin and it's the line in which the projection of the points (red dots) is the most spread out. Or mathematically speaking, it's the line that maximizes the variance



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. Because sometimes, variables are highly correlated in such a way that they contain redundant information. So, in order to identify these correlations, we compute the covariance matrix.

The covariance matrix is a $p \times p$ symmetric matrix (where p is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables. For example, for a 3-dimensional data set with 3 variables x, y, and z, the covariance matrix is a 3×3 data matrix of this from:

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