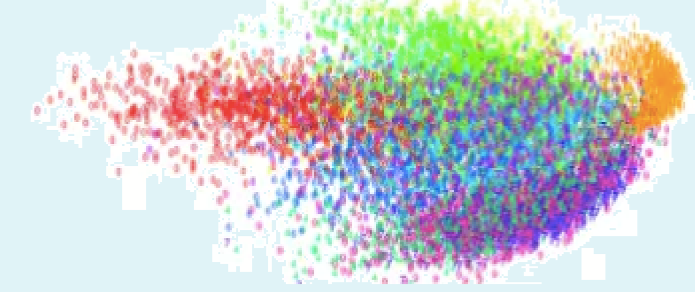
Principal Component Analysis (PCA)



What is PCA?

- Technique used to simplify a dataset
- A linear transformation that chooses a new coordinate system for the data set of linearly uncorrelated variables (principle components) such that greatest variance by any projection of the data set comes to lie on the first axis

How does it work?

Given: m x n matrix D. There are m data points and n variables (dimensions)

$$D = \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \\ a_4 & b_4 & c_4 \end{bmatrix}$$

Where

$$\vec{d}_1 = \begin{bmatrix} a_1 \\ b_1 \\ c_1 \end{bmatrix}^T, \vec{d}_2 = \begin{bmatrix} a_2 \\ b_2 \\ c_2 \end{bmatrix}^T, \vec{d}_3 = \begin{bmatrix} a_3 \\ b_3 \\ c_3 \end{bmatrix}^T, \vec{d}_4 = \begin{bmatrix} a_4 \\ b_4 \\ c_4 \end{bmatrix}^T$$

Data in Context:

Let each $\vec{d}$ represents a student. And the numbers a_i, b_i, and c_i represent age, height, and weight respectively. In this case, there are 4 students (data points) and 3 variables.

1. Centralize the data

Transform the data so that the data points are centered at origin.

$$\vec{\mu} = \frac{1}{4}(\vec{d}_1 + \vec{d}_2 + \vec{d}_3 + \vec{d}_4) = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$$

$\vec{\mu}$ is the mean vector of all the column vectors of D

By subtracting the mean, we centralize the data.

$$X = \begin{bmatrix} a_1 - \mu_1 & b_1 - \mu_2 & c_1 - \mu_3 \\ a_2 - \mu_1 & b_2 - \mu_2 & c_2 - \mu_4 \\ a_3 - \mu_1 & b_3 - \mu_2 & c_3 - \mu_3 \\ a_4 - \mu_1 & b_4 - \mu_2 & c_4 - \mu_3 \end{bmatrix}$$

Data in Context:

In the example, μ_1 will be mean age, μ_2 mean height, and μ_3 mean weight. We subtract the mean age from the age of each student and same process for height and weight.

2. Find Covariance Matrix

Recall: Covariance Matrix in n dimension is calculated

$$C = \frac{1}{n} X^T X$$

It tells the relationship between variables and measures how changes in one variable are associated with changes in a second variable.

$$C = \begin{bmatrix} Cov(X_1,X_1) & Cov(X_1,X_2) & Cov(X_1,X_3) \\ Cov(X_2,X_1) & Cov(X_2,X_2) & Cov(X_2,X_3) \\ Cov(X_3,X_1) & Cov(X_3,X_2) & Cov(X_3,X_3) \end{bmatrix}$$

Data in Context:

In this example, X_1 is age, X_2 height, and X_3 weight. $Cov(X_1,X_2)$ represents the relationship between the variable age and height and vice versa. Therefore, $Cov(X_1,X_2) = Cov(X_2,X_1)$ and the covariance matrix is symmetric.

3. Find Eigenvectors and Eigenvalues

Given Covariance Matrix C, calculate λ Eigenvalues and $\vec{v}$ eigenvectors,

so that

$$C\vec{v} = \lambda\vec{v}$$

An eigenvector is a direction, and an eigenvalue is a magnitude, telling you how much variance there is in the data in that direction.

4. Reduce dimensions

Given a matrix W of l eigenvectors that represent directions of each variation of the data

$$[\vec{v}_1, \vec{v}_2, \vec{v}_3, \ldots, \vec{v}_l]^T$$

we pick k (where $k < l$) eigenvectors with the k largest eigenvalues, since the largest eigenvalues account for most of the variation. Usually, we pick 1 or 2 eigenvectors. In this example, we will pick 2.

$$W = [\vec{v}_1, \vec{v}_2]^T$$

5. Compute Principal Component

Each principal component is a linear combination of the variables from original data ($U = [X_1, X_2, X_3]^T$) with coefficients from the k eigenvectors.

$$\underset{k \times 1}{Y} = \underset{k \times n}{W} \; \underset{n \times 1}{U}$$

Now, $Y = \left[Y_1, Y_2\right]^T$ since $k = 2$ and each Y_j is a linear combination of X_1, X_2 and X_3.

For example, Y_1 might look like

$$Y_1 = 0.3X_1 + 3.98X_2 + 3.21X_3$$

By plugging values from each row of X, we get a new data point.

Data in Context:

By finding eigenvectors and eigenvalues of the covariance matrix, we no longer understand the data by age, height, and weight. We are transforming the data in a set of data with linearly uncorrelated variables.