Principal Component Analysis (PCA)

Sourav Karmakar

souravkarmakar29@gmail.com

Variance And Covariance

- Variance and Covariance are the measures of "spread" of a set of points around their center of mass (mean).
- **Variance:** measures the "spread" of a feature around its center of mass (mean). For a feature X which takes the values $x_1, x_2, x_3, \dots, x_n$ the variance is calculated as:

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \quad , where \ \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Covariance:** between two features how much each of the feature vary from the mean with respect to each other. For example consider the following table which records the values of two features X_1 & X_2

X_1	$x_1^{(1)}$	$x_1^{(2)}$	$x_1^{(3)}$	. . .	$x_1^{(n-2)}$	$x_1^{(n-1)}$	$x_1^{(n)}$
X_2	$x_2^{(1)}$	$x_2^{(2)}$	$x_2^{(3)}$	. . .	$x_2^{(n-2)}$	$x_2^{(n-1)}$	$x_2^{(n)}$

$$Cov(X_1, X_2) = \frac{1}{n}\sum_{i=1}^{n}(x_1^{(i)} - \mu_1)(x_2^{(i)} - \mu_2), where \ \mu_j = \frac{1}{n}\sum_{i=1}^{n} x_j^{(i)}, j = 1 \ or \ 2$$

- **Observation:** covariance of one feature with itself is nothing but its variance. i.e. $Cov(X, X) = Var(X)$

Covariance Matrix

- If we have a three dimensional dataset (X_1, X_2, X_3) then we can measure the covariance between X_1 and X_2, X_2 and X_3, and X_1 and X_3. Measuring the covariance between X_1 and X_1, X_2 and X_2, and X_3 and X_3 would give us the variances of X_1, X_2 and X_3 respectively.

- We can construct the following matrix of dimension 3×3 which records the values of these variances and covariances. This is known as Covariance Matrix.

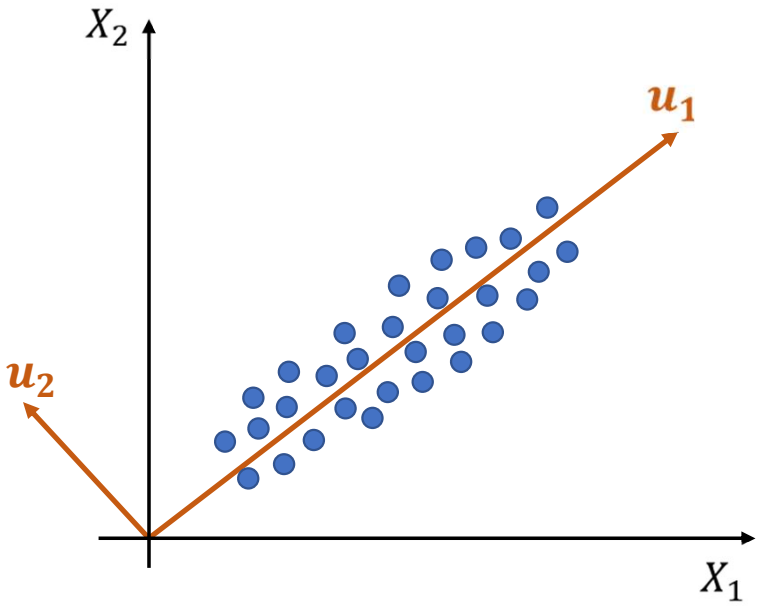
$$Covariance\ Matrix\ (\Sigma) = \begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & Cov(X_1, X_3) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & Cov(X_2, X_3) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Cov(X_3, X_3) \end{bmatrix}$$

  Similarly for D dimensional data we shall have a $D \times D$ covariance matrix.

- **Properties of Covariance Matrix:**
  - Covariance matrix is a symmetric matrix. This is because $Cov(X_i, X_j) = Cov\ (X_j, X_i)$
  - Covariance matrix is positive semi-definite i.e. all its eigen values are non-negative.

Principal Component Analysis

Intuition:



- Consider the data points as shown in the figure beside.
- Now let's look at the dataset from a different perspective.
- All the datapoints now shall be projected into the tilted orthogonal axes denoted by $\boldsymbol{u}_1$ & $\boldsymbol{u}_2$.
- What is the uniqueness about the (u_1, u_2) system of coordinates as compared to the original (x_1, x_2) system of coordinates?
- The variance of the datapoints is maximum along $\boldsymbol{u}_1$ and the axis $\boldsymbol{u}_1$ contains the maximum information about the dataset. This is called the *First Principal Component* of the dataset. Similarly $\boldsymbol{u}_2$ which contains relatively less information is known as *Second Principal Component*.

- Hence the dataset can be represented by only one feature $\boldsymbol{u}_1$ without much loss of information. This is how PCA acts as a dimension reduction technique.
- Note that the principal components are linear combinations of original basis of the dataset. Hence, it PCA is a linear feature extraction technique.

Standardization Of Data

Why Standardization or Normalization is required for PCA?

- Usually the dataset contains variables / feature which are in different units of measurement. For example weight of a person measured in kg and blood sugar level in mg/dL.

- Now if the unit of weight is changed to gram then the weight column would now contain values which is simply thousand times the weights in kg.

- Now $Var(aX) = a^2\ Var(X), where\ "a"\ is\ scalar$. Hence changing the scale of measurement will change the variance significantly, however the inter-relationship between the variables will not change.

- Thus few variables whose values are high in range will now dominate principal component and give the misleading directions of principal components.

- Hence, it is always advisable to perform normalization on the dataset before applying PCA. Normalization/ Standardization will make sure that all the data falls in the same scale.

Standardization Of Data

Steps for standardizing the data

X_1	X_2	...	X_D
$x_1^{(1)}$	$x_2^{(1)}$	...	$x_D^{(1)}$
$x_1^{(2)}$	$x_2^{(2)}$	...	$x_D^{(2)}$
:	:	...	:
$x_1^{(n)}$	$x_2^{(n)}$	...	$x_D^{(n)}$

- Consider the D-dimensional dataset as shown beside
- Let $\vec{\boldsymbol{\mu}} = [\mu_1, \mu_2, \ldots, \mu_D]$ denotes the mean of the columns. Where, $\mu_j = \mathbb{E}[X_j],\ j = 1, 2, \ldots, D$
- Let $\vec{\boldsymbol{\sigma}} = [\sigma_1, \sigma_2, \ldots, \sigma_D]$ denotes the standard deviation of the columns. Where, $\sigma_j = \sqrt{\mathrm{Var}(X_j)},\ j = 1, 2, \ldots, D$

- **Make the dataset zero centred:**

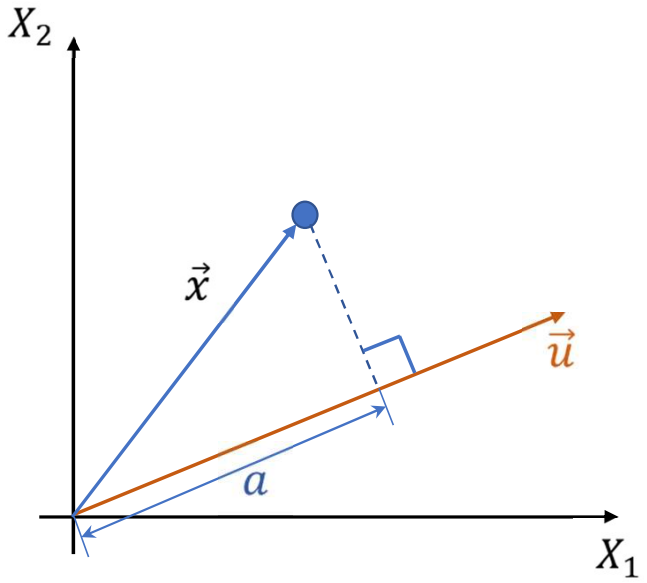
  That means $X_j := X_j - \mu_j, \forall j \in \{1,2, \ldots, D\}$. Now the mean of each column of the transformed data is zero.

- **Make each column unit variance:**

  This is done by $X_j := \frac{X_j}{\sigma_j}, \forall j \in \{1,2, \ldots, D\}$. Now the variance of each column of the transformed data is one.

After normalization the covariance matrix of the normalized data simply becomes *correlation matrix*.

Variance Probe



- Assume zero centred data. i.e. $\mathbb{E}[\vec{x}] = 0$
- $\vec{u}$ is the unit vector along the direction of principal component.

$$||\vec{u}|| = 1 \Rightarrow ||\vec{u}||^2 = 1 \Rightarrow \vec{u}^T\vec{u} = 1.$$

- Projection of $\vec{x}$ on $\vec{u}$ is a. It is obtained as following:

$$a = \vec{x}^T\vec{u} = \vec{u}^T\vec{x}$$

- Now let's see few statistical properties of a.

- $\mathbb{E}[a] = \mathbb{E}[\vec{u}^T\vec{x}] = \vec{u}^T\mathbb{E}[\vec{x}] = 0$ (as $\vec{u}$ is fixed in space wrt. $\vec{x}$)
- $Var\,(a) = \mathbb{E}[a^2] - (\mathbb{E}[a])^2 = \mathbb{E}[a^2]$ (as $\mathbb{E}[a] = 0$)

$\Rightarrow Var\,(a) = \mathbb{E}[a^2] = \mathbb{E}[a.a] = \mathbb{E}[\vec{u}^T\vec{x}.\vec{x}^T\vec{u}]$

$\Rightarrow Var\,(a) = \vec{u}^T\,\mathbb{E}[\vec{x}.\vec{x}^T]\,\vec{u} = \vec{u}^T\Sigma\,\vec{u}$ (because $\mathbb{E}[\vec{x}.\vec{x}^T] = \Sigma$, which the covariance matrix)

Thus $Var(a)$ is a function of $\vec{u}$. $Var(a) = \psi(\vec{u}) = \vec{u}^T\Sigma\,\vec{u}$. This $\psi(\vec{u})$ is also known as **variance probe**.

PCA as Optimization Problem

- Our objective is to find the unit vector $(\vec{u})$ along which the variance of the dataset is maximum.
- As the variance probe $\left(\psi(\vec{u})\right)$ estimates the variance of the dataset along the direction of unit vector $\vec{u}$. We can reformulate PCA as the following optimization problem:

$$Find\ the\ vector\ \vec{u}\ which\ will\ maximize\ \psi(\vec{u})\ subject\ to\ the\ constraint\ \left|\left|\vec{u}\right|\right| = 1$$

$$\therefore \underset{\vec{u}}{maximize}\ \ \vec{u}^T \Sigma\, \vec{u}\ ;\ \ subject\ to\ the\ constraint\ \ \vec{u}^T \vec{u} = 1$$

- This is a constraint optimization problem and we shall use *Lagrange's Multiplier Method* to solve this.
- Let λ is the *Lagrange's Multiplier* (also known as auxiliary variable). Then our modified objective function:

$$L(\vec{u}, \lambda) = \vec{u}^T \Sigma\, \vec{u} - \lambda\, \vec{u}^T \vec{u}$$

- We have to maximize this modified objective function (also known as *Lagrangian*) wrt. $\vec{u}$

$$\frac{\partial}{\partial \vec{u}} L(\vec{u}, \lambda) = 0$$

PCA as Optimization Problem

$$\frac{\partial}{\partial \vec{u}} L(\vec{u}, \lambda) = 0 \Rightarrow \frac{\partial}{\partial \vec{u}} (\vec{u}^T \Sigma \vec{u} - \lambda \vec{u}^T \vec{u}) = 0$$

$$\therefore \frac{\partial}{\partial \vec{u}} (\vec{u}^T \Sigma \vec{u} - \lambda \vec{u}^T \vec{u}) = (\Sigma + \Sigma^T)\vec{u} - 2\lambda \vec{u} = 0$$

These equations are obtained from standard formula of Matrix Calculus

Now as Σ is a symmetric matrix. Hence, $\Sigma^T = \Sigma$.

Hence, $2\Sigma\vec{u} - 2\lambda\vec{u} = 0 \Rightarrow \boxed{\Sigma\vec{u} = \lambda\vec{u}}$

- Thus, Principal Component Analysis simply reduced to an *Eigenvalue problem*. Here we have to find the eigenvalues and corresponding eigenvectors of covariance matrix of the normalized dataset.
- We can find D many eigenvalues and corresponding eigenvectors as the dimension of the dataset is D.
- We rank the eigenvalues in descending order (Note: all the eigenvalues are non-negative)
- Let the eigenvalues in descending order are: $\lambda_1 > \lambda_2 > \lambda_3 > \ ... > \lambda_D$ and the corresponding eigenvectors are $\vec{u}_1, \vec{u}_2, \vec{u}_3, ..., \vec{u}_D$. Note that $\vec{u}_j$ is unit vector.

Principal Components

- As per the ordering $\vec{u}_1$ is the unit vector along the direction of the first principal component of the dataset. Similarly $\vec{u}_2$ points towards second principal component of the dataset and so goes on.
- λ_j is the variance along j^{th} principal component. Because, $\Sigma\vec{u}_j = \lambda_j\vec{u}_j \Rightarrow \vec{u}_j^T\Sigma\vec{u}_j = \lambda_j\vec{u}_j^T\vec{u}_j = \lambda_j \ (\vec{u}_j^T\vec{u}_j = 1)$
- One interesting property of Eigenvectors / Principal Components:

  Let, λ_i and λ_j are two eigenvalues $(\lambda_i \neq \lambda_j)$ and corresponding eigenvectors are $\vec{u}_i$ and $\vec{u}_j$ respectively.

  $\therefore \Sigma\vec{u}_i = \lambda_i\vec{u}_i \ and \ \Sigma\vec{u}_j = \lambda_j\vec{u}_j$

  $\therefore (\Sigma\vec{u}_i)^T = (\lambda_i\vec{u}_i)^T \Rightarrow \vec{u}_i^T\Sigma^T = \lambda_i\vec{u}_i^T \Rightarrow \vec{u}_i^T\Sigma = \lambda_i\vec{u}_i^T$, as Σ is a symmetric matrix.

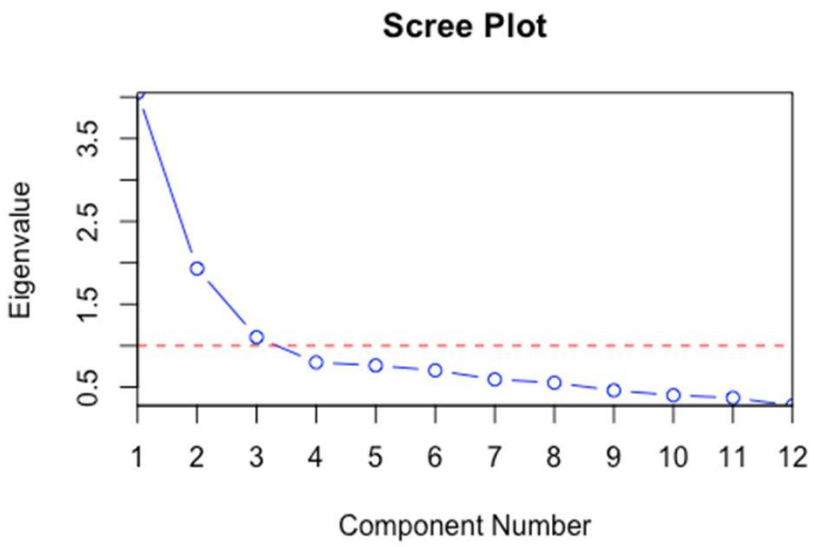
  $\therefore \vec{u}_i^T\Sigma\vec{u}_j = \lambda_i\vec{u}_i^T\vec{u}_j \Rightarrow \vec{u}_i^T\Sigma\vec{u}_j = \lambda_i\vec{u}_i^T\vec{u}_j$

  $\therefore \vec{u}_i^T\lambda_j\vec{u}_j = \lambda_i\vec{u}_i^T\vec{u}_j \Rightarrow \lambda_j\vec{u}_i^T\vec{u}_j = \lambda_i\vec{u}_i^T\vec{u}_j$, as $\Sigma\vec{u}_j = \lambda_j\vec{u}_j$

  $\therefore (\lambda_j - \lambda_i)\vec{u}_i^T\vec{u}_j = 0 \Rightarrow \vec{u}_i^T\vec{u}_j = 0$, as $\lambda_i \neq \lambda_j \Rightarrow$ Principal components are orthogonal to each other.

Scree Plot

- Scree plot is a line plot of eigenvalues of correlation matrix in descending order.



- Usually the along the first principal component the dataset shows maximum variance.
- Scree plot is a monotonically decreasing plot and helps us to determine how many principal components are enough to faithfully represent the data without much loss of information.
- Usually scree plot shows an elbow point. We consider those many components to represent our dataset, the rest of the components are discarded.

- Hence, using Scree plot one can identify the best rank-k representation of the dataset. $(k < D)$ where D is the dimension of the dataset.

Steps of Principal Component Analysis

Steps at a glance:

- Standardize the dataset.
- Calculate the covariance matrix of the standardized dataset or correlation matrix of the dataset.
- Calculate the eigenvalues and eigenvectors of the covariance matrix (Σ).
- Order the eigenvectors by decreasing order of eigenvalues.
- Using scree plot determine how many principal components to retain (say k).
- Number of principal components to retain would be the dimension of the transformed dataset.
- Form the matrix U with the selected k eigenvectors by stacking them by columns. i.e. $U = [\vec{u}_1, \vec{u}_2, \dots, \vec{u}_k]$
- The transformed dataset X' is obtained by pre-multiplying matrix U with the normalized dataset X. $X' = XU$. Then we do further analysis on the transformed dataset X'.

Limitations of Principal Component Analysis

Limitations:

- PCA works only for quantitative variables. For categorical variable it doesn't work well.
- PCA transforms the data from higher dimensional space to lower dimensional space using linear projections. If the data points lie on the complex manifold in higher dimensional space then PCA fails to work. Then we have to apply non-linear projection techniques.
- If the model is trained on principal component and not on the direct features, the explainability / interpretability of the model takes a toll.

Despite its limitations PCA works like charm in many cases. Hence, we shall use it often for feature extraction.

Thank You