Principal Component Analysis (PCA)

Principal Component Analysis is one of the main algorithms used for **dimensionality reduction** and can be thought off as a **statistical interpretation** of Singular Value Decomposition (SVD) .

On it's core the **PCA** is a maximization problem where we wanna reduce dimensions of our dataset matrix while keeping as much **variance** from the original dimension, before going on the details a quick review on **covariance** and Vector Projections :

Covariance

Simply it's a measure how correlated the variables are , which also allow us to construct **Covariance matrix**.

The typical **variance** formula in statistics is :

$$\text{Var}(X) = \frac{\sum(X_i - \bar{X})^2}{N}$$

- With $\bar{X}$ is the mean of X vector

The **covariance** is the joint **variance**:

$$\text{Cov}(X,Y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{N}$$

- $\bar{X}, \bar{Y}$ is the mean of X and Y values

For a matrix closed form of covariance , imagine both of X and Y being a rows in a matrix called A, then the covariance of them is simply the **mean centered** data times it's **transpose** :

$$A_c = A - \bar{A}$$

$$\text{Cov}(A) = A_c A_c^\intercal$$

$$\text{Cov}(A) = \begin{bmatrix} \sigma_X & \text{Cov}(X,Y) \\ \text{Cov}(X,Y) & \sigma_Y \end{bmatrix}$$

- This will results in a **covariance matrix**
- The diagonal elements are just the variance of X and Y
- The off-diagonal elements are the **covariance** between the variables

Intuition

Restating the goal of the **PCA** will help us identify the underline logic and its relation to linear algebra :

The goal of PCA is to identify the most meaningful basis to re-express a data set .

and that new basis should be a **Linear Combination** of the original basis, to put that into a algebraic linear equation will be :

$$\mathbf{PX} = \mathbf{Y}$$

- Matrix $\mathbf{X}$ is the original data set with each column is a single sample
- Matrix $\mathbf{Y}$ is a matrix related to $\mathbf{X}$ by a linear transformation $\mathbf{P}$, a new representation

The linear equation have many interpretations :

- $\mathbf{P}$ is a matrix that transforms $X \to Y$
- The rows of $\mathbf{P}$ are set of new basis vectors to express $\mathbf{X}$

- Geometrically $\mathbf{P}$ is a rotation and a stretch to transforms X into Y
  (Reminder that matrices can be thought of as a linear transformations)

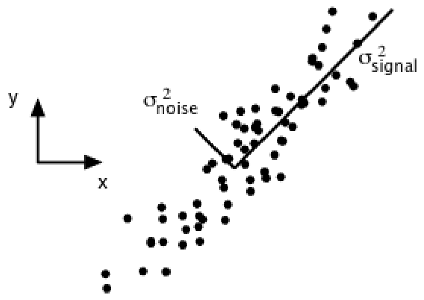
By assuming linearity to problem at hand becomes finding the appropriate *change of basis*, the rows vectors of $\{\mathbf{p_1, \dots p_m}\}$ will become the **principal components of $\mathbf{X}$**.
Now the question becomes **What is the best way to re-express $\mathbf{X}$** ?

Noise & Variance

The noise in any data set must be low or else no matter the analysis technique no information about the data can be extracted, the noise is quantified relative to the signal(**variance**), ratio of variances σ^2

$$\text{signal to noise ratio} = SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}$$

- a higher $SNR \gg_1$ indicates a high precision while lower very noisy data



Diagonalize the Covariance Matrix

The **covariance** measure the degree of linear relationship between two variables

- $\sigma^2_{AB} = 0$ if both A and B are uncorrelated
- $\sigma^2_{AB} = \sigma^2_A$ if $A = B$

As it was shown above the matrix form of covariance is :

$$C_X = \frac{1}{n}\mathbf{X}\mathbf{X}^\intercal$$

- C_X Square Symmetric matrix

if we the option to manipulate C_X it will result in C_Y **covariance matrix** of the transformed data X

To summarize the main two goals for C_Y :

- Minimize Redundancy (noise)
- Maximize Variance(signal)

To achieve that on C_Y it would be optimized to :

- All off-diagonal terms in C_Y should be zero $\implies$ C_Y must be diagonal and no correlation exists between variables
- Each dimension in Y should be $rank$ ordered by **variance**

The **PCA** selects the simplest method, by assuming that all basis vectors of matrix $\mathbf{P}$ are $orthonormal$ which means they are **perpendicular** to each other and **unit length**

$$p_i . p_j = 0$$

The **PCA** steps are :

1. Select normalized direction along which **variance** of $\mathbf{X}$ is maximized $\rightarrow$ that would be the $first\ principal\ component$
2. Find another direction along which **variance** is maximized , while following the orthonormality rule to search to all directions orthogonal to the previous selected direction
3. Repeat this procedure until m vectors are selected

PCA Assumptions

This is the assumptions behind **PCA** and might also explain when it performs poorly :

1. **Linearity** : Since the new basis is a **linear combination** of the old basis , but there is methods to expand the **PCA** to non-linear territory Kernel PCA
2. **Mean and Variance are sufficient statistics** : The mean and the variance entirely describe class of probability distribution, in order for this assumption to hold the probability distribution of $\mathbf{X}$ should be **exponential** distributed
3. **Large Variance have important dynamics** : This assumption simply assumes that the data has a high **SNR** , that's why principal components with large variance represent the *interesting part*
4. **The principal components are orthogonal** : This assumption links and makes **PCA** solvable with linear algebra decomposition Singular Value Decomposition (SVD)

Solving PCA

The principal components for the data matrix $\mathbf{X}$ are the eigenvectors of it's covariance matrix $\mathbf{C_X}$, and to obtain those there is two methods **Eigenvectors Decomposition** and Singular Value Decomposition (SVD).

Eigenvectors Decomposition

- $\mathbf{X} \in \mathbb{R}^{m \times n}$
- $n \rightarrow$ number of features
- $m \rightarrow$ number of samples

We can derive a solution for the **PCA** using linear algebra and the concept of eigenvectors, the goal is to find some $\mathrm{orthonormal\ matrix}\ \mathbf{P}$ where :

$$\mathbf{Y} = \mathbf{PX}$$

So that the **covariance matrix** of $\mathbf{C_Y} = \frac{\mathbf{1}}{\mathbf{n-1}}\mathbf{YY^T}$ is diagonalized, and the rows of $\mathbf{P}$ are the principal components of $\mathbf{X}$

It's start by rewriting the covariance matrix $\mathbf{C_Y}$ in terms of P :

$$\begin{aligned}\mathbf{C_Y} &= \frac{1}{n-1}\mathbf{YY^T} \\ &= \frac{1}{n-1}(\mathbf{PX})(\mathbf{PX})^\mathbf{T} \\ &= \frac{1}{n-1}\mathbf{P}(\mathbf{XX^T})\mathbf{P^T} \\ &= \frac{1}{n-1}\mathbf{PAP^T}\end{aligned}$$

With :

- $\mathbf{A} \equiv \mathbf{XX^T}$ a symmetric matrix, which is **diagonalized** by an orthogonal matrix of it's eigenvectors(proven in side notes), results in $\mathbf{A}$:

$$\mathbf{A} = \mathbf{EDE^T}$$

- $\mathbf{E}$ is a matrix made of the eigenvectors of $\mathbf{A}$
- $\mathbf{D}$ is a diagonal matrix which contains the **eigenvalues** of said eigenvectors

Now the trick which shows that the selection of the matrix $\mathbf{P}$ can diagonalize $\mathbf{Y}$ covariance matrix $\mathbf{C_Y}$, remember the goal of us wanting to diagonalize $\mathbf{C_Y}$ is find the directions of **maximum variance in the data** which are uncorrelated to capture the maximum amount of information with minimum noise or redundancy.

Since $\mathbf{E}$ is the eigenvectors of $\mathbf{XX^T}$ as columns , we select $\mathbf{P}$ to be a matrix with each row p_i being an eigenvector of $\mathbf{XX^T}$ which allow us to make the substitution:

$$\mathbf{A = P^T D P}$$

With $\mathbf{P} \equiv \mathbf{E}^\intercal$

$$\begin{aligned}\mathbf{C_Y} &= \frac{1}{n-1}\mathbf{PAP^T}\\ &= \frac{1}{n-1}\mathbf{P(P^TDP)P^T}\\ &= \frac{1}{n-1}\mathbf{(PP^T)D(PP^T)}\\ &= \frac{1}{n-1}\mathbf{(PP^{-1})D(PP^{-1})}\\ &= \mathbf{C_Y} = \frac{1}{n-1}\mathbf{D}\end{aligned}$$

Note : Since $\mathbf{XX^T}$ is a symmetric matrix it's eigenvectors are orthogonal that mean the matrix $\mathbf{E}$ is an orthogonal matrix so does $\mathbf{P}$, which explains why $\mathbf{P^T = P^{-1}}$

- This shows that the selection of the matrix $\mathbf{P}$ diagonalize the covariance matrix $\mathbf{C_Y}$, $\mathbf{D}$ is a diagonal matrix which contain the eigenvalues of XX^T

To summary this the principal components of $\mathbf{X}$ are the eigenvectors $XX^T \rightarrow \mathbf{E} \equiv \mathbf{P^T}$

Singular Value Decomposition (SVD)

Let $\mathbf{X}$ be out data matrix of $n \times m$ dimensions, $\mathbf{X}^\intercal\mathbf{X}$ with rank r (independent columns):

- $\{\mathbf{\hat{v}_1}, \ldots, \mathbf{\hat{v}_r}\}$ is the set of **orthonormal** eigenvectors with $\{\lambda_1, \ldots, \lambda_r\}$ for symmetric matrix $\mathbf{X}^\intercal\mathbf{X}$

$$(\mathbf{X}^\intercal\mathbf{X})\mathbf{\hat{v}_i} = \lambda_i\mathbf{\hat{v}_i}$$

- $\{\mathbf{\hat{u}_1}, \ldots, \mathbf{\hat{u}_r}\}$ is the set of **orthonormal** eigenvectors for symmetric matrix $\mathbf{XX}^\intercal$

$$(\mathbf{XX}^\intercal)(\mathbf{X\hat{v}_i}) = \mathbf{X(X^\intercal X)\hat{v}_i}$$

- With $(\mathbf{X}^\intercal\mathbf{X})\mathbf{\hat{v}_i} = \lambda_i\mathbf{\hat{v}_i}$

$$\begin{aligned}\mathbf{(XX^\intercal)(X\hat{v}_i)} &= \mathbf{X\lambda_i\hat{v}_i}\\ &= \lambda_i\mathbf{X\hat{v}_i}\end{aligned}$$

- Concluding that $\mathbf{X\hat{v}_i}$ are the eigenvectors for $\mathbf{XX}^\intercal$

Following that we have $\mathbf{\hat{u}_i}$ a normalized of the projected data $\mathbf{X}$ on the principal component vectors $\mathbf{\hat{v}_i}$

$$\mathbf{\hat{u}_i} = \frac{1}{\sigma_i}\mathbf{X\hat{v}_i}$$

With : $\sigma_i = \sqrt{\lambda_i}$,and λ_i are the eigenvalues for $\mathbf{\hat{v}_i}$

$$\mathbf{X\hat{v}_i} = \sigma_i\mathbf{\hat{u}_i}$$

$\mathbf{X}$ multiplied by the eigenvectors of the covariance matrix $X^\intercal X$ equal to a scalar σ_i times a vector $\mathbf{\hat{u}_i}$

Transforming this into a vectorized version:

$$\mathbf{XV = U\Sigma}$$

$$\mathbf{V} = [\hat{\mathbf{v}}_\mathbf{1}, \ldots, \hat{\mathbf{v}}_\mathbf{m}]$$
$$\mathbf{U} = [\hat{\mathbf{u}}_\mathbf{1}, \ldots, \hat{\mathbf{u}}_\mathbf{n}]$$

- Σ is a diagonal matrix containing the values σ_i ordered by rank
  Note : we have appended an additional $(m - r)$ and $(n - r)$ orthonormal vectors to *fill up* the matrices $\mathbf{V}, \mathbf{U}$

Since $\mathbf{V}$ is orthonormal we can multiply both sides by $\mathbf{V}^{-1} = \mathbf{V}^\intercal$

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal$$

- This means any matrix can be decomposed into **orthogonal matrix** , a **diagonal matrix** and another **orthogonal matrix**.
- More details on Singular Value Decomposition (SVD)

So the **SVD** offers numerically stable and interpretable method to calculate the eigenvectors of the covariance matrix $\mathbf{C}_\mathbf{x}$

Calculating SVD **un-normalized** covariance matrix of $\mathbf{X}$:

$$\mathbf{X}^\intercal\mathbf{X} = (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal)^\intercal\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal$$

$$\mathbf{X}^\intercal\mathbf{X} = \mathbf{V}\mathbf{\Sigma}^\mathbf{2}\mathbf{V}^\intercal$$

- With $\mathbf{V}$ being the eigenvectors of $\mathbf{X}^\intercal\mathbf{X}$

By taking the covariance matrix $\mathbf{C}_\mathbf{X}$

$$\mathbf{C}_\mathbf{X} = \frac{1}{n-1}\mathbf{X}^\intercal\mathbf{X}$$

$$\mathbf{C}_\mathbf{X} = \frac{1}{n-1}\mathbf{V}\mathbf{\Sigma}^\mathbf{2}\mathbf{V}^\intercal$$

- The **principal components** are given by the columns of $\mathbf{V}$ (right singular values)

Solving **PCA** with **SVD** is simply done by :

1. Perform **SVD** on the centered data matrix $\mathbf{X}$
2. Take $\mathbf{V}$ as the principal axes (direction of maximum variance)
3. The projected data (principal component scores) is given by :

   $$\mathbf{Y} = \mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{\Sigma}$$

   With $\mathbf{X}\mathbf{V}$ being the eigenvectors of $\mathbf{X}\mathbf{X}^\intercal$ and the coordinates of all samples along the principal axis (useful when plotting PCA)

PCA Summary

In practice **PCA** is quite simple.

1. Organize the data set in a matrix form with n observations and p features
2. Subtract off the mean for each observation $X - \bar{X}$
3. Calculate the **SVD** or the eigenvectors of the covariance matrix C_X
4. Calculate the Principal Components Scores $\mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{\Sigma}$

$$\text{PCA(X)} = \mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{\Sigma}$$

PCA as an Optimization Problem

PCA is fundamentally a constrained optimization problem, Since we want the find the direction with the maximum variance (principal components) .

We projected data (Principal Components Scores) $\mathbf{z} = \mathbf{X}\mathbf{V}$, the variance of this projection :

$$\mathrm{Var(z)} = \frac{1}{n}\mathbf{z}^\intercal\mathbf{z} = \frac{1}{n}\mathbf{V}^\intercal\mathbf{X}^\intercal\mathbf{X}\mathbf{V}$$

Let's make $S = \frac{1}{n}\mathbf{X}^\intercal\mathbf{X}$ which is the covariance matrix of the data set, we want to maximize the variance :

$$\max_{\mathbf{V}} \mathbf{V}^\intercal S\mathbf{V}$$

With the constrain $V^\intercal V = 1$.

This constrain follows the **SVD** rules which solves the **PCA** problem by making all the principal components **perpendicular** since they are eigenvectors of a symmetric matrix $X^\intercal X$

Lagrangian

Combining the objective function and the constraint into a single function :

$$\mathcal{L}(\mathbf{V}, \lambda) = \mathbf{V}^\intercal S\mathbf{V} - \lambda(\mathbf{V}^\intercal\mathbf{V} - \mathbf{1})$$

Differentiate w.r.t. $\mathbf{V}$ and set it to zero :

$$\nabla_{\mathbf{V}}\mathcal{L} = 2S\mathbf{V} - 2\lambda\mathbf{V} = 0$$

Simplify :

$$S\mathbf{V} = \lambda\mathbf{V}$$

- This also resamples the eigenvector formula with $\mathbf{V}$ being the eigenvector for the matrix S which is the covariance matrix for $\mathbf{X}$
  substitute in $\mathcal{L}(\mathbf{V}, \lambda)$

  $$\mathcal{L}(\mathbf{V}, \lambda) = \lambda$$

  $$\max_{\lambda} \mathcal{L}(\mathbf{V}, \lambda) = \max_{\lambda} \lambda$$

  which means to maximize the variance we need to pick the biggest λ value which is the amount of variance captured

Extending PCA