Why and Where we should use the PCA ?

- When we want to reduce the traning time we need to use the PCA by reducing the dimensions.
- If we want to decrese the model complexity that we shold have to use the PCA.
- It is unsupervised machine learning and it is good to use to compress the imformation.
- It is reduceses the space requirement to our data where we will reduces the memory cost for storing the data.

- Principal Component Analysis (PCA) is an **unsupervised, non-parametric statistical technique primarily used for dimensionality reduction in machine learning.**

- High dimensionality means that the dataset has a large number of features. The primary problem associated with high-dimensionality in the machine learning field is model overfitting, which reduces the ability to generalize beyond the examples in the training set. Richard Bellman described this phenomenon in 1961 as the Curse of Dimensionality where "Many algorithms that work fine in low dimensions become intractable when the input is high-dimensional. "

- Models also become more efficient as the reduced feature set boosts learning rates and diminishes computation costs by removing redundant features.

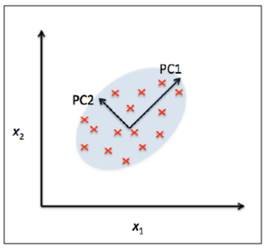
- PCA can also be used to filter noisy datasets, such as image compression. The first principal component expresses the most amount of variance. Each additional component expresses less variance and more noise, so representing the data with a smaller subset of principal components preserves the signal and discards the noise.

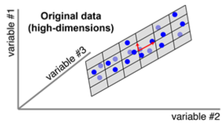
$$PC1 = w_{1,1}(\text{Feature A}) + w_{2,1}(\text{Feature B}) + w_{3,1}(\text{Feature C}) \ldots + w_{n,1}(\text{Feature N})$$
$$PC2 = w_{1,2}(\text{Feature A}) + w_{2,2}(\text{Feature B}) + w_{3,2}(\text{Feature C}) \ldots + w_{n,2}(\text{Feature N})$$
$$PC3 = w_{1,3}(\text{Feature A}) + w_{2,3}(\text{Feature B}) + w_{3,3}(\text{Feature C}) \ldots + w_{n,3}(\text{Feature N})$$

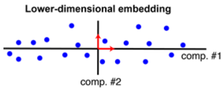
- In the graphic above, the first of the principal components (PC1) is a synthetic variable constructed as a linear combination to determine the magnitude and the direction of the maximum variance in the dataset. This component has the highest variability of all the components and therefore the most information. The second principal component (PC2) is also a synthetic linear combination which captures the remaining variance in the data set and is not correlated with PC1. The following principal components similarly capture the remaining variation without being correlated with the previous component.

- PCA is an unsupervised learning algorithm as the directions of these components is calculated purely from the explanatory feature set without any reference to response variables.

- The number of feature combinations is equal to the number of dimensions of the dataset and in general set the maximum number of PCAs which can be constructed.



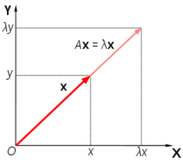




- Each blue point corresponds to an observation, and each principal component reduces the three dimensions to two. The algorithm finds a pair of orthogonal vectors (red arrows) that define a lower-dimensional space (grey plane) to capture as much variance as possible from the original dataset.

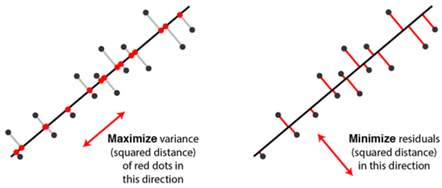
Measurement.

- Eigenvectors and eigenvalues are measures used to quantify the direction and the magnitude of the variation captured by each axis. Eigenvector describes the angle or direction of the axis through the data space, and the eigenvalue quantifies the magnitude of the variance of the data on the axis.



- A is an x n matrix, λ is the eigenvalue, and the X is the eigenvector.
- The number of feature combinations is equal to the number of dimensions of the dataset. For example, a dataset with ten features will have ten eigenvalues/eigenvector combinations.

- The correlation between each principal component should be zero as subsequent components capture the remaining variance. Correlation between any pair of eigenvalue/eigenvector is zero so that the axes are orthogonal, i.e., perpendicular to each other in the data space.
- The line which maximizes the variance of the data once it is projected into the data space is equivalent to finding the path which minimizes the least-squares distance of the projection.



Assumptions.

- PCA is based on the Pearson correlation coefficient framework and inherits similar assumptions.
- **Sample size:** Minimum of 150 observations and ideally a 5:1 ratio of observation to features (Pallant, 2010)
- **Correlations:** The feature set is correlated, so the reduced feature set effectively represents the original data space.
- **Linearity:** All variables exhibit a constant multivariate normal relationship, and principal components are a linear combination of the original features.
- **Outliers:** No significant outliers in the data as these can have a disproportionate influence on the results.
- **Large variance implies more structure:** high variance axes are treated as principal components, while low variance axes are treated as noise and discarded.

Steps Taken During The PCA Workflow.

- **Normalize the data.**
  - PCA is used to identify the components with the maximum variance, and the contribution of each variable to a component is based on its magnitude of variance. It is best practice to normalize the data before conducting a PCA as unscaled data with different measurement units can distort the relative comparison of variance across features.
- **Create a covariance matrix for Eigen decomposition.**
  - A useful way to get all the possible relationship between all the different dimensions is to calculate the covariance among them all and put them in a covariance matrix which represents these relationships in the data. Understanding the cumulative percentage of variance captured by each principal component is an integral part of reducing the feature set.
- **Select the optimal number of principal components.**
  - The optimal number of principal components is determined by looking at the cumulative explained variance ratio as a function of the number of components. The choice of PCs is entirely dependent on the tradeoff between dimensionality reduction and information loss. The graphical representation of the cumulative variance below shows that nearly 75% of the variance can be attributed to just 100/784 features and 95% to 300/784 indicating high feature redundancy.

PCA Limitations.

- **Model performance:** PCA can lead to a reduction in model performance on datasets with no or low feature correlation or does not meet the assumptions of linearity.
- **Classification accuracy:** Variance based PCA framework does not consider the differentiating characteristics of the classes. Also, the information that distinguishes one class from another might be in the low variance components and may be discarded.
- **Outliers:** PCA is also affected by outliers, and normalization of the data needs to be an essential component of any workflow.
- **Interpretability:** Each principal component is a combination of original features and does not allow for the individual feature importance to be recognized.