What is PCA

Principal component analysis (PCA) is one of the most commonly used dimensionality reduction techniques in the industry. Converting large data sets into smaller ones containing fewer variables helps improve model performance, visualise complex data sets, and in many more areas.

Common Interview Questions

1. What is the curse of dimensionality?
2. Define Principal component analysis (PCA)?
3. Can PCA be used in feature selection, if yes then how?
4. How to select the first principal component axis?
5. What does a Principal Component Analysis's major component represent?
6. What are the disadvantages of dimension reduction?
7. Why do we standardize before using Principal Component Analysis?
8. What happens when the eigenvalues are nearly equal?
9. What happens if the PCA components are not rotated?
10. Can we implement PCA for Regression?
11. Can PCA be used on Large Datasets?
12. How is PCA used to detect anomalies?

Applications of PCA:

- Dimensionality reduction
- Data Visualisation and EDA
- For creating uncorrelated features that can be input to a prediction model
- Finding latent themes in the data
- Noise Reduction

Why PCA?

- **Feature Selection:** Iteratively removing features takes time and leads to information loss.
- **Data Visualisation:** It is not possible to visualise more than two variables at the same time using any 2-D plot. Therefore, finding relationships between the observations in a data set with several variables through visualisation is tricky.

PCA is fundamentally a dimensionality reduction technique, and it helps manipulate a data set to one with fewer variables. In simple terms, dimensionality reduction is the exercise of dropping the unnecessary variables, i.e., the ones that add no useful information. Now, this is something that you must have done in the previous modules.

In EDA, you dropped columns with many nulls or duplicate values, and so on. In linear and logistic regression, you dropped columns based on their p-values and VIF scores in the feature elimination step. Similarly, what PCA does is that it converts the data by creating new features from old ones, where it becomes easier to decide which features to consider and which not to

What is PCA

- **Definition:** PCA is a statistical procedure to convert observations of possibly correlated variables to ‘principal components’ such that:
- They are uncorrelated with each other
- They are linear combinations of the original variables.
- They help in capturing maximum information in the data set. Now, the aforementioned definition introduces some new terms, such as ‘linear combinations’ and ‘capturing maximum information’, for which you will need some knowledge of linear algebra concepts and other building blocks of PCA.
- Basis: The first fundamental building block of PCA is Basis. Essentially, ‘basis’ is a unit in which we express the vectors of a matrix. For example, we describe the weight of an object in terms of the kilogram, gram and so on; to describe length, we use a meter, centimetre.

1. **Eigenvalues:** In PCA, eigenvalues represent the variance of the data along the principal components. Each eigenvalue corresponds to a principal component and indicates the amount of variance explained by that component. Higher eigenvalues signify that the corresponding principal component carries more information from the original data.
2. **Eigenvectors:** Eigenvectors are the directions or axes along which the data varies the most. They are associated with the eigenvalues and determine the principal components. Each eigenvector points in a direction that maximizes the variance of the data when projected onto that direction. In PCA, the eigenvectors are orthogonal to each other.
3. **Principal Components:** Principal components are the transformed variables that result from PCA. They are linear combinations of the original variables, where each component is a weighted sum of the original variables. The first principal component corresponds to the eigenvector with the highest eigenvalue and explains the most variance in the data. Subsequent principal components explain the remaining variance in decreasing order.
4. **Scree Plot:** A scree plot is a graphical tool used in Principal Component Analysis (PCA) to visualize the eigenvalues of the principal components. It helps in determining the number of principal components to retain for further analysis.

In the scree plot, the x-axis represents the number of principal components, and the y-axis represents the corresponding eigenvalues. Each eigenvalue is represented by a point or a dot on the plot. The scree plot is typically sorted in descending order of eigenvalues. The first principal component (PC1) has the highest eigenvalue, followed by PC2 with the second-highest eigenvalue, and so on. The scree plot helps in visualizing the "elbow" or "knee" point, which indicates the point at which the eigenvalues start to level off. The eigenvalues before this point contribute significantly to the variance explained by the principal components, while the eigenvalues after this point contribute less. The elbow point is considered as a cutoff for selecting the number of principal components to retain.

By examining the scree plot, you can make an informed decision about how many principal components to retain based on the eigenvalues. Retaining a sufficient number of principal components ensures that you capture a significant amount of variance in the data while reducing dimensionality.

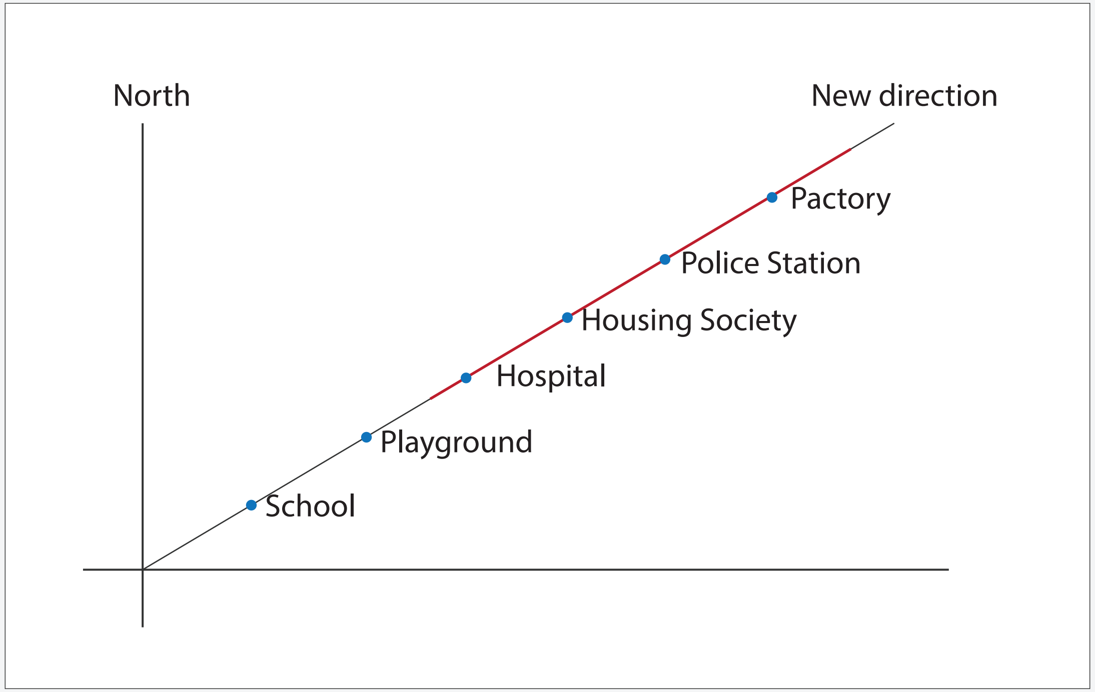
What is PCA

Introduction:

Using the analogy of basis as a unit of representation, different basis vectors can be used to represent the same observations, just like you can represent the weight of a patient in kilograms or pounds.

Demonstration:

Suppose you make a list of places for your friend to visit on a Roadmap on 2-D cartesian, with 2 directions North and east where x-direction is East and y-direction is North. Every single point on the map is represented by 2 dimensions, for example, Factory can be (10, 8) as 10 units East and 8 units North. For successive movements, like from Hospital to Housing Society, it will be 3 unit North, 4 unit east.



Now, your friend sees all the points are on a single line and can be represented in this new direction without north or east like Hospital to Housing Society is 5 units in the new direction. So, with this new direction, we have reduced the dimensions from 2 to 1 without losing information. This is PCA, better representation was in a new direction, so by dimension reduction, we changed the basis.

What is PCA

Calculation: When you have one dimension, the calculations for the change of basis are pretty straightforward. All you need to do here is to multiply the factor M which gives you the method of transforming from one basis to another.

NEW BASIS REPRESENTION (FT) = M x OLD BASIS REPRESENTATION (FT)

M^{-1}x NEW BASIS REPRESENTION (FT) = OLD BASIS REPRESENTATION (FT)

CHANGE OF BASIS 2-DIMENSION

BASIS: $\left\{ \begin{bmatrix} \text{1 ft} \\ \text{0 lbs} \end{bmatrix}, \begin{bmatrix} \text{0 cm} \\ \text{1 lbs} \end{bmatrix} \right\}$

BASIS: $\left\{ \begin{bmatrix} \text{1 cm} \\ \text{0 kg} \end{bmatrix}, \begin{bmatrix} \text{0 cm} \\ \text{1 kg} \end{bmatrix} \right\}$

SAME AS: $\left\{ \begin{bmatrix} \text{30.48 cm} \\ \text{0 kg} \end{bmatrix}, \begin{bmatrix} \text{0 cm} \\ \text{0.453kg} \end{bmatrix} \right\}$

$$\begin{bmatrix} 30.48 & 0 \\ 0 & 0.45 \end{bmatrix} \text{x} \begin{bmatrix} 5.4 \\ 121.3 \end{bmatrix} = \begin{bmatrix} 165 \\ 55 \end{bmatrix}$$

5.4 ft 5.4 x 30.48 cm 165 cm
121.3 lbs 121.3 x 0.45-55kg

New basis Represention Mx Old basis Representation
Mis a representation of old basis in new basis

CHANGE OF BASIS

$$\begin{bmatrix} 30.48 & 0.0 \\ 0.0 & 0.45 \end{bmatrix}\begin{bmatrix} 5.4 \\ 121.3 \end{bmatrix}$$

Height (ft)	Weight (lbs)
54	121.3
5.1	156.5
54	194.0
5.2	231.5
5.2	207.2

Height (ft)	Weight (lbs)
165	55
155	71
165	88
160	105
160	94

$$\begin{bmatrix} 0.0328 & 0.0 \\ 0.0 & 2.22 \end{bmatrix}\begin{bmatrix} 165 \\ 55 \end{bmatrix}$$

$$\boxed{M = B_2^{-1} * B_1}$$

And finally the transformation is given as

$$\boxed{v_2 = M * v_1}$$

PCA

The ideal basis vectors required has the following properties:

- They explain the directions of maximum variance
- When used as the new set of basis vectors, the transformed dataset is now suitable for dimensionality reduction.
- These directions explaining the maximum variance are called the Principal Components of our data.

Implementation of PCA:

1. After basic data cleaning procedures, standardize your data

2. Once standardization has been done, you can go ahead and perform PCA on the dataset. For doing this you import the necessary libraries from sklearn.decomposition.

```
from sklearn.decomposition import PCA
```

3. Instantiate the PCA function and set the random state to some specific number so that you get the same result every time you execute that code.

```
pca = PCA(random_state=42)
```

4. Perform PCA on the dataset by using the pca.fit function. pca.fit(x) The above function does both the steps: finding the covariance matrix and doing an eigen decomposition of it to obtain the eigenvectors, which are nothing but the Principal Components of the original dataset.

5. The Principal Components can be accessed using the following code.

```
pca.components_
```

6. Variance is being explained by each Principal Component using the following code.

```
pca.explained_variance_ratio_
```

PCA : Practical Considerations

Important points to remember while using PCA:

- Most software packages use SVD to compute the principal components and assume that the data is scaled and centered, so it is important to do standardization/normalization.
- PCA is a linear transformation method and works well in tandem with linear models such as linear regression, logistic regression etc., though it can be used for computational efficiency with non-linear models as well.
- It should not be used forcefully to reduce dimensionality (when the features are not correlated).

Important shortcomings of PCA :

- PCA is limited to linearity, though we can use non-linear techniques such as t-SNE as well (you can read more about t-SNE in the optional reading material below).
- PCA needs the components to be perpendicular, though in some cases, that may not be the best solution. The alternative technique is to use Independent Components Analysis.
- PCA assumes that columns with low variance are not useful, which might not be true in prediction setups (especially classification problems with a high-class imbalance).

List of useful functions that use after importing the PCA function from sklearn libraries.

- **pca.fit()** - Perform PCA on the dataset.
- **pca.components_** - Explains the principal components in the data
- **pca.explained_variance_ratio_** - Explains the variance explained by each component
- **pca.fit(n_components = k)** - Perform PCA and choose only k components
- **pca.fit_transform** - Transform the data from original basis to PC basis.
- **pca(var)** - Here 'var' is a number between 0-1. Perform PCA on the dataset and choose the number of components automatically such that the variance explained is (100*var) %.