**Principal Component Analysis (PCA) – Unsupervised - Decomposition**

* **PCA is that it is an Unsupervised dimensionality reduction technique**
* **Principal components represent the directions of the data that explain a maximal amount of variance.**
* **It tries to preserve the data that have more variation(variance)**
* **The relationship between variance and information here, is that, the larger the variance carried by a line/features, the larger the dispersion of the data points along it, and the larger the dispersion along a line, the more the information it has**
* **In simple terms principal components as new axes that provide the best angle to see and evaluate the data, so that the differences between the observations are better visible**
* **Principal components do not correlate with each other.**
* **Principal components have both direction and magnitude.** 
  + **The direction represents across which *principal axes* the data is mostly spread out or has most variance.**
  + **The magnitude signifies the amount of variance that Principal Component captures of the data when projected onto that axis**
* **The principal components are a straight line, and the first principal component holds the most variance in the data.**
* **The PCA class contains explained\_variance\_ratio\_ which returns the variance captured by each of the principal components.**
* **The explained\_variance variable is now a float type array which contains variance ratios for each principal component. The values for the explained\_variance variable looks like this:**

1. **0.722265**
2. **0.239748**
3. **0.0333812**
4. **0.0046056**

**It can be seen that first principal component is responsible for 72.22% variance. Similarly, the second principal component causes 23.9% variance in the dataset. Collectively we can say that (72.22 + 23.9) 96.21% percent of the classification information contained in the feature set is captured by the first two principal components.**

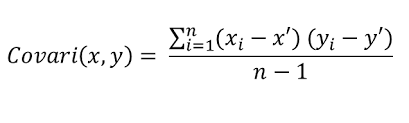
**Steps:**

**Outliers Treatment, Missing values, EDA, type conversion**

**Standardization – Normalization or Scaling**

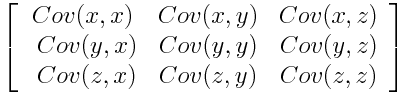
* 1. **The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis. The dataset on which PCA technique is to be used must be scaled.**
  2. **That is, if there are large differences between the ranges of initial variables, those variables with larger ranges will dominate over those with small ranges (For example, a variable that ranges between 0 and 100 will dominate over a variable that ranges between 0 and 1), which will lead to biased results.**
  3. **Categorical features are required to be converted into numerical features before PCA can be applied**

1. **Calculate Covariance Matrix:**
   1. **Covariance is a measure of how much two random variables gets changed together – direction of their linear relationship. It is actually used for computing the covariance in between every column of data matrix**

****

* 1. **The covariance matrix is a *p* × *p* symmetric matrix (where *p*is the number of dimensions) that has as entries the covariance’s associated with all possible pairs of the initial variables.**

**For example, for a 3-dimensional data set with 3 variables *x*, *y*, and *z*, the covariance matrix is a 3×3 matrix of this from:**

****

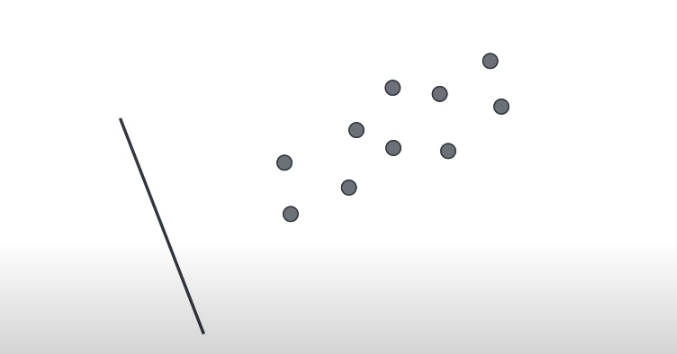
**if positive then: the two variables increase or decrease together (correlated)**

**if negative then : One increases when the other decreases (Inversely correlated)**

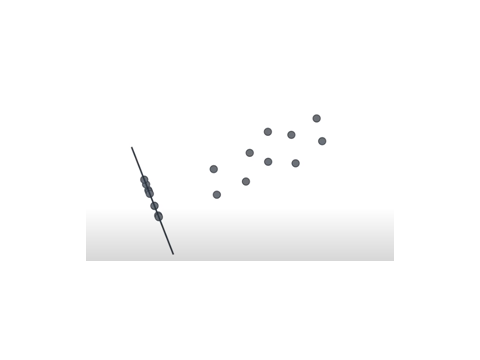
* 1. **“Covariance” indicates the direction of the linear relationship between variables.**

**“Correlation” on the other hand measures both the strength and direction of the linear relationship between two variables**

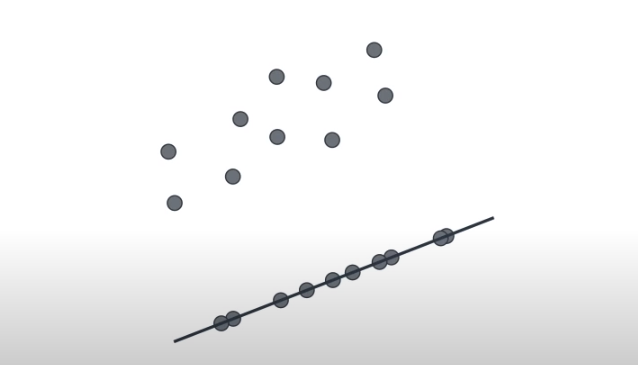
1. **Eigen values and Eigen vectors of covariance matrix are computed to identify Principal components** 
   1. **Eigenvectors of the Covariance matrix are actually *the directions of the axes where there is the most variance* (most information) and that we call Principal Components 1**
   2. **Eigenvalues are simply the coefficients attached to eigenvectors, which give the *amount of variance carried in each Principal Component*.**
   3. **By ranking your eigenvectors in order of their eigenvalues (variance/magnitude), highest to lowest, you get the principal components in order of significance.**
   4. **first principal component accounts for the largest possible variance in the data set, the line that maximizes the variance (the average of the squared distances from the projected points**
   5. **The second principal component is calculated in the same way, with the condition that it is uncorrelated with (i.e., perpendicular to) the first principal component and that it accounts for the next highest variance.**
2. **Projecting data on to Y axis – it wont be able to explain for maximum of variance of Data**

****

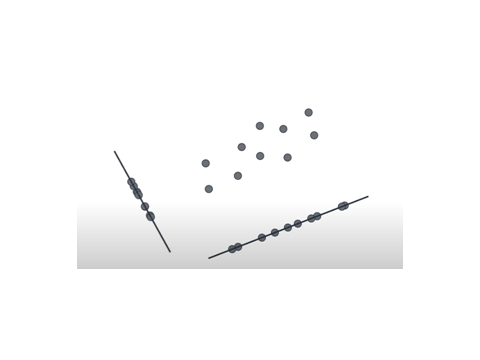
1. **Projecting Data on to Y-axis**

****

1. **Projecting Data on to X-axis – it caters for the maximum variance of the data**

****

1. **Graph having data projected – and their respective variances captured**

****

**Import Data**

**Preprocessing**

**Scaling**

**PCA – Decompose --- X-train , test decomposed , y\_orignal**

**Sklearn Algo training --- X-decomposed , y\_orignal**

**evaluate**