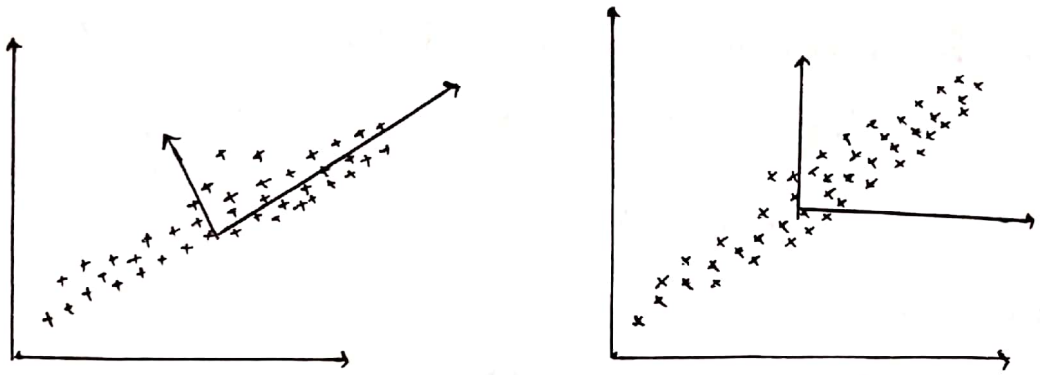
Principle Component Analysis (PCA)



Goal: PCA finds a new set of dimensions such that all the dimensions are orthogonal (and hence linearly independent) and ranked according to the variance of data along them.

Find a transformation such that:

- ⇒ The transformed features are linearly independent.
- ⇒ Dimentionality can be reduced by taking on the dimensions with the ~~height~~ highest importance.
- ⇒ those newly found dimensions should minimize the projection error.
- ⇒ the projected points should have maximum spread i.e. maximum variance

Variance: How much variance or spread the data has.

$$Var(x) = \frac{1}{n} \Sigma (x_i - \bar{x})^2$$

Covariance Matrix: Indicates the level to which two variables vary together

$$Cov(x, Y) = \frac{1}{n} \Sigma (x_i - \bar{x})(Y_i - \bar{Y})^T$$

$$Cov(x, x) = \frac{1}{n} \Sigma (x_i - \bar{x})(x_i - \bar{x})^T$$

Eigenvector, Eigenvalues: The eigenvectors point in the direction of the maximum variance, and the corresponding signvalues indicates the importance of its corresponding eigen vector.

$$A\vec{v} = \lambda\vec{v}$$

Approach:

* Substract the mean from X
* Calculate Cov(X, X)
* Calculate eigenvectors and eigenvalues of covariance matrix
* Sort the eigenvectors according to their eigenvalues in decreasing order.
* Choose first K eigenvectors and that will be the new K dimensions
* Transform the original n dimensional data points into K dimensions (= Projections with dot product)