PRINCIPAL COMPONENT ANALYSIS.

It is hard to plot data with a great no. of dimensions therefore we try to reduce the dimensions of the data without losing the relevant information (or minimizing the loss).

The question arises, how to reduce the dimension of our data. (slide no. 7)

In Principal Component analysis:

→ we project our data (~~to~~ higher dimensional data) to a new coordinate system.

→ then we chose only a few of those new coordinates / axes or dimensions that describe the maximum

* no. of principal components can be less than or equal to attributes.

* PCA solves the prob of overfitting.

Date

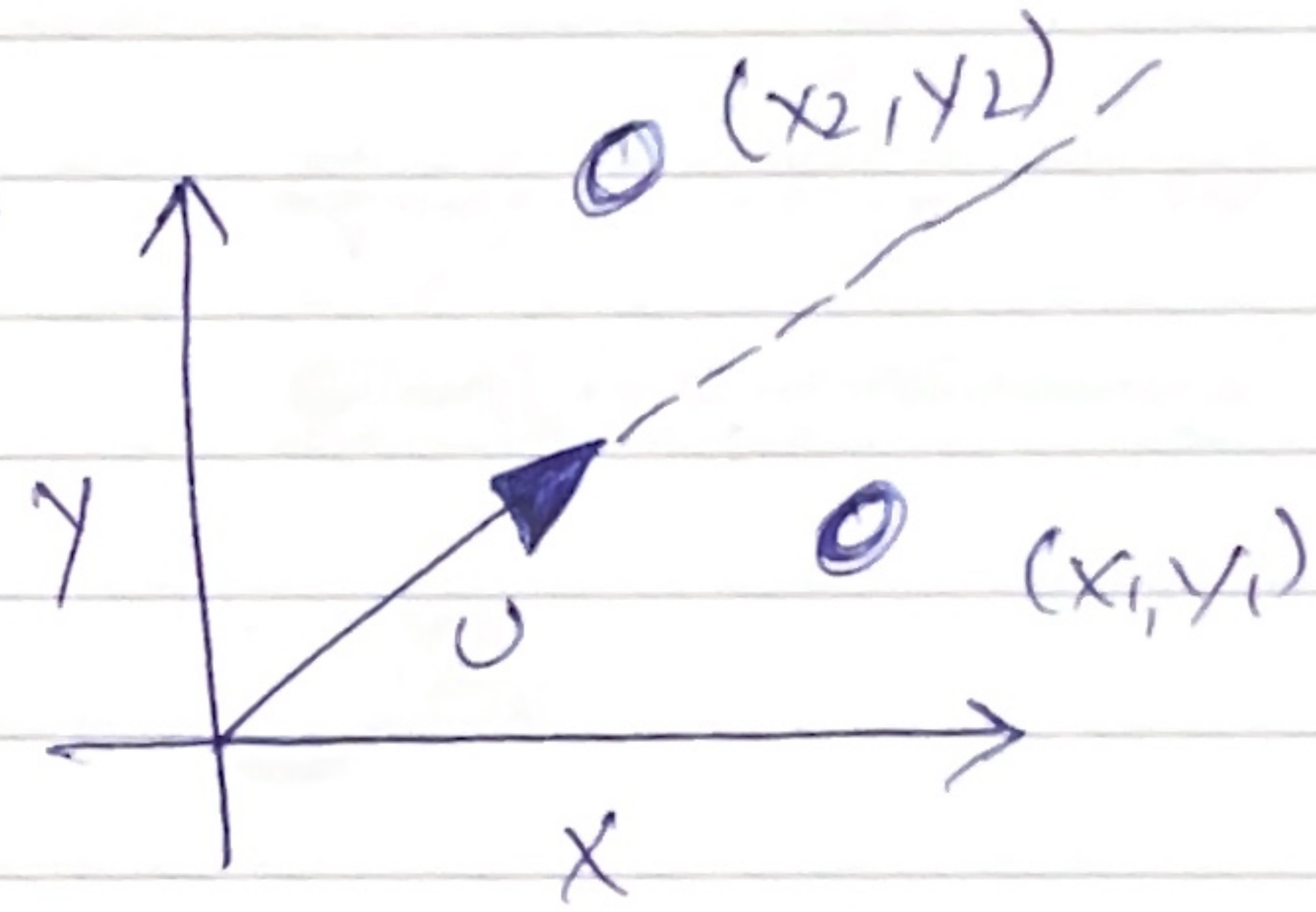
variation of our data.

→ These axes are orthogonal to eachother.
linearly independant (have little covariance b/w eachother)

Our main targets in this projection are:-

(i) Project the data (X) on a vector (u) such that:-
- variance of the projected data is max.
- u is a unit vector.

Consider 2 dimensional data.



$$X = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \end{bmatrix}$$

u is a unit vector.

$$\|u\| = 1.$$

Projecting X1 onto u.
- we can project X1 vector on u using the vector projection formula.

Datapoint 1:
$$X_1 = [x_1 \; y_1]$$

$$\text{proj}_u X = (X.u)u.$$

vector Projection Formula is used to project a data vector onto the direction of a principal component. This helps transform the data into a new coordinate system defined by the principal components.

$$\text{proj}_u(X) \Rightarrow (X \cdot u)u.$$
dotproduct.

'X' be the original vector.
'u' be the principal component vector (eigen vector of covariance matrix, normalizied to unit length).
$\text{proj}_u(X)$ be the projection of X on u.

u_i corresponds to eigen vectors of the co-variance matrix which represent the direction of maximum variance.

Date

The $(x \cdot u)$ is a scalar due to dot product and is the coordinate of x in the new basis defined by u.

* The projection formula is applied iteratively for each principal component to transform the data into principal component space.

PCA Example

Consider the following data

Sample	x_1	x_2
1	4	11
2	8	4
3	13	5
4	7	14

1. Centered the matrix X

$$\bar{x}_1 = \frac{4+8+13+7}{4} = 8$$

$$\bar{x}_2 = \frac{11+4+5+14}{4} = 8.5$$

$$\text{Centered Data} = \begin{bmatrix} x_1 - \bar{x}_1 & x_2 - \bar{x}_2 \\ 4-8 & 11-8.5 \\ 8-8 & 4-8.5 \\ 13-8 & 5-8.5 \\ 7-8 & 14-8.5 \end{bmatrix} \quad \begin{bmatrix} -4 & 2.5 \\ 0 & -4.5 \\ 5 & -3.5 \\ -1 & 5.5 \end{bmatrix}$$

PCA Example

2. Covariance matrix

$$\Sigma = \frac{1}{n-1} X^T X$$

$$\Sigma = \frac{1}{4-1} \begin{bmatrix} -4 & 0 & 5 & -1 \\ 2.5 & -4.5 & -3.5 & 5.5 \end{bmatrix} \begin{bmatrix} -4 & 2.5 \\ 0 & -4.5 \\ 5 & -3.5 \\ -1 & 5.5 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} (-4)^2 + 0^2 + 5^2 + (-1)^2 & (-4)(2.5) + 0(-4.5) + 5(-3.5) + (-1)(5.5) \\ (2.5)(-4) + (-4.5)(0) + (-3.5)(5) + (5.5)(-1) & (2.5)^2 + (-4.5)^2 + (-3.5)^2 + (5.5)^2 \end{bmatrix}$$

$$\Sigma = \frac{1}{3} \begin{bmatrix} 42 & -33 \\ -33 & 69 \end{bmatrix} = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

PCA Example

3. Eigenvalues and Eigenvectors

Eigenvalues – t o find the eigenvalues λ, solve the characteristic equation

$$\det(\Sigma - \lambda I) = 0$$

$$\det \begin{bmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{bmatrix} = 0$$

$$(14-\lambda)(23-\lambda) - (-11)(-11) = 0$$

$$(14-\lambda)(23-\lambda) - 121 = 0$$

$$\lambda^2 - 37\lambda + 322 - 121 = 0$$

$$\lambda^2 - 37\lambda + 201 = 0$$

Solve this quadratic equation:

PCA Example

3. Eigenvalues and Eigenvectors

Eigenvalues – solve this quadratic equation:

$$\lambda = \frac{-(-37) \pm \sqrt{(-37)^2 - 4(1)(201)}}{2(1)}$$

$$\lambda = \frac{37 \pm \sqrt{1369 - 804}}{2}$$

$$\lambda = \frac{37 \pm \sqrt{565}}{2}$$

$$\lambda = \frac{37 \pm 23.74}{2}$$

$$\lambda_1 = \frac{37 + 23.74}{2} = 30.37$$

$$\lambda_2 = \frac{37 - 23.74}{2} = 6.62$$

PCA Example

3. Eigenvalues and Eigenvectors

Eigenvector – need to solve the equation:

$$(\Sigma - \lambda_1 I)v_1 = 0$$

$$\Sigma - \lambda_1 I = \begin{bmatrix} 14 - 30.38 & -11 \\ -11 & 23 - 30.38 \end{bmatrix} = \begin{bmatrix} -16.38 & -11 \\ -11 & -7.38 \end{bmatrix}$$

$$\begin{bmatrix} -16.38 & -11 \\ -11 & -7.38 \end{bmatrix} \begin{bmatrix} v_{1_1} \\ v_{1_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$-16.38 v_{1_1} - 11 v_{1_2} = 0$$

$$-11 v_{1_1} - 7.38 v_{1_2} = 0$$

PCA Example

3. Eigenvalues and Eigenvectors

Eigenvector – need to solve the equation:

$$-16.38v_{1_1} - 11v_{1_2} = 0$$

$$-11v_{1_1} - 7.38v_{1_2} = 0$$

$$v_{1_1} = \frac{-11}{-16.38}v_{1_2} = 0.6736v_{1_2}$$

$$-11(0.6736v_{1_2}) - 7.38v_{1_2} = 0$$

$$-7.4096v_{1_2} - 7.38v_{1_2} = 0$$

$$(-7.4096 - 7.38)v_{1_2} = 0$$

$$-14.7896v_{1_2} = 0$$

This means $v1_2$ can be any non-zero value, so let's take $v1_2=1$.

PCA Example

3. Eigenvalues and Eigenvectors

Eigenvector – need to solve the equation:

$$v_1 = \begin{bmatrix} -0.6736 \\ 1 \end{bmatrix}$$

Normalize the eigen vector.

$$\|v_1\| = \sqrt{(-0.6736)^2 + 1^2} \approx \sqrt{0.4537 + 1} \approx \sqrt{1.4537} \approx 1.2065$$

$$v_1 = \frac{1}{1.2065} \begin{bmatrix} -0.6736 \\ 1 \end{bmatrix} \approx \begin{bmatrix} -0.558 \\ 0.830 \end{bmatrix}$$

PCA Example

3. Eigenvalues and Eigenvectors

Eigenvector – follow the same procedure for λ_2 as we did for v1:

$$v_2 = \begin{bmatrix} 1.49 \\ 1 \end{bmatrix}$$

$$\|v_2\| = \sqrt{(1.49)^2 + 1^2} \approx \sqrt{2.2201 + 1} \approx \sqrt{3.2201} \approx 1.795$$

$$v_2 = \frac{1}{1.795} \begin{bmatrix} 1.49 \\ 1 \end{bmatrix} \approx \begin{bmatrix} 0.830 \\ 0.558 \end{bmatrix}$$

PCA Example

4. Select the Subset of Eigenvectors

Eigenvector – Sort eigenvalues in descending order:

$$\lambda_1 = \frac{37 + 23.74}{2} = 30.37$$

$$\lambda_2 = \frac{37 - 23.74}{2} = 6.62$$

- $\lambda 1$ is the largest, large eigen values captures large
- variance.
- We may keep both eigenvalues and vectors or can select one, capturing the more variance
- We may get eigenvalues variance distribution as well by dividing all eigenvalues with the sum of all eigen values

$$\frac{\lambda_1}{\sum_{i=1}^{d} \lambda_i} > \frac{\lambda_2}{\sum_{i=1}^{d} \lambda_i} > \cdots > \frac{\lambda_d}{\sum_{i=1}^{d} \lambda_i}$$

PCA Example

5. Projection using $u\lambda_1 u^T$ - Dimensionality Reduction

$$u_1\lambda_1 = \begin{bmatrix} -0.558 \\ 0.830 \end{bmatrix} \times 30.38 = \begin{bmatrix} -16.97 \\ 25.22 \end{bmatrix}$$

$$u_1^T = \begin{bmatrix} -0.558 & 0.830 \end{bmatrix}$$

$$P_1 = \begin{bmatrix} -16.97 \\ 25.22 \end{bmatrix} \times \begin{bmatrix} -0.558 & 0.830 \end{bmatrix}$$

$$P_1 = \begin{bmatrix} 9.47 & -14.09 \\ -14.07 & 20.95 \end{bmatrix}$$

- Where P_1 is the 1st principle component projection
- The projection matrix P1 is used to project the original (centered) data Xcentered onto the new subspace defined by the eigenvector u1.

- This allows you to reduce the data dimensionality while retaining the maximum variance.

PCA Example

6. Project the Centered Data onto the Subspace

$$X_{\text{projected}} = X_{\text{centered}} \cdot P$$

$$X_{\text{projected}} = \begin{bmatrix} -2.615 & -3.805 \\ -63.405 & -94.455 \\ -2.515 & -3.115 \\ 68.035 & 101.375 \end{bmatrix}$$

- Now, the projected data $X_{\text{projected}}$ is in the space defined by the first principal component.

- In this case, by projecting the data onto the first principal component (PC1), we reduce the dimensionality from 2D to 1D

PCA Example

Direct Projection Using the Eigenvector

- Directly projecting the data onto a single eigenvector (usually corresponding to the first principal component)

- Instead of constructing a full projection matrix, you compute the dot product between the centered data and the eigenvector.

$$\text{Projection} = X_{\text{centered}} \cdot u_1$$

- Use this when you need to project on a single component.

- Use earlier method when reducing to multiple dimensions and that is mathematically consistent for any number of dimensions