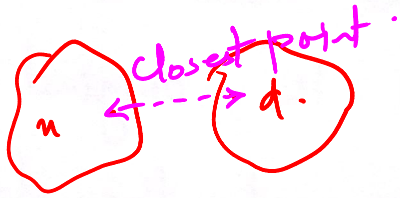
# Dimensionality Reductions (PCA)

Consider examples $x_i$ in a d-dimensional space $x_i \in \mathbb{R}^d$

A matrix $W \in \mathbb{R}^{n,d}$ can project the examples $x_i$ to a n-dimensional space.




$$\underset{n\times 1}{y_i} = \underset{n\times d}{W}\, \underset{d\times 1}{x_i}$$

The vector $y_i$ is the projection of $x_i$ onto an n-dimensional linear subspace.

We can also have a linear transformation defined by matrix $U \in \mathbb{R}^{d,n}$ to project $y_i$ back to the d-dim subspace. The mapped point in the d-dimensional space is denoted as $\hat{x}$. It is the reconstruction of the original vector $x$.

The Principal Component Analysis (PCA) problem can be defined as that of finding matrices $W_{n\times d}$ and $U_{d\times n}$ such that the total reconstruction error for all the examples is minimized.

$$U, W = \underset{U, W}{\arg\min} \sum_{i=1}^{m} \left\| x_i - \hat{x}_i \right\|^2$$

$$= \underset{U\ W}{\arg\min} \sum_{i=1}^{m} \left\| x_i - UWx_i \right\|^2$$

$\boxed{y_i = Wx_i}$

$\hat{x}_i = Uy_i$

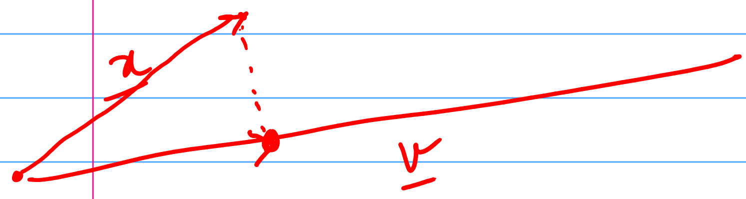
$\hat{x}_i = UWy_i$

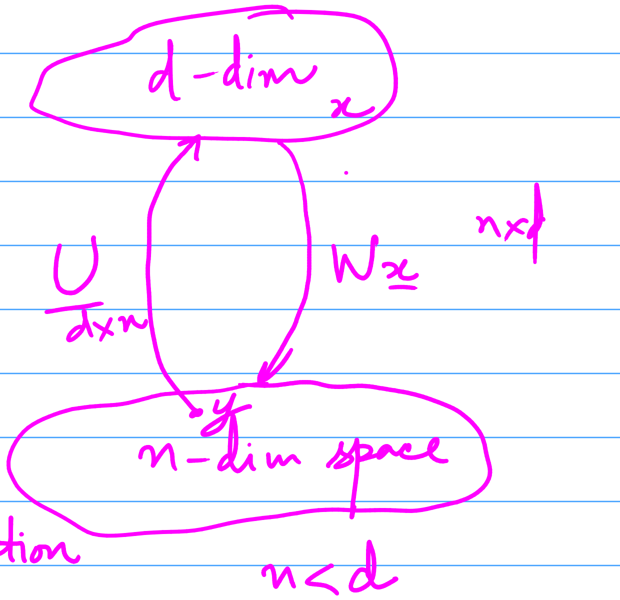
This is the first interpretation of the PCA problem.

$$y_{n\times 1} = W_{n\times d}\, x_{d\times 1}$$

$\underline{W}^{n\times d}$ has mapped a d-dim vector to an n-dim vector

linear mapping



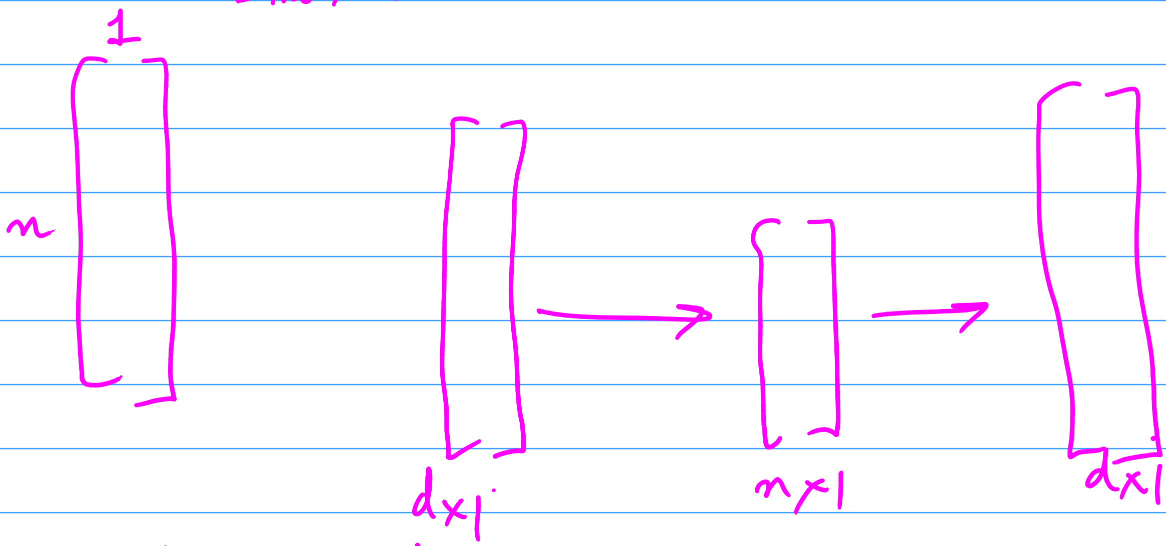




$$\underline{x} \xrightarrow{W} y \xrightarrow{U} \hat{x}$$

reconstruction

Dim reduction

$y_{n\times 1}$ column vector

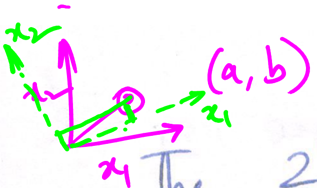



Feature Selection

Dim Reduction

$$\underset{x}{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}} \xrightarrow{FS} \begin{bmatrix} x_1 \\ x_4 \\ x_7 \\ \vdots \end{bmatrix} n<d.$$

$$\underline{\text{Dim}}\ \text{Reduction} \quad \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \underset{U}{\overset{W}{\rightleftarrows}} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$




The $2^{nd}$ interpretation of the PCA problem is to find new axis directions (basis vectors) which are rotated versions of the original axes directions such that if we represent the data ($\underline{x}_i$ $i \in [1, m]$) in terms of the new feature axes then the correlations and redundancies in the transformed feature values $\underline{y}_i$ are removed.

PCA can be used to automatically determine the $\boxed{\text{correlation removing}}$ axis system.

$\equiv$ Variance maximizing direction

To apply PCA, we first need to create a d data matrix

$X_{m\times d}$

data matrix

$$\underline{X} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_m^T \end{bmatrix} = \begin{array}{l} \text{ex } 1 \rightarrow \\ \\ m \\ \\ \text{ex } m \rightarrow \end{array} \begin{bmatrix} x_1^1 & x_1^2 & x_1^3 & \cdots & x_1^d \\ x_2^1 & x_2^2 & x_2^3 & \cdots & x_2^d \\ \vdots & & & & \\ x_m^1 & x_m^2 & x_m^3 & \cdots & x_m^d \end{bmatrix}$$

If the vectors $\underline{x}_i$ are mean centered, then we can compute the $\boxed{\text{covariance matrix}}$ as $\frac{1}{m} \sum_{i=1}^{m} \underline{x}_i \underline{x}_i^T$

m examples

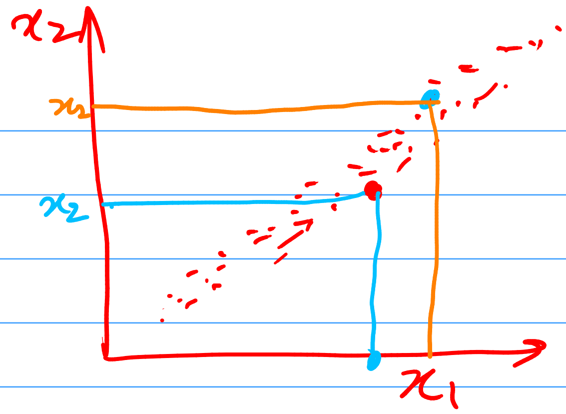
dx1 1xd
dxd
outer product matrix

If the data is not mean centered, then the covariance matrix is given as

$$\frac{1}{m} \sum_{i=1}^{m} (\underline{x}_i - \underline{\mu})(\underline{x}_i - \underline{\mu})^T$$

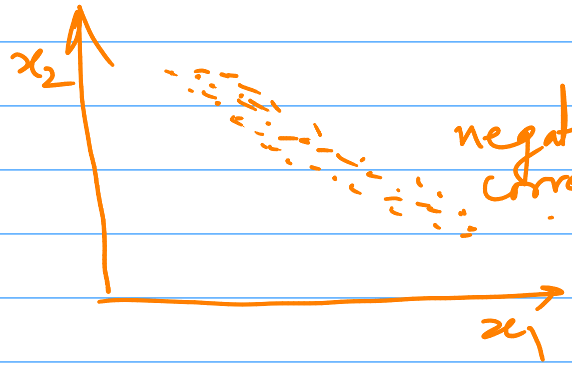
where $\underline{\mu}$ is a dx1 mean vector of the data set.

$$= \frac{1}{m} \sum_{i=1}^{m} (\underline{x}_i - \underline{\mu})(\underline{x}_i^T - \underline{\mu}^T)$$

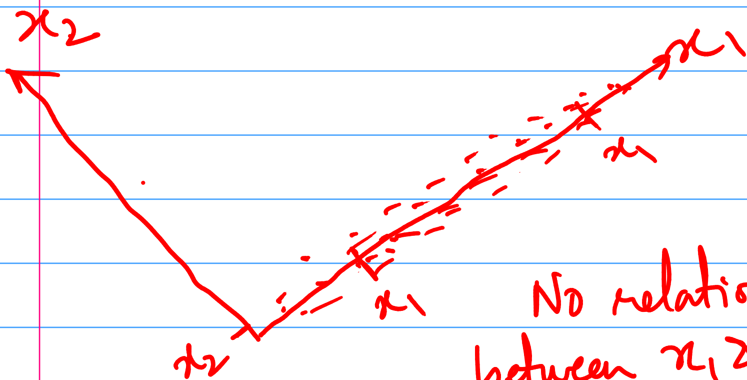



$x_1$ and $x_2$ are positively correlated.

rotation of the axis system

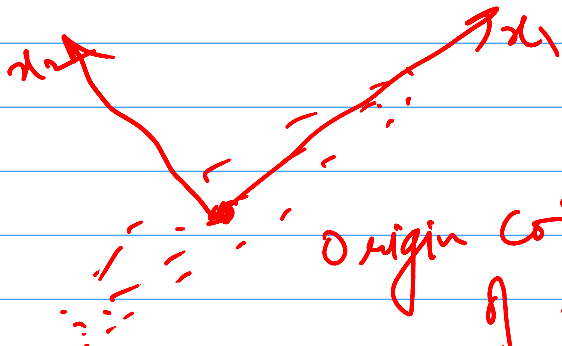



negative correlation




No relation between $x_1$ & $x_2$

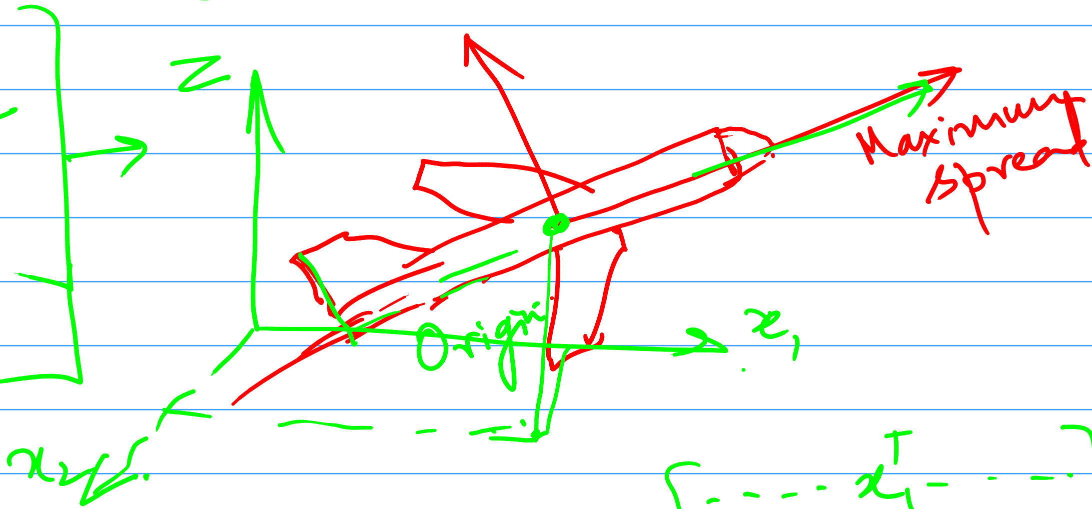
Features $x_1$ & $x_2$ have been decorrelated.




Origin coincides with the mean of the data.

4D → 3D.

$$\begin{bmatrix} \text{Name} \\ \text{Aadhar} \\ \text{DoB} \\ \text{Age} \\ \text{DoB} \end{bmatrix}$$




$$\underline{X} = \begin{bmatrix} \text{---} & \underline{x}_1^T & \text{---} \\ \text{---} & \underline{x}_2^T & \text{---} \\ & \vdots & \end{bmatrix} \quad X^T \begin{bmatrix} x_1 & x_2 & \cdots \\ \vdots & \vdots & \end{bmatrix}$$

$X^T X =$

$$= \frac{1}{m} \sum_i \left( \underline{x}_i \underline{x}_i^T - \underline{x}_i \underline{\mu}^T - \underline{\mu} \underline{x}_i^T + \underline{\mu} \underline{\mu}^T \right)$$

$$= \frac{1}{m} \left( \sum_i \underline{x}_i \underline{x}_i^T - \underbrace{\left( \sum_i \underline{x}_i \right)} \underline{\mu}^T - \underline{\mu} \underbrace{\left( \sum_i \underline{x}_i^T \right)} + \sum_{i=1}^{m} \underline{\mu} \underline{\mu}^T \right)$$

$$= \frac{1}{m} \left( \sum_i \underline{x}_i \underline{x}_i^T - \underbrace{m \underline{\mu}} \underline{\mu}^T - \cancel{\underbrace{m \underline{\mu} \underline{\mu}^T}} + \cancel{\underbrace{m \underline{\mu} \underline{\mu}^T}} \right)$$

$$= \boxed{\frac{1}{m} \sum_{i=1}^{m} \underline{x}_i \underline{x}_i^T - \underline{\mu} \underline{\mu}^T} \qquad \underline{C} \equiv \frac{X^T X}{m} - \underline{\mu} \underline{\mu}^T$$

We denote the covariance matrix as $\underline{C} \equiv \frac{1}{m} \sum_i \underline{x}_i \underline{x}_i^T - \underline{\mu} \underline{\mu}^T$

The covariance matrix is a square $d \times d$ matrix.

The diagonal entries of $\underline{C}$ give the variance values along the $d$ feature axes.

The off diagonal entries give the covariance values for the pairs of feature components.

The covariance matrix $\underline{C}$ is positive semi-definite. $\underbrace{1 \times d \;\; d \times d \;\; d \times 1}_{1 \times 1 \text{ scalar}}$

We find that for any vector $\underline{v}$, the value of $\underbrace{\underline{v}^T \underline{C} \underline{v}}_{\text{scalar}} \geq 0.$

To show this we write the covariance matrix as

$$\underline{C} = \frac{1}{m} X^T X - \underline{\mu} \underline{\mu}^T$$

where $X$ is $m \times d$ data matrix

$$\underline{X}: \begin{array}{cc} & d \\ m & \begin{bmatrix} \cdots \underline{x}_1^T \cdots \\ \cdots \underline{x}_2^T \cdots \\ \cdots \underline{x}_m^T \cdots \end{bmatrix} \end{array}$$

$$\underline{v}^T \underline{C} \underline{v} = \frac{\underline{v}^T \underline{X}^T \underline{X} \underline{v}}{m} - \underbrace{\underline{v}^T \underline{\mu}} \underbrace{\underline{\mu}^T \underline{v}}$$

$(AB)^T = B^T A^T$

Variance $= \frac{1}{m}\sum x_i^2 - \mu^2$

$v^T C v \geq 0$

$$= \left[ \frac{(\underline{X}\,\underline{v})^T (\underline{X}\,\underline{v})}{m} - (\underline{\mu}^T \underline{v})^T (\underline{\mu}^T \underline{v}) \right] \geq 0$$

Variance is always non neg

$\Rightarrow \underline{C}$ is psd.

$\underline{X}\,\underline{v}$ $\rightarrow v$

row 1, 2, 3 — col vec

1D projection

variance of the projection of the row vectors in $\underline{X}$ on the column vector $\underline{v}$. $\geq 0$

1D proj — variance $\underline{\mu}^T\underline{v}$

$\underline{\mu}^T \underline{v}$ is the projection of the mean vector on the direction $\underline{v}$. This will also be the mean of the 1-D projections of the examples $\underline{x}_i$'s on $\underline{v}$.

$\underline{\mu}^T v$

The goal of PCA is to one-by-one determine the orthonormal vectors $\underline{v}$ maximizing $\underline{v}^T \underline{C}\, \underline{v}$. variance along direction $\underline{v}$

Because the covariance matrix is symmetric and positive semi definite, we can obtain its eigen value decomposition

$$C = V \Lambda V^T$$

$V$: orthonormal eigen vectors of $C$. as columns.

$\Lambda$: diagonal matrix of eigen values

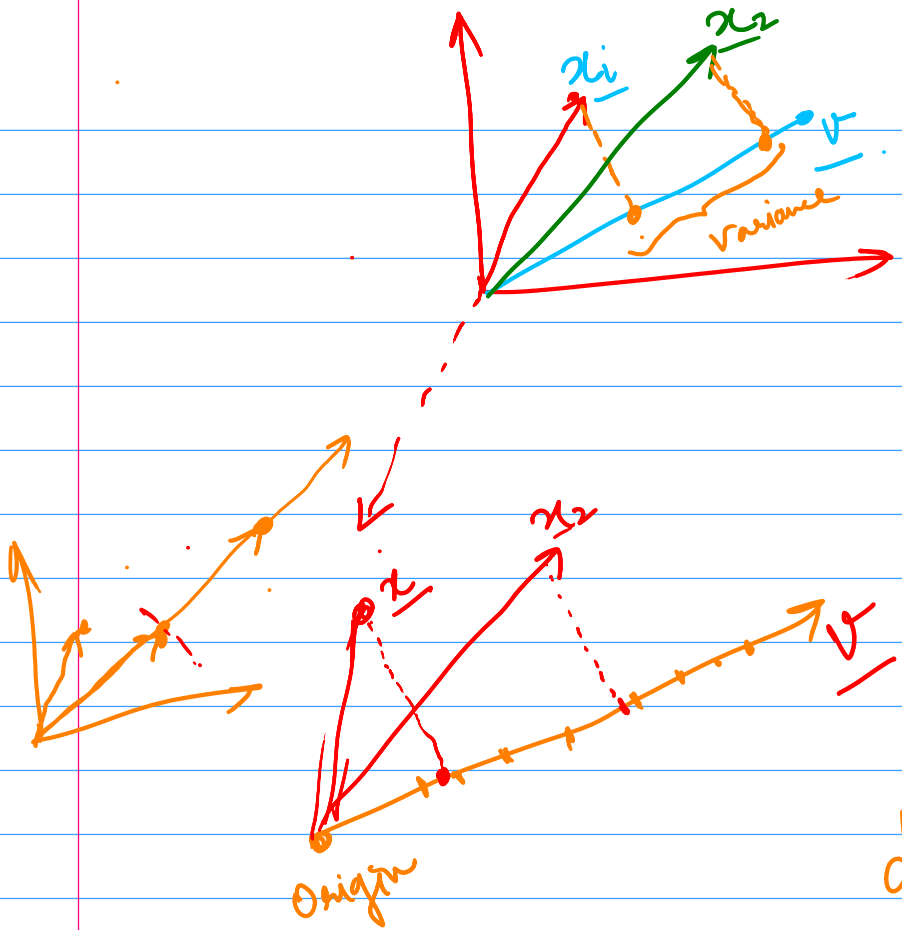
$\Lambda_{ii}$: eigen value corresponding to $i^{th}$ eigenvector of $V$.

Objective is to maximize $\underline{v}^T \underline{C}\, \underline{v}$ subject to $\underline{v}$ being a unit vector.

Formulating the Lagrangian $\underline{v}^T \underline{C}\, \underline{v} - \lambda(\|\underline{v}\|^2 - 1)$

Taking derivative w.r.t. $\underline{v}$ and equating it to zero gives

# 1D projection




$v^T C v$

variance of all the points (examples) along direction $\underline{v}$

$\underline{v}$ = ( - - - )

direction

$\underline{v}$ defines a one-dim space

projecting any vector on $\underline{v}$ $\Rightarrow$ 1-D projection