BITS Pilani
Hyderabad Campus

# BITS Pilani

Prof.Aruna Malapati
Professor
Department of CSIS

innovate achieve lead

BITS Pilani
Hyderabad Campus

# Dimension Reduction using PCA

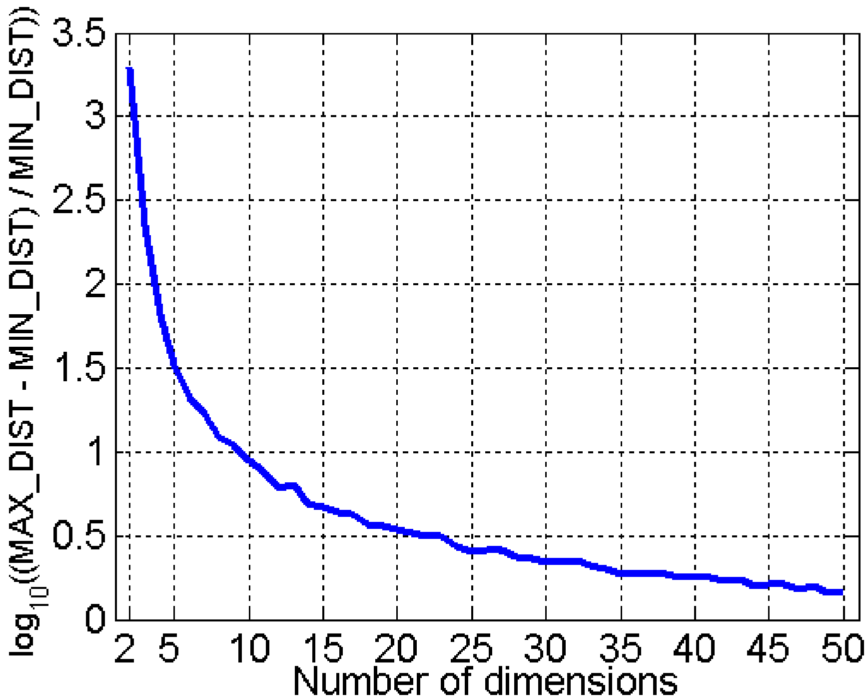
# Today's Agenda

- Curse of Dimensionality
- Introduction to Dimension Reduction
- Motivation for PCA
- PCA

# Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies

- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful




- **Randomly generate 500 points**
- **Compute difference between max and min distance between any pair of points**

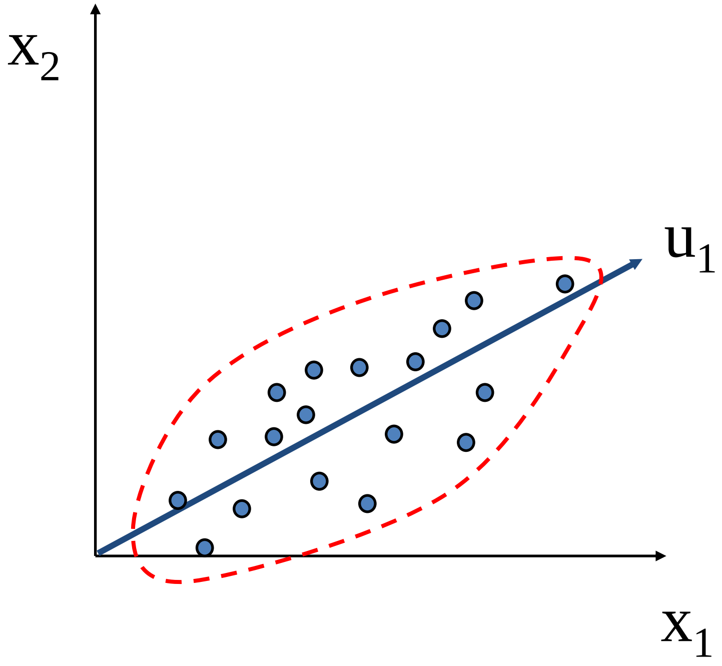
# Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

- Techniques
  - Principle Component Analysis
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

# Dimensionality Reduction: PCA

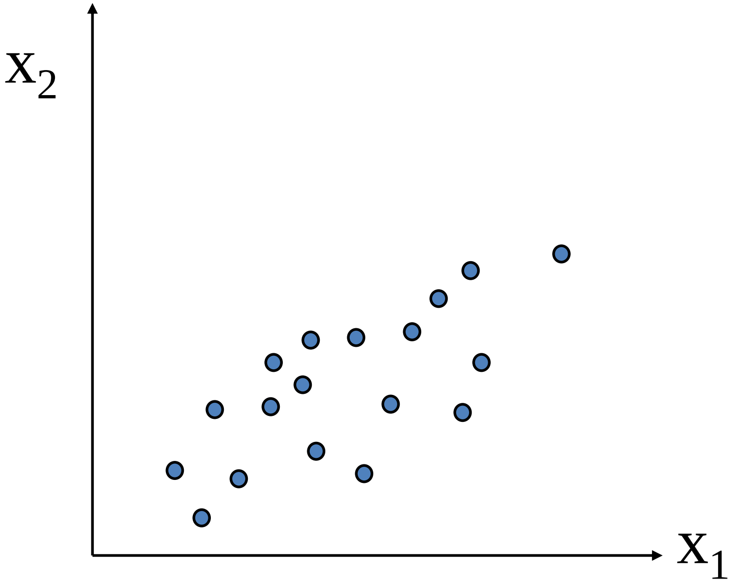
- Goal is to find a projection that captures the largest amount of variation in data



# PCA

- Reduce High Dimensional data into something that can be explained in fewer dimensions.
- We need PCA since we suspect that in our interesting data set not all measures are independent i.e there exist correlations.

Assume the data set represent height and weight of people in a region.



# Principal Component Analysis (PCA)

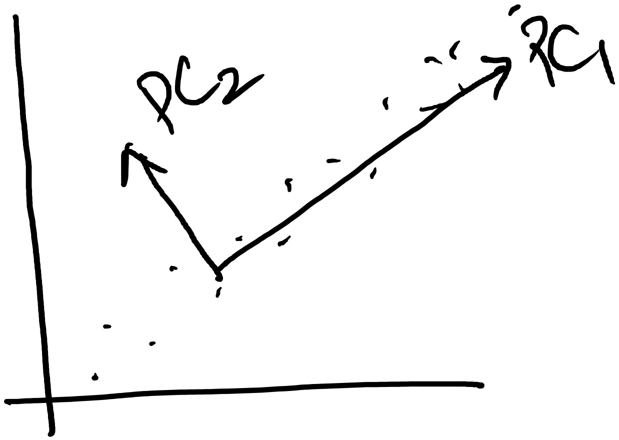
- Reduce higher dimension data into something that can be explained in fewer dimensions and gain an understanding of the data.
- We need PCA since we suspect that in our data set not all measures are independent and there exist correlations or structures or patterns.

# Motivation for PCA

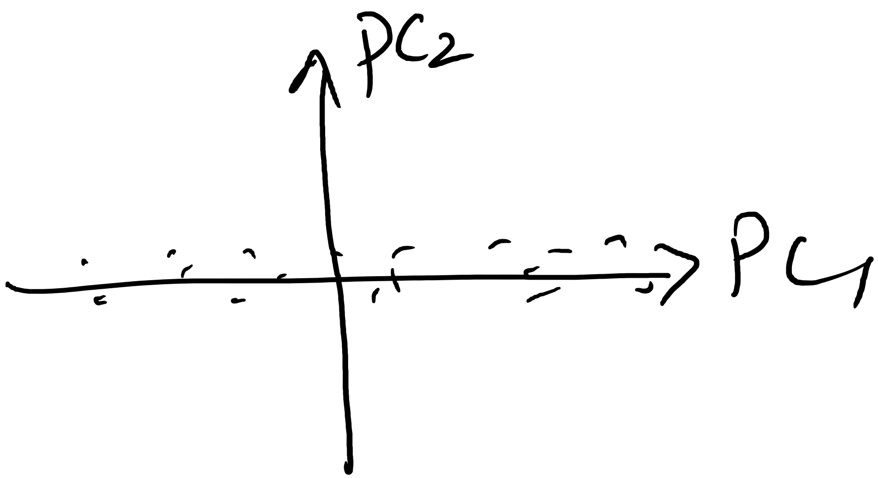

The data is not random, as $x_1$ increases $x_2$ also increases and hence +vely Co-related

Hence we can reduce from 2D → 1D

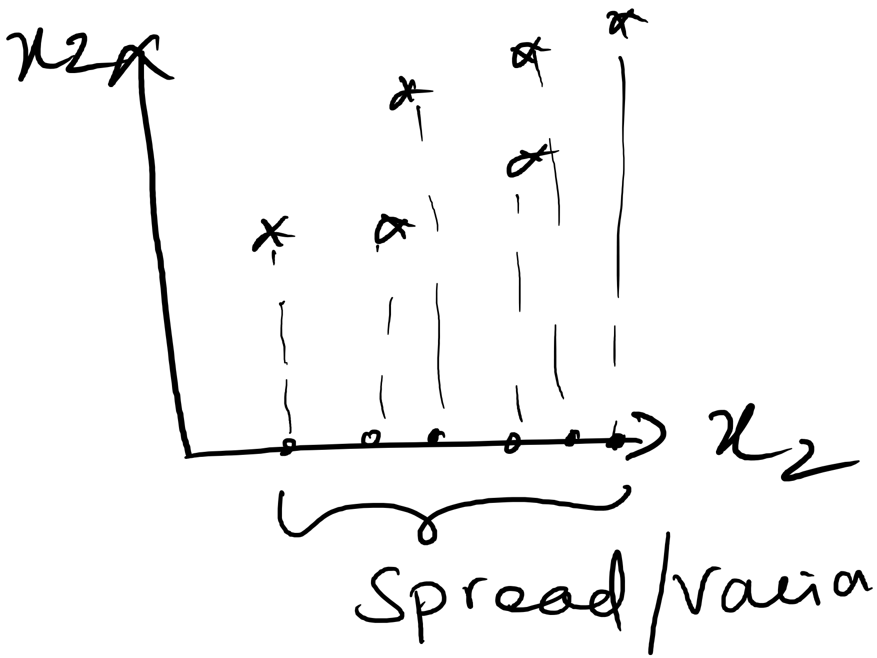
# Projection

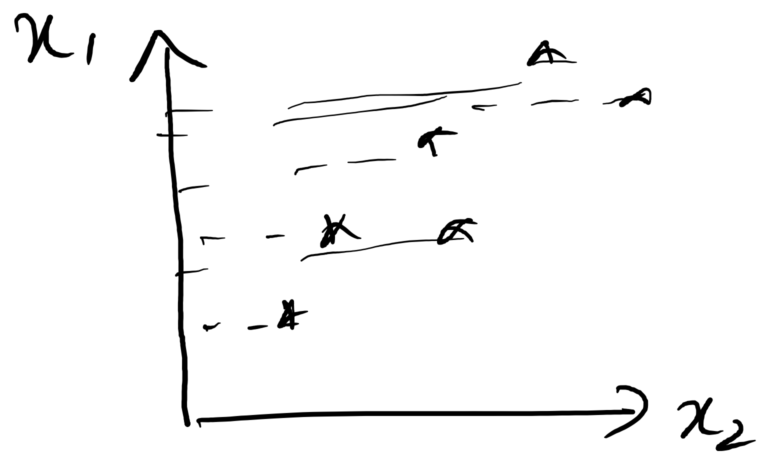



PC1 is the direction of the data, so that if we transform the data by projection.



# Projection (Contd..)



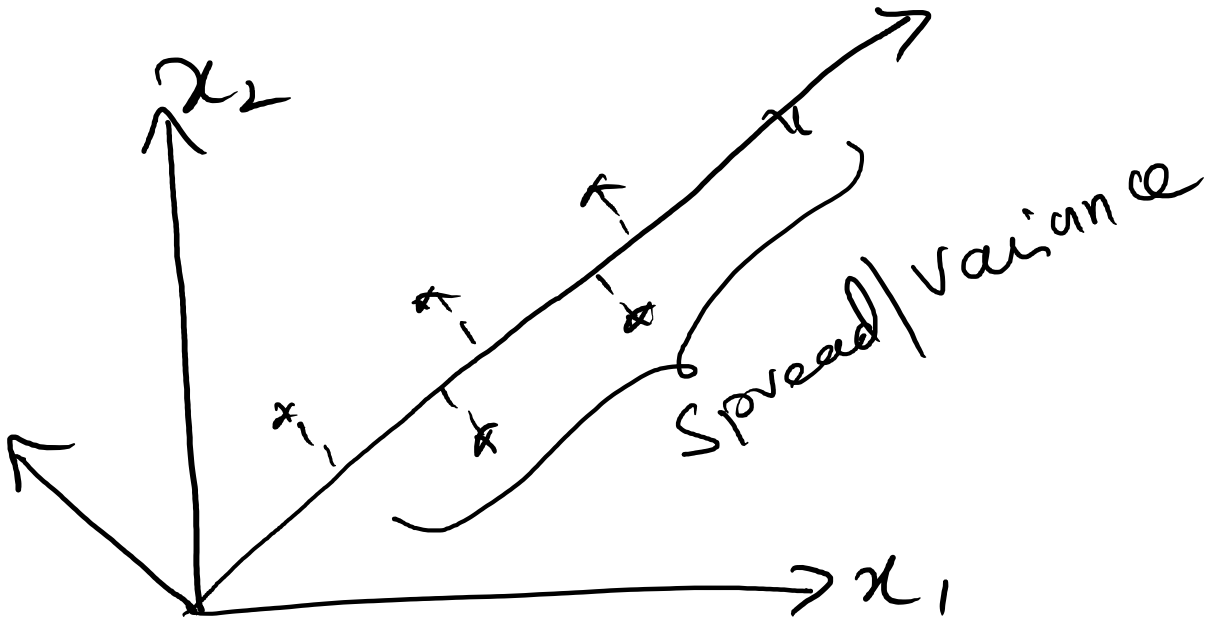




If data is projected on $x_1$ or $x_2$ we loose one dimension.

# Principal Component Analysis

PCA helps us in identifying the best projection.

The goal is to find a lower dimensional surface on which to project such that the sum of squares errors are minimal.



# Principal Component Analysis (Contd..)


We can project data on multiple vectors $(v_1, \ldots, v_n)$

But which is the best suitable vector for projection?

# Salient features of PCA

- Directions are in the order of % of variance explained.
- Every PC is orthogonal.

- PCA can be solved using
- Maximum Variance
- Minimum Error

# PCA formulation

Data Set $= \{x_i, y_i\}_{i=1}^{S}$ $S$ = no of Samples

Each $x_i \in \mathbb{R}^N$ $N$ = no of dimensions

Goal: Project data onto a space having M dim < D while Maximizing the Variance of projected data.

# Steps for PCA

1. Mean center the data
2. Compute covariance matrix
3. Compute the eigen value decomposition of covariance matrix.
4. Find the best eigen vector by using eigen values.

# Example

Suppose you have a 2D matrix

Compute $X^T X$ covariance matrix $2 \times 2$

$\Downarrow$

eigen decomposition

$\Downarrow$

2 eigen values

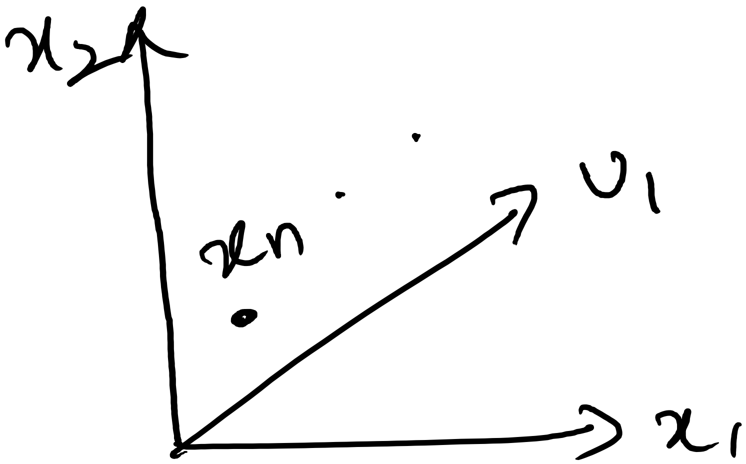
2 eigen vectors

# PCA formulation



Suppose our $M=1$ and let $v_1$ be a unit vector so that $v^T v = 1$

So that we are intrested in direction of $v_1$ not magnitude.




Each $x_n$ is projected on $v_1$ by taking dot product $v_1^T \cdot x_n$

# PCA formulation (Contd..)

Mean of projected data will be

$U_1^T \cdot \bar{x}$ where $\bar{x}$ is the set mean

$$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

$$cov(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

# PCA formulation (Contd..)

Variance $\sigma^2 = \frac{1}{N} \sum_{n=1}^{N} \left( u_1^T x_n - u_1^T \bar{x} \right)^2$

$$\sigma^2 = \frac{1}{N} \sum_{n=1}^{N} \left( u_1^T \left( x_n - \bar{x} \right) \right)^2$$

using linear algebra

$$\underbrace{(A.B)^2}_{N} = (A.B)(A.B)^T$$

$$= (A.B)\, B^T \cdot A^T$$

# PCA formulation (Contd..)

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}\left(v_1^T(x_n-\bar{x})\right)\left(v_1^T(x_n-\bar{x})\right)^T$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(v_1^T(x_n-\bar{x})\left(v_1(x_n-\bar{x})\right)^T\right)$$

Since $v_1$ & $v_1^T$ has nothing to do with data points

$$= v_1^T\left(\frac{1}{N}\sum_{n=1}^{N}(x_n-\bar{x})(x_n-\bar{x})^T\right)v_1$$

# PCA formulation (Contd..)

$$= u_1^T S u_1 \quad — (1)$$

where $S$ is the covariance Matrix

$$S = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T \quad — (2)$$

# How to derive S?

$$x = \begin{bmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{nD} \end{bmatrix} \quad \text{and} \quad \bar{x} = [\bar{x}_1 \;\; \bar{x}_2 \;\; \ldots \;\; \bar{x}_D]$$

$\bar{x}_1$ → mean of 1st feature

$$x_n - \bar{x} = \begin{bmatrix} x_{n1} - \bar{x}_1 \\ x_{n2} - \bar{x}_2 \\ \vdots \\ x_{nD} - \bar{x}_D \end{bmatrix}$$

$$(x_n - \bar{x})^T = [x_{n1} - \bar{x}_1, \; x_{n2} - \bar{x}_2 \ldots \; x_{nD} - \bar{x}_D]$$

# How to derive S?

$$(x_n-\bar{x})(x_n-\bar{x})^T =$$

$$\begin{matrix} (x_{n1}-\bar{x}_1)(x_{n1}-\bar{x}_1) & (x_{n1}-\bar{x}_1)(x_{n2}-\bar{x}_2) & \cdots & (x_{n1}-\bar{x}_1)(x_{nD}-\bar{x}_D) \\ (x_{n2}-\bar{x}_2)(x_{n1}-\bar{x}_1) & (x_{n2}-\bar{x}_2)(x_{n2}-\bar{x}_2) & \cdots & (x_{n2}-\bar{x}_2)(x_{nD}-\bar{x}_D) \\ \vdots & & & \end{matrix}$$

$$= \begin{bmatrix} (x_{n1}-\bar{x}_1)^2 & - & - & - & - & - \\ - & (x_{n2}-\bar{x}_2)^2 & - & - & - & \end{bmatrix}$$

Substitute this matrix in S eqn 2

# How to derive S?

$$= \frac{1}{N} \sum_{n=1}^{N} \begin{bmatrix} (x_{n1} - \bar{x}_1)^2 & & \\ & (x_{n2} - \bar{x}_2)^2 & \\ & & \ddots \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{N} \sum_{n=1}^{N} (x_{n1} - \bar{x}_1)^2 & \frac{1}{N} \sum_{n=1}^{N} (x_{n1} - \bar{x}_1)(x_{n2} - \bar{x}_2) & \cdots \\ & & \end{bmatrix}$$

# How to derive S?

innovate achieve lead

The resolt is a co-variance matrix of the form.

$$\begin{pmatrix} \text{Var}\, x_{n1} & \text{Co-var}(x_{n1}, x_{n2}) \cdots & \text{Cov}(x_{n1}, x_{nn}) \\ \text{Cov}(x_{n1}, x_{n2}) & & \\ & & \end{pmatrix}$$

# PCA overview

Once Co-variance matrix is derived

$$X^T X$$

$$\Downarrow$$

$$\text{eign}(X^T X)$$

$$\Downarrow$$

$$(\lambda_1\ \lambda_2\ --\ \lambda_D)\ (w_1\ w_2\ --\ w_D)$$

eigen values 'gen vectors

# PCA Limitations

- Covariance is extremely sensitive to large values
  - Multiply some dimension by 1000
  - Dominates covariance
  - Becomes principal component
- Normalize each dimension to zero mean and unit variance.

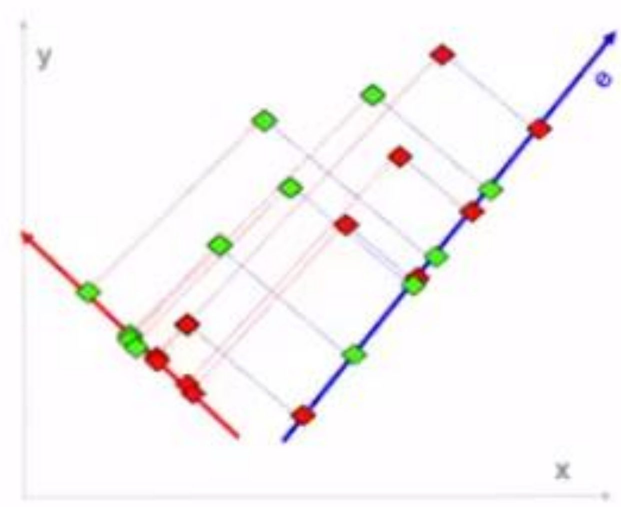
  X'=(X-mean)/standard-deviation

# PCA Limitations

- PCA assumes underlying subspace is linear.
- 1D –line
- 2D - plane

# PCA and classification

- PCA is unsupervised
- Maximize overall variance of the data along a small set of directions
- Does not known anything about class labels
- Can pick direction that makes it hard to separate classes



# Take home message

- As the number of dimensions increases, the complexity and computational power required to build the model also increases.
- Dimension reduction methods are employed to find the best representation of data.
- PCA finds the best vectors on which the maximum variance in the data can be preserved.