Principal Component Analysis (PCA)

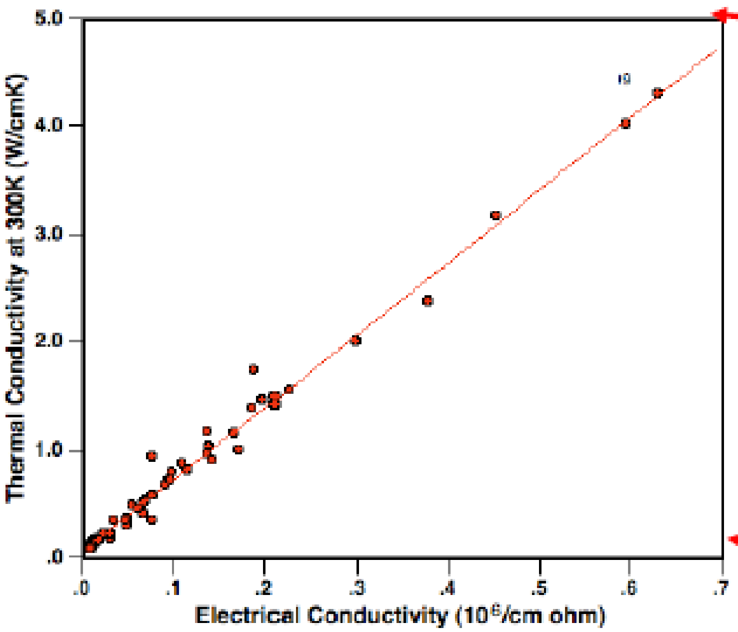
Dr. M. Shiple

Motivation

- Visualization
- Clustering
  - One way to summarize a complex real-valued data point with a single categorical variable
- Dimensionality reduction
  - Another way to simplify complex high-dimensional data
  - Summarize data with a lower dimensional real valued vector

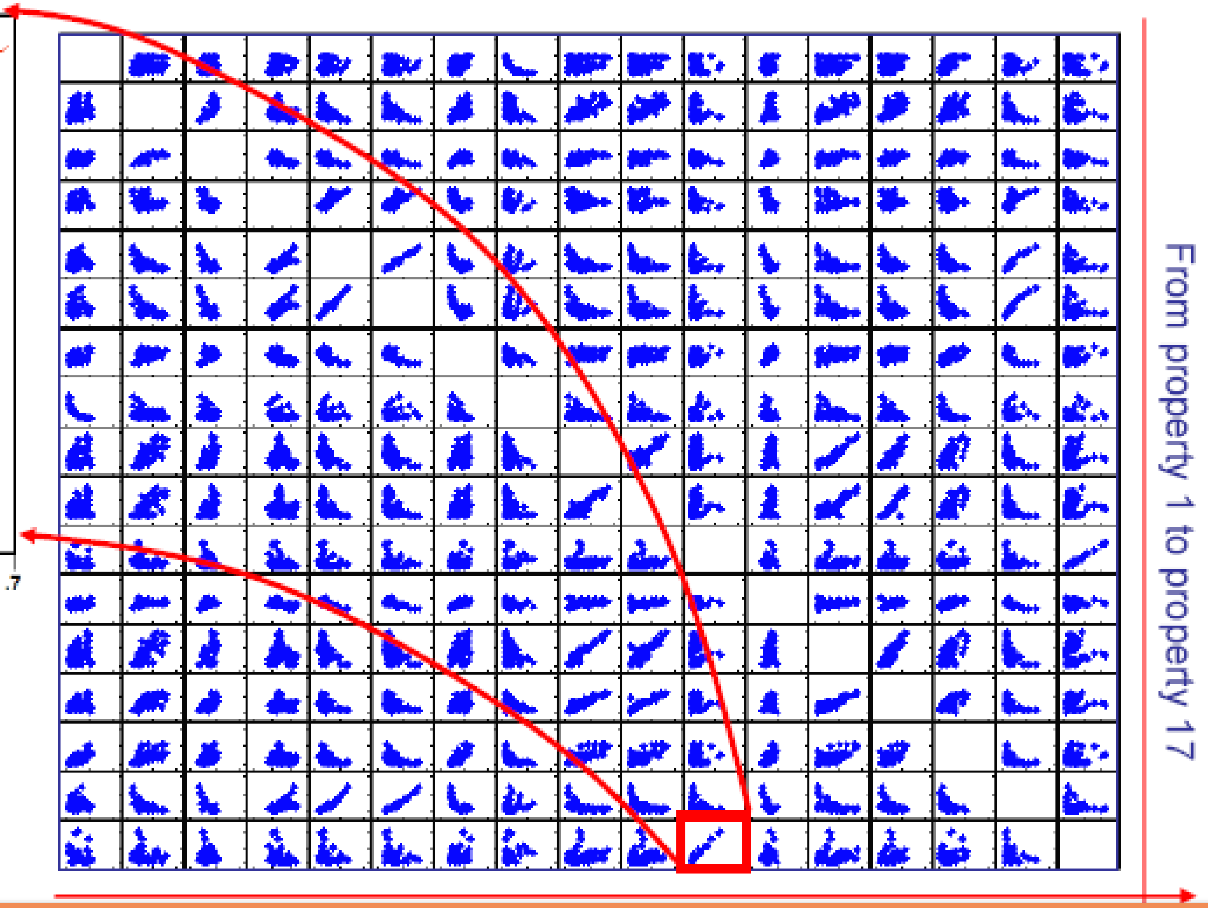
Traditional way of handling multivariate data set

- Bivariate Plots of Elemental Properties



Some relationships can be found from the bivariate plots of raw data

Ex. **Wiedemann-Franz Law**



When we use traditional techniques,

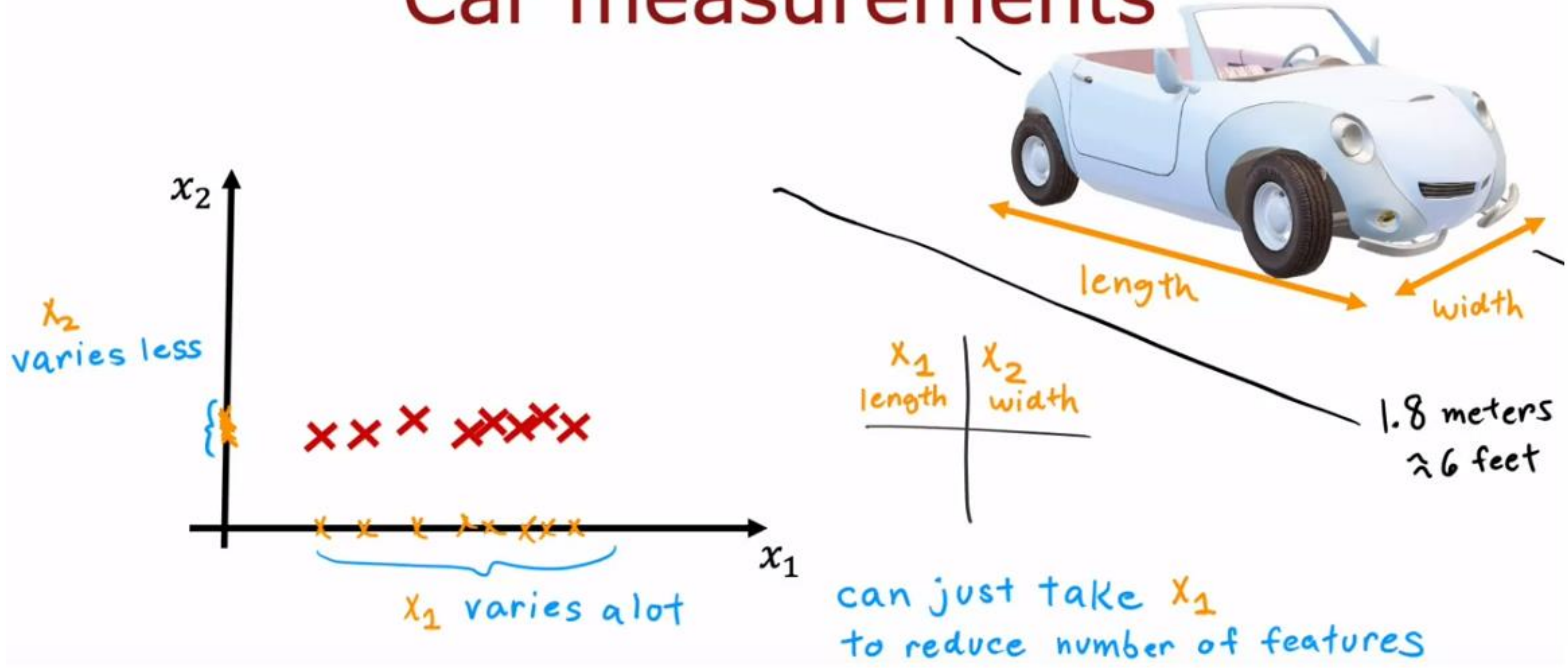
- 1. Not easy to extract useful information from the multivariate data
- 1) Many bivariate plots are needed
- 2) Bivariate plots, however, mainly represent correlations between variables (not samples).

Motivation

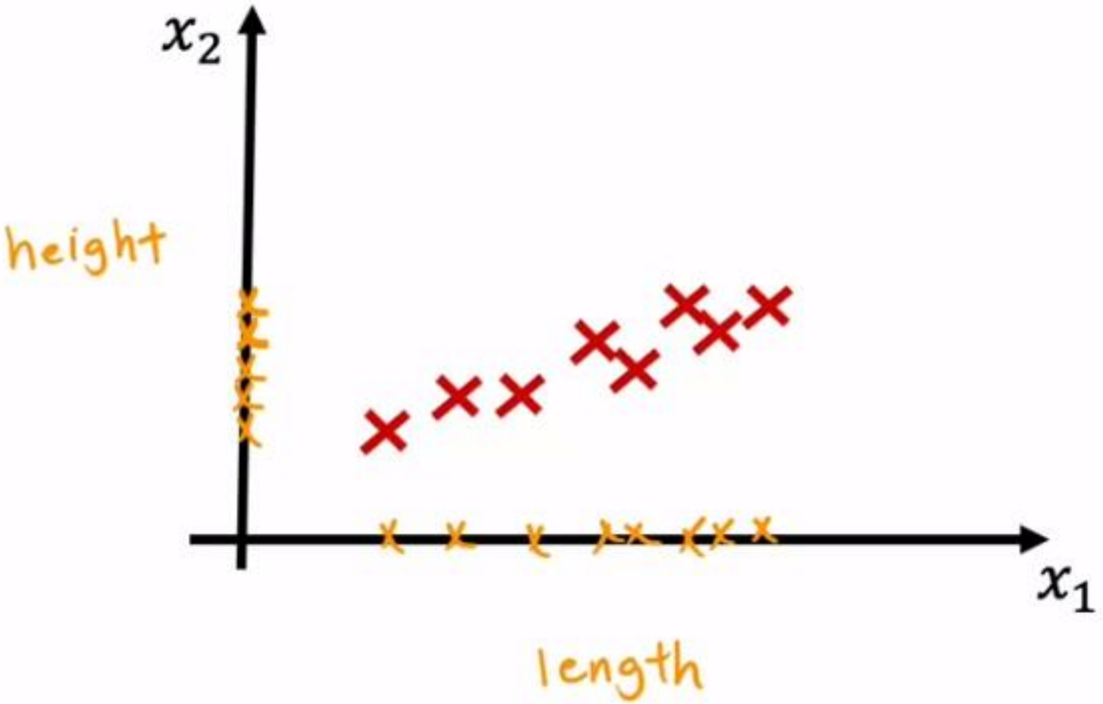
- Visualization
- Clustering
  - One way to summarize a complex real-valued data point with a single categorical variable
- Dimensionality reduction
  - Another way to simplify complex high-dimensional data
  - Summarize data with a lower dimensional real valued vector

- Given data points in *d* dimensions
- Convert them to data points in $r<d$ dimensions
- With minimal loss of information

Car measurements

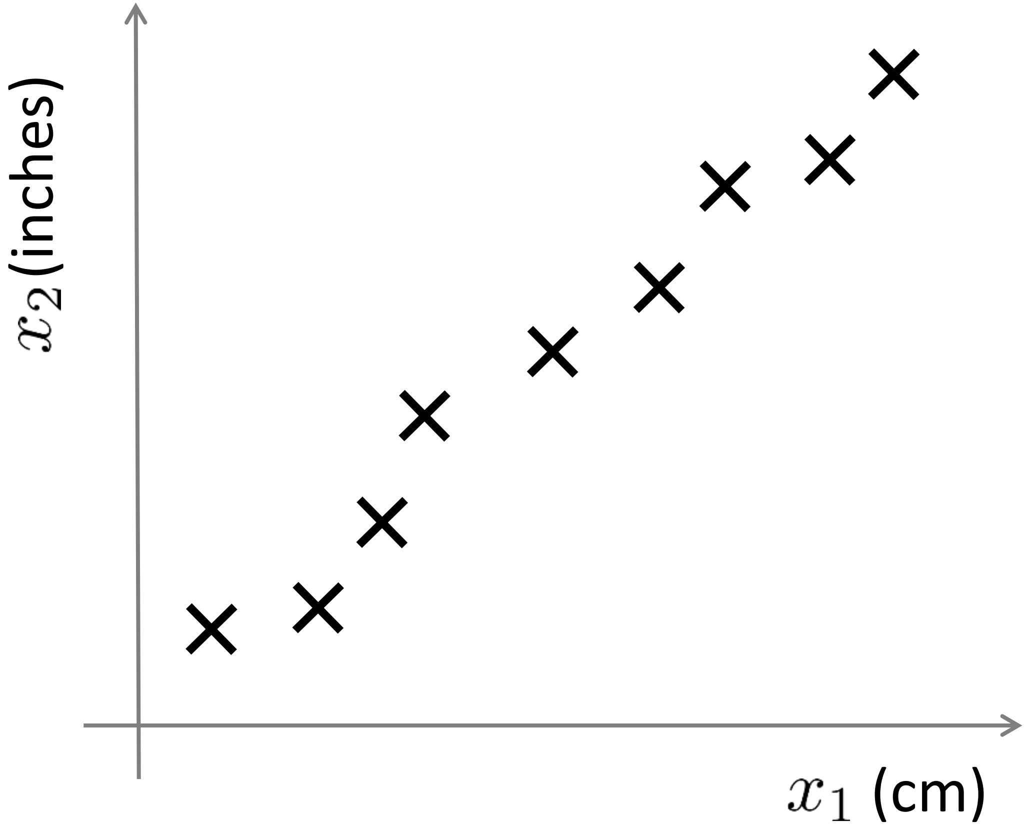


Size



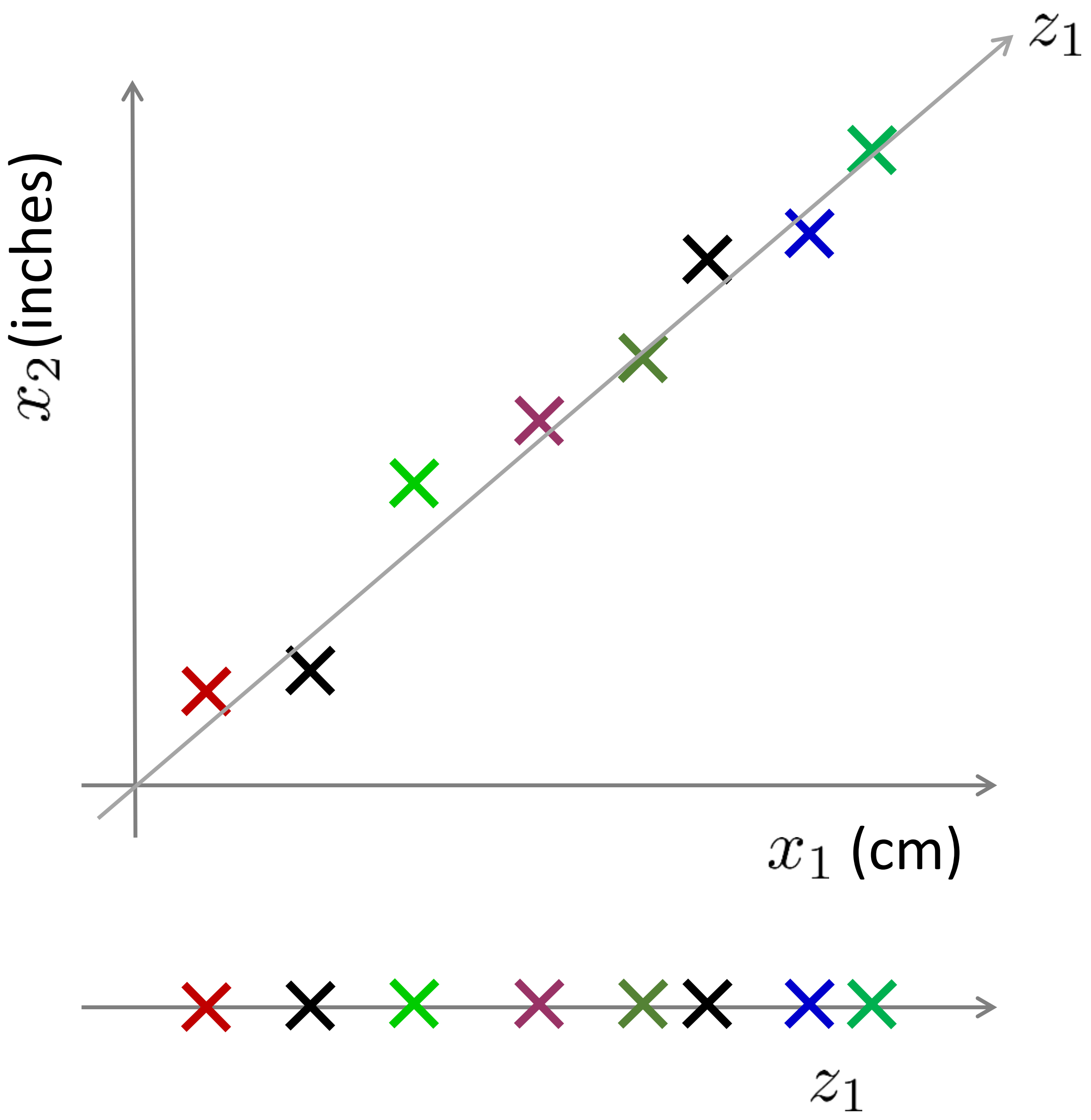
x_1	x_2
length	height

Data Compression



Reduce data from 2D to 1D

Data Compression



Reduce data from 2D to 1D

$$x^{(1)} \rightarrow z^{(1)}$$

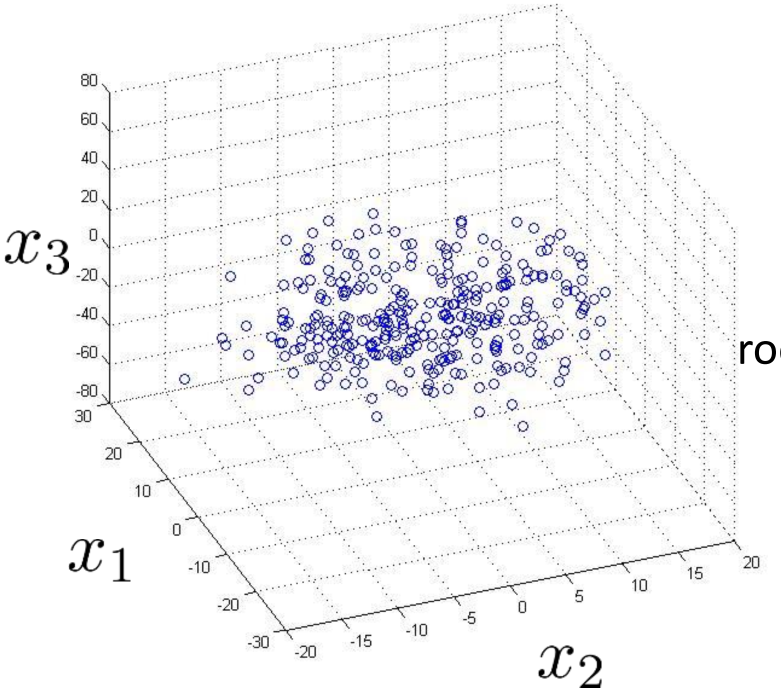
$$x^{(2)} \rightarrow z^{(2)}$$

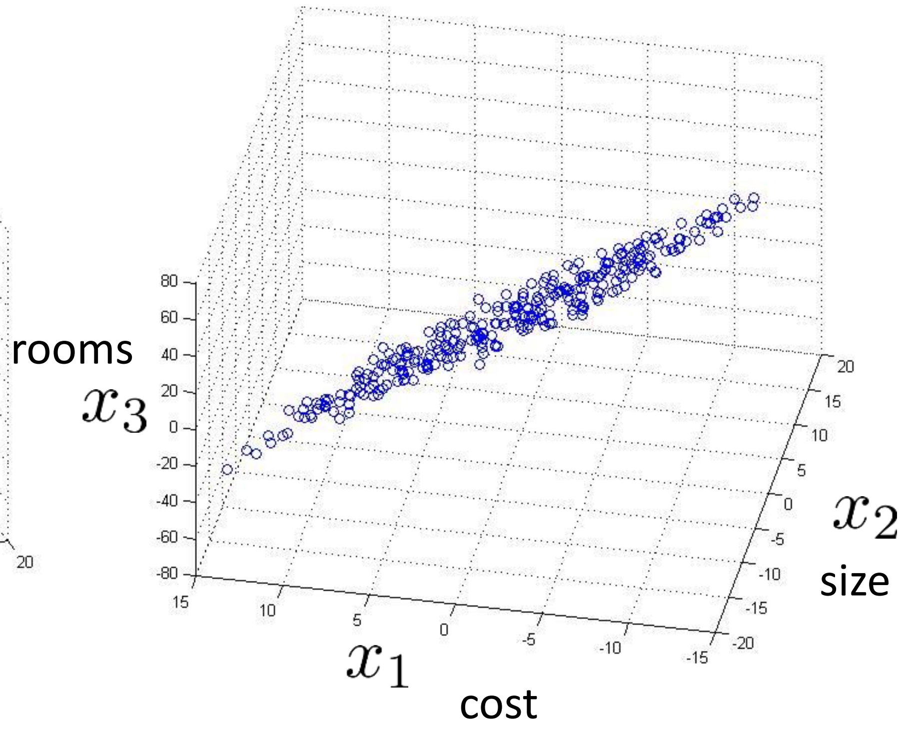
$$\vdots$$

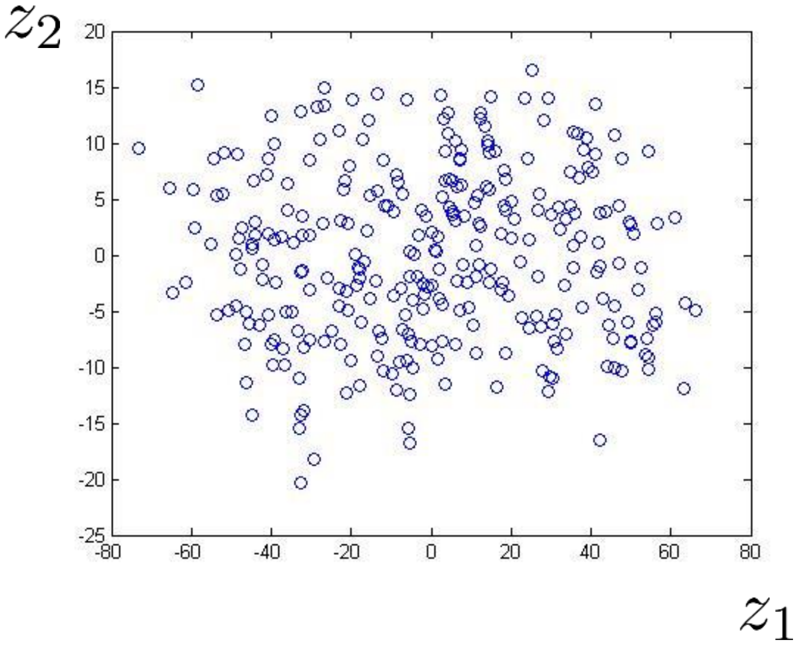
$$x^{(m)} \rightarrow z^{(m)}$$

Data Compression

Reduce data from 3D to 2D

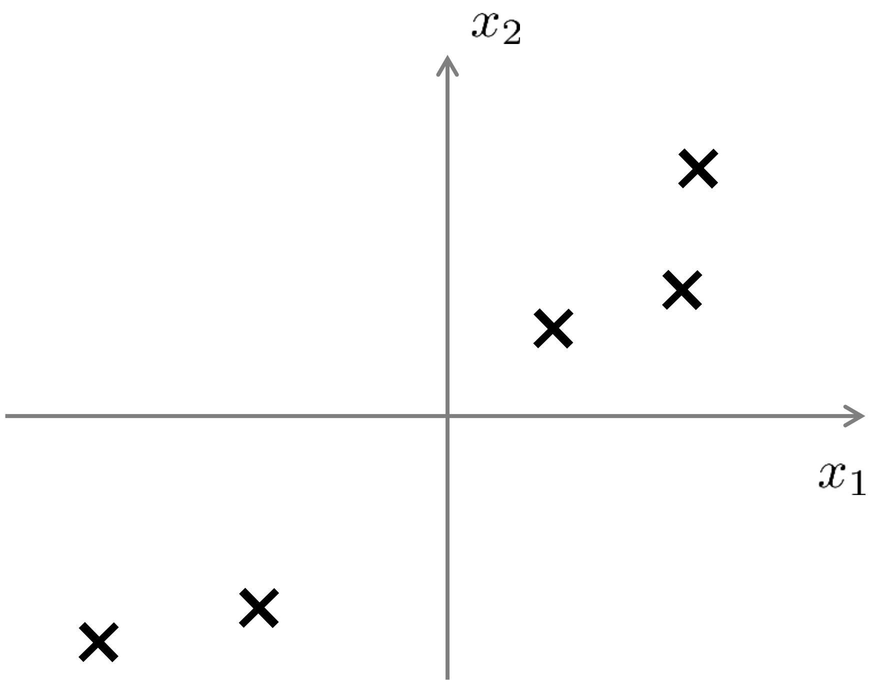




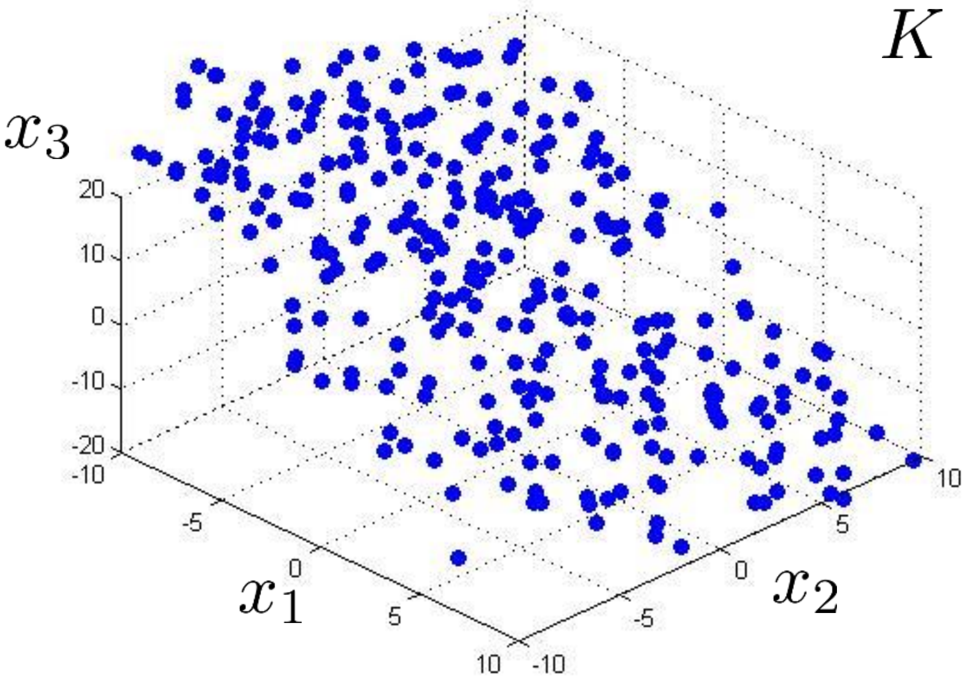


1- normalization
2- normalize mean

Principal Component Analysis (PCA) problem formulation



$3D \rightarrow 2D$
$K = 2$



Reduce from 2-dimension to 1-dimension: Find a direction (a vector $u^{(1)} \in \mathbb{R}^n$) onto which to project the data so as to minimize the projection error.

Reduce from n-dimension to k-dimension: Find $k$vectors $u^{(1)}, u^{(2)}, \ldots, u^{(k)}$ onto which to project the data, so as to minimize the projection error.

Principal Component Analysis

Goal: Find r-dim projection that best preserves variance

1. Compute mean vector μ and covariance matrix Σ of original points
2. Compute eigenvectors and eigenvalues of Σ
3. Select top r eigenvectors
4. Project points onto subspace spanned by them:

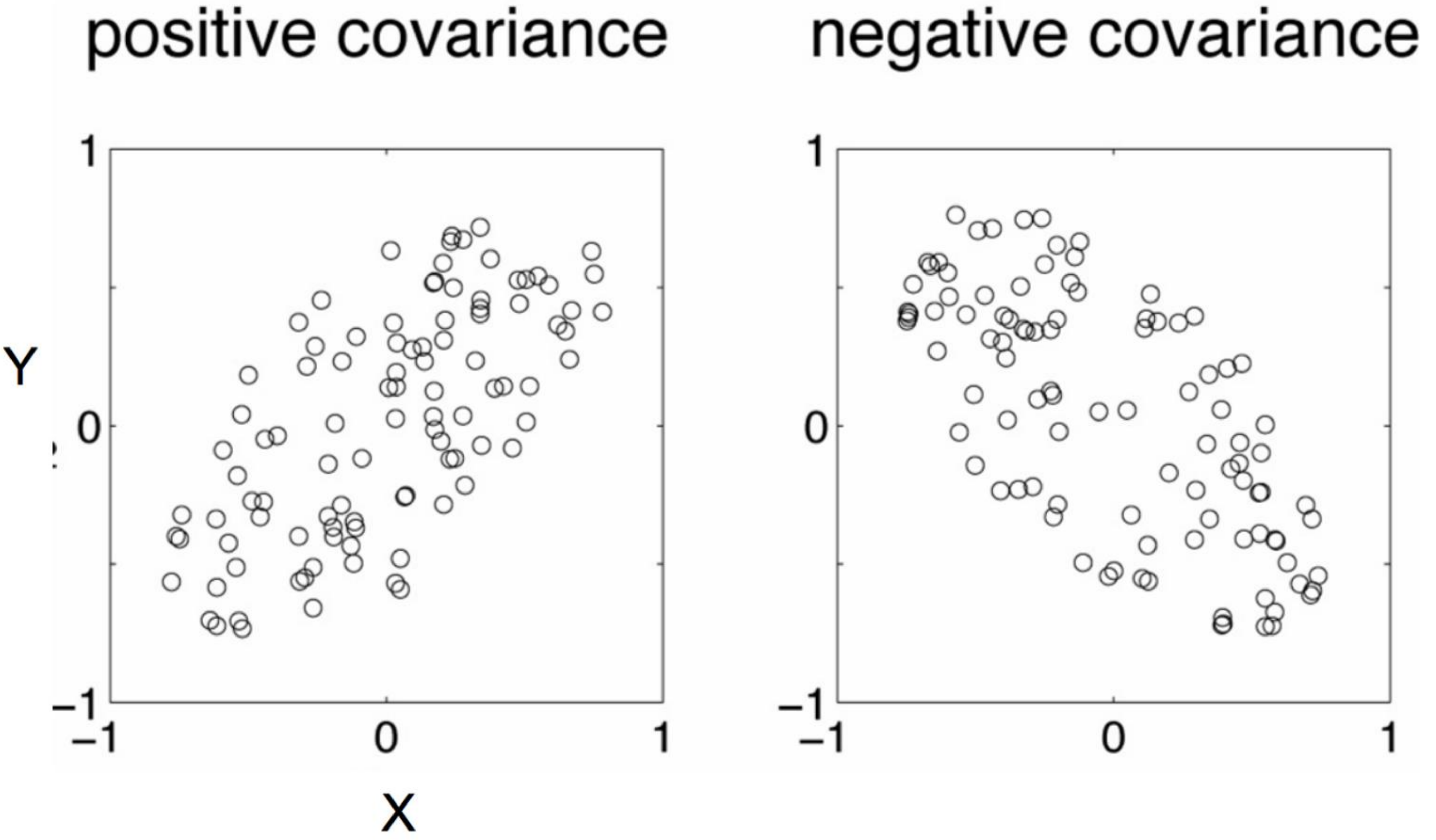
$$y = A(x - \mu)$$

   where y is the new point, x is the old one, and the rows of A are the eigenvectors

Covariance

- Variance and Covariance:
  - Measure of the “spread” of a set of points around their center of mass(mean)
- Variance:
  - Measure of the deviation from the mean for points in one dimension
- Covariance:
  - Measure of how much each of the dimensions vary from the mean with **respect to each other**

- **Covariance is measured between two dimensions**
- **Covariance sees if there is a relation between two dimensions**
- **Covariance between one dimension is the variance**



Positive: Both dimensions increase or decrease together

Negative: While one increase the other decrease

Standard Deviation

The average distance from the mean of the data set to a point

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}}$$

MEAN: $$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Example:

Measurement 1: 0,8,12,20
Measurement 2: 8,9,11,12

M1	M2
Mean 10	Mean 10
SD 8.33	SD 1.83

Variance

Variance is another measure of the spread of data in a data set.

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}$$

Example:

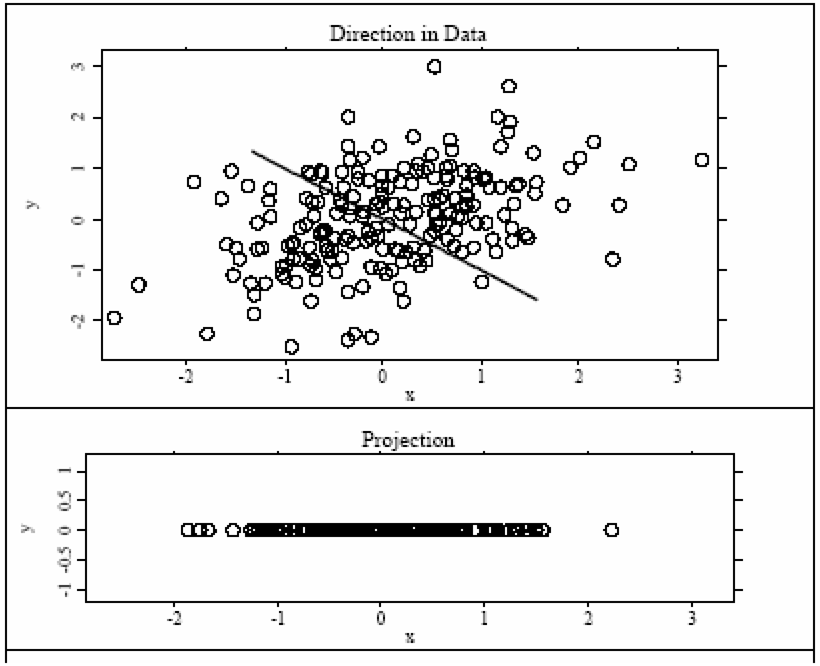
Measurement 1: 0,8,12,20
Measurement 2: 8,9,11,12

M1	M2
Mean 10	Mean 10
SD 8.33	SD 1.83
Var 69.33	Var 3.33

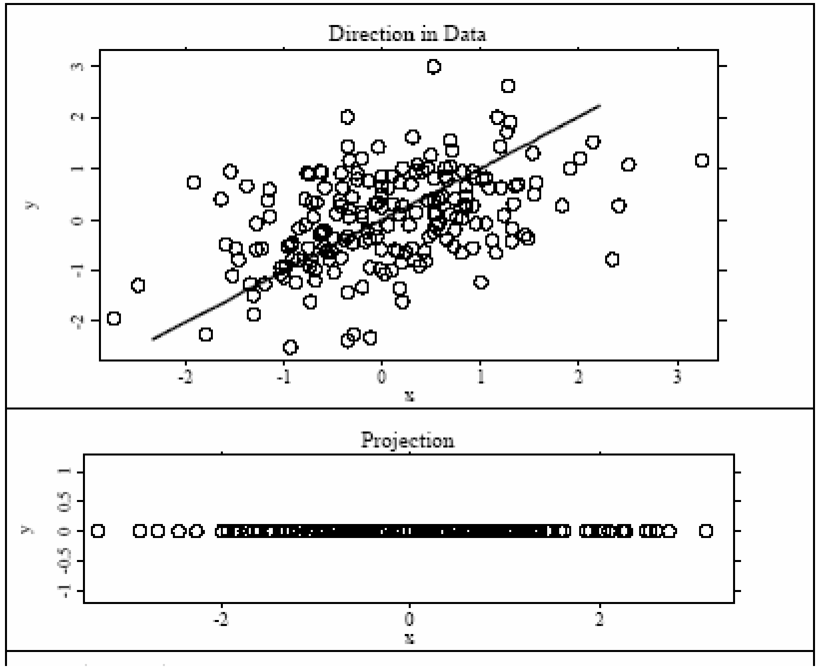
Transformation

Can we intuitively see that in a picture?

Good



Better



Covariance

Standard Deviation and Variance are 1-dimensional

How much do the dimensions vary from the mean with respect to each other ?

Covariance measures between 2 dimensions

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

We easily see, if X=Y we end up with variance

Covariance Matrix

Let X be a random vector.

Then the covariance matrix of X, denoted by Cov(X), is $\{Cov(X_i, X_j)\}$

The diagonals of Cov(X) are $Cov(X_i, X_i) = Var[X_i]$

In matrix notation,

$$\mathbf{Cov(X)} = \begin{pmatrix} Var[X_1] & \cdots & Cov(X_1, X_n) \\ \vdots & & \vdots \\ Cov(X_n, X_1) & \cdots & Var[X_n] \end{pmatrix}.$$

The covariance matrix is symmetric

Eigenvector and Eigenvalue

$$Ax = \lambda x$$

A: Square Matirx

λ: Eigenvector or characteristic vector

X: Eigenvalue or characteristic value

- ***The zero vector can not be an eigenvector***
- ***The value zero can be eigenvalue***

Eigenvector and Eigenvalue

Example 1: Find the eigenvalues of

$$A = \begin{bmatrix} 2 & -12 \\ 1 & -5 \end{bmatrix}$$

$$|\lambda I - A| = \begin{vmatrix} \lambda - 2 & 12 \\ -1 & \lambda + 5 \end{vmatrix} = (\lambda - 2)(\lambda + 5) + 12$$

$$= \lambda^2 + 3\lambda + 2 = (\lambda + 1)(\lambda + 2)$$

two eigenvalues: –1, – 2

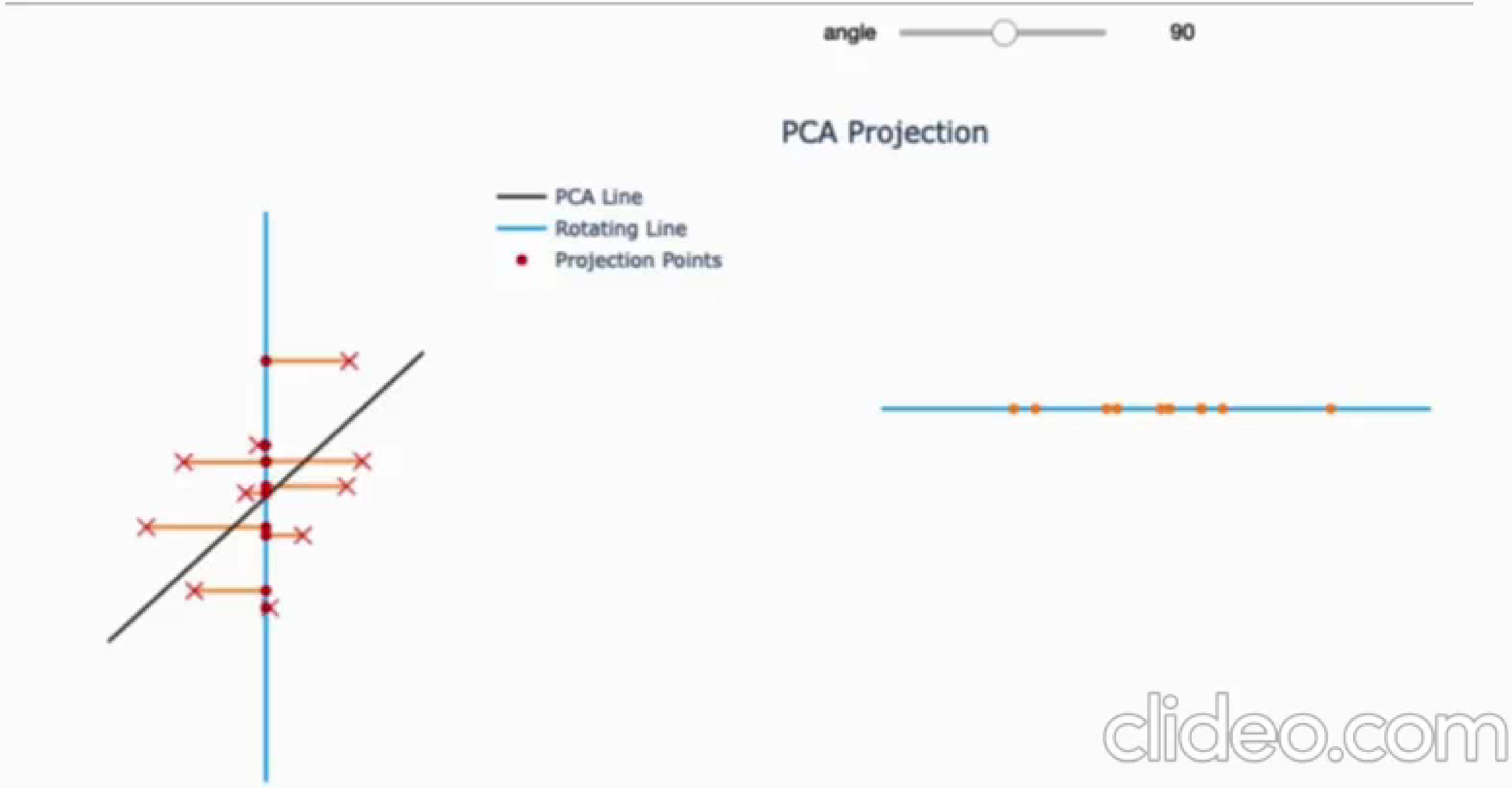
Note: The roots of the characteristic equation can be repeated. That is, $\lambda_1 = \lambda_2 = \ldots = \lambda_k$. If that happens, the eigenvalue is said to be of multiplicity k.

Example 2: Find the eigenvalues of

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

$$|\lambda I - A| = \begin{vmatrix} \lambda - 2 & -1 & 0 \\ 0 & \lambda - 2 & 0 \\ 0 & 0 & \lambda - 2 \end{vmatrix} = (\lambda - 2)^3 = 0$$

$\lambda = 2$ is an eigenvector of multiplicity 3.


angle
90
PCA Projection
PCA Line
Rotating Line
Projection Points
clideo.com