Principal Component Analysis

JAYANT RANGI(MA17BTECH11006)
RITESH YADAV(MA17BTECH11009)

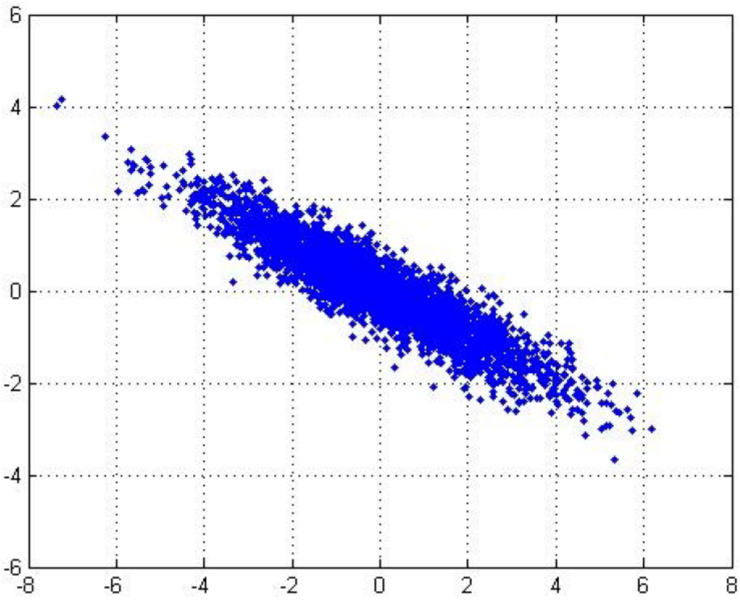
February 27, 2019

Definition :

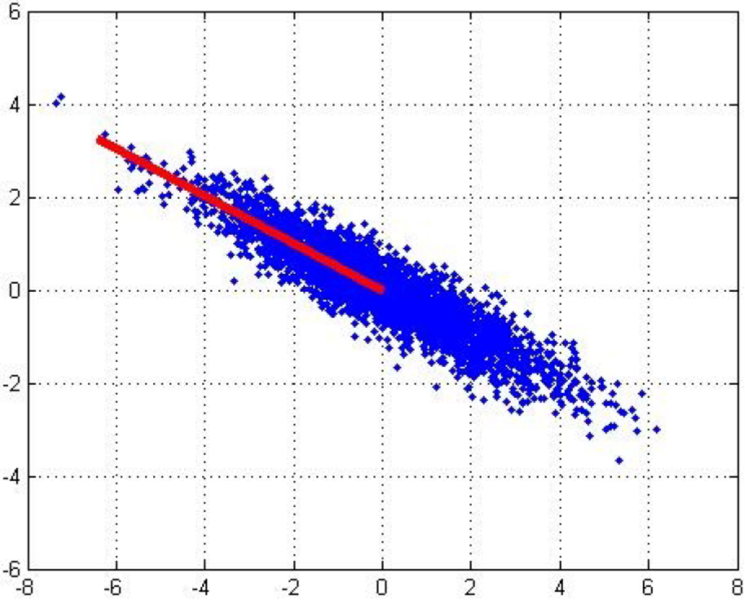
Orthogonal projection of data onto lower-dimension linear space that...

- maximizes variance of projected data
- minimizes mean squared distance between
  - data points and
  - projections

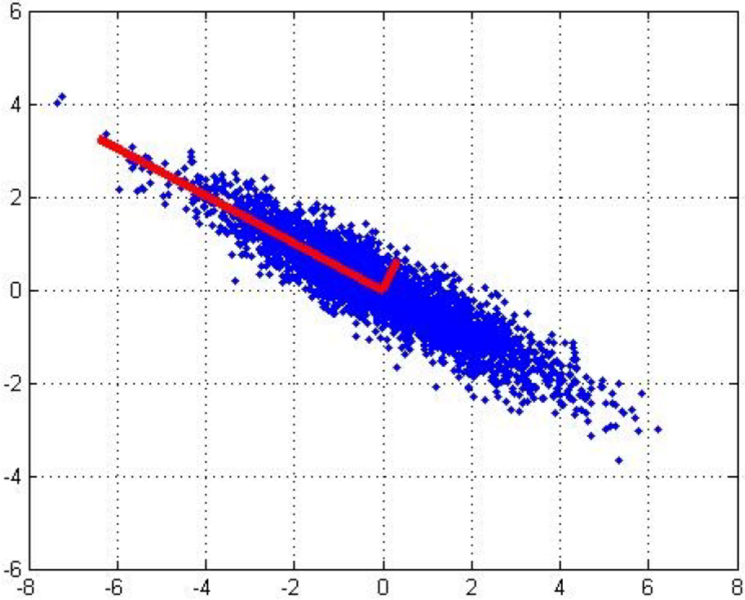
Random 2D Gaussian Data set :



1st PCA axis :



2st PCA axis :



Idea :

- Given data points in a **d-dimensional** space, project into **lower dimensional** space while preserving as much information as possible

- In particular, choose projection that **minimizes squared error** in reconstructing original data

Applications :

- **Data Visualization**
- **Data Compression**
- **Noise Reduction**
- **Data Classification**
- **Trend Analysis**
- **Factor Analysis**

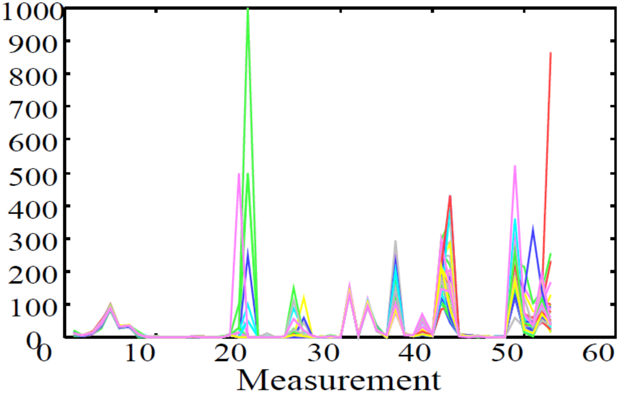
Example :

Let's take an example :

- You are given data of 53 different features from 65 people.
- **How can you visualize such larger measurements?**

Now consider the given graph :

Spectral Format :



Difficult to compare different features by the above Graph .

Conclusion From The Example :

- Is there a representation better than the coordinate axes?
- Is it really necessary to show all the 53 dimensions?
  what if there are strong correlations between the features?

Principal Component Analysis helps us in dealing such situations by removing the irrelevant information and keeping the strongly relevant information .

Mathematics behind Principal Components Analysis:

The principal components of a set of data in $|R^p$ provide a sequence of best linear approximations to that data, of all ranks q $\leq$ p

Denote the observations by $x_1, x_2, ..., x_N$, and consider the rank-q linear model for representing them.

$f(\lambda) = \mu + \mathbf{V}_q\lambda$

μ is a location vector in R^p.

V_q is a p$\times$q matrix with q orthogonal unit vectors as columns, and λ is a q vector of parameters. This is the parametric representation of an affine hyperplane of rank q.

To fit this model to data we will use least squares approximation .

Mathematics behind Principal Components Analysis:

$$\min_{\mu,\lambda_i,\mathbf{V}_q} \sum_{n=1}^{N} ||x_i\text{-}\mu\text{-}\mathbf{V}_q\lambda_i||^2$$

$\mu = \overline{x}$
$\lambda_i = \mathbf{V}_q^T(x_i - \overline{x})$
This leaves us to find the orthogonal matrix $\mathbf{V}_q$:

$$\min_{\mathbf{V}_q} \sum_{n=1}^{N} ||(x_i - \overline{x}) - \mathbf{V}_q\mathbf{V}_q^T(x_i - \overline{x})||^2$$

For convenience we assume $\overline{x} = 0$
The pmatrix $\mathbf{H}_q = \mathbf{V}_q\mathbf{V}_q^T$ is a a projection matrix, and maps each point x_i onto its rank-q recontruction $\mathbf{H}_q x_i$, the orthogonal projection of x_i onto the subspace spanned by columns of $\mathbf{V}_q$.