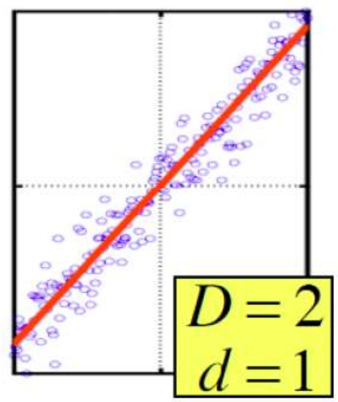
Dimensionality Reduction

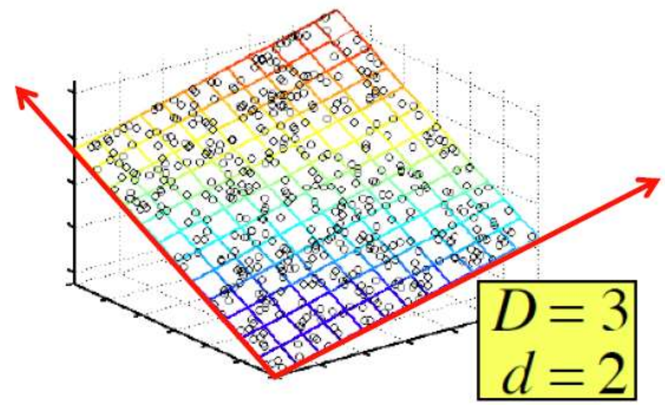
Curse of Dimensionality

- Most of the real-world datasets are having thousands, or millions of dimensions.
- Problems of having high dimensional data
  - The error increases with he increase in the number of features
  - The computational cost of data mining/machine learning techniques increases exponentially.
  - The data becomes very sparse in high dimensional dataset, making the machine learning/data mining algorithms in effective.
  - Overfitting problem in the predictive models.

Dimensionality Reduction

- Usually, the data can be described with fewer dimensions, without losing much of the meaning of the data.
  - The data reside in a space of lower dimensionality





Why to Reduce Dimension?

- Visualization: Projection of high dimensional data onto 2D or 3D.
- Data Compression: Efficient storage and retrieval.
- Noise Removal: Positive effect on accuracy of the built model.
- Remove Redundant Features: Positive effect on the performance of the model.
- Hidden Correlations: May find hidden correlations among features.

Covariance

- Variance: measure of the deviation from the mean for points in one dimension e.g., heights
- Covariance: measure of how much each of the dimensions vary from the mean with respect to each other.
- Covariance is measure between two dimension to see if there is a relationship between the 2 dimensions e.g., number of hours studied, and marks obtained.
- The covariance between one dimension and itself is the variance.

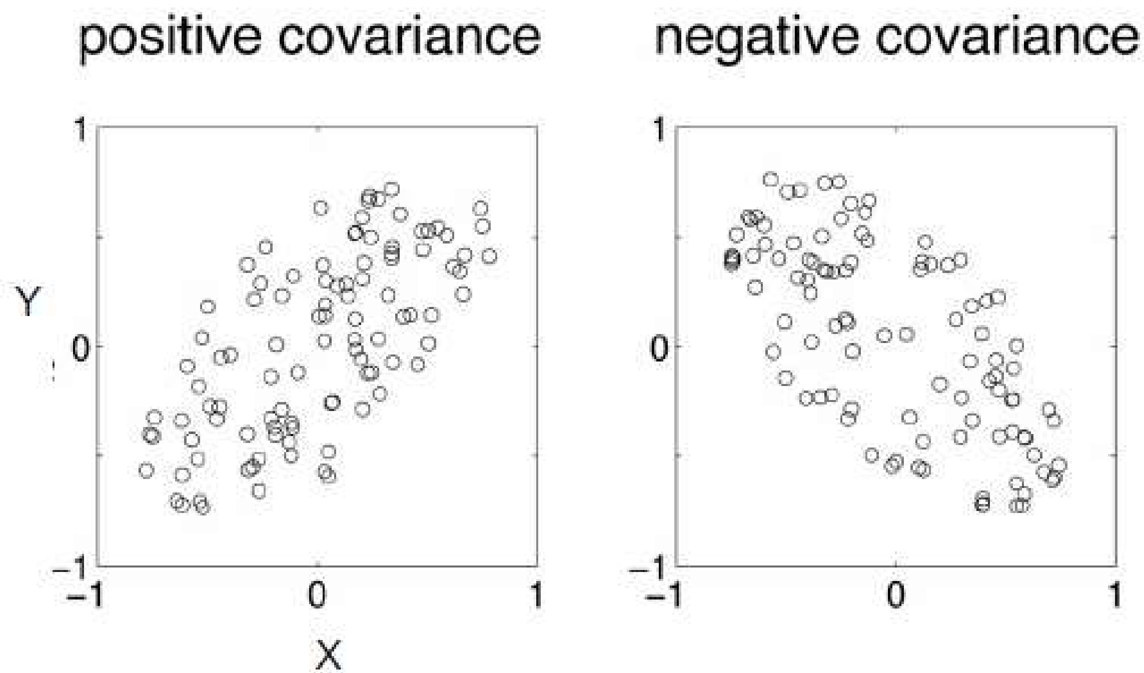
Covariance Matrix

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

$$C = \begin{bmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{bmatrix} \quad \textbf{Variances}$$

- Diagonal is the variances of x, y and z.
- cov (x,y) = cov (y,x) hence matrix is symmetrical about the diagonal.
- N-dimensional data will result in N x N covariance matrix.

Covariance Examples



Covariance

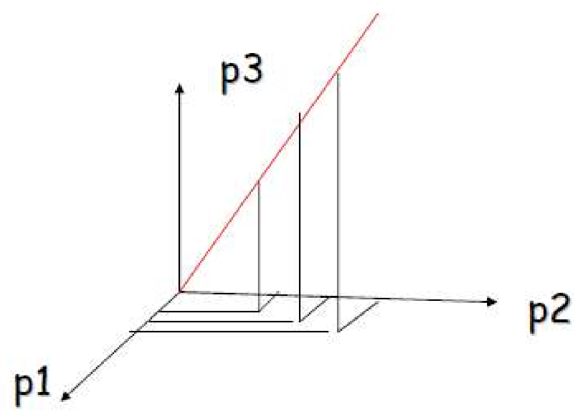
- A positive value of covariance indicates both dimensions increase or decrease together, e.g. as the number of hour studied increases, the marks in that subject increase.
- A negative value indicates while one increases the other decreases, or vice versa.
- If covariance is zero: the two dimensions are independent of each other e.g., height of students vs marks obtained in a subject.

Principal Component Analysis

- PCA is a technique to reduce the dimension of a dataset without affecting the information.
- It is a linear transformation that chooses a new coordinate system for the dataset such that:
  - The greatest variance by any projection of the data set comes to lie on first axis (called the first principal component)
  - The second greatest variance on the second axis and so on.
- PCA can be used for reducing dimensionality by eliminating the later principal components.

Geometrical Interpretation

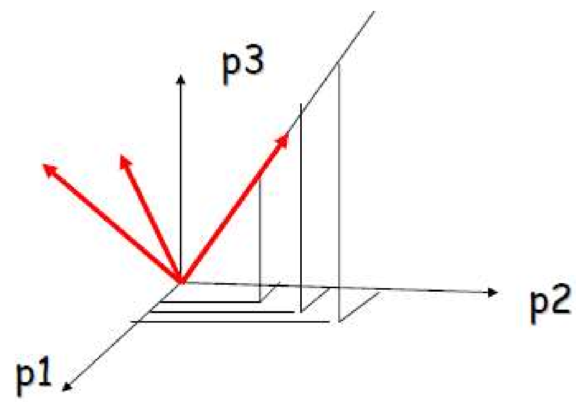
View each point in 3D space.



In this example, all the points happen to belong to a line: a 1D subspace of the original 3D space.

Geometrical Interpretation

Consider a new coordinate system where one of the axes is along the direction of the line.



Here every point has only one non-zero coordinate.

PCA-Concept

- Given a set of points, how do we know if they can be compressed like in the previous example?
  - We have to look into the correlation between the points
  - By finding the eigenvalues and eigenvectors of the covariance matrix, we find that the eigenvectors with the largest eigenvalues correspond to the dimensions that the strongest correlation in the dataset.
  - This is the principal component.

PCA-Theorem

Let x_1 x_2 … x_n be a set of n N x 1 vectors and let $\bar{x}$ be their average:

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iN} \end{bmatrix} \qquad \bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{i=n} \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iN} \end{bmatrix}$$

PCA-Theorem

Let X be the N x n matrix with columns
$x_1 - \bar{x}$, $x_2 - \bar{x}$,… $x_n - \bar{x}$:

$$X = \left[\ \mathbf{x}_1 - \bar{\mathbf{x}} \quad \mathbf{x}_2 - \bar{\mathbf{x}} \quad \cdots \quad \mathbf{x}_n - \bar{\mathbf{x}}\ \right]$$

Note: subtracting the mean is equivalent to translating the coordinate system to the location of the mean.

PCA-Theorem

Let Q = X X^T be the N x N matrix:

$$Q = XX^T = \left[\begin{array}{cccc} \mathbf{x}_1 - \bar{\mathbf{x}} & \mathbf{x}_2 - \bar{\mathbf{x}} & \cdots & \mathbf{x}_n - \bar{\mathbf{x}} \end{array} \right] \left[\begin{array}{c} (\mathbf{x}_1 - \bar{\mathbf{x}})^T \\ (\mathbf{x}_2 - \bar{\mathbf{x}})^T \\ \vdots \\ (\mathbf{x}_n - \bar{\mathbf{x}})^T \end{array} \right]$$

Generally:

1. Q is square
2. Q is symmetric
3. Q is the covariance matrix

PCA-Theorem

Each x_j can be written as: $x_j = \bar{x} + \sum_{i=1}^{n} g_{ji} e_i$

Where, e_i are the n eigenvectors of Q with non-zero eigenvalues.

Note:

1. The eigenvectors $e_1, e_2, \ldots, e_n$ span an eigenspace.
2. These are N x 1 orthogonal vectors (directions in N-Dimensional space.
3. The scalars g_{ji} are the coordinates of x_j in the space.

$$g_{ji} = (x_j - \bar{x}). e_i$$

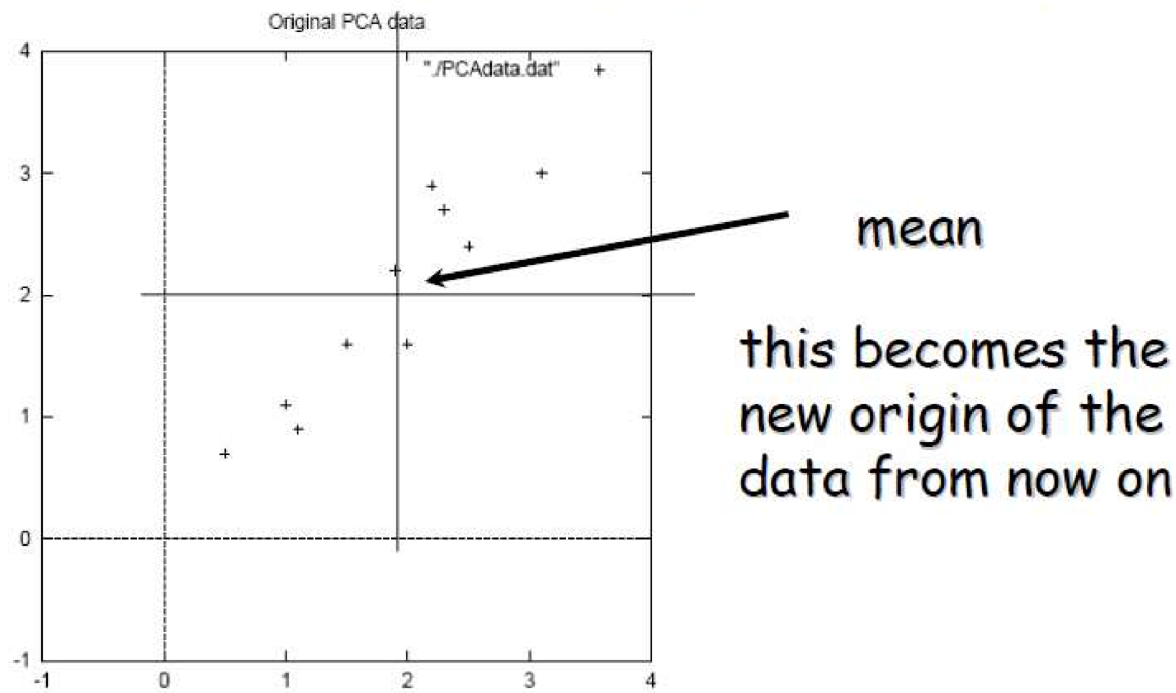
Using PCA to Compress Data

- Expressing x in terms of $e_1 \ldots e_n$ has not changed the size of the data.
- If the points are highly correlated many of the coordinates of x will be zero or close to zero.
- Sort the eigenvectors e_i according to their eigenvalue.

PCA Example-Step 1

- DATA:

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9



PCA Example-Step 2

Calculate the covariance matrix.

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

Since the non-diagonal elements in this covariance matrix are positive, we should expect that both x and y variable increase together.

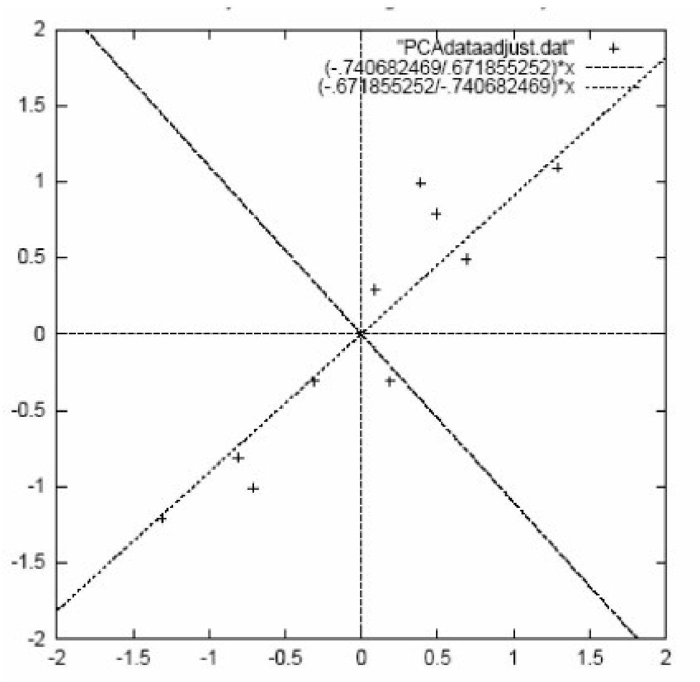
PCA Example-Step 3

Calculate the eigenvectors and eigenvalues of the covariance matrix.

$$\text{eigenvalues} = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$\text{eigenvectors} = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -735178656 \end{pmatrix}$$

PCA Example-Step 3



- Eigenvectors are plotted as diagonal dotted lines on the plot.
- They are perpendicular to each other.
- One of the eigenvectors goes through the middle of the points, like drawing a line of best fit.

PCA Example-Step 4

- Feature Vector

  $$FeatureVector = (eig_1\ eig_2\ eig_3\ \ldots\ eig_n)$$

  We can either form a feature vector with both of the eigenvectors:

  $$\begin{pmatrix} -.677873399 & -.735178656 \\ -.735178656 & .677873399 \end{pmatrix}$$

  or, we can choose to leave out the smaller, less significant component and only have a single column:

  $$\begin{pmatrix} -.677873399 \\ -.735178656 \end{pmatrix}$$

PCA Example-Step 5

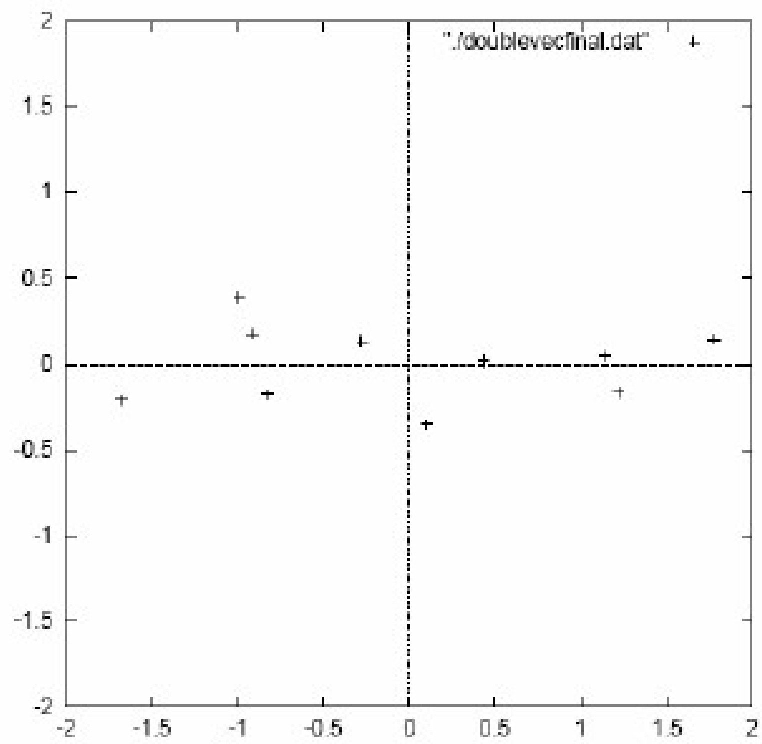
- Deriving new data coordinates

FinalData = RowFeatureVector x RowZeroMeanData

RowFeatureVector is the matrix with the eigenvectors in the columns *transposed* so that the eigenvectors are now in the rows, with the most significant eigenvector at the top

RowZeroMeanData is the mean-adjusted data *transposed*, ie. the data items are in each column, with each row holding a separate dimension.

PCA Example-Step 5



Thank You