

Data Mining :: Unit-2

(Dimensionality Reduction – PCA)

Er. Dinesh Baniya Kshatri
(Lecturer)

Department of Electronics and Computer Engineering
Institute of Engineering, Thapathali Campus

The Curse of Dimensionality

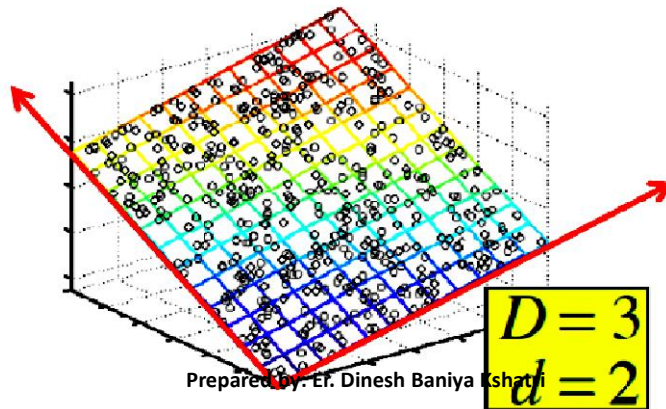
- **Real data usually have thousands, or millions of dimensions**
 - Web documents, where the dimensionality is the vocabulary of words
 - Facebook graph, where the dimensionality is the number of users
- **Huge number of dimensions causes problems:**
 - Data becomes very sparse, some algorithms become meaningless
 - E.g. Density based Clustering
 - The complexity of several algorithms depends on the dimensionality and they become infeasible
 - E.g. Nearest Neighbor Search

Prepared by: Er. Dinesh Baniya Kshatri

2

Dimensionality Reduction – 1

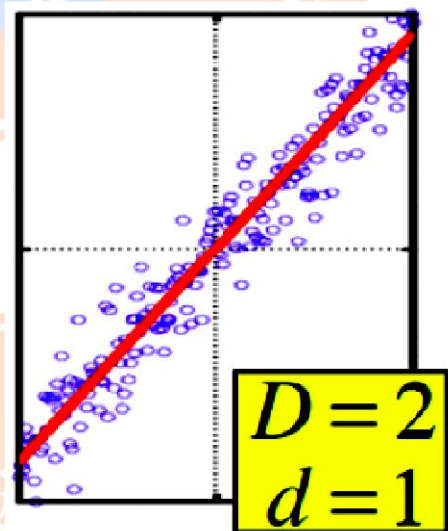
- Usually data can be described with fewer dimensions, without losing much of the meaning of the data
 - The data reside in a space of lower dimensionality



3

Dimensionality Reduction – 2

- **Goal of Dimensionality Reduction**
 - Discover the axis of data
- **Example:**
 - Rather than representing every point with two coordinates, each point is represented with one coordinate
 - Corresponding to the position of the point on the solid red line
 - This results in a bit of error as all the points do not exactly lie on the line



Prepared by: Er. Dinesh Baniya Kshatri

4

Dimensionality Reduction – 3

- **Find the “true dimension” of the data**
 - In reality things are never as clear and simple, but we can still reduce the dimension
- **Some of the data is useful signal and some data is noise**
 - Approximate the useful part with a lower dimensionality space
 - Dimensionality reduction does not just reduce the amount of data, it often brings out the useful part of the data

Prepared by: Er. Dinesh Baniya Kshatri

5

Example – 1

TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

- **In this data matrix the dimension is essentially three:**
 - There are three types of products and three types of users

Prepared by: Er. Dinesh Baniya Kshatri

6

Example – 2

Day	We	Th	Fr	Sa	Su
	7/10/96	7/11/96	7/12/96	7/13/96	7/14/96
Customer					
ABC Inc.	1	1	1	0	0
DEF Ltd.	2	2	2	0	0
GHI Inc.	1	1	1	0	0
KLM Co.	5	5	5	0	0
Smith	0	0	0	2	2
Johnson	0	0	0	3	3
Thompson	0	0	0	1	1

- The matrix is really “Two-dimensional”

- All rows can be reconstructed by scaling
[1 1 1 0 0] and
[0 0 0 1 1]

Prepared by: Er. Dinesh Baniya Kshatri

7

Why Reduce Dimensions?

- **Discover hidden correlations/topics**
 - In documents, words that occur commonly together
- **Remove redundant and noisy features**
 - In documents, not all words are useful
- **Interpretation and visualization**
- **Easier storage and processing of the data**

Prepared by: Er. Dinesh Baniya Kshatri

8

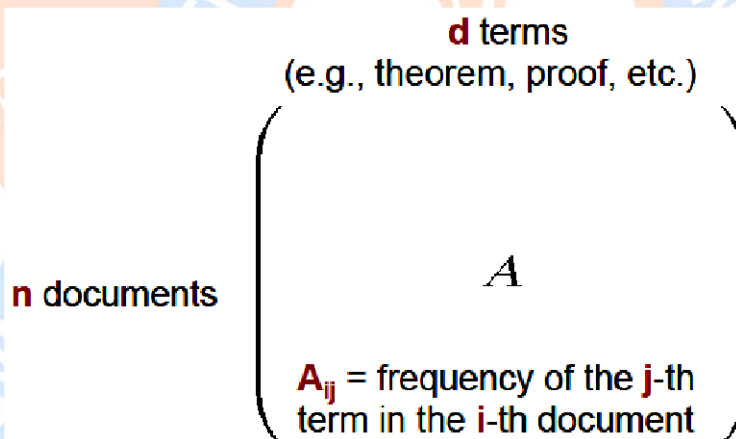
Data in Matrix Form

- Suppose there are (n) objects and (d) attributes describing the data
 - Represent the matrix as a (n x d) matrix (A)
- Goal is to produce a new (n x k) matrix (B) such that:
 - It preserves as much of the information in the original matrix (A) as possible
 - It reveals something about the structure of the data in (A)

Prepared by: Er. Dinesh Baniya Kshatri

9

Example: Document Types



Find subsets of terms that bring documents together

Prepared by: Er. Dinesh Baniya Kshatri

10

Example: Recommendation systems

$$\begin{matrix} & d \text{ movies} \\ n \text{ customers} & \left(\begin{matrix} A \\ A_{ij} = \text{rating of } j\text{-th} \\ & \text{product by the } i\text{-th} \\ & \text{customer} \end{matrix} \right) \end{matrix}$$

Find subsets of movies that capture the behavior of the customers

Prepared by: Er. Dinesh Baniya Kshatri

11

Background Material (Linear Algebra)

- **Assume vector (V) is a column vector**
 - Use $(V)^T$ for the transpose of vector (V)
 - Then $(V)^T$ becomes a row vector
- **Example of row vector multiplied by column vector**
 - Row and column vector sizes are: $(1 \times n)$ and $(n \times 1)$ respectively
 - The output is of size (1×1)

$$[1, 2, 3] \begin{bmatrix} 4 \\ 1 \\ 2 \end{bmatrix} = 12$$

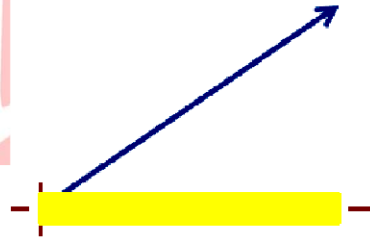
Prepared by: Er. Dinesh Baniya Kshatri

12

Background Material (Linear Algebra)

- **Dot product is represented as: $(U)^T(V)$**
 - Dot product is the projection of vector (V) on (U) (and vice versa)

$$u^T v = \|v\| \|u\| \cos(u, v)$$



- If $\|u\| = 1$ (unit vector) then $u^T v$ is the **projection length** of v on u
- If $\|u\| = \|v\| = 1$ then $u^T v$ is the **cosine similarity** of v and u

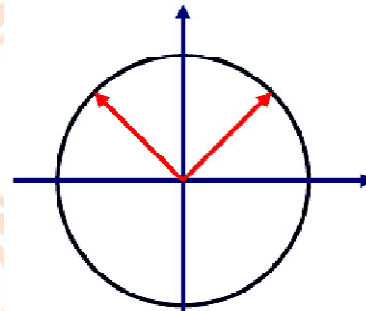
Prepared by: Er. Dinesh Baniya Kshatri

13

Background Material (Orthogonal and Orthonormal Vectors)

- **Orthogonal Vectors: Dot product between vectors evaluates to zero**

$$[-1, 2, 3] \begin{bmatrix} 4 \\ -1 \\ 2 \end{bmatrix} = 0$$



- **Orthonormal Vectors**
 - Two unit vectors that are orthogonal

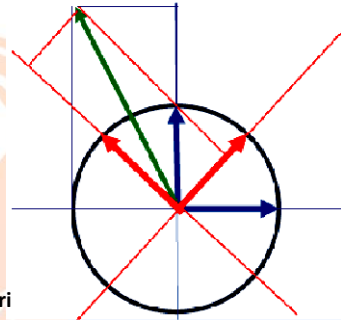
Prepared by: Er. Dinesh Baniya Kshatri

14

Background Material (Change of Basis)

- **By default a vector is expressed in the axis-aligned basis:**
 - For example, for vector $[-1, 2]$ we have:
 - $\begin{bmatrix} -1 \\ 2 \end{bmatrix} = -1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$
- **With a projection we can change the basis over which a vector is expressed:**

$$\begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{3\sqrt{2}}{2} \end{bmatrix}$$



Prepared by: Er. Dinesh Baniya Kshatri

15

Background Material (Rank of Matrix)

- **Rank of a matrix is the number of linearly independent row (or column) vectors**
 - These vectors define a basis for the row (or column) space of the matrix
 - All vectors in the row (or column) space are linear combinations of the basis vectors

Matrix $\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix}$ has rank $r=2$

- **Why?** The first two rows are linearly independent, so the rank is at least 2, but all three rows are linearly dependent (the first is equal to the sum of the second and third) so the rank must be less than 3.

Prepared by: Er. Dinesh Baniya Kshatri

16

Background Material (Eigenvalue Problem)

- The eigenvalue problem is any problem having the following form:

$$\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$$

\mathbf{A} : $m \times m$ matrix

\mathbf{v} : $m \times 1$ non-zero vector

λ : scalar

- Any value of (λ) for which this equation has a solution is called Eigenvalue of (A) and the vector (V) corresponding to this value is called the Eigenvector of (A)

Prepared by: Er. Dinesh Baniya Kshatri

17

Background Material (Eigenvectors)

- A square symmetric matrix having rank (r) has (r) number of orthonormal Eigenvectors
- Eigenvectors define an orthonormal basis for the column space of the matrix

Prepared by: Er. Dinesh Baniya Kshatri

18

Background Material (Eigenvectors – Example)

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4 \times \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

$\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$

- (3,2) is an Eigenvector of the square matrix (A) and (4) is an Eigenvalue of (A)

Prepared by: Er. Dinesh Baniya Kshatri

19

Background Material (Calculating Eigenvectors)

- Simple matrix algebra shows that:

$$\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$$

$$\Leftrightarrow \mathbf{A} \cdot \mathbf{v} - \lambda \cdot \mathbf{I} \cdot \mathbf{v} = \mathbf{0}$$

$$\Leftrightarrow (\mathbf{A} - \lambda \cdot \mathbf{I}) \cdot \mathbf{v} = \mathbf{0}$$

- Finding the roots of $|\mathbf{A} - \lambda \cdot \mathbf{I}|$ will give the eigenvalues and for each of these eigenvalues there will be an eigenvector

Prepared by: Er. Dinesh Baniya Kshatri

20

Background Material

(Calculating Eigenvectors – Example)

- Let

$$A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$$

- Then:

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

- And setting the determinant to 0, we obtain 2 eigenvalues:

$$\lambda_1 = -1 \text{ and } \lambda_2 = -2$$

Prepared by: Er. Dinesh Baniya Kshatri

21

Background Material

(Calculating Eigenvectors – Example)

- For λ_1 the eigenvector is:

$$(A - \lambda_1 I)v_1 = 0$$

$$\begin{bmatrix} 1 & 1 \\ -2 & -2 \end{bmatrix} \begin{bmatrix} v_{1:1} \\ v_{1:2} \end{bmatrix} = 0$$

$$v_{1:1} + v_{1:2} = 0 \quad \text{and} \quad -2v_{1:1} - 2v_{1:2} = 0$$

$$v_{1:1} = -v_{1:2}$$

- Therefore the first eigenvector is any column vector in which the two elements have equal magnitude and opposite sign.

Prepared by: Er. Dinesh Baniya Kshatri

22

Background Material

(Calculating Eigenvectors – Example)

- Therefore eigenvector v_1 is

$$v_1 = k_1 \begin{bmatrix} +1 \\ -1 \end{bmatrix}$$

where k_1 is some constant.

- Similarly we find that eigenvector v_2

$$v_2 = k_2 \begin{bmatrix} +1 \\ -2 \end{bmatrix}$$

where k_2 is some constant.

Prepared by: Er. Dinesh Baniya Kshatri

23

Background Material

(Properties of Eigenvectors)

- Eigenvectors can only be found for square matrices and not every square matrix has eigenvectors.
- Given an $m \times m$ matrix (with eigenvectors), we can find n eigenvectors.
- All eigenvectors of a symmetric* matrix are perpendicular to each other, no matter how many dimensions we have.
- In practice eigenvectors are normalized to have unit length.

Prepared by: Er. Dinesh Baniya Kshatri

24

Example Problem

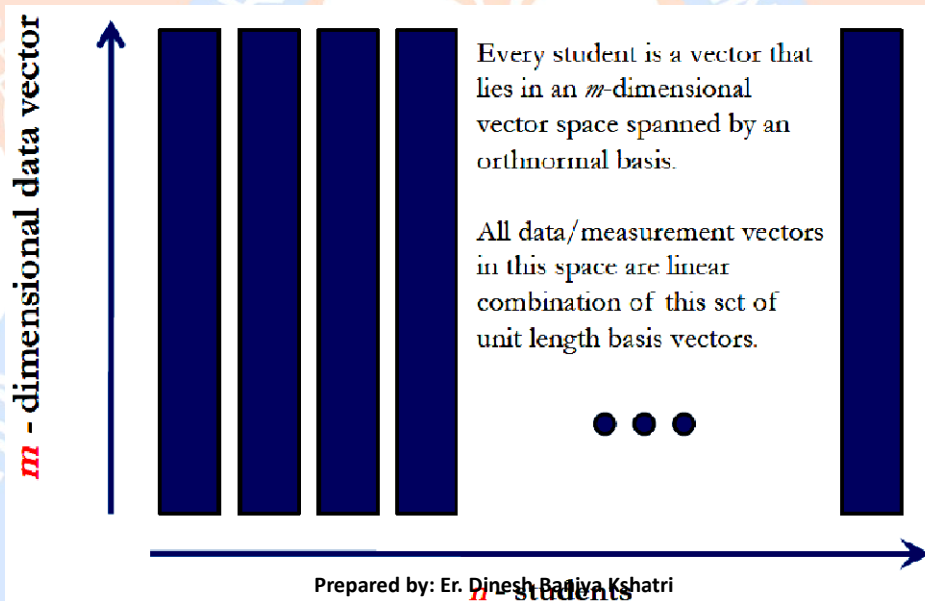
- We collected m parameters about 100 students:
 - Height
 - Weight
 - Hair color
 - Average grade
 - ...
- We want to find the most important parameters that best describe a student.



Prepared by: Er. Dinesh Baniya Kshatri

25

Example Problem



Prepared by: Er. Dinesh Baniya Kshatri

26

त्रिभुवन विश्वविद्यालय Example Problem – 3

(Which parameters can be ignored?)

- **Constant** parameter (number of heads)
 - 1,1,...,1.
- Constant parameter with some **noise** - (thickness of hair)
 - 0.003, 0.005, 0.002, ..., 0.0008 → low variance
- Parameter that is **linearly dependent** on other parameters (head size and height)
 - $Z = aX + bY$

Prepared by: Er. Dinesh Baniya Kshatri

27

त्रिभुवन विश्वविद्यालय Example Problem – 4

(Which parameters do we want to keep?)

- Parameter that doesn't depend on others (e.g. eye color), i.e. uncorrelated → low covariance.
- Parameter that changes a lot (grades)
 - The opposite of noise
 - High variance

Prepared by: Er. Dinesh Baniya Kshatri

28

Questions

- How can the “most important” features be described using math ?
 - Use variance
- How to represent the data so that the “most important” features can be extracted easily?
 - Perform a change of basis

Prepared by: Er. Dinesh Baniya Kshatri

29

Change of Basis – 1

- Let \mathbf{X} and \mathbf{Y} be $m \times n$ matrices related by a linear transformation \mathbf{P} .
- \mathbf{X} is the original recorded data set and \mathbf{Y} is a re-representation of that data set.

$$\mathbf{PX} = \mathbf{Y}$$

- What does this mean?
 - \mathbf{P} is a matrix that transforms \mathbf{X} into \mathbf{Y} .
 - Geometrically, \mathbf{P} is a rotation and a stretch (scaling) which again transforms \mathbf{X} into \mathbf{Y} .
 - The rows of \mathbf{P} , $\{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_m\}$ are a set of new basis vectors for expressing the columns of \mathbf{X} .

Prepared by: Er. Dinesh Baniya Kshatri

30

Change of Basis – 2

- The explicit dot product of PX is: $PX = \begin{bmatrix} p_1 \\ \vdots \\ p_m \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}$
- Note the form of each column of Y :
 - Each coefficient of (Y_i) is a dot product of (X_i) with the corresponding row in (P)

$$Y = \begin{bmatrix} p_1 \cdot x_1 & \cdots & p_1 \cdot x_n \\ \vdots & \ddots & \vdots \\ p_m \cdot x_1 & \cdots & p_m \cdot x_n \end{bmatrix}$$

In other words, the j^{th} coefficient of y_i is a projection onto the j^{th} row of P .

Therefore, the rows of P are a new set of basis vectors for representing the columns of X .

Prepared by: Er. Dinesh Baniya Kshatri

31

Change of Basis – 3

- Changing the basis doesn't change the data – only its representation.
- Changing the basis is actually projecting the data vectors on the basis vectors.
- Geometrically, P is a rotation and a stretch of X .
 - If P basis is orthonormal (length = 1) then the transformation P is only a rotation

Prepared by: Er. Dinesh Baniya Kshatri

32

Principle Component Analysis (PCA)

- Goal: reduce the dimensionality while preserving the “information in the data”.
- In the new space we want to:
 - **Maximize** the amount of **information**
 - **Minimize redundancy** – remove the redundant dimensions
 - **Minimize** the **noise** in the data.

Prepared by: Er. Dinesh Baniya Kshatri

33

Variability – 1

- **Information in data implies variability in data**

- Measure variability using the covariance matrix

- Variance for variable (X) is:

$$\sigma_X^2 = \frac{1}{N-1} \sum_i (x_i - \mu_X)(x_i - \mu_X) = \frac{1}{N-1} (x - \mu_X)^T (x - \mu_X)$$

- Covariance of variables (X) and (Y)

$$\sigma_{XY}^2 = \frac{1}{N-1} \sum_i (x_i - \mu_X)(y_i - \mu_Y) = \frac{1}{N-1} (x - \mu_X)^T (y - \mu_Y)$$

Prepared by: Er. Dinesh Baniya Kshatri

34

Variability – 2

- **High variance implies high information in dimension (attribute)**
 - Need to maximize the signal-to-noise ratio: $\frac{\sigma_{signal}^2}{\sigma_{noise}^2}$
- **High covariance means high correlation between attributes and thus high redundancy**
 - Ideally want covariance to be zero

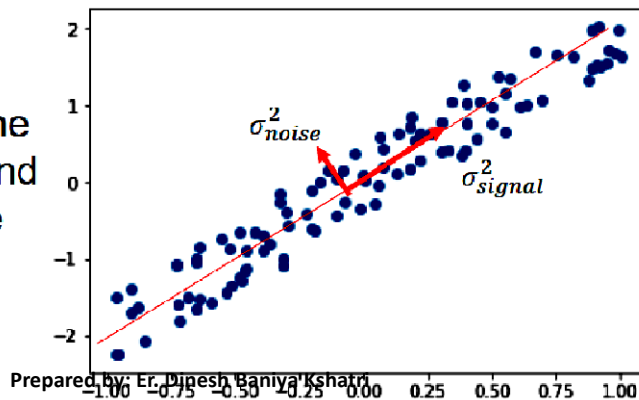
Prepared by: Er. Dinesh Baniya Kshatri

35

Variability – Example

- In the data below the data are essentially one-dimensional, but what is the axis we should use?
 - The direction in which the variance is **maximized**.

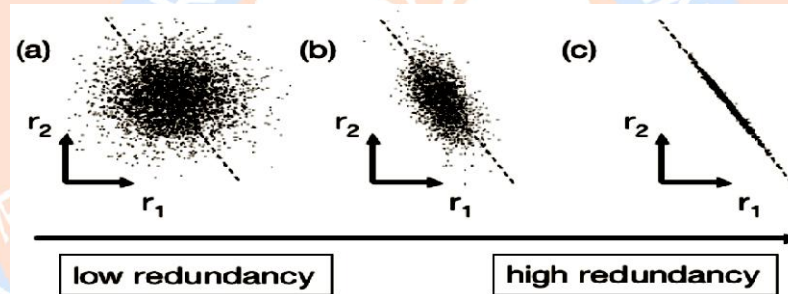
The variance along the direction **orthogonal** to the main direction is small and captures the noise in the data



Prepared by: Er. Dinesh Baniya Kshatri

36

Redundancy



- (a) depicts two recordings (r_1) and (r_2) that are uncorrelated
- (c) depicts two recordings (r_1) and (r_2) that are strongly related
 - One recording can be expressed in terms of the other

Prepared by: Er. Dinesh Baniya Kshatri

37

Covariance Matrix – 1

- **Consider the following:**
 - Each row of matrix (X) corresponds to all measurements of a particular measurement type
 - Each column of matrix (X) corresponds to set of measurements corresponding to a particular time instant
- **The covariance is calculated as follows:**

$$S_X \equiv \frac{1}{n-1} X X^T \quad \text{where} \quad X = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$$

Prepared by: Er. Dinesh Baniya Kshatri

38

Covariance Matrix – 2

- S_X is a square symmetric ($m \times m$) matrix
- The diagonal terms of S_X are the variance of particular measurement types
- The off-diagonal terms of S_X are the covariance between measurement types

Prepared by: Er. Dinesh Baniya Kshatri

39

Covariance Matrix – 3

Suppose, we have the option of manipulating S_X . We will suggestively define our manipulated covariance matrix S_Y .

What features do we want to optimize in S_Y ?

Prepared by: Er. Dinesh Baniya Kshatri

40

Features to be Optimized (Diagonalize the Covariance Matrix)

Our goals are to find the covariance matrix that:

1. Minimizes redundancy, measured by covariance. (off-diagonal), i.e. we would like each variable to co-vary as little as possible with other variables.
2. Maximizes the signal, measured by variance. (the diagonal)

Prepared by: Er. Dinesh Baniya Kshatri

41

Diagonalize the Covariance Matrix (PCA Assumptions – 1)

- PCA assumes that all basis vectors $\{\mathbf{p}_1, \dots, \mathbf{p}_m\}$ are orthonormal (i.e. $\mathbf{p}_i \cdot \mathbf{p}_j = \delta_{ij}$).
- Hence, in the language of linear algebra, PCA assumes \mathbf{P} is an orthonormal matrix.
- Secondly, PCA assumes the directions with the largest variances are the most “important” or in other words, most principal.

Prepared by: Er. Dinesh Baniya Kshatri

42

Diagonalize the Covariance Matrix (PCA Assumptions – 2)

- By the variance assumption PCA first selects a normalized direction in m -dimensional space along which the variance in \mathbf{X} is maximized - it saves this as \mathbf{p}_1 .
- Again it finds another direction along which variance is maximized, however, because of the orthonormality condition, it restricts its search to all directions perpendicular to all previous selected directions.
- This could continue until m directions are selected. The resulting ordered set of \mathbf{p} 's are the *principal components*.
- The variances associated with each direction \mathbf{p}_i quantify how principal each direction is. We could thus rank-order each basis vector \mathbf{p}_i according to the corresponding variances.

Prepared by: Er. Dinesh Baniya Kshatri

43

Eigenvectors of Covariance Matrix – 1

- The data set is given by matrix \mathbf{X}
- The goal is summarized as follows:
 - Find some orthonormal matrix \mathbf{P} where $\mathbf{Y} = \mathbf{PX}$ such that $\mathbf{S}_Y \equiv \frac{1}{n-1}\mathbf{Y}\mathbf{Y}^T$ is diagonalized. The rows of \mathbf{P} are the *principal components* of \mathbf{X} .

Prepared by: Er. Dinesh Baniya Kshatri

44

Eigenvectors of Covariance Matrix – 2

- Rewriting (S_Y) in terms of the variable of choice (P)

$$\begin{aligned} S_Y &= \frac{1}{n-1} Y Y^T \\ &= \frac{1}{n-1} (P X) (P X)^T \\ &= \frac{1}{n-1} P X X^T P^T \\ &= \frac{1}{n-1} P (X X^T) P^T \\ S_Y &= \frac{1}{n-1} P A P^T \end{aligned}$$

Prepared by: Er. Dinesh Baniya Kshatri

45

Eigenvectors of Covariance Matrix – 3

- Note that we defined a new matrix $A = X X^T$, where A is *symmetric*
- Our roadmap is to recognize that a symmetric matrix (A) is diagonalized by an orthogonal matrix of its eigenvectors
- A symmetric matrix A can be written as $V D V^T$ where D is a diagonal matrix and V is a matrix of eigenvectors of A arranged as columns.
- The matrix A has $r \leq m$ orthonormal eigenvectors where r is the rank of the matrix.

Prepared by: Er. Dinesh Baniya Kshatri

46

Eigenvectors of Covariance Matrix – 4

- **Trick:** Select matrix (P) such that each row (P_i) is an Eigenvector of (XX^T)
- By this selection, $P = V^T$. Hence $A = P^T D P$.
- With this relation and the fact that $P^{-1} = P^T$ since the inverse of orthonormal matrix is its transpose, we can finish evaluating S_Y as follows;

$$\begin{aligned} S_Y &= \frac{1}{n-1} P A P^T \\ &= \frac{1}{n-1} P (P^T D P) P^T \\ &= \frac{1}{n-1} (P P^T) D (P P^T) \\ &= \frac{1}{n-1} (P P^{-1}) D (P P^{-1}) \\ S_Y &= \frac{1}{n-1} D \end{aligned}$$

It is evident that the choice of P diagonalizes S_Y . This was the goal for PCA.

Prepared by: Er. Dinesh Baniya Kshatri

47

PCA Process – STEP 1 :: [1]

- **Subtract the mean from each of the dimensions**
 - This produces a data set whose mean is zero
 - Subtracting the mean makes variance and covariance calculation easier by simplifying their equations
 - The variance and covariance values are not affected by the mean
- **Suppose there are two measurement types X_1 and X_2 , so $m = 2$, and ten samples each, hence $n = 10$**

Prepared by: Er. Dinesh Baniya Kshatri

48

PCA Process – STEP 1 :: [2]

X_1	X_2		X'_1	X'_2		
2.5	2.4		0.69	0.49		
0.5	0.7		-1.31	-1.21		
2.2	2.9		0.39	0.99		
1.9	2.2		0.09	0.29		
3.1	3.0	\Rightarrow	$\overline{X_1} = 1.81$	\Rightarrow	1.29	1.09
2.3	2.7		$\overline{X_2} = 1.91$		0.49	0.79
2.0	1.6				0.19	-0.31
1.0	1.1				-0.81	-0.81
1.5	1.6				-0.31	-0.31
1.2	0.9				-0.71	-1.01

Prepared by: Er. Dinesh Baniva Kshatri

Prepared by: Er. Dinesh Baniya Kshatri

49

PCA Process – STEP 2

- Calculate the covariance matrix

$$S_X = \begin{bmatrix} 0.616555556 & 0.615444444 \\ 0.615444444 & 0.716555556 \end{bmatrix}$$

- Since the non-diagonal elements in this covariance matrix are positive, we should expect that both the X_1 and X_2 variables increase together.
- Since it is symmetric, we expect the eigenvectors to be orthogonal.

Prepared by: Er. Dinesh Baniya Kshatri

50

PCA Process – STEP 3 (1)

- Calculate the eigen vectors **V** and eigen values **D** of the covariance matrix

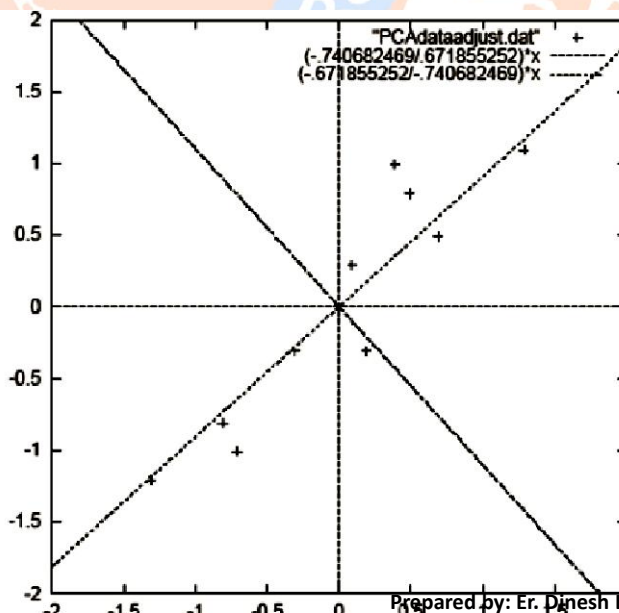
$$D = \begin{bmatrix} 0.0490833989 \\ 1.28402771 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}$$

$$V = \begin{bmatrix} -0.735178656 & -0.677873399 \\ 0.677873399 & -0.735178656 \end{bmatrix}$$

Prepared by: Er. Dinesh Baniya Kshatri

51

PCA Process – STEP 3 (2)



Eigenvectors are plotted as diagonal dotted lines on the plot. (note: they are perpendicular to each other).

One of the eigenvectors goes through the middle of the points, like drawing a line of best fit.

The second eigenvector gives us the other, less important, pattern in the data, that all the points follow the main line, but are off to the side of the main line by some amount.

Prepared by: Er. Dinesh Baniya Kshatri

52

PCA Process – STEP 4 (1)

- **The Eigenvector with the highest Eigenvalue is the Principal Component of the dataset**
 - In the previous example, the Eigenvector with the largest Eigenvalue is the one that points down the middle of the data
- **Once Eigenvectors are found from the covariance matrix, the next step is to order them by Eigenvalue, highest to lowest**
 - This gives the components in order of significance

Prepared by: Er. Dinesh Baniya Kshatri

53

PCA Process – STEP 4 (2)

- **It is possible to ignore the components of lesser significance:**
 - Causes some information to be lost, however if the Eigenvalues are very small, the loss is negligible
 - For (m) dimensions in the data set, calculate (m) Eigenvectors and Eigenvalues
 - Choose only the first (r) Eigenvectors
 - The final data set will have only (r) dimensions

Prepared by: Er. Dinesh Baniya Kshatri

54

PCA Process – STEP 4 (2)

- When the λ_i 's are sorted in descending order, the proportion of variance explained by the r principal components is:

$$\frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^m \lambda_i} = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_r}{\lambda_1 + \lambda_2 + \dots + \lambda_p + \dots + \lambda_m}$$

- If the dimensions are highly correlated, there will be a small number of eigenvectors with large eigenvalues and r will be much smaller than m .
- If the dimensions are not correlated, r will be as large as m and PCA does not help.

Prepared by: Er. Dinesh Baniya Kshatri

55

PCA Process – STEP 4 (2)

FeatureVector = $(\lambda_1 \lambda_2 \lambda_3 \dots \lambda_r)$

(take the eigenvectors to keep from the ordered list of eigenvectors, and form a matrix with these eigenvectors in the columns)

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

$$\begin{bmatrix} -0.677873399 & -0.735178656 \\ -0.735178656 & 0.677873399 \end{bmatrix}$$

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

$$\begin{bmatrix} -0.677873399 \\ -0.735178656 \end{bmatrix}$$

Prepared by: Er. Dinesh Baniya Kshatri

56

PCA Process – STEP 5 (1)

- Derive the new data

$$\text{FinalData} = \text{RowFeatureVector} \times \text{RowZeroMeanData}$$

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*, i.e., the data items are in each column, with each row holding a separate dimension.

Prepared by: Er. Dinesh Baniya Kshatri

57

PCA Process – STEP 5 (2)

- FinalData** is the final data set, with data items in columns, and dimensions along rows.
- What does this give us?

The original data *solely in terms of the vectors we chose.*

- We have changed our data from being in terms of the axes X_1 and X_2 , to now be in terms of our 2 eigenvectors.

Prepared by: Er. Dinesh Baniya Kshatri

58

PCA Process – STEP 5 (3)

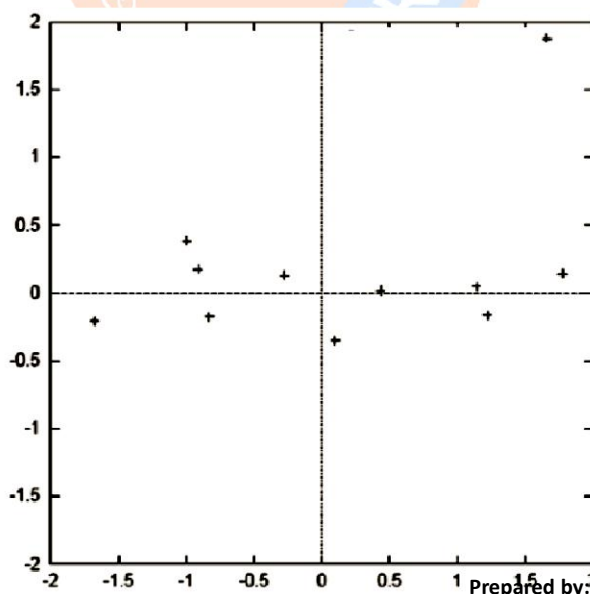
FinalData (transpose: dimensions along columns)

Y_1	Y_2
-0.827870186	-0.175115307
1.77758033	0.142857227
-0.992197494	0.384374989
-0.274210416	0.130417207
-1.67580142	-0.209498461
-0.912949103	0.175282444
0.0991094375	-0.349824698
1.14457216	0.0464172582
0.438046137	0.0177646297
1.22382956	0.16295287

Prepared by: Er. Dinesh Baniya Kshatri

59

PCA Process – STEP 5 (4)



- Plot of the new data points by applying the PCA analysis using both Eigenvectors

Prepared by: Er. Dinesh Baniya Kshatri

60

PCA Application (Face Recognition)

The objectives of using PCA in face recognition are:

1. Data Reduction
2. Feature Selection



Prepared by: Er. Dinesh Baniya Kshatri

61

Converting an Image to a vector – 1

In image recognition an input image with n pixels can be treated as a point in an n -dimensional space called the image space. The coordinates of this point represent the values of each pixel of the image and form a vector

$$\mathbf{p}_x = (i_1, i_2, i_3, \dots, i_n)$$


obtained by concatenating the rows (or columns) of the image matrix.

Prepared by: Er. Dinesh Baniya Kshatri

62

Converting an Image to a vector – 2

Given a greyscale image of, for example, 128 by 128 pixels:


$$= \begin{bmatrix} 150 & 152 & \cdot & 151 \\ 131 & 133 & \cdot & 72 \\ \cdot & \cdot & \cdot & \cdot \\ 144 & 171 & \cdot & 67 \end{bmatrix} \quad 128 \times 128$$

We concatenate each row to make a 16384 vector

$$[150, 152, \dots, 151, 131, 133, \dots, 72, \dots, 144, 171, \dots, 67]_{16K}$$

Prepared by: Er. Dinesh Baniya Kshatri

63

Redundant Information

Face images are highly redundant:

- All the background pixels are the same
- Each subject has the same facial features



Prepared by: Er. Dinesh Baniya Kshatri

64

Dimension Reduction

In the data space each pixel is a variable, so the dimension of the space is very high (min 16K).

Dimension reduction is achieved by PCA. Let an $N \times n$ data matrix D be composed of N input face images with n pixels. Each row is one image of our data set.

$$D = \begin{bmatrix} 150 & 152 & \dots & 254 & 255 & \dots & 252 \\ 131 & 133 & \dots & 221 & 223 & \dots & 241 \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ 144 & 171 & \dots & 244 & 245 & \dots & 223 \end{bmatrix} \quad N \times n$$

Prepared by: Er. Dinesh Baniya Kshatri

65

Practical Face recognition

In face recognition the eigenvectors are often called eigenfaces.

As an example we will find the eigenface basis of a set of fourteen faces images of resolution 384×256 . This initial data set D is sometimes called the training data set.



Prepared by: Er. Dinesh Baniya Kshatri

66

What do the Eigenfaces look like?

Mean:



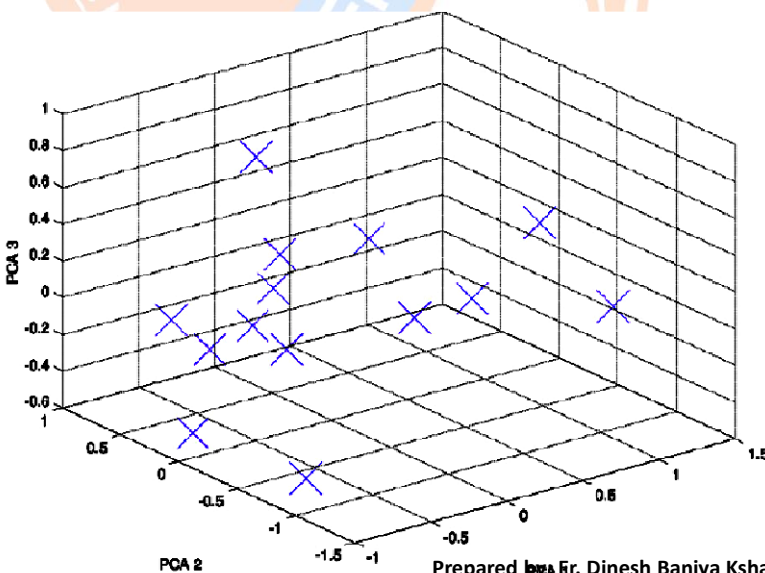
The four eigenfaces with the largest eigenvalues:



Prepared by: Er. Dinesh Baniya Kshatri

67

How different are the face images?



- This is a plot of the fourteen face images on the first three principal components:

Prepared by: Er. Dinesh Baniya Kshatri

68

Reconstructing the Face Images: (Example – 1)

Original:



3 PCs



5



8



11



13



Prepared by: Er. Dinesh Baniya Kshatri

69

Reconstructing the Face Images: (Example – 2)

Original:



3 PCs



5



8



11



13



Prepared by: Er. Dinesh Baniya Kshatri

70

How many principal components are needed?

There is no definitive answer to this question. We want to minimise m (the number retained) but maximise the accuracy of the data representation.

One approach is to see how much of the variance each principal component accounts for. We can express the percentage of the total variance accounted for by the i^{th} eigenvector as:

$$r_i = 100 \times \frac{\lambda_i}{\sum_{j=1}^n \lambda_j}$$

One heuristic method would be to discard components where r_i falls below a threshold, say 1%.

Prepared by: Er. Dinesh Baniya Kshatri

71