

Principal Component Analysis

Dr. Avinash Kumar Singh

Founder, Robotics and Artificial Intelligence Training Academy

Senior Researcher Montpellier University France



Discussion Points

- What is a feature ?
- Principal Component Analysis
 - Objective
 - Steps
 - Use Case – Face Recognition

What is a Feature

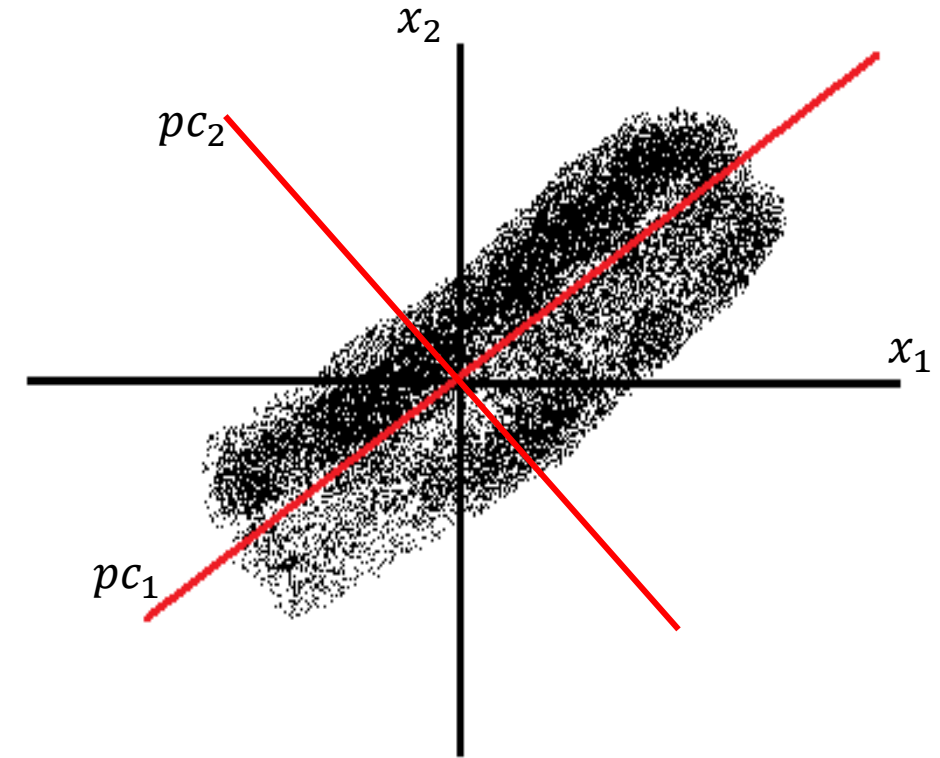
- Feature is a unit to represent the information.

Objective

- Used to reduce the dimension of multivariate data, while preserving as much of the relevant information as possible.
 - If we have samples represented in the m dimensional space, $\{x_1, x_2, x_3, x_4, x_5, x_6, \dots, x_M\}$, our intension is to find out k features, such that $k \leq m$ that preserve most of the variance present in the data.
- Good tool for feature extraction.

Summary

- We should select those directions where variance is maximum.
- Each direction will give one principal component.
- First principal component has maximum variance, Second has 2nd maximum variance, and orthogonal to first, Third has 3rd maximum variance and orthogonal to first and second one, like this we will select k component.



Derivation

- In many physical, biological, and statistical convention it is desirable to represent a system of points with the help of a line or plane.
- Therefore, we can represent the equation of line with the linear combination of these points(variables).

$$y_k = \sum_{i=1}^m a_{ki} x_i \quad \text{Where } K=1,2,3, \dots, p$$

y_1	a_{11}	a_{12}	a_{13}	.	a_{1m}	x_1
y_2	a_{21}	a_{22}	a_{23}	.	a_{2m}	x_2
y_3	a_{31}	a_{32}	a_{33}	.	a_{3m}	x_3
.
y_p	a_{p1}	a_{p2}	a_{p3}	.	a_{pm}	x_m

Steps used in PCA

- Mean of the data (along each feature):

Let us assume, we have samples having m features. We have stored all these in a matrix called training samples T having N rows and M columns (where N represents the population and M represents the Features). So in that regard we have data like.

$a1_1$	$a1_2$	$a1_3$	$.$	$a1_m$
$a2_1$	$a2_2$	$a2_3$	$.$	$a2_m$
$a3_1$	$a3_2$	$a3_3$	$.$	$a3_m$
$.$	$.$	$.$	$.$	$.$
ap_1	ap_2	ap_3	$.$	ap_m
μ_1	μ_2	μ_3	μ_{m-1}	μ_m

$$(\mu)_{1 \times m} = \frac{\sum_{i=1}^m \sum_{j=1}^p T(i, j)}{p}$$

Here

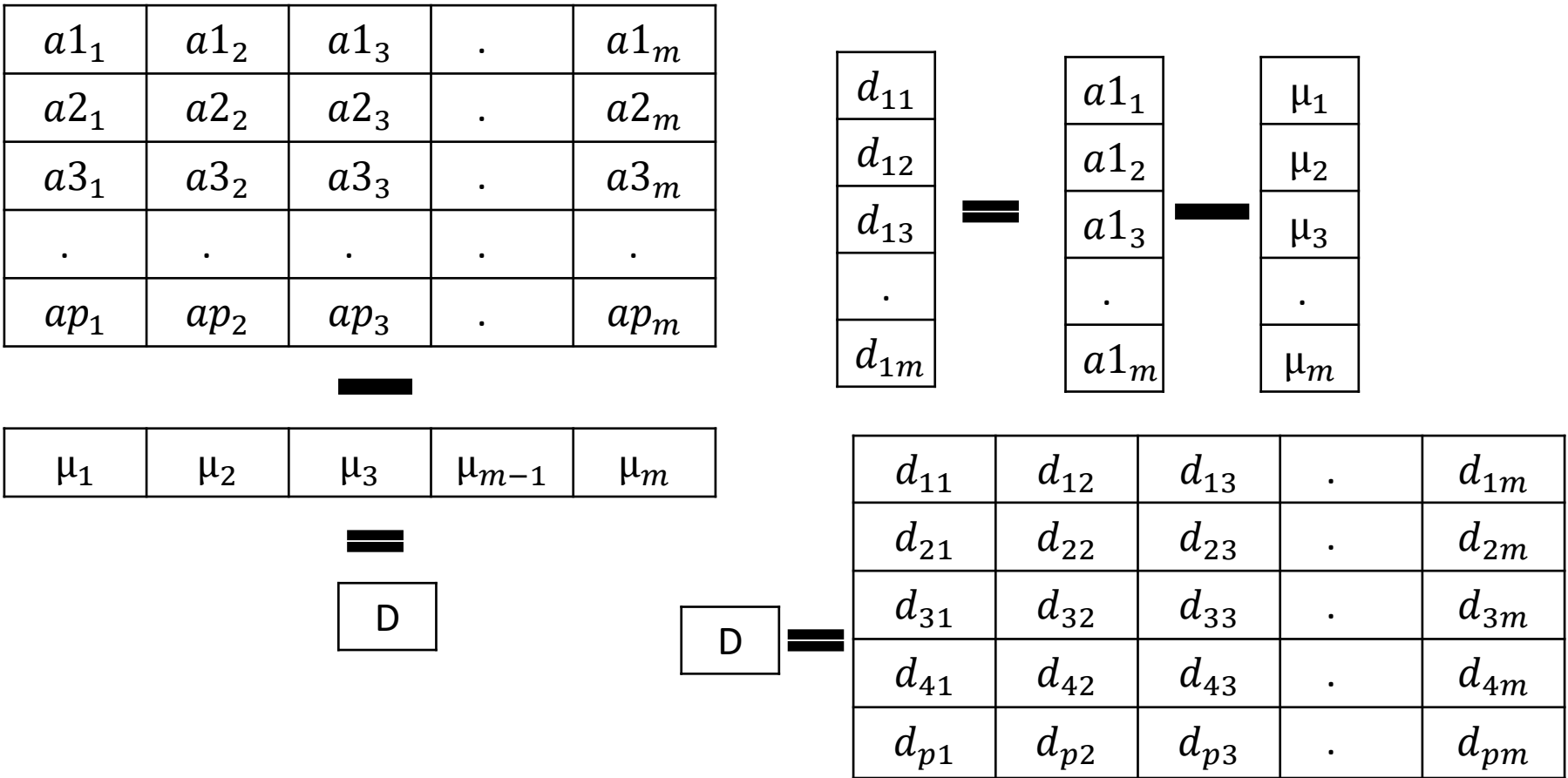
: i represents features

: j represents observation.

: $T(i, j)$ represents the value of i^{th} feature at j^{th} observation

Step-1: Mean Aligned Data

- Subtract the from each sample to do the mean zero of the data.



Step-2: Calculate Variance, Covariance

- Variance (σ) denotes how the data is distributed along its mean, while Covariance (Σ) shows how the data is relate to other.

The diagram illustrates the calculation of variance through mean alignment and matrix multiplication.

Mean Alignment: Three vertical columns are shown, separated by equals and minus signs. The first column contains data points $d_{11}, d_{21}, d_{31}, d_{41}, d_{p1}$. The second column contains alignment values $a_{11}, a_{21}, a_{31}, \cdot, a_{p1}$. The third column contains the mean μ_1 repeated for each row. A red bracket under the first column is labeled "Mean Alignment". A green arrow points from this section to the variance calculation.

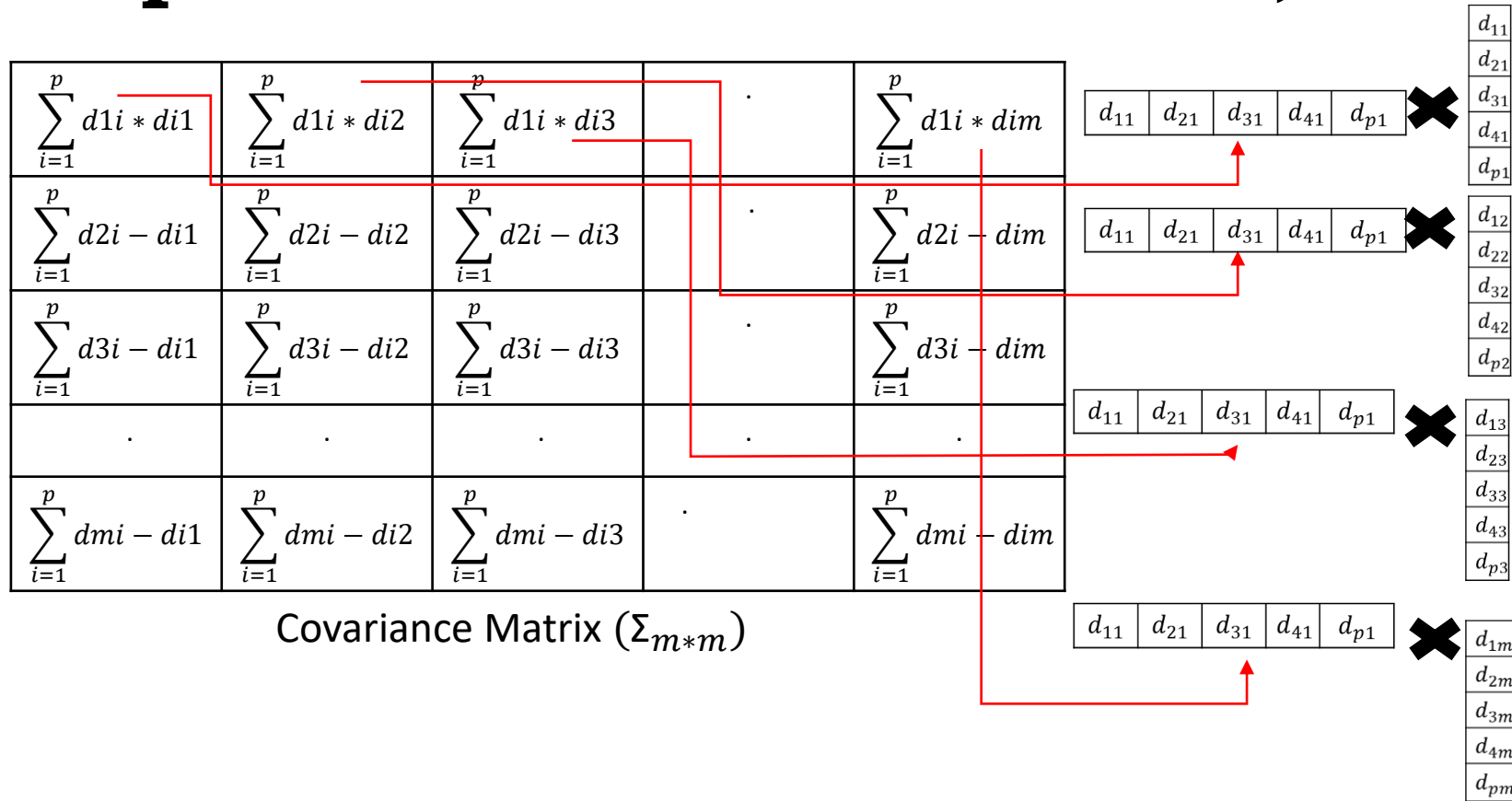
Variance Calculation: The formula for variance is given as
$$\text{Variance } (\sigma) = \sum_{i=1}^p d_{1i} * d_{i1}$$

This is expanded to show the sum of products:
$$d_{11} * d_{11} + d_{12} * d_{21} + d_{13} * d_{31} + d_{14} * d_{41} + d_{1p} * d_{p1}$$

This sum is then represented as a matrix multiplication:
$$\begin{bmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{1p} \end{bmatrix} \times \begin{bmatrix} d_{11} \\ d_{21} \\ d_{31} \\ d_{41} \\ d_{p1} \end{bmatrix}$$

The first matrix is labeled D^T and the second matrix is labeled D .

Step-2: Calculate Variance, Covariance

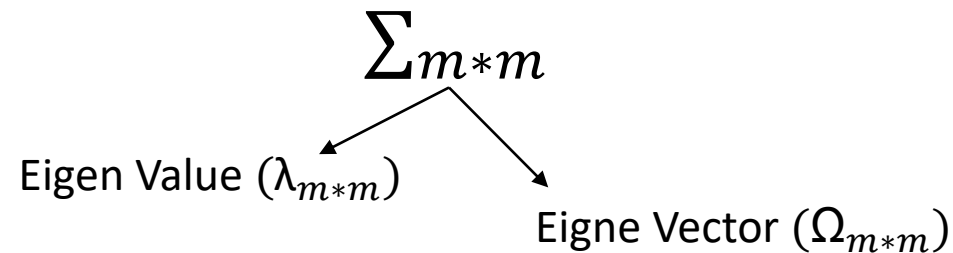


d_{11}	d_{12}	d_{13}	.	d_{1m}
d_{21}	d_{22}	d_{23}	.	d_{2m}
d_{31}	d_{32}	d_{33}	.	d_{3m}
d_{41}	d_{42}	d_{43}	.	d_{4m}
d_{p1}	d_{p2}	d_{p3}	.	d_{pm}

Lookup table

Step-3: Eigen Vector and Eigen Value Decomposition

- From the above equation we can say that co-variance can be calculated as $(\bar{Y} \times \bar{Y}^T)$, resultant $(m * m)$ dimension. Let's say the covariance matrix is Σ
- Principal components are the Eigen values and Eigen vectors, those are computed on the basis of co-variance matrix, calculated in the previous step.



Example of Eigen value and vector

Eigen Vector →		1	2	3	4
	1	0.5797	0.6532	0.4545	0.1749
	2	-0.8049	0.3701	0.4154	0.2063
	3	0.1222	-0.6605	0.6830	0.2867
	4	0.0322	-0.0014	-0.3929	0.9190

Eigen Values →		1	2	3	4
	1	0.5264	0	0	0
	2	0	2.0388	0	0
	3	0	0	24.7450	0
	4	0	0	0	40.2399

Diagonal elements

Step:4 Eigen Vector Selection

- Sort the eigen values in the descending order and the eigen vector as well.
- We should select those λ 's which have maximum values, because this shows the variance. Hence for selecting the best principal components (k), we should define a threshold above which we can select all principal components such that ($\lambda \geq TH$).
- What should be the optimal value for selecting k ?

$$\text{Variance} = \sum \lambda_i$$

- Combination of these eigenvectors are known as feature vectors, say “W”

Step 5: Projecting the data to principal directions

$$W = \begin{bmatrix} w1_1 & w1_2 & w1_3 & & w1_m \\ w2_1 & w2_2 & w2_3 & & w2_m \\ w3_1 & w3_2 & w3_3 & & w3_m \\ . & . & . & & . \\ wm_1 & wm_2 & wm_3 & & wm_m \end{bmatrix} \quad m \times m$$

$$W_k = \begin{bmatrix} w1_1 & w1_2 & w1_3 \\ w2_1 & w2_2 & w2_3 \\ w3_1 & w3_2 & w3_3 \\ . & . & . \\ wm_1 & wm_2 & wm_3 \end{bmatrix} \quad m \times k$$

$$ProjectedSamples_{p \times k} = D_{p \times m} * W_{m \times k}$$

$$D = \begin{bmatrix} d_{11} & d_{12} & d_{13} & . & d_{1m} \\ d_{21} & d_{22} & d_{23} & . & d_{2m} \\ d_{31} & d_{32} & d_{33} & . & d_{3m} \\ d_{41} & d_{42} & d_{43} & . & d_{4m} \\ d_{p1} & d_{p2} & d_{p3} & . & d_{pm} \end{bmatrix}$$

×

$$W_k = \begin{bmatrix} w1_1 & w1_2 & w1_3 \\ w2_1 & w2_2 & w2_3 \\ w3_1 & w3_2 & w3_3 \\ . & . & . \\ wm_1 & wm_2 & wm_3 \end{bmatrix}$$

Constraint

- Linearity:

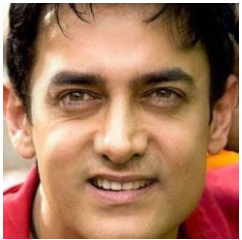
We can apply PCA to classify the data which is linearly separable, to deal with non-linear data we can use kernel PCA.

- Only suits for the Gaussian distribution:

In PCA we took assumption like mean should be zero and maximum variance will be 1, this assumption only holds when the distribution followed by the data is Gaussian.

Face Recognition – use case

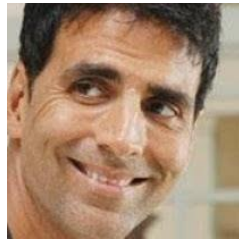
- Information Representation



Aamir



Ajay



Akshay



Amithabh



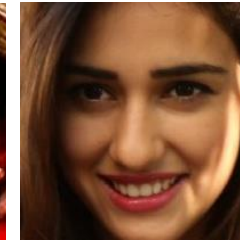
Farhan



Aaliya



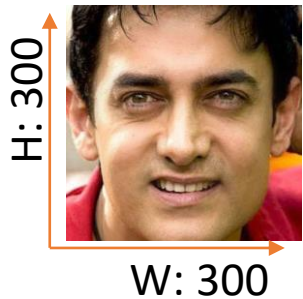
Deepika



Disha



Ileana



<https://csfieldguide.org.nz/en/interactives/pixel-viewer/>