Dimensionality Reduction

Computer Science- Pattern Recognition

Prof. Dr. Dhahir A. Abdullah

Dimensionality Reduction

- We can **reduce dimensionality** by combining features.
- **Principle Component Analysis** seeks a projection that best represents the data in a least square sense.

Principal Component Analysis

By far the most commonly used feature extraction technique

PCA assumes that the information is carried in the variance of the features: the higher the variance in one dimension (feature), the higher the information carried by that feature

- The transformation is based on preserving the most variance in the data using the least number of dimensions.
- The data is projected onto a lower dimensional space where the new features best represent the old features in the ***least squares sense.***

Principal Component Analysis

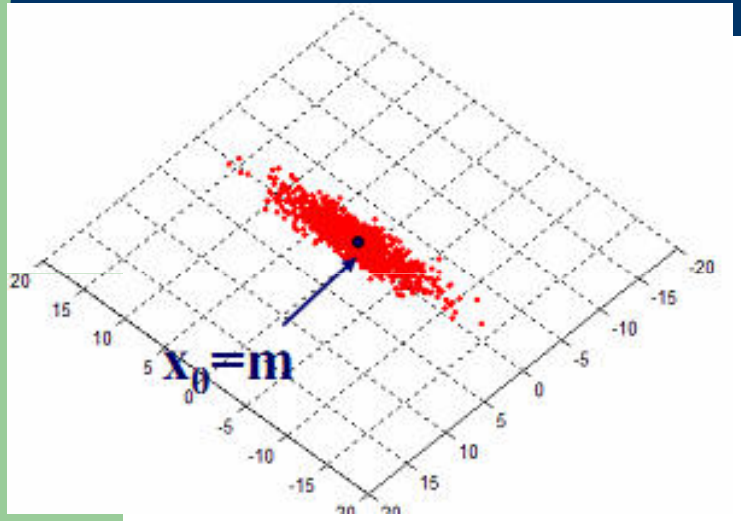
Assume we wish to represent *n d-dimensional* vectors $\mathbf{x}_1 \ldots \mathbf{x}_n$ with only one such vector, $\mathbf{x}_0$, such that the sum of squared distances between $\mathbf{x}_0$ and each of the $\mathbf{x}_k$ determined by the criterion function J_0 is minimum

$$J_0(\mathbf{x}_0) = \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{x}_0\|^2$$

This function is minimized if x_o is equal to mean →

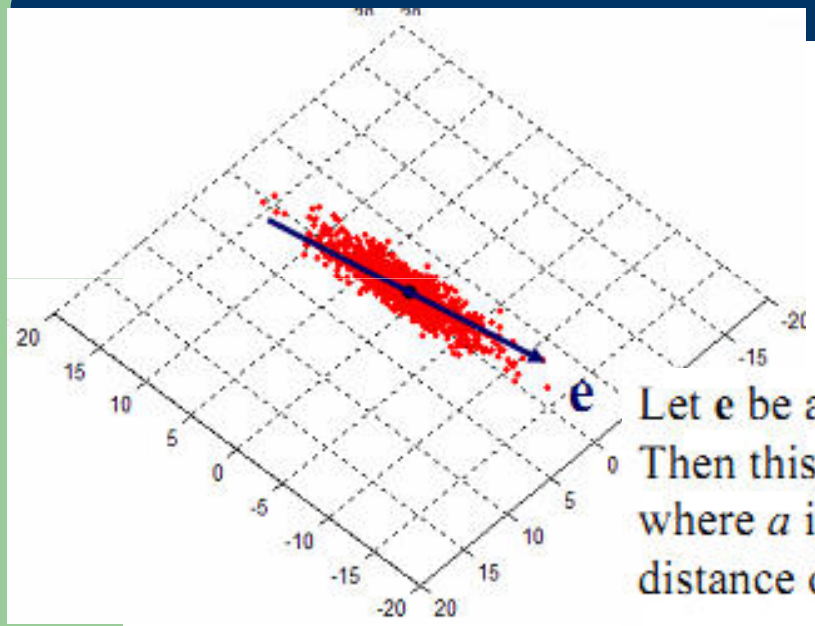
$$\mathbf{m} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k$$

Principal Component Analysis



- The sample mean is the zero-degree representation of the entire dataset.
- Simple, but provides no information about the variability in the data.
- A better representation can be obtained with **a projection**, a line, through the sample mean

Principal Component Analysis



Let **e** be a unit vector in the direction of this line. Then this line can be represented as follows, where a is a constant coefficient that indicate the distance of any point **x** from **m**

$$\mathbf{x} = \mathbf{m} + a\mathbf{e}$$

Principal Component Analysis

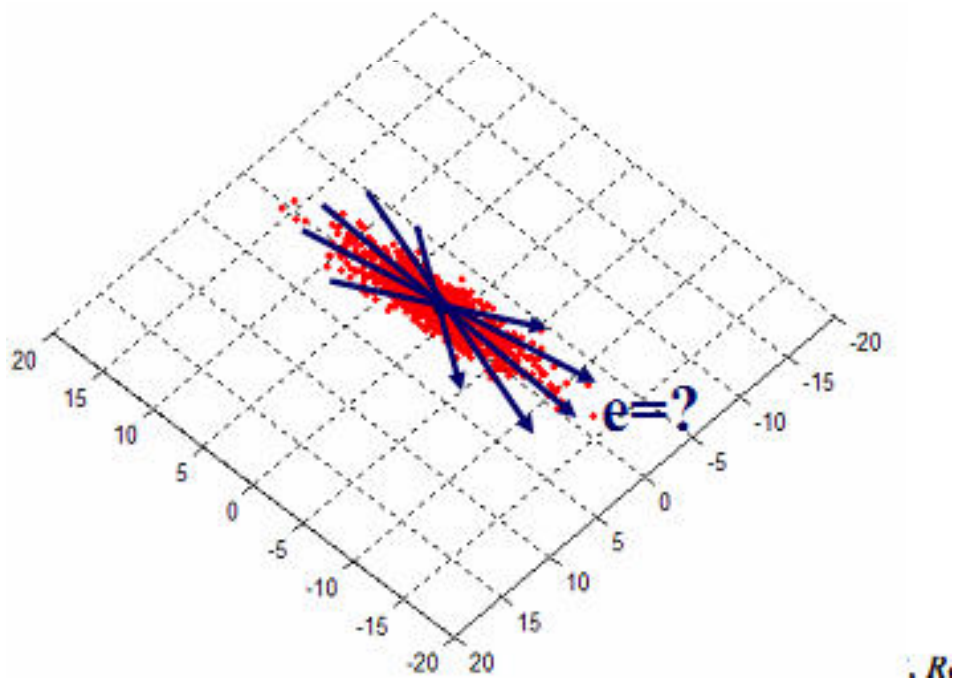
In general, we can represent any $\mathbf{x}_k$ by $\mathbf{m}+a_k\mathbf{e}$, where the optimal coefficients a_k can be obtained by minimizing the "squared error criterion function"

$$J_1(a_1,\cdots,a_n,\mathbf{e})=\sum_{k=1}^{n}\left\|(\mathbf{m}+a_k\mathbf{e})-\mathbf{x}_k\right\|^2$$

which yields $a_k=\mathbf{e}^T(\mathbf{x}_k-\mathbf{m})$, that is, we obtain the least square error coefficients by projecting the data vector $\mathbf{x}_k$ onto a line $\mathbf{e}$ that passes through the sample mean $\mathbf{m}$.

Principal Component Analysis

What is the best direction for **e**?



Principal Component Analysis

We look closer to the criterion function J_1 to determine the best direction of **e**:

$$J_1(a_1,\cdots,a_n,\mathbf{e}) = \sum_{k=1}^{n}\left\|(\mathbf{m}+a_k\mathbf{e})-\mathbf{x}_k\right\|^2 \qquad a_k = \mathbf{e}^T(\mathbf{x}_k-\mathbf{m})$$

$$= \sum_{k=1}^{n} a_k^2\cdot 1 - 2\sum_{k=1}^{n} a_k\mathbf{e}^T(\mathbf{x}_k-\mathbf{m}) + \sum_{k=1}^{n}\left\|\mathbf{x}_k-\mathbf{m}\right\|^2$$

$$= \sum_{k=1}^{n}\left[\mathbf{e}^T(\mathbf{x}_k-\mathbf{m})\right]^2 - 2\sum_{k=1}^{n}\left[\mathbf{e}^T(\mathbf{x}_k-\mathbf{m})\right]^2 + \sum_{k=1}^{n}\left\|\mathbf{x}_k-\mathbf{m}\right\|^2$$

$$= -\sum_{k=1}^{n}\left[\mathbf{e}^T(\mathbf{x}_k-\mathbf{m})\right]^2 + \sum_{k=1}^{n}\left\|\mathbf{x}_k-\mathbf{m}\right\|^2$$

$$= -\sum_{k=1}^{n}\mathbf{e}^T(\mathbf{x}_k-\mathbf{m})(\mathbf{x}_k-\mathbf{m})^T\mathbf{e} + \sum_{k=1}^{n}\left\|\mathbf{x}_k-\mathbf{m}\right\|^2$$

$$= -\mathbf{e}^T\mathbf{S}\mathbf{e} + \sum_{k=1}^{n}\left\|\mathbf{x}_k-\mathbf{m}\right\|^2 \qquad \mathbf{S} = \sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m})(\mathbf{x}_k-\mathbf{m})^T$$

(Total) Scatter matrix

Note that to minimize J_1 we need to maximize $\mathbf{e}^T\mathbf{S}\mathbf{e}$, subject to the constraint that $\|\mathbf{e}\|=1$

Principal Component Analysis

$$L(\mathbf{e},\lambda) = \mathbf{e}^T\mathbf{S}\mathbf{e} + \lambda(1 - \mathbf{e}^T\mathbf{e})$$

$$\frac{\partial L}{\partial \mathbf{e}} = 2\mathbf{S}\mathbf{e} - 2\lambda\mathbf{e}$$

$$\mathbf{S}\mathbf{e} = \lambda\mathbf{e}$$

Ring any bells?

Principal Component Analysis

The $\mathbf{e}$ that maximize $\mathbf{e}^T\mathbf{S}\mathbf{e}$ are the eigenvectors of $\mathbf{S}$ with corresponding eigenvalues λ

$$\mathbf{S}\mathbf{e} = \lambda\mathbf{e}$$

Principal Component Analysis

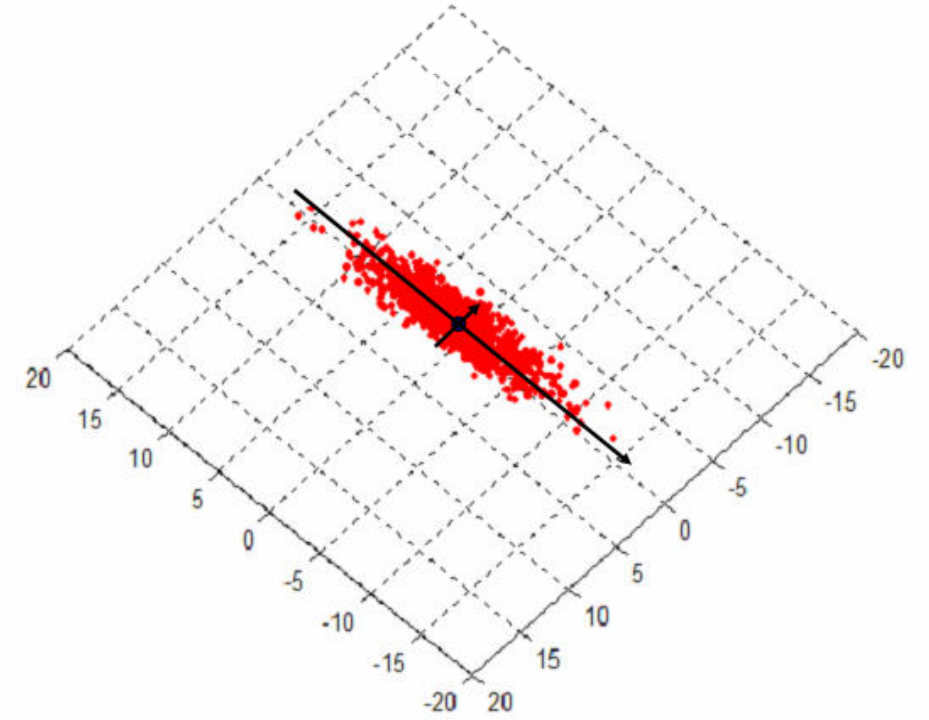
If we want the "best" line that represents the data, then we need to project the data onto a single line, for which we need to pick only one of the eigenvectors of **S**. To ensure that $\mathbf{e}^T\mathbf{S}\mathbf{e}$ is maximized, we pick the eigenvector corresponding to largest eigenvalue λ_{max}.

This can be readily extended to larger dimensions:

- If we want to project ***d-dimensional*** data onto a ***d'-dimensional*** subspace (***d'<d***), we project the data onto the ***d' eigenvectors*** of the scatter matrix **S** (which is really a constant multiplier of the covariance matrix), corresponding to largest ***d'-eigenvalues***.

Principal Component Analysis

From a geometrical stand point, the eigenvectors represent the ***principal axes***, along which the data (and hence the covariance matrix) show largest variance. The weight coefficients a_i are called the ***principal components.***



PCA - Example

Compute the principal components for the following two-dimensional dataset

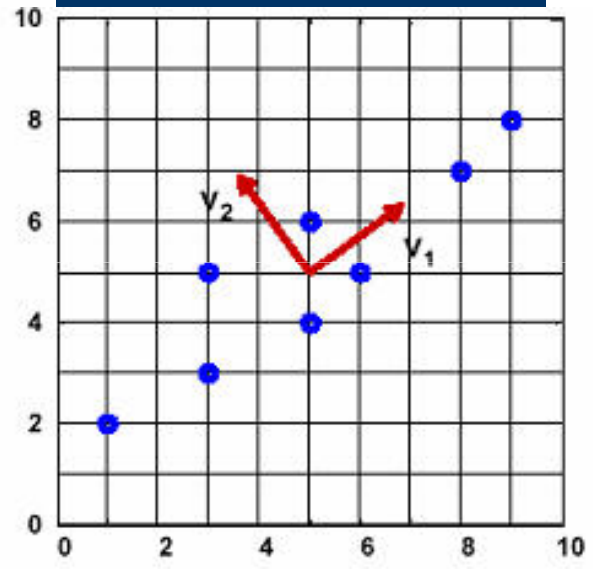
- $X=(x_1,x_2)=\{(1,2),(3,3),(3,5),(5,4),(5,6),(6,5),(8,7),(9,8)\}$
  - Let's first plot the data to get an idea of which solution we should expect

SOLUTION (by hand)

- The (biased) covariance estimate of the data is

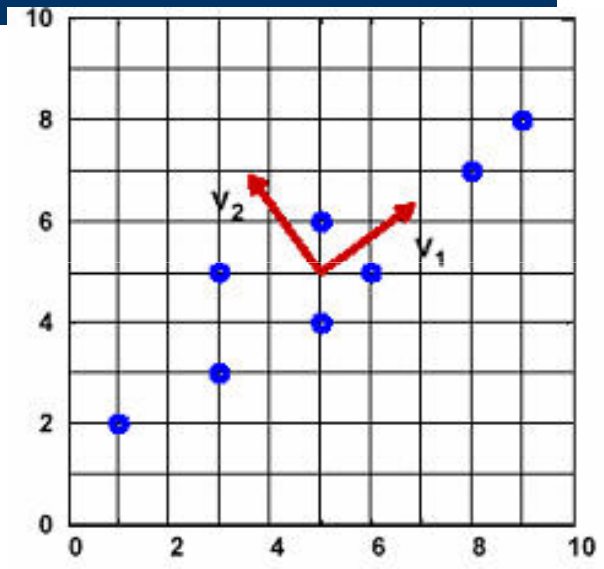
$$\Sigma_x = \begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix} \qquad S=\Sigma_x$$

HOW ?



PCA - Example

	x1	x2
	1	2
	3	3
	3	5
	5	4
	5	6
	6	5
	8	7
	9	8
m=	5	5



$$\Sigma_x = \frac{1}{8}\sum_{i=1}^{8}(x_i - m)(x_i - m)^T$$

HOW ?

$$\Sigma_x = \begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix}$$

PCA – Example (cont.)

x1	x2
1	2
3	3
3	5
5	4
5	6
6	5
8	7
9	8
5	5

m

x1 - m1	x2 - m2
-4	-3
-2	-2
-2	0
0	-1
0	1
1	0
3	2
4	3

16	12	12	9
4	4	4	4
4	0	0	0
0	0	0	1
0	0	0	1
1	0	0	0
9	6	6	4
16	12	12	9
50	34	34	28
6,25	4,25	4,25	3,5

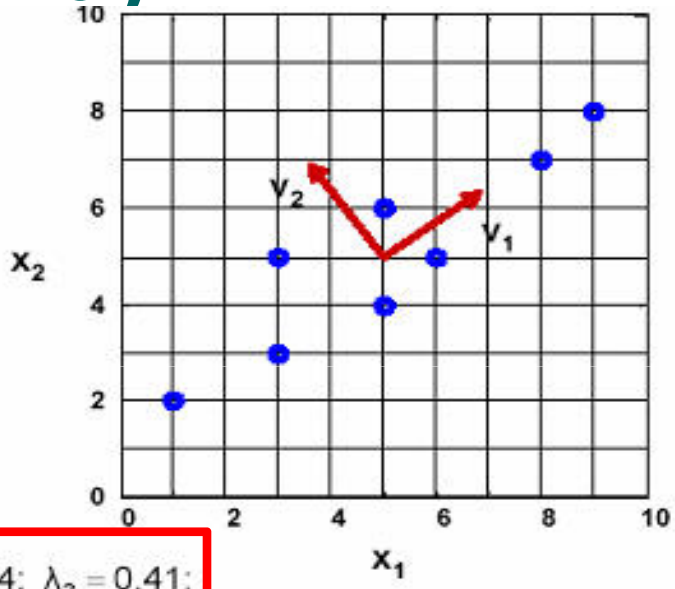
$$\Sigma_x = \frac{1}{8}\sum_{i=1}^{8}(x_i - m)(x_i - m)^T$$

$$\Sigma_x = \begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix}$$

PCA – Example (cont.)

Compute the principal components for the following two-dimensional dataset

- $X=(x_1,x_2)=\{(1,2),(3,3),(3,5),(5,4),(5,6),(6,5),(8,7),(9,8)\}$
  - Let's first plot the data to get an idea of which solution we should expect



SOLUTION (by hand)

- The (biased) covariance estimate of the data is:

$$\Sigma_x = \begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix} \qquad \mathbf{S=\Sigma_x}$$

- The eigenvalues are the zeros of the characteristic equation

$$\Sigma_x v = \lambda v \Rightarrow |\Sigma_x - \lambda I| = 0 \Rightarrow \begin{vmatrix} 6.25-\lambda & 4.25 \\ 4.25 & 3.5-\lambda \end{vmatrix} = 0 \Rightarrow \lambda_1 = 9.34;\ \lambda_2 = 0.41;$$

- The eigenvectors are the solutions of the system

$$\begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} \lambda_1 v_{11} \\ \lambda_1 v_{12} \end{bmatrix} \Rightarrow \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} 0.81 \\ 0.59 \end{bmatrix}$$

$$\begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} \lambda_2 v_{21} \\ \lambda_2 v_{22} \end{bmatrix} \Rightarrow \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} -0.59 \\ 0.81 \end{bmatrix}$$

Continue and find principle components.

- HINT: To solve each system manually, first assume that one of the variables is equal to one (i.e. $v_{i1}=1$), then find the other one and finally normalize the vector to make it unit-length