# Lecture 11: PCA

Andreas Wichert
Department of Computer Science and Engineering
Técnico Lisboa

1

## Feature selection

- **Entropy** of a variable $$E(y_j) = -\sum_{v \in y_j} P(v)\log_2 P(v)$$
- Variable-**conditional entropy** $$E(z|y_j) = \sum_{v \in y_j} P(v)E(z|v)$$
- **Information gain** $$IG(z|y_j) = E(z) - E(z|y_j)$$

| | Hair | Height | Weight | Lotion | Result |
|---|---|---|---|---|---|
| $i_1$ | 1 | 2 | 1 | 0 | 1 |
| $i_2$ | 1 | 3 | 2 | 1 | 0 |
| $i_3$ | 2 | 1 | 2 | 1 | 0 |
| $i_4$ | 1 | 1 | 2 | 0 | 1 |
| $i_5$ | 3 | 2 | 3 | 0 | 1 |
| $i_6$ | 2 | 3 | 3 | 0 | 0 |
| $i_7$ | 2 | 2 | 3 | 0 | 0 |
| $i_8$ | 1 | 1 | 1 | 1 | 0 |

Rank(Hair)=IG(Result|Hair)=0.45

Rank(Height)=IG(Result|Height)=0.26

Hair variable has higher IG than Height, hence is more important and less susceptible to removal

2

## The Karhunen-Loève Transform

- The Karhunen-Loève transform is a linear transform that maps possibly correlated variables into a set of values of linearly uncorrelated variables
- This transformation is defined in such a way that the first principal component has the largest possible variance

3

## The Covariance

- The sample size denoted by n, is the number of data items in a sample of a population.
- The goal is to make inferences about a population from a sample.
- Sample covariance indicates the relationship between two variables of a sample $$cov(X,Y) = \frac{\sum_{k=1}^{n}(x_k - \overline{x}) \cdot (y_k - \overline{y})}{n-1}$$
  - The sample covariance has n − 1 in the denominator rather than n due to Bessel's correction.
- For the whole population, the covariance is

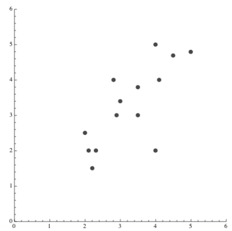
$$cov(X,Y) = \frac{\sum_{k=1}^{n}(x_k - \overline{x}) \cdot (y_k - \overline{y})}{n}.$$

4

- The sample covariance relies on the difference between each observation and the sample mean.
- In computer science, the sample covariance is usually used and
- In statistics (Bishop) population covariance

- In a linear relationship, either the high values of one variable are paired with the high values of another variable or the high values of one variable are paired with the low values of another variable

5



- For example, for a list of two variables, *(X, Y ),*

$$\Sigma = \{(2.1, 2), (2.3, 2), (2.9, 3), (4.1, 4), (5, 4.8), (2, 2.5), (2.2, 1.5), (4, 5), (4, 2), (2.8, 4), (3, 3.4), (3.5, 3.8), (4.5, 4.7), (3.5, 3)\}$$

- represents the data set *Σ*. The sample covariance of the data set is *0.82456*. Ordering the list by *X*, we notice that the **ascending** X values are matched by **ascending** *Y* values

6

- The covariance matrix measures the tendency of two features, $x_i$ and $x_j$ , to vary in the same direction. The covariance between features $x_i$ and $x_j$ is estimated for *n* vectors as

$$c_{ij} = \frac{\sum_{k=1}^{n}(x_{k,i} - \overline{x_i}) \cdot (y_{k,j} - \overline{y_j})}{n-1}$$

$$c_{ij} = \frac{\sum_{k=1}^{n}\left(x_i^{(k)} - m_i\right)\left(x_j^{(k)} - m_j\right)}{n-1}$$

- with $x_i$ and $y_j$ being the arithmetic mean of the two variables of the sample. Covariances are symmetric; $c_{ij} = c_{ji}$

7

## Correlation

- Covariance is related to correlation

$$r_{ij} = \frac{\sum_{k=1}^{n}\left(x_i^{(k)} - m_i\right)\left(x_j^{(k)} - m_j\right)}{(n-1)s_i s_j} = \frac{c_{ij}}{s_i s_j} \in [-1,1]$$

8

- The resulting covariance matrix C is symmetric and positive-definite,

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mm} \end{pmatrix}$$

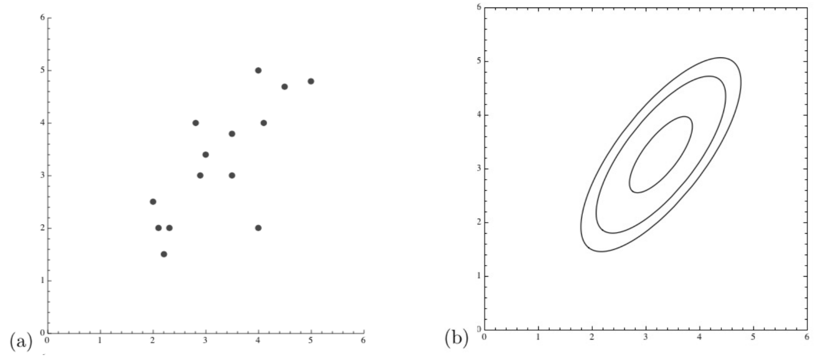
9

$$\Sigma = \{(2.1, 2), (2.3, 2), (2.9, 3), (4.1, 4), (5, 4.8), (2, 2.5), (2.2, 1.5),$$
$$(4, 5), (4, 2), (2.8, 4), (3, 3.4), (3.5, 3.8), (4.5, 4.7), (3.5, 3)\}$$

$$c_{ij} = \frac{\sum_{k=1}^{n} (x_{k,i} - \overline{x_i}) \cdot (y_{k,j} - \overline{y_j})}{n-1}$$

$$C = \begin{pmatrix} 0.912582 & 0.82456 \\ 0.82456 & 1.34247 \end{pmatrix}$$

10




- (a) The data points of the set *Σ* (b) The two dimensional distribution of *Σ* can be described by three ellipse that divide the data points in four equal groups.



# The Karhunen-Loève Transform

A real matrix $M$ is positive definite if $\mathbf{z}^\top \cdot M \cdot \mathbf{z}$ is positive for any non-zero column vector $\mathbf{z}$ of real numbers. A symmetric and positive-definite matrix can be diagonalized. It follows that

$$U^{-1} \cdot C \cdot U = \Lambda = diag(\lambda_1, \lambda_2, \cdots .\lambda_m)$$

$U$ is an orthonormal matrix of the dimension $m \times m$,

$$U^\top \cdot U = I$$

$$U^\top \cdot C \cdot U = \Lambda = diag(\lambda_1, \lambda_2, \cdots .\lambda_m)$$

and

$$U \cdot \Lambda = C \cdot U.$$

$$U \cdot \Lambda = C \cdot U.$$

There are $m$ eigenvalues and eigenvectors with

$$(\lambda_i \cdot I - C) \cdot \mathbf{u}_i = 0$$

and

$$C \cdot \mathbf{u}_i = \lambda_i \cdot \mathbf{u}_i$$

An eigenvector can have two directions, it is either $\mathbf{u}_i$ or $-\mathbf{u}_i$.

$$C \cdot (-\mathbf{u}_i) = \lambda_i \cdot (-\mathbf{u}_i)$$

13

The eigenvectors are always orthogonal, and their length is arbitrary. The normalized eigenvectors define the orthonormal matrix $U$ of dimension $m \times m$. Each normalized eigenvector is a column of $U$ with

$$U^\top \cdot U = I.$$

The matrix $U$ defines the Karhunen-Loève transform. The Karhunen-Loève transform rotates the coordinate system in such a way that the new covariance matrix will be diagonal

$$\mathbf{y} = U^\top \cdot \mathbf{x}$$

14

- The squares of the eigenvalues represent the variances along the eigenvectors. The eigenvalues corresponding to the covariance matrix of the data set Σ are

$$\lambda_1 = 1.97964, \quad \lambda_2 = 0.275412$$

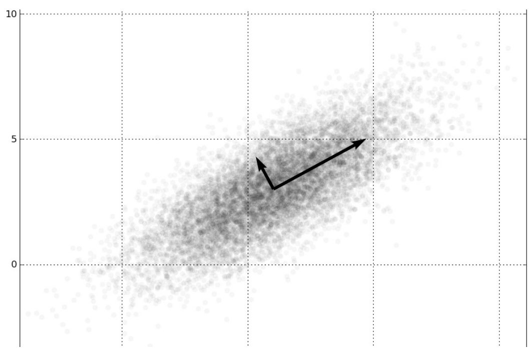
and the corresponding normalized eigenvectors are

$$\mathbf{u}_1 = \begin{pmatrix} 0.611454 \\ 0.79128 \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} -0.79128 \\ 0.611454 \end{pmatrix}.$$

The define the matrix $U$ with

$$U = \begin{pmatrix} 0.611454 & -0.79128 \\ 0.79128 & 0.611454 \end{pmatrix}.$$

15




16

The define the matrix $U$ with

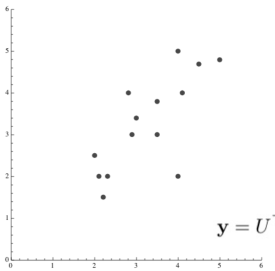
$$U = \begin{pmatrix} 0.611454 & -0.79128 \\ 0.79128 & 0.611454 \end{pmatrix}.$$

The Karhunen-Loève transform for the data set $\Sigma$ is given by

$$\mathbf{y} = U^\top \cdot \mathbf{x} = \begin{pmatrix} 0.611454 & 0.79128 \\ -0.79128 & 0.611454 \end{pmatrix} \cdot \mathbf{x},$$

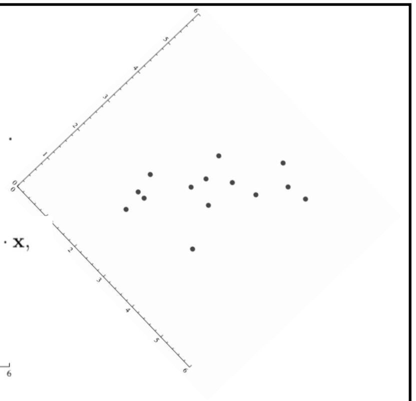
it rotates the coordinate system in such a way that the new covariance matrix will be diagonal

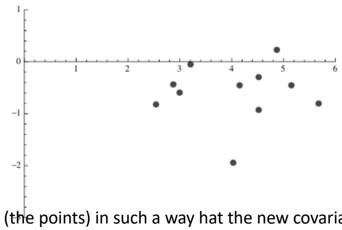
17



$$U = \begin{pmatrix} 0.611454 & -0.79128 \\ 0.79128 & 0.611454 \end{pmatrix}.$$

$$\mathbf{y} = U^\top \cdot \mathbf{x} = \begin{pmatrix} 0.611454 & 0.79128 \\ -0.79128 & 0.611454 \end{pmatrix} \cdot \mathbf{x},$$





It rotates the system (the points) in such a way hat the new covariance matrix will be diagonal.

18

# Principal component analysis

- Principal component analysis (PCA) is a technique that is useful for the compression of data.
- The purpose is to reduce the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables, that nonetheless retains most of the sample’s information.
- The first principal component corresponds to the normalized eigenvector with the highest variance.
- In principal component analysis (PCA), the significant eigenvectors define the principal components.

19

- Accordingly to the **Kaiser criterion**, the eigenvectors whose eigenvalues are below 1 are discarded
- Each of the *s* non-discarded eigenvectors is a column of the matrix W of dimension $s \times m$ with the linear mapping from $\mathbf{R}^m \to \mathbf{R}^s$,

$$\mathbf{z} = W^\top \cdot \mathbf{x}$$

- The Principal component analysis for the data set $\Sigma$ is given by

$$\mathbf{z} = W^\top \cdot \mathbf{x} = \begin{pmatrix} 0.611454 & 0.79128 \end{pmatrix} \cdot \mathbf{x}$$

20

$$\lambda_1 = 1.97964, \quad \lambda_2 = 0.275412$$

and the corresponding normalized eigenvectors are

$$\mathbf{u}_1 = \begin{pmatrix} 0.611454 \\ 0.79128 \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} -0.79128 \\ 0.611454 \end{pmatrix}.$$

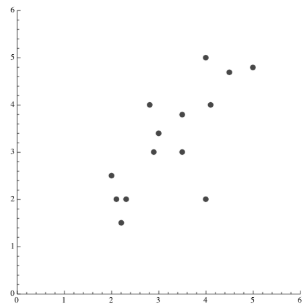
The define the matrix $U$ with

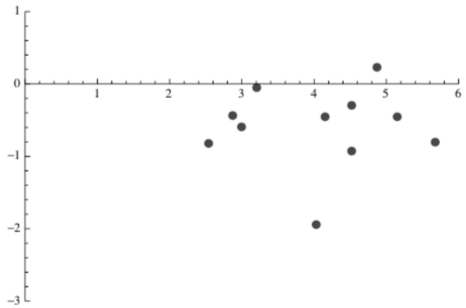
$$U = \begin{pmatrix} 0.611454 & -0.79128 \\ 0.79128 & 0.611454 \end{pmatrix}.$$

$$\mathbf{y} = U^\top \cdot \mathbf{x} = \begin{pmatrix} 0.611454 & 0.79128 \\ -0.79128 & 0.611454 \end{pmatrix} \cdot \mathbf{x},$$

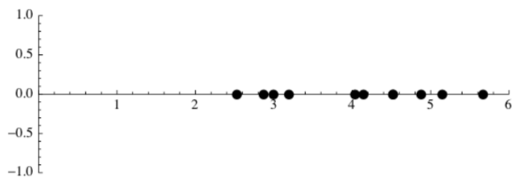
$$\mathbf{z} = W^\top \cdot \mathbf{x} = \begin{pmatrix} 0.611454 & 0.79128 \end{pmatrix} \cdot \mathbf{x}$$

21





$$\mathbf{y} = U^\top \cdot \mathbf{x} = \begin{pmatrix} 0.611454 & 0.79128 \\ -0.79128 & 0.611454 \end{pmatrix} \cdot \mathbf{x},$$



$$\mathbf{z} = W^\top \cdot \mathbf{x} = \begin{pmatrix} 0.611454 & 0.79128 \end{pmatrix} \cdot \mathbf{x}$$

22

- Suppose we have a covariance matrix

$$C = \begin{pmatrix} 3 & 1 \\ 1 & 21 \end{pmatrix}$$

- What is the corresponding matrix of the K-L transformation?
- First, we have to compute the eigenvalues.
- The system has to become linear depend-able (singular).
- The determinant has to become zero.

$$|\lambda \cdot I - C| = 0.$$

23

$$|\lambda \cdot I - C| = 0.$$

Solving the Equation

$$\lambda^2 - 24 \cdot \lambda + 62 = 0$$

we get the two eigenvalues

$$\lambda_1 = 2.94461, \quad \lambda_2 = 21.05538.$$

24

$$\lambda_1 = 2.94461, \quad \lambda_2 = 21.05538.$$

To compute the eigenvectors we have to solve two singular, dependent systems

$$|\lambda_1 \cdot I - C| = 0$$

and

$$|\lambda_2 \cdot I - C| = 0.$$

25

For $\lambda_1 = 2.94461$ we get

$$\left( \begin{pmatrix} 2.94461 & 0 \\ 0 & 2.94461 \end{pmatrix} - \begin{pmatrix} 3 & 1 \\ 1 & 21 \end{pmatrix} \right) \cdot \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = 0$$

and we have to find a nontrivial solution for

$$\begin{pmatrix} -0.05538 & -1 \\ -1 & -18.055 \end{pmatrix} \cdot \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = 0$$

Because the system is linear dependable, the left column is a multiple value of the right column, and there are infinitely many solution. We only have to determine the direction of the eigenvectors; if we simply suppose that $u_1 = 1$,

26

$u_1 = 1$,

$$\begin{pmatrix} -0.05538 & -1 \\ -1 & -18.055 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ u_2 \end{pmatrix} = 0$$

and

$$\begin{pmatrix} -0.05538 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ 18.055 \end{pmatrix} \cdot u_2$$

with

$$\mathbf{u}_1 = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -0.05539 \end{pmatrix}.$$

27

For $\lambda_2 = 21.05538$ we get

$$\begin{pmatrix} 18.055 & -1 \\ -1 & 0.05538 \end{pmatrix} \cdot \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = 0$$

with

$$\mathbf{u}_2 = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 18.055 \end{pmatrix}.$$

The two normalized vectors $\mathbf{u}_1$, $\mathbf{u}_2$ define the columns of the matrix $U$

$$U = \begin{pmatrix} 0.998469 & 0.0553016 \\ -0.0553052 & 0.99847 \end{pmatrix}.$$

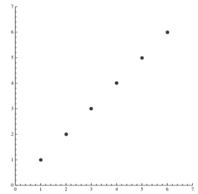
Because $\lambda_1 = 2.94461 < \lambda_2 = 21.05538$ the second eigenvector is more significant, however we can not apply the Kaiser criterion.

28

$$\Psi = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$$

the covariance matrix is

$$C = \begin{pmatrix} 3.5 & 3.5 \\ 3.5 & 3.5 \end{pmatrix}.$$



The two two eignvalues are

$$\lambda_1 = 7, \quad \lambda_2 = 0$$

and the two normalized eigenvectors are

$$\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}.$$

29

The matrix that describes the K-L transformation is given by

$$U = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}. \end{pmatrix} = \frac{1}{\sqrt{2}} \cdot \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \sqrt{2} \cdot \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix}. \tag{3.124}$$

The K-L transformation maps the two dimensional data set $\Psi$ in one dimension because $\lambda_2$ is zero (see Figure 3.35). For example, the data point $(1,1)$ is mapped on the $x - axis$

$$\begin{pmatrix} \sqrt{2} \\ 0 \end{pmatrix} = \begin{pmatrix} \sqrt{2} \cdot \frac{1+1}{2} \\ \sqrt{2} \cdot \frac{1-1}{2} \end{pmatrix} = \sqrt{2} \cdot \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{3.125}$$

with the value $\sqrt{2} \approx 1.4142$ corresponding to the length of the vector $(1,1)$.

30

## Principal component analysis

- The variance in the direction of the $k^{\text{th}}$ singular vector (or principal component) is given by the singular value $\lambda_k$
- Singular values can be used to estimate how many components to keep
- ***Rule of thumb:*** keep enough to explain 85% of the variation
- $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{n} \lambda_j} \approx 0.85$$

if $k = n$, we preserve 100% of the original variation

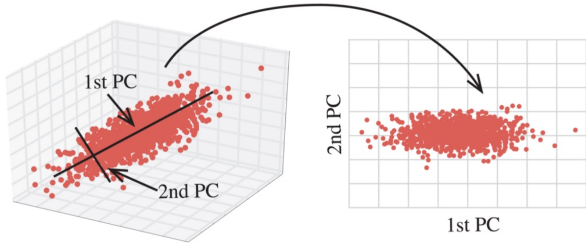
31

## Problems

- Principal components are linear transformation of the original features
- It is difficult to attach any semantic meaning to principal components
- For new data which is added to the dataset, the PCA has to be recomputed

32

# Space transformation




Axes of greater variance given by *eigenvectors* of *covariance matrix*

33

# Reconstruction

- This PCA matrix W corresponds to a linear mapping from $R^m \rightarrow R^s$

$$\mathbf{z} = W^{\top} \cdot \mathbf{x}$$

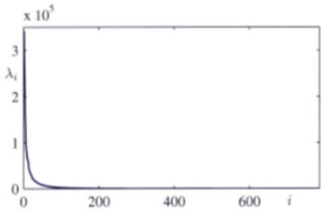
- Reconstruction

$$W \cdot \boldsymbol{z} = \hat{\boldsymbol{x}}$$

- PCA minimizes the reconstruction error: $||\mathbf{x} - \hat{\mathbf{x}}||$
- It can be shown that the reconstruction error is: $error = 1/2 \sum_{i=k+1}^{m} \lambda_i$

34

## PCA: Only the images of three




- Eigenvalues


- Projection of the Eigenvectors, blue positive values, yellow negative values

35

## Singular Value Decomposition

- We can use SVD to perform PCA
  - SVD **is more numerically stable** if the columns are close to collinear
  - Factorize a Covariance Matrix *A:=C*
    - Difference: compute the eigenvectors out of $C\,C^T = C^T C = C\,C$, use *U* as before....

Any matrix $A$ can be factorized as

$$A = U \cdot S \cdot V^T$$

$U$ is a orthogonal matrix with orthonormal eigenvectors from $A \cdot A^T$

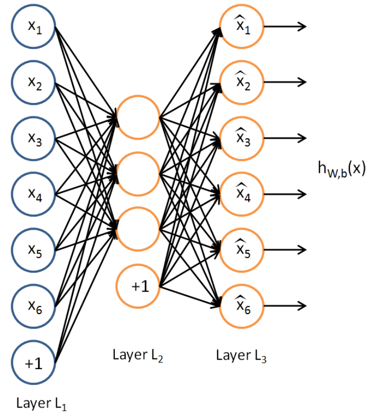
$V$ a orthogonal matrix with orthonormal eigenvectors from $A^T \cdot A$

$S$ is a diagonal matrix with $r$ elements equal to the root of the positive eigenvalues of $A \cdot A^T$ or $A^T \cdot A$

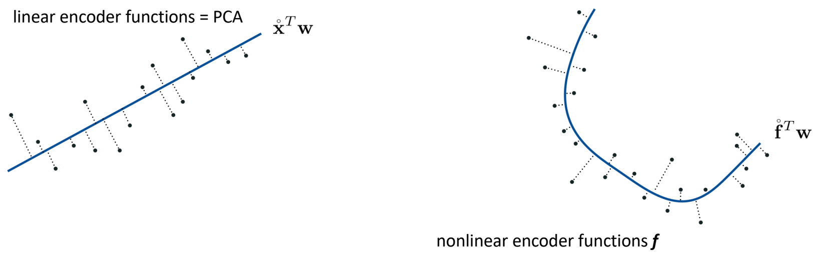
36

## Other Methods for Dimension Reduction

- Unsupervised Learning
  - Data: no labels!
  - Goal: Learn the structure of the data
- Traditionally, autoencoders were used for dimensionality reduction or feature learning.
- Undercomplete autoencoder (less units in the hidden layer)

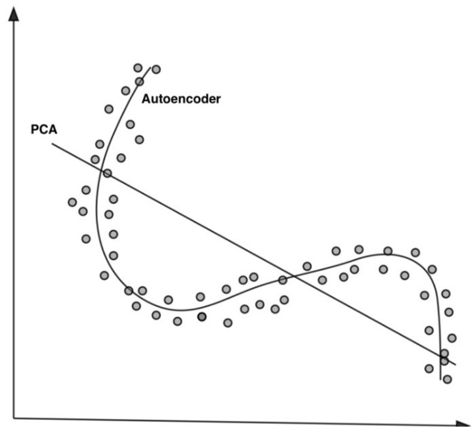



37




- Autoencoders with nonlinear encoder functions ***f*** and nonlinear decoder functions ***g*** can thus learn a more powerful nonlinear generalization of PCA

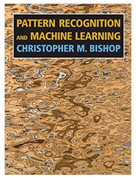
38



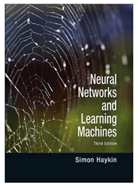

When the decoder is linear and L is the mean squared error, an undercomplete autoencoder learns to span the same subspace as PCA.
Autoencoders with nonlinear encoder functions f and nonlinear decoder functions g can learn a more powerful nonlinear generalization of PCA, the structure of **the manifold**.

39

# Literature



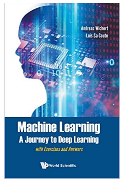

- Christopher M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer 2006
  - Chapter 12



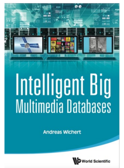

- Simon O. Haykin, Neural Networks and Learning Machine, (3rd Edition), Pearson 2008
  - Chapter 10

40

# Literature




- Machine Learning - A Journey to Deep Learning, A. Wichert, Luis Sa-Couto, World Scientific, 2021
  - Chapter 15




- Intelligent Big Multimedia Databases, A. Wichert, World Scientific, 2015
  - *Chapter 3, Section 3.3*

41