Lecture 16: PCA, ICA

Andreas Wichert

Department of Computer Science and Engineering

Técnico Lisboa

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

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

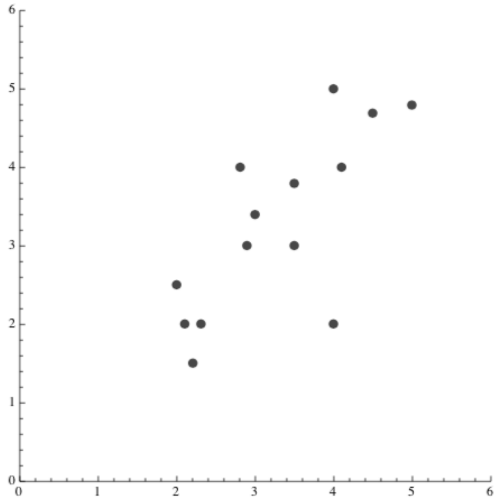
$$cov(X,Y) = \frac{\sum_{k=1}^{n}(x_k - \overline{x}) \cdot (y_k - \overline{y})}{n-1}$$

  - The sample covariance has n − 1 in the denominator rather than n due to Bessel’s correction.
- For the whole population, the covariance is

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

- 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



- 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

- 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; $c_{ij} = c_{ji}$

Correlation

- Covariance is related to correlation

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

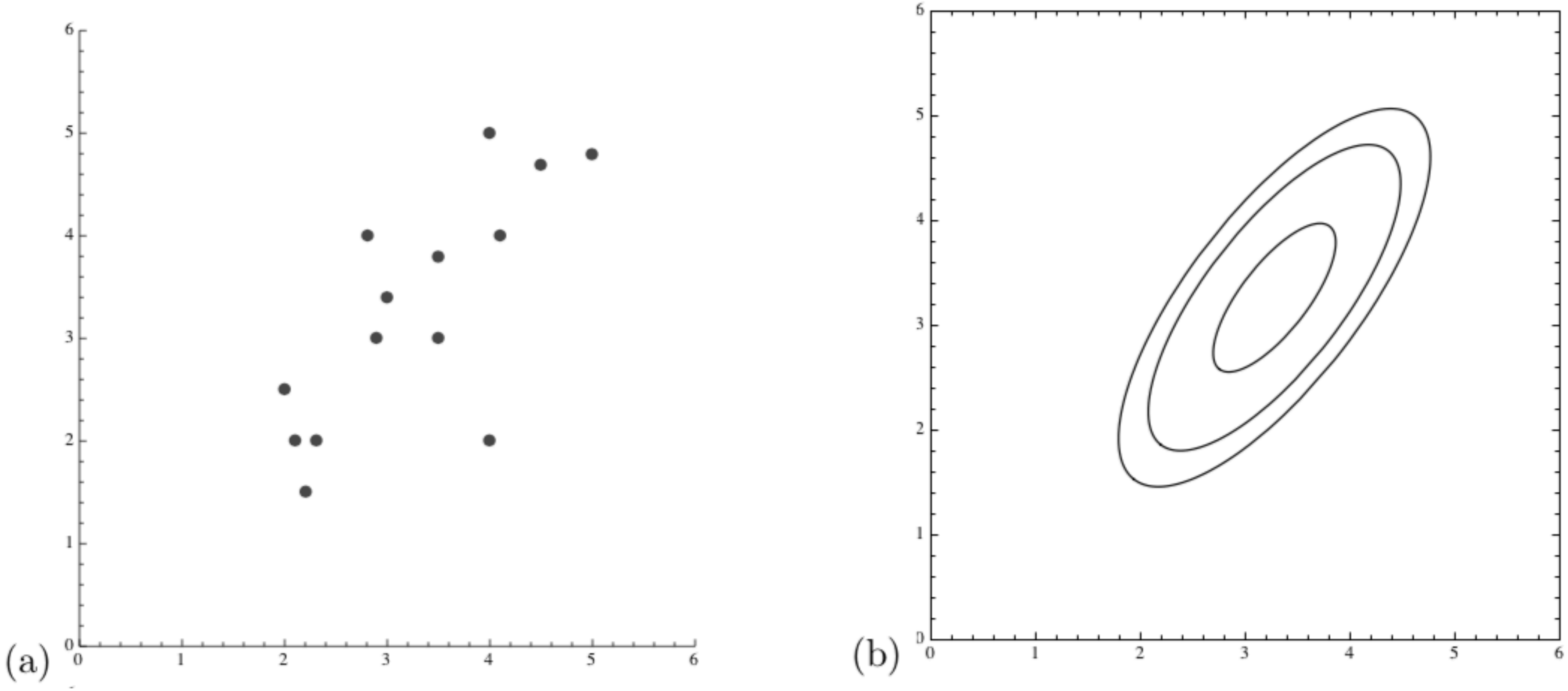
- The resulting covariance matrix C is symmetric and positive-definite,

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

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

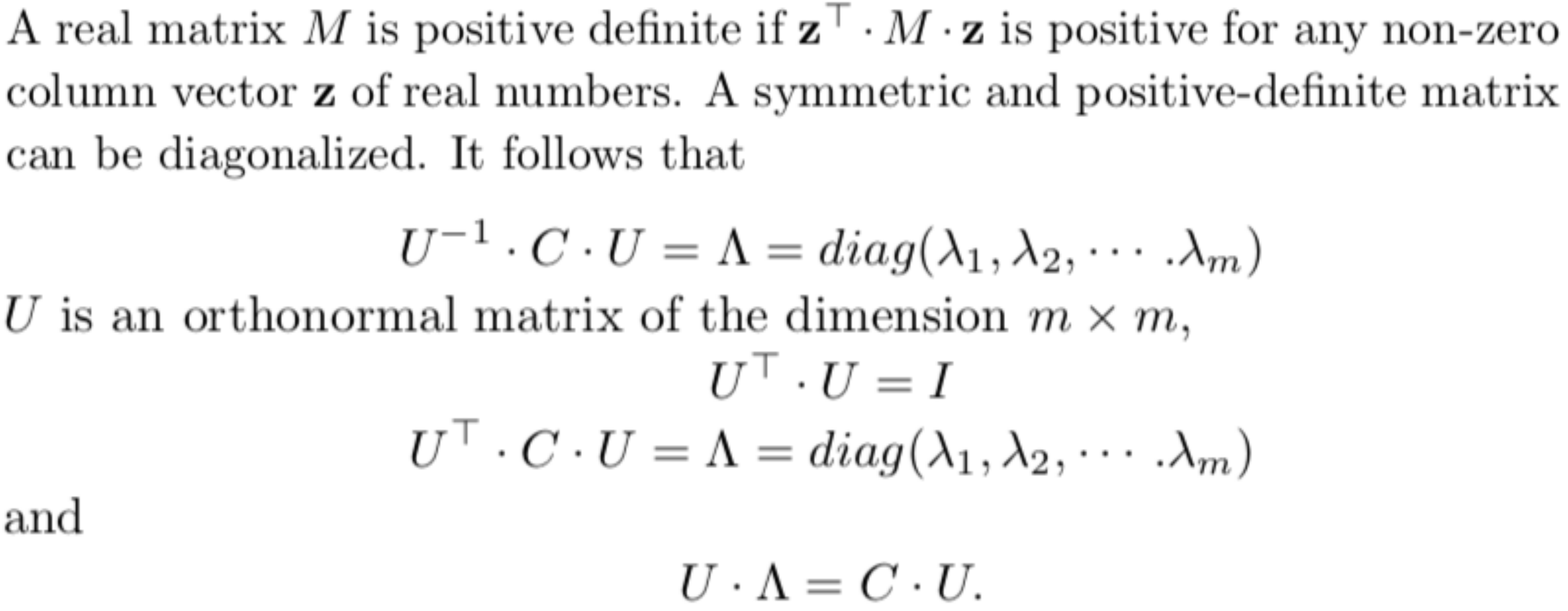
$$c_{ij} = \frac{\sum_{k=1}^{n} (x_{k,i} - \overline{x_i}) \cdot (y_{k,j} - \overline{y_j})}{n-1}$$

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



- (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 $-\mathbf{u}_i$.

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

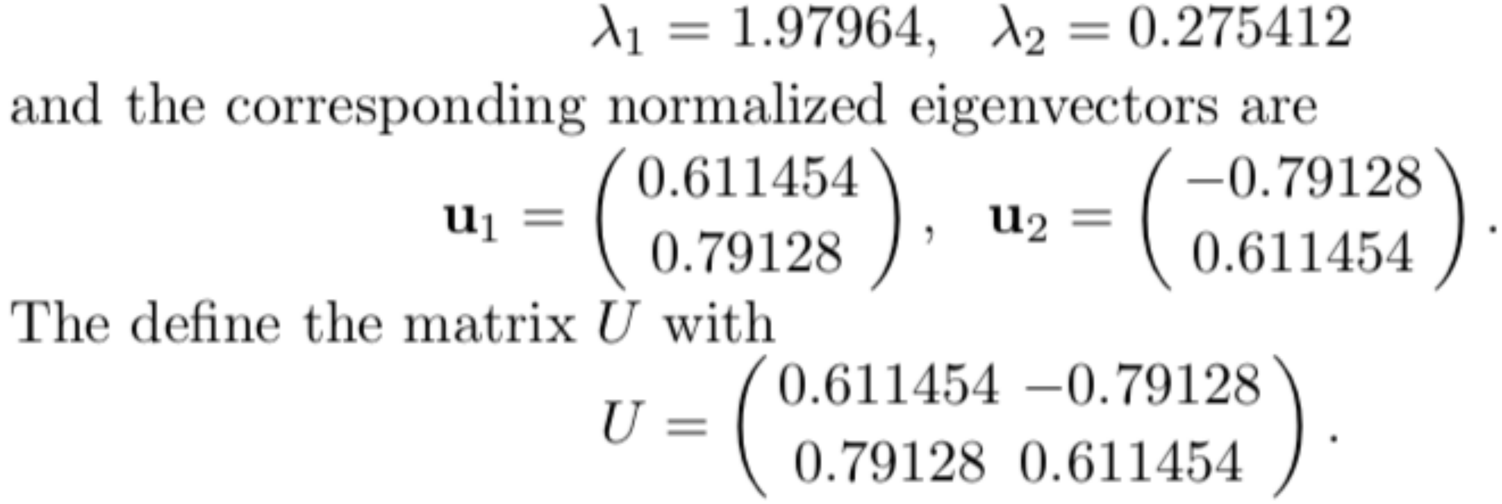
The eigenvectors are always orthogonal, and their length is arbitrary. The normalized eigenvectors define the orthonormal matrix U of dimension $m \times m$. Each normalized eigenvector is a column of U with

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

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

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

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



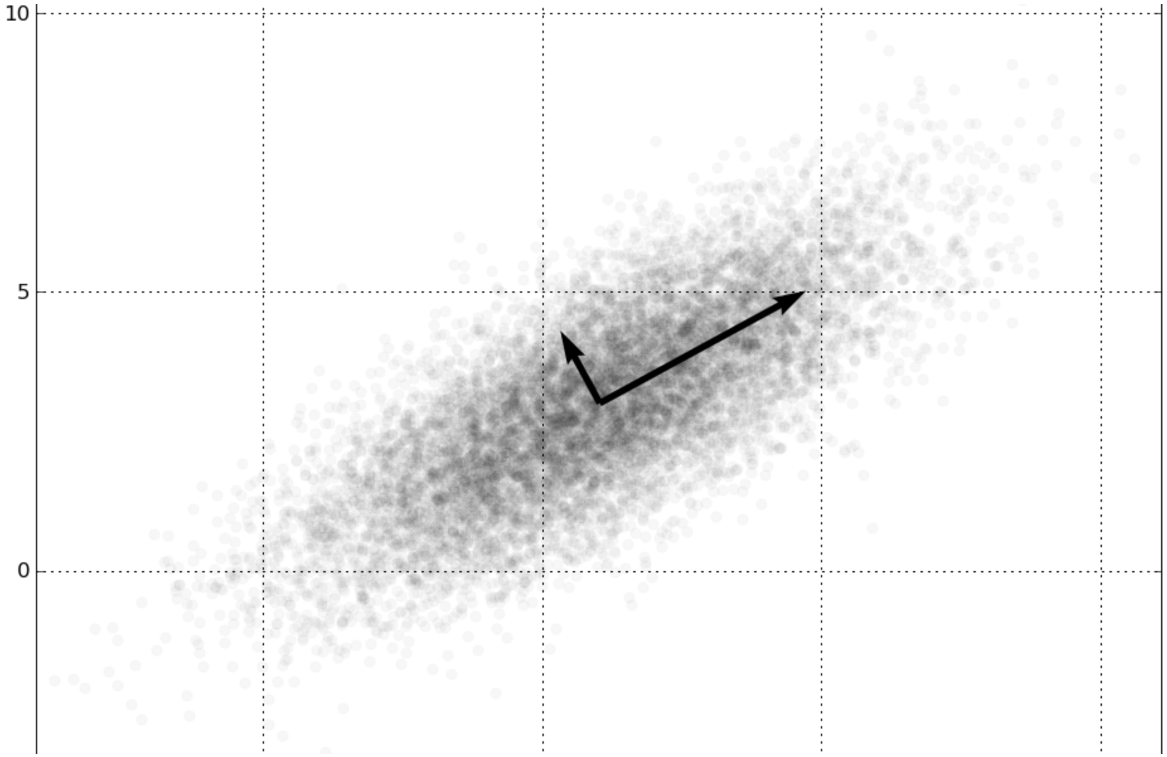
$$\lambda_1 = 1.97964, \quad \lambda_2 = 0.275412$$

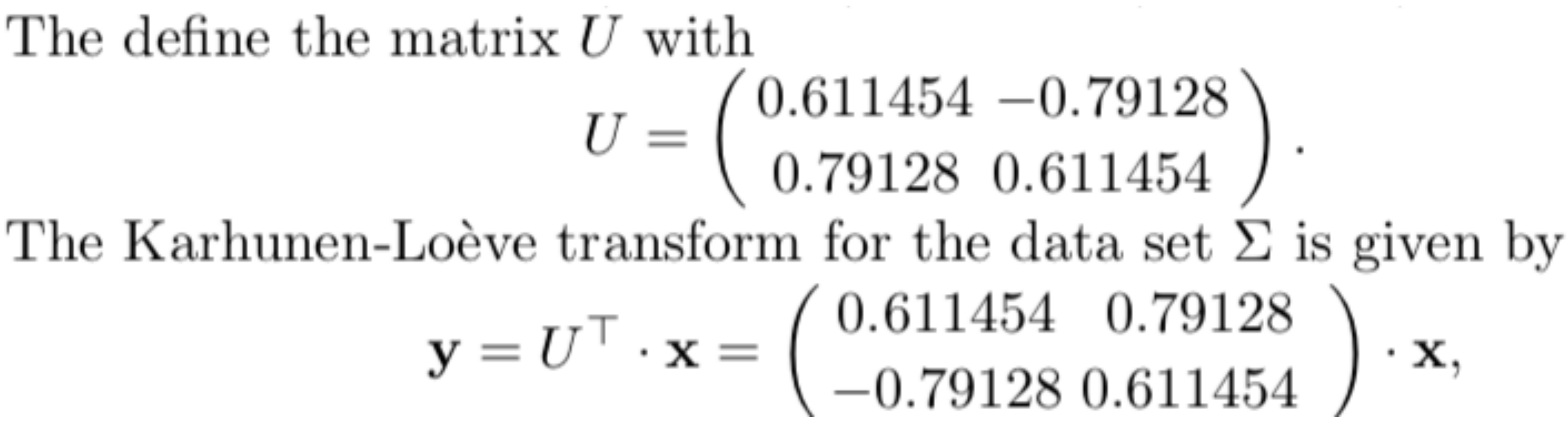
and the corresponding normalized eigenvectors are

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

The define the matrix U with

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


10
5
0



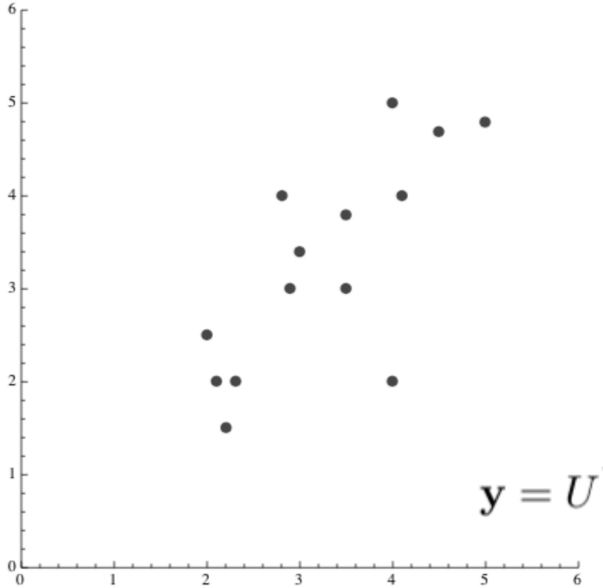
The define the matrix U with

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

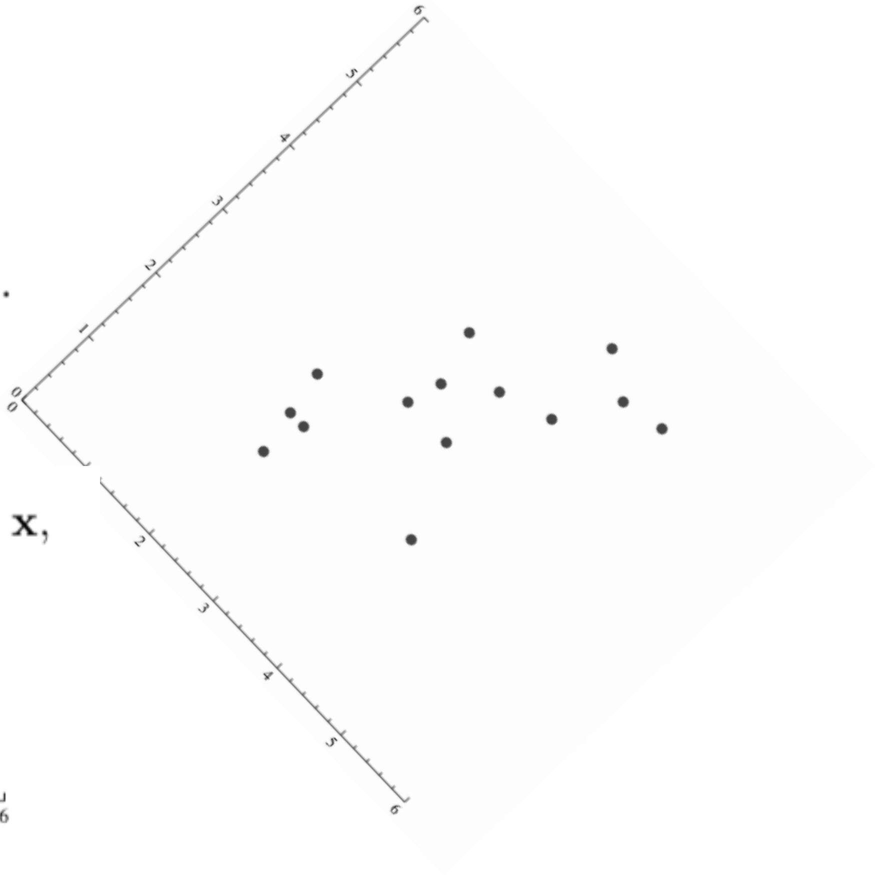
The Karhunen-Loève transform for the data set Σ is given by

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

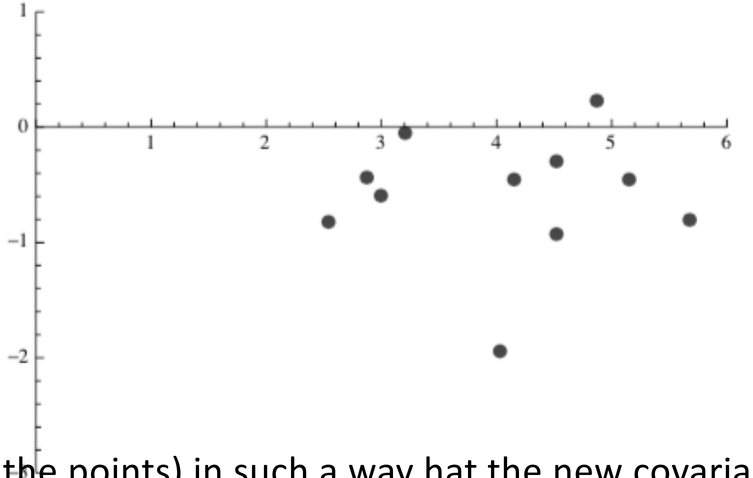
it rotates the coordinate system in such a way that the new covariance matrix will be diagonal



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



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



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

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.

- 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* × *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 *Σ* is given by

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

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

and the corresponding normalized eigenvectors are

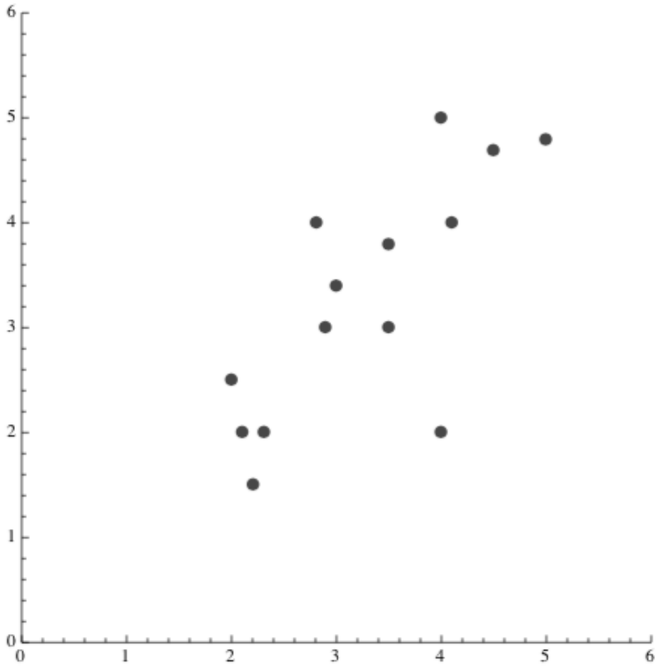
$$\mathbf{u}_1 = \begin{pmatrix} 0.611454 \\ 0.79128 \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} -0.79128 \\ 0.611454 \end{pmatrix}.$$

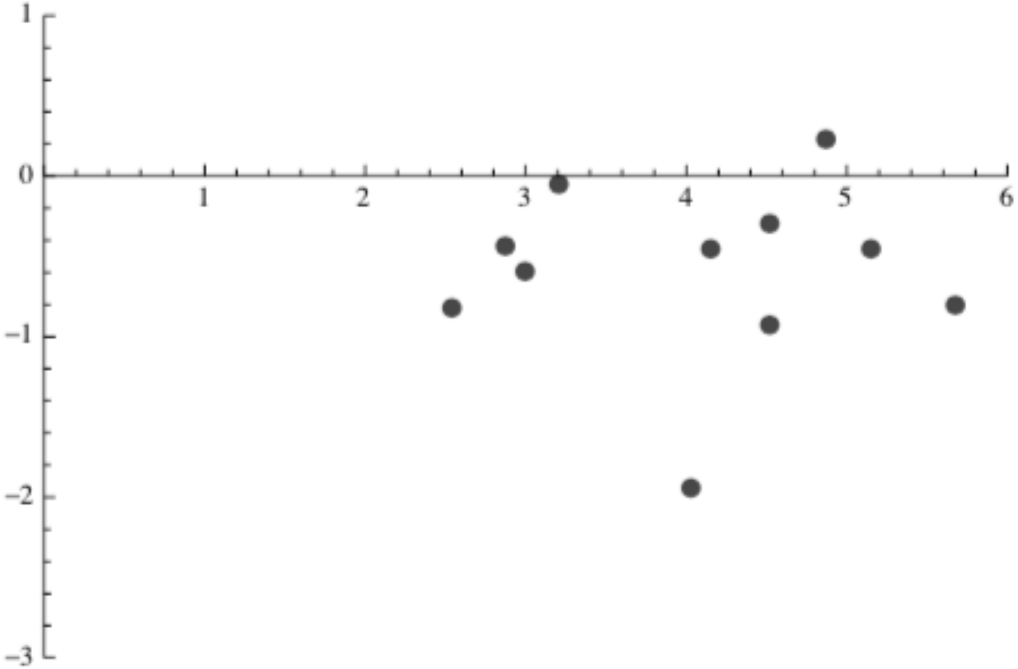
The define the matrix U with

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

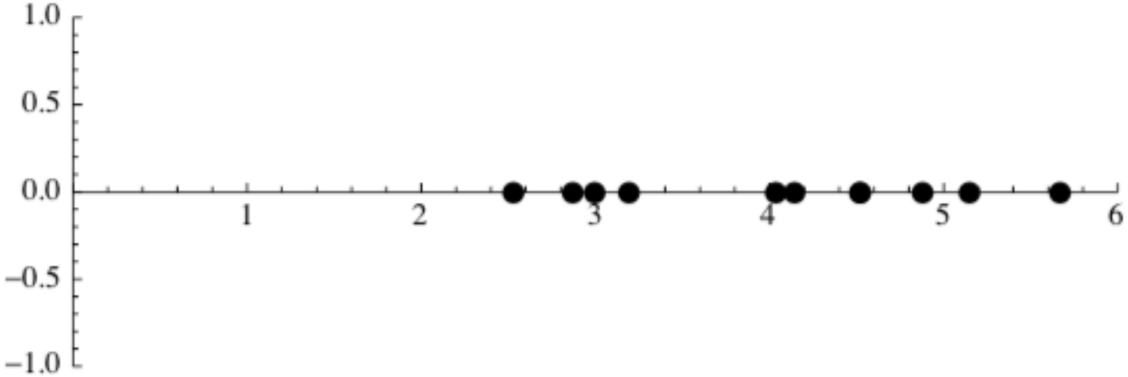
$$\mathbf{y} = U^\top \cdot \mathbf{x} = \begin{pmatrix} 0.611454 & 0.79128 \\ -0.79128 & 0.611454 \end{pmatrix} \cdot \mathbf{x},$$

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





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



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

- 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.

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

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

Solving the Equation

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

we get the two eigenvalues

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

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

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

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

and

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

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 $u_1 = 1$,

$u_1 = 1,$

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

and

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

with

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

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

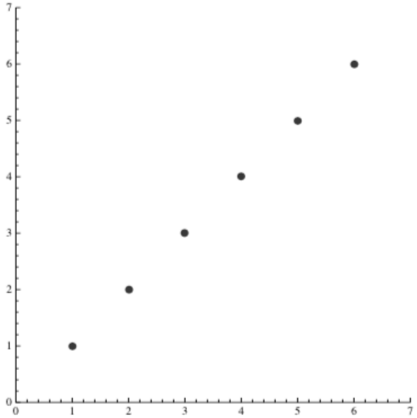
$$U = \begin{pmatrix} 0.998469 & 0.0553016 \\ -0.0553052 & 0.99847 \end{pmatrix}.$$

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

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

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

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

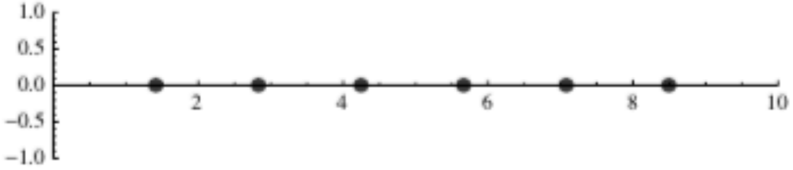
The matrix that describes the K-L transformation is given by

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

The K-L transformation maps the two dimensional data set Ψ in one dimension because λ_2 is zero (see Figure 3.35). For example, the data point $(1,1)$ is mapped on the $x - axis$

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

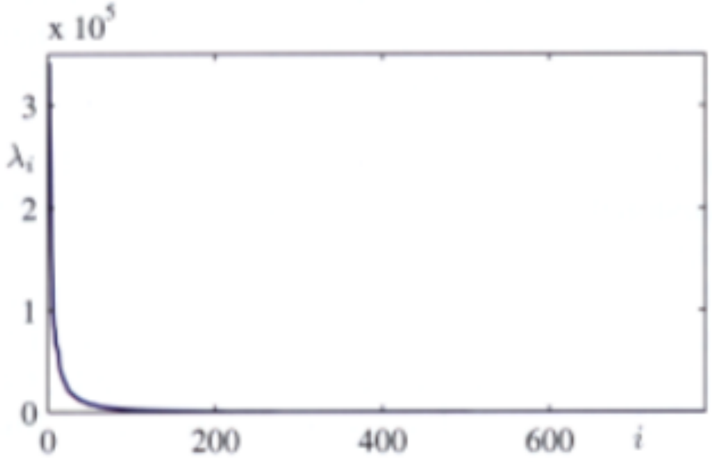
with the value $\sqrt{2} \approx 1.4142$ corresponding to the length of the vector $(1,1)$.



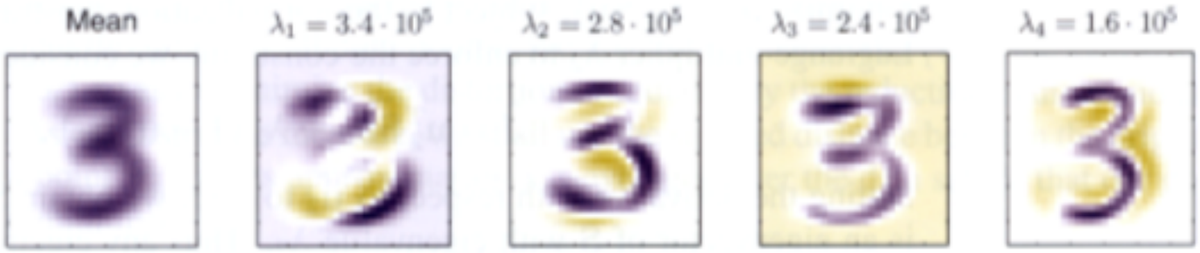
Problems

- 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

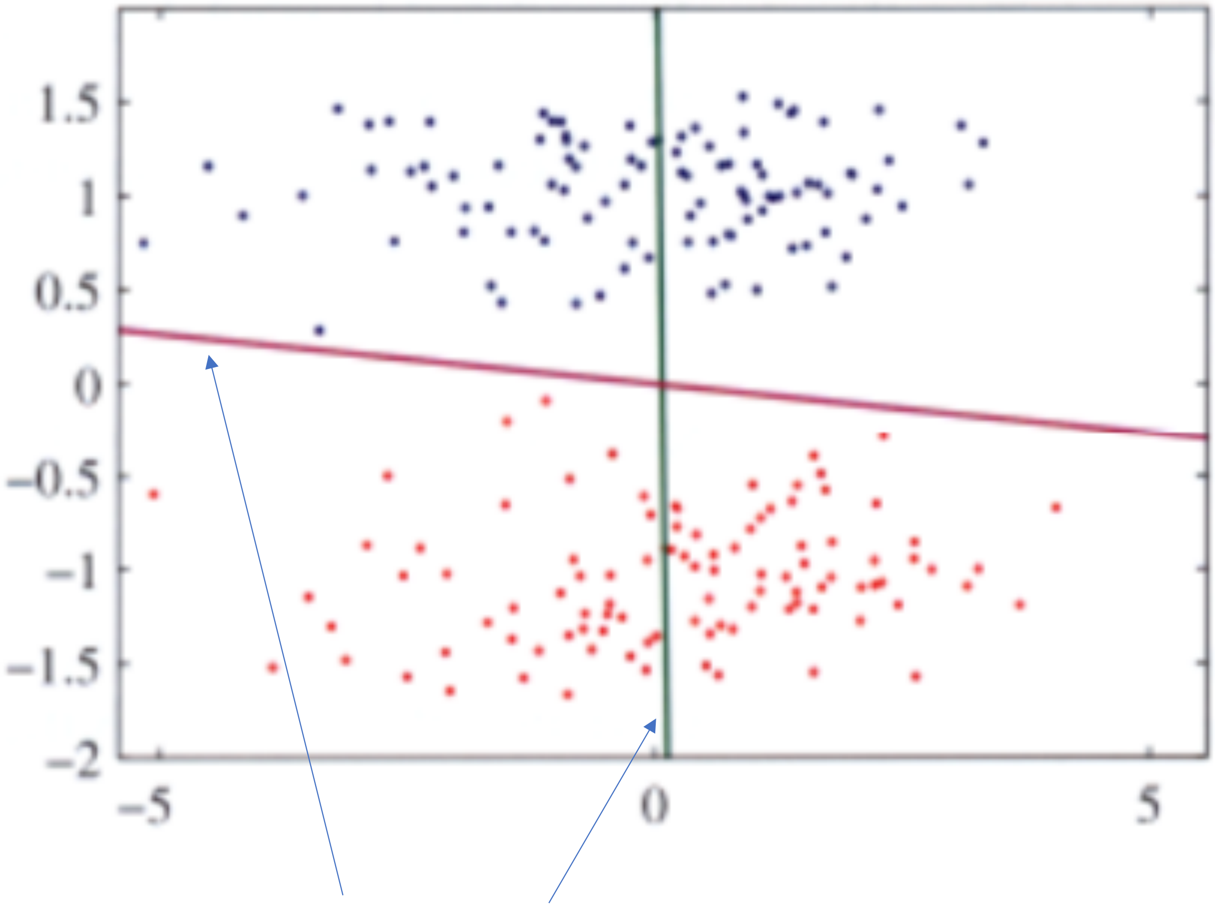
PCA: Only the images of three



- Eigenvalues


Mean
$\lambda_1 = 3.4 \cdot 10^5$
$\lambda_2 = 2.8 \cdot 10^5$
$\lambda_3 = 2.4 \cdot 10^5$
$\lambda_4 = 1.6 \cdot 10^5$

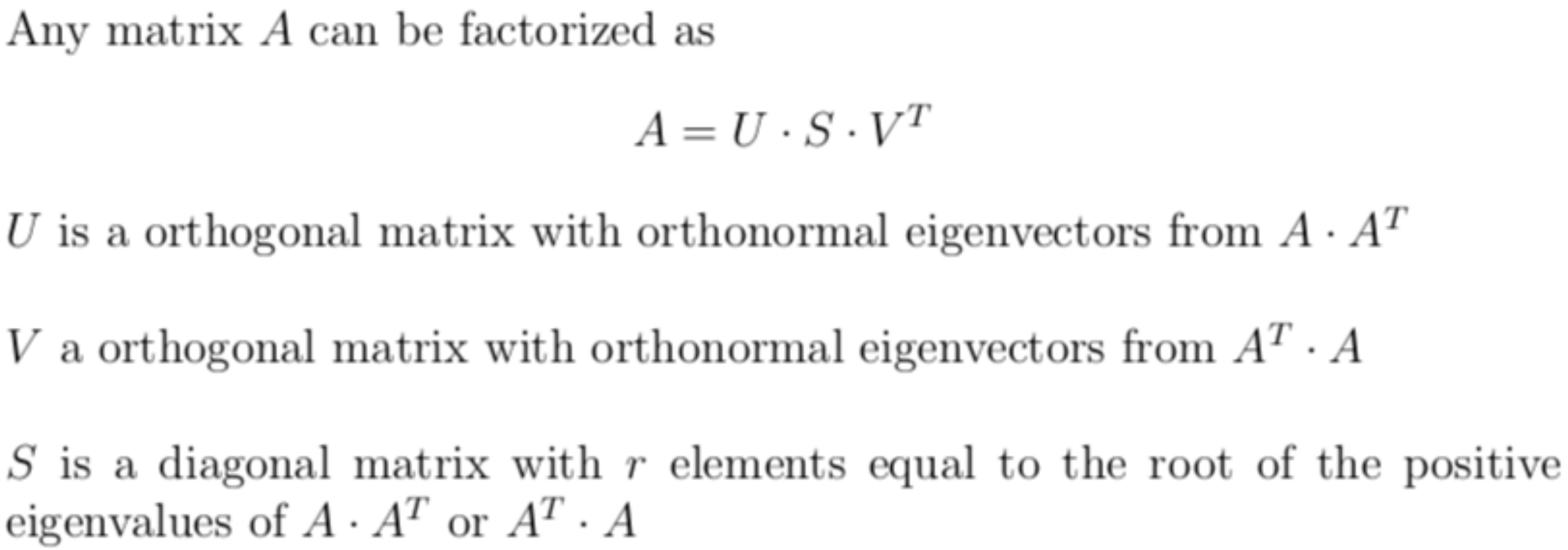
- Projection of the Eigenvectors, blue positive values, yellow negative values



- A comparison of PCA to Fisher’s linear discriminant

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

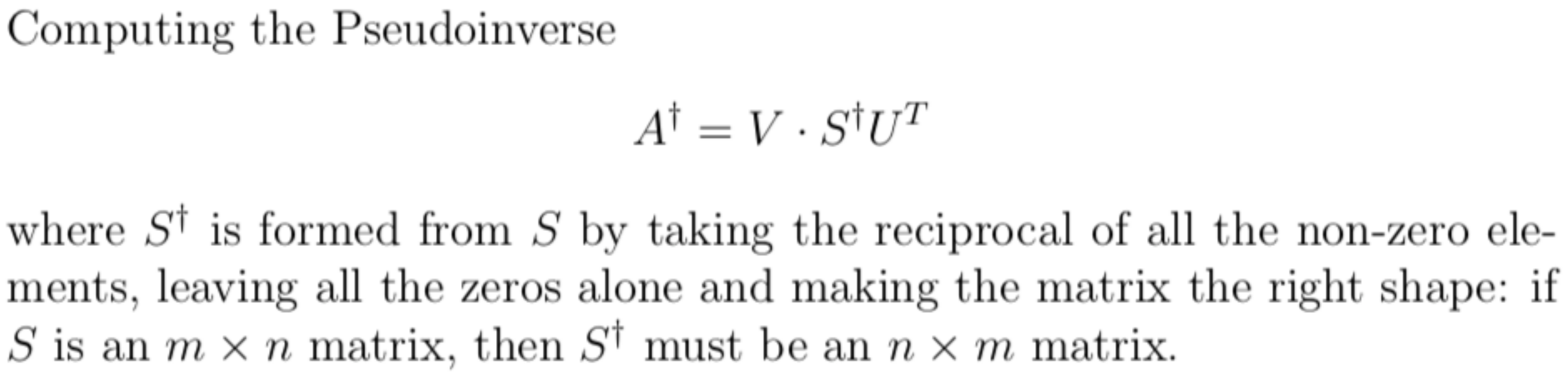
$$A = U \cdot S \cdot V^T$$

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 $A^T \cdot A$

SVD



Computing the Pseudoinverse

$$A^{\dagger} = V \cdot S^{\dagger} U^T$$

where $S^{\dagger}$ is formed from S by taking the reciprocal of all the non-zero elements, leaving all the zeros alone and making the matrix the right shape: if S is an $m \times n$ matrix, then $S^{\dagger}$ must be an $n \times m$ matrix.

$$A_{m\times n} = U_{m\times m} S_{m\times n} V^{T}_{n\times n}$$

$$\underset{m\times n}{\overset{\text{A}}{\begin{pmatrix} x_{11} & x_{12} & & x_{1n} \\ & & \ddots & \\ x_{m1} & & & x_{mn} \end{pmatrix}}} = \underset{m\times m}{\overset{\text{U}}{\begin{pmatrix} u_{11} & & u_{m1} \\ & \ddots & \\ u_{1m} & & u_{mm} \end{pmatrix}}} \underset{m\times n}{\overset{\text{S}}{\begin{pmatrix} \sigma_1 & & 0 \\ & \ddots \; \sigma_r \; \ddots & \\ 0 & & 0 \end{pmatrix}}} \underset{n\times n}{\overset{\text{V}^{\text{T}}}{\begin{pmatrix} v_{11} & & v_{1n} \\ & \ddots & \\ v_{n1} & & v_{nn} \end{pmatrix}}}$$

$$\boldsymbol{\sigma}_1 \geq \boldsymbol{\sigma}_2 \geq \ldots \geq \boldsymbol{\sigma}_{\mathbf{m}}$$

$$\begin{pmatrix} u_1 & \ldots & u_m \end{pmatrix}$$

$$A = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix}$$

$$AA^T = \begin{pmatrix} 17 & 8 \\ 8 & 17 \end{pmatrix}$$

eigenvalues: $\lambda_1 = 25$, $\lambda_2 = 9$

eigenvectors

$$u_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \quad u_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$

$$A^T A = \begin{pmatrix} 13 & 12 & 2 \\ 12 & 13 & -2 \\ 2 & -2 & 8 \end{pmatrix}$$

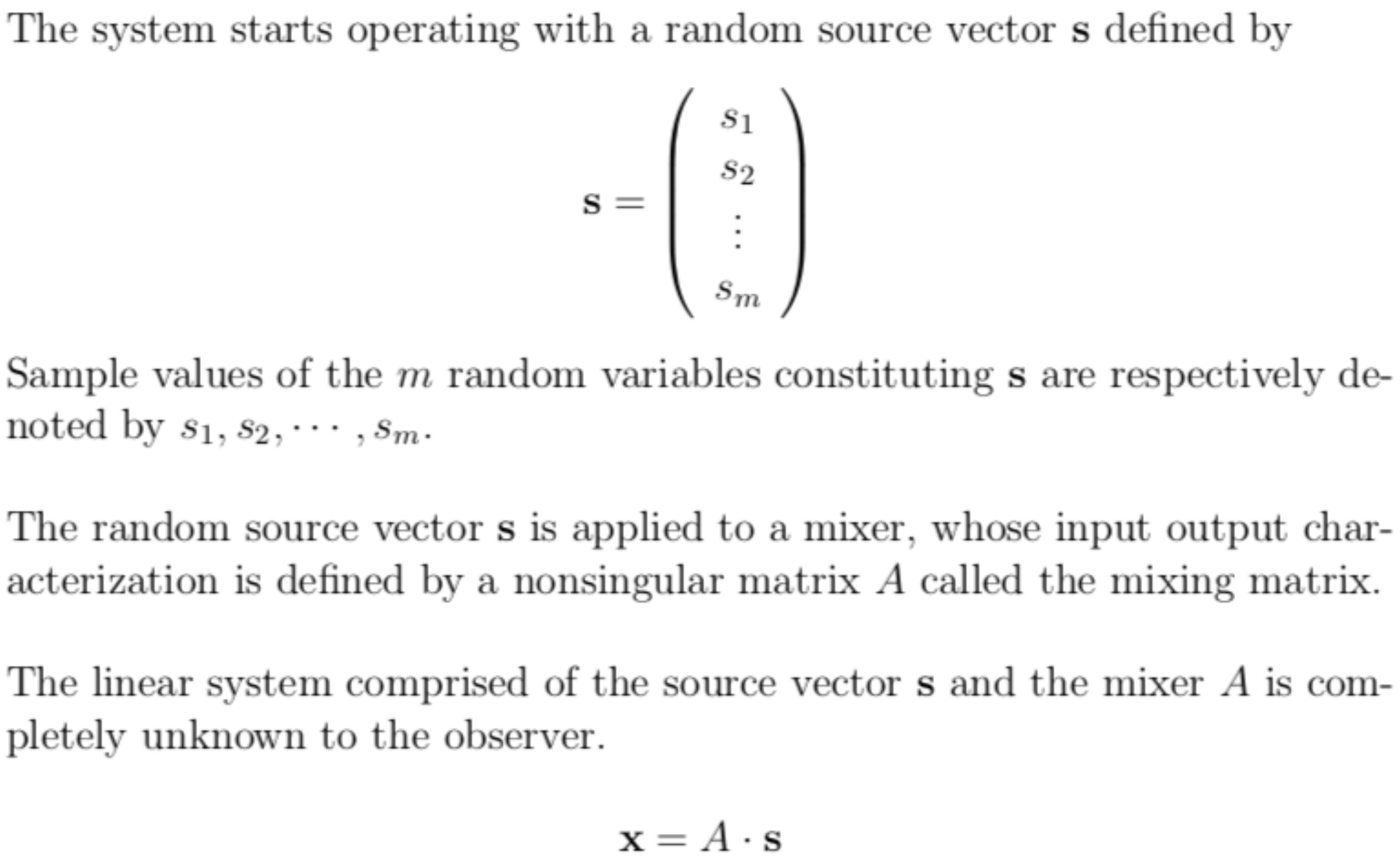
eigenvalues: $\lambda_1 = 25$, $\lambda_2 = 9$, $\lambda_3 = 0$

eigenvectors

$$v_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{pmatrix} \quad v_2 = \begin{pmatrix} 1/\sqrt{18} \\ -1/\sqrt{18} \\ 4/\sqrt{18} \end{pmatrix} \quad v_3 = \begin{pmatrix} 2/3 \\ -2/3 \\ -1/3 \end{pmatrix}$$

$$A = USV^T = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{18} & -1/\sqrt{18} & 4/\sqrt{18} \\ 2/3 & -2/3 & -1/3 \end{pmatrix}$$

Independent-Components Analysis ICA



The system starts operating with a random source vector $\mathbf{s}$ defined by

$$\mathbf{s} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{pmatrix}$$

Sample values of the m random variables constituting $\mathbf{s}$ are respectively denoted by $s_1, s_2, \cdots, s_m$.

The random source vector $\mathbf{s}$ is applied to a mixer, whose input output characterization is defined by a nonsingular matrix A called the mixing matrix.

The linear system comprised of the source vector $\mathbf{s}$ and the mixer A is completely unknown to the observer.

$$\mathbf{x} = A \cdot \mathbf{s}$$

The linear system comprised of the source vector $\mathbf{s}$ and the mixer A is completely unknown to the observer.

$$\mathbf{x} = A \cdot \mathbf{s}$$

with

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

The model is called a generative model, in the sense that it is responsible for generating the random variables $x_1, x_2, \cdots, x_m$, only they are known.

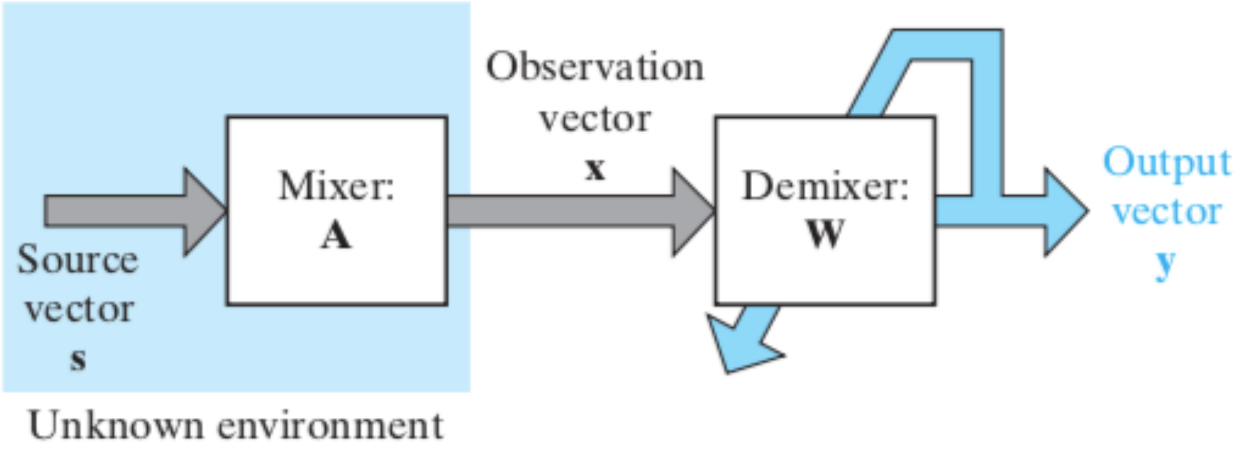
The random variables $s_1, s_2, \cdots, s_m$ representing the source are called latent variables.

The Blind Source Separation Problem

A demixer, described by an $m \times m$ demixing matrix W.

In response to the observation vector $\mathbf{x}$, the demixer produces an output defined by the random vector

$$\mathbf{y} = W \cdot \mathbf{x}$$



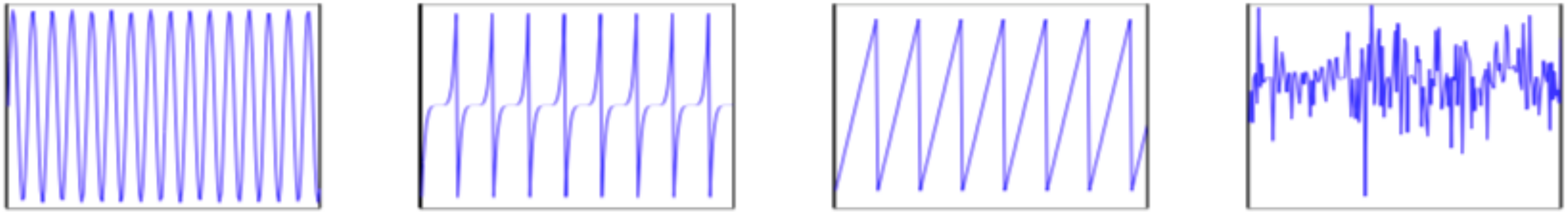
Given a set of independent realizations of the observation vector $\mathbf{x}$ resulting from an unknown linear mixing of the latent (source) variables $s_1, s_2, \cdots, s_m$, estimate the demixing matrix W such that the components of the resulting output vector $\mathbf{y}$ are as statistically independent as possible; here, the term independence should be understood in its strong statistical sense.

The demixing matrix W is carried out in an unsupervised manner.

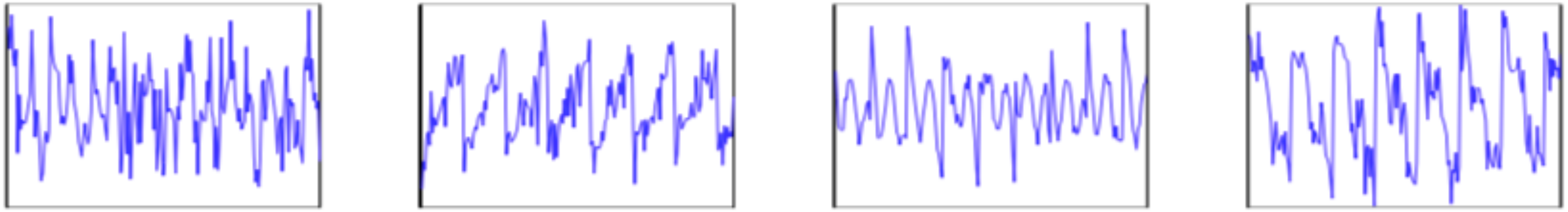
Moreover, the only information used to recover the original source vector $\mathbf{s}$ is contained in the observation vector $\mathbf{x}$

This problem is called Independent-Components Analysis ICA

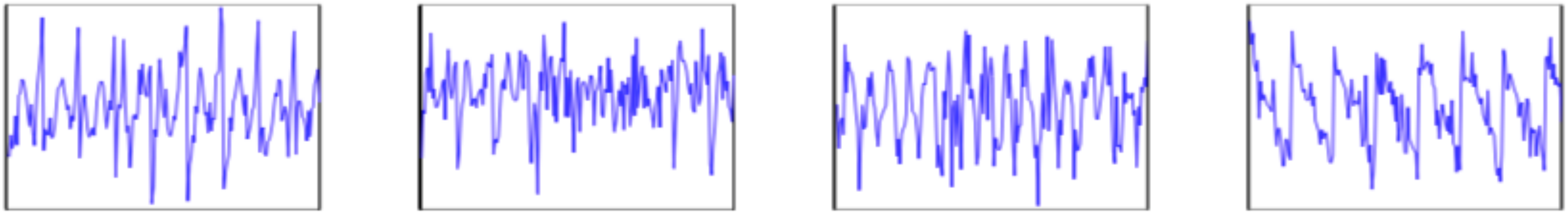
- There is a number of “source signals”:



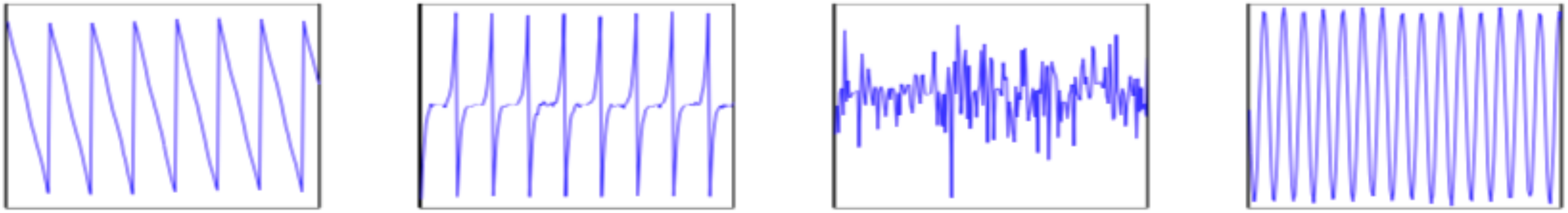
only linear mixtures of the source signals are observed:



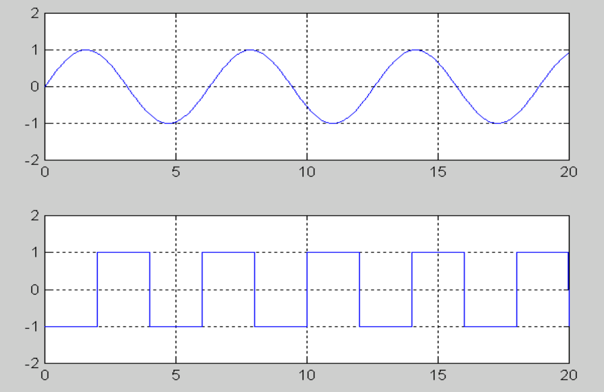
- Estimate (separate) original signals!



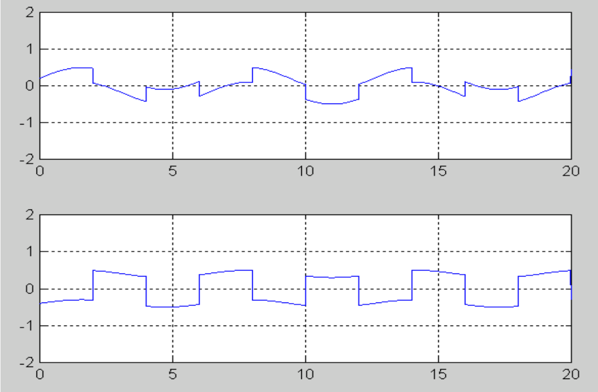
- Use information on statistical independence to recover:



Motivation



Two Independent Sources



Mixture at two Mics

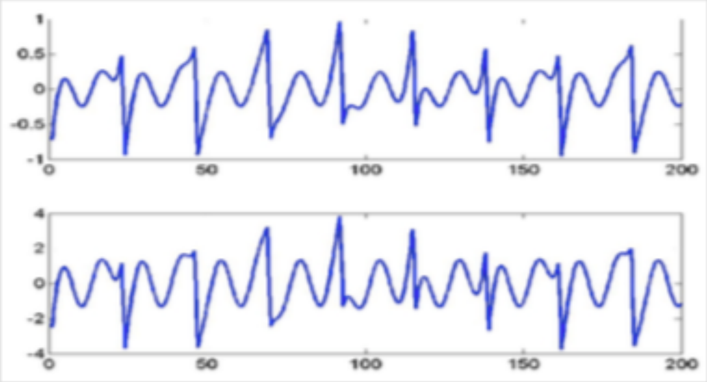
$$x_1(t) = a_{11}s_1 + a_{12}s_2$$

$$x_2(t) = a_{21}s_1 + a_{22}s_2$$

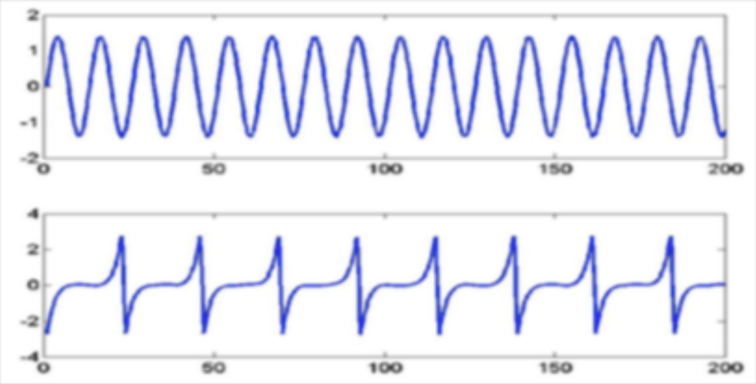
a_{IJ} ... Depend on the distances of the microphones from the speakers

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) \end{aligned}$$

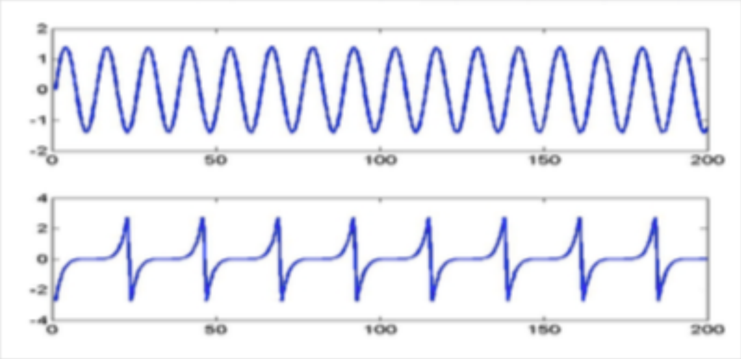
Model



Observations (Mixtures)



ICA estimated signals



original signals

Model

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) \end{aligned}$$

We observe

$$\begin{pmatrix} x_1(1) \\ x_2(1) \end{pmatrix}, \begin{pmatrix} x_1(2) \\ x_2(2) \end{pmatrix}, \ldots, \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}$$

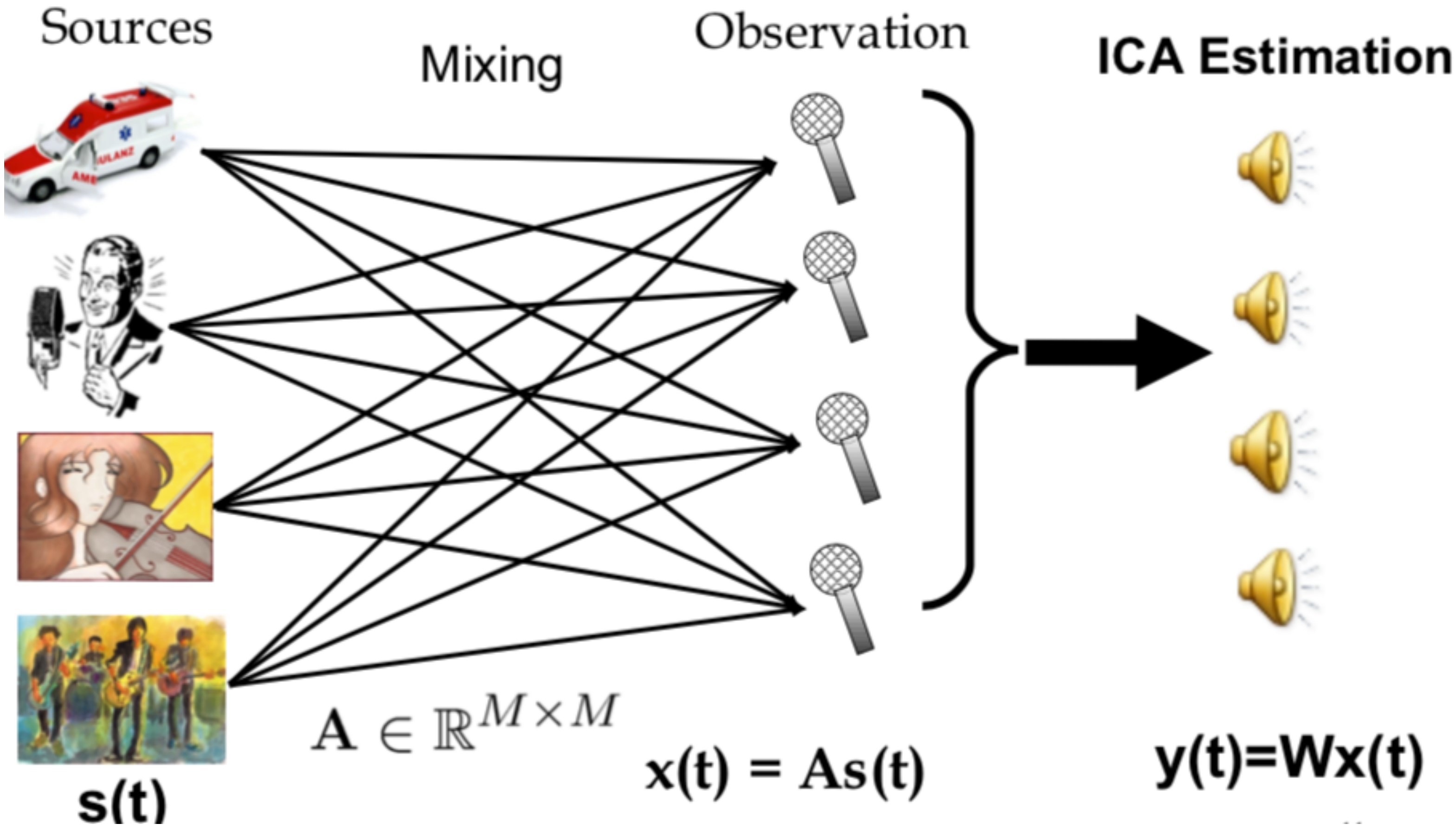
We want

$$\begin{pmatrix} s_1(1) \\ s_2(1) \end{pmatrix}, \begin{pmatrix} s_1(2) \\ s_2(2) \end{pmatrix}, \ldots, \begin{pmatrix} s_1(t) \\ s_2(t) \end{pmatrix}$$

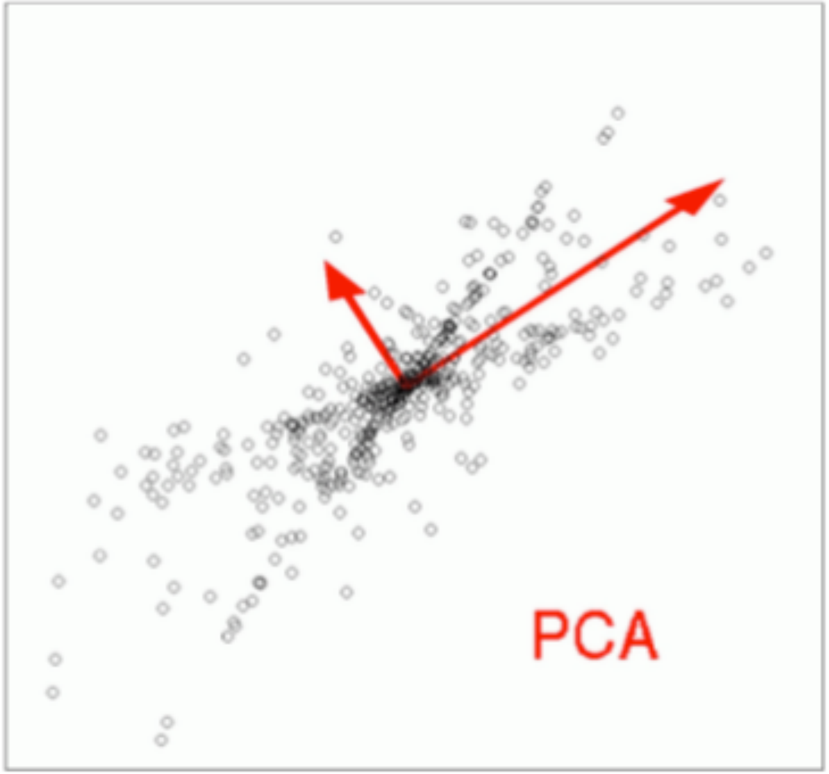
But we don't know $\{a_{ij}\}$, nor $\{s_i(t)\}$

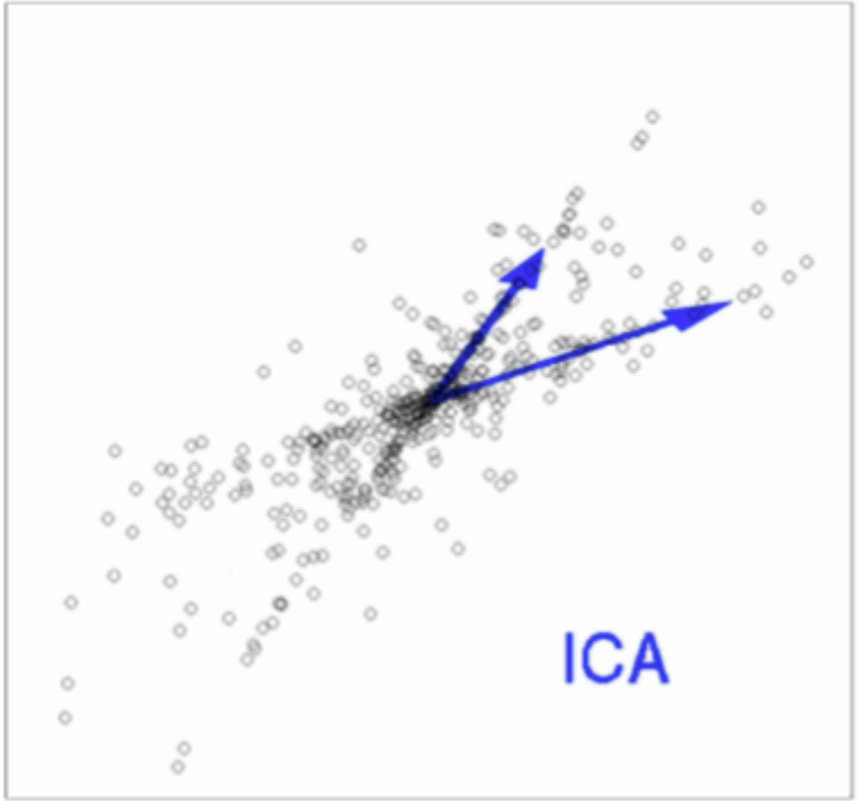
Goal: Estimate $\{s_i(t)\}$, (and also $\{a_{ij}\}$)

The Cocktail Party Problem

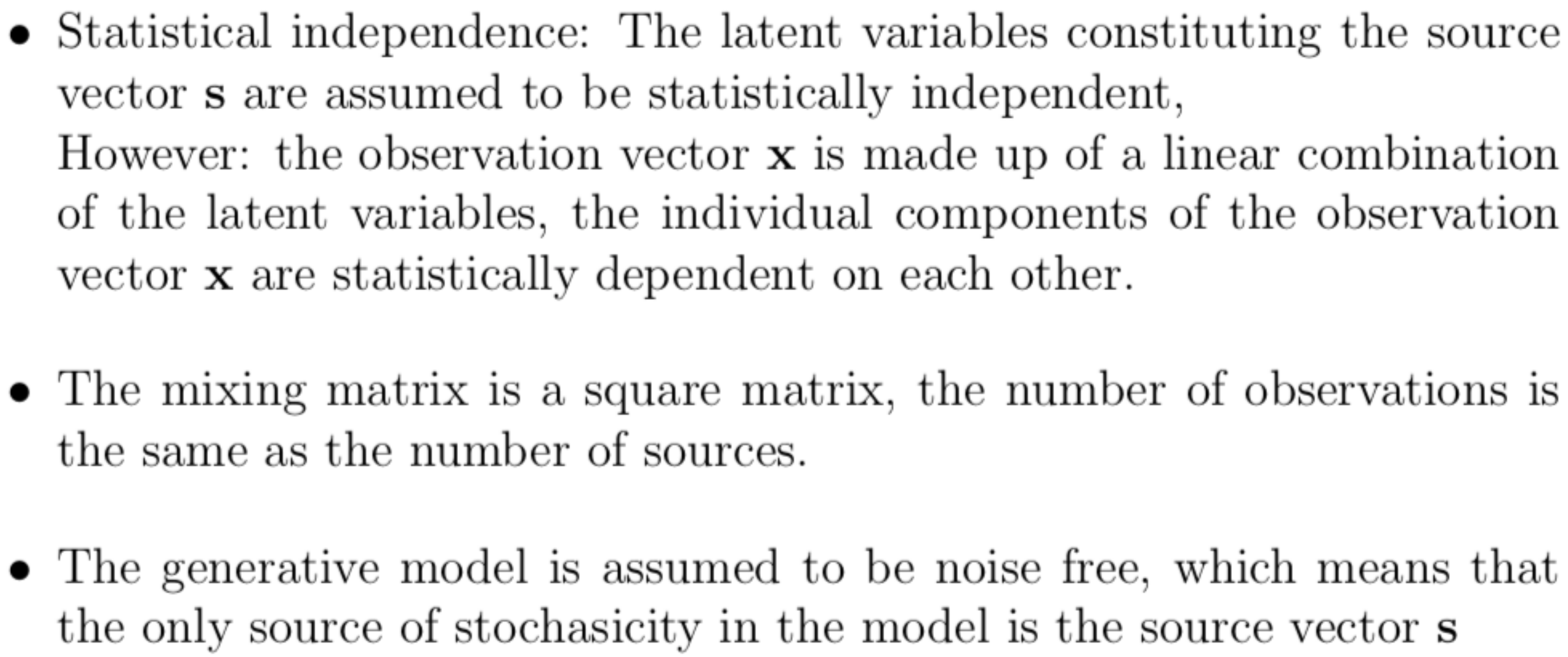


PCA vs ICA





Assumptions

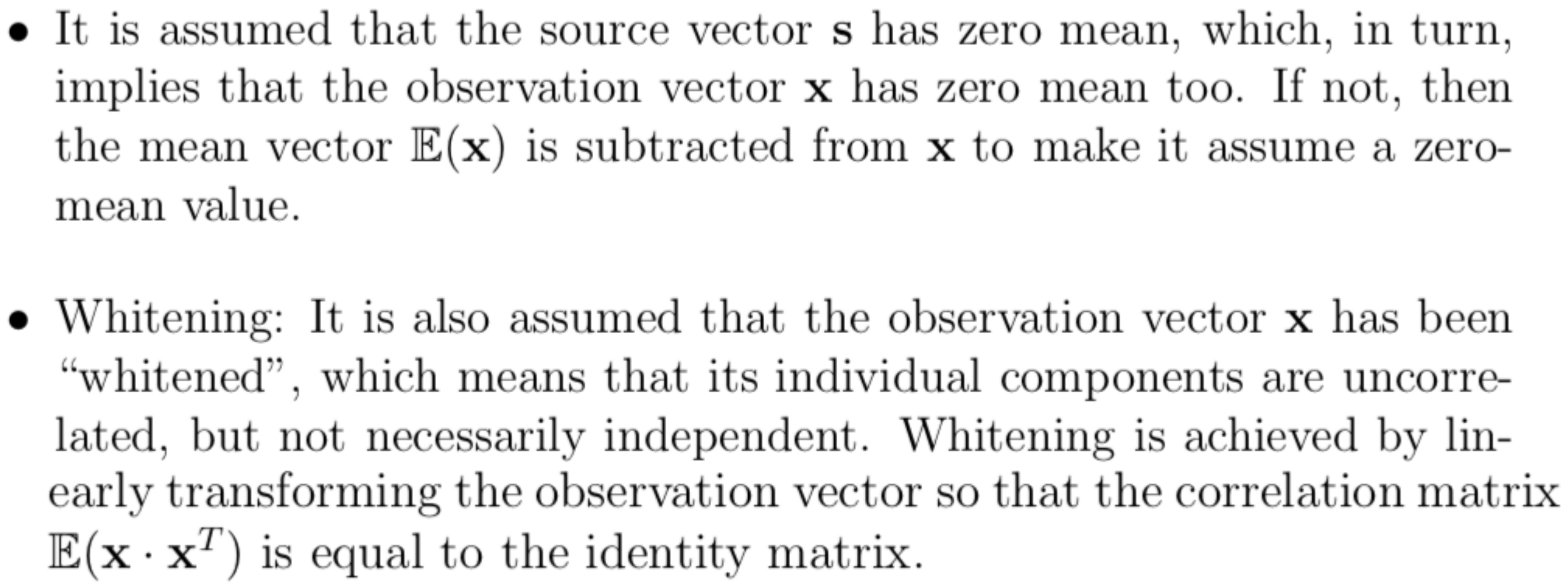


- Statistical independence: The latent variables constituting the source vector **s** are assumed to be statistically independent,
  However: the observation vector **x** is made up of a linear combination of the latent variables, the individual components of the observation vector **x** are statistically dependent on each other.

- The mixing matrix is a square matrix, the number of observations is the same as the number of sources.

- The generative model is assumed to be noise free, which means that the only source of stochasicity in the model is the source vector **s**

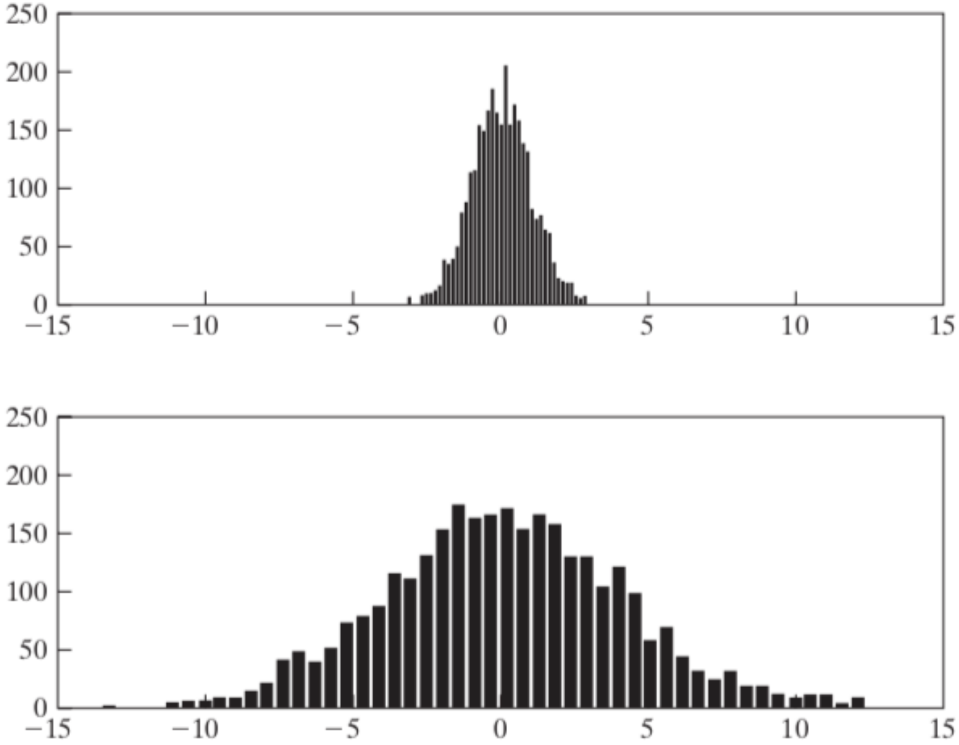
Assumptions

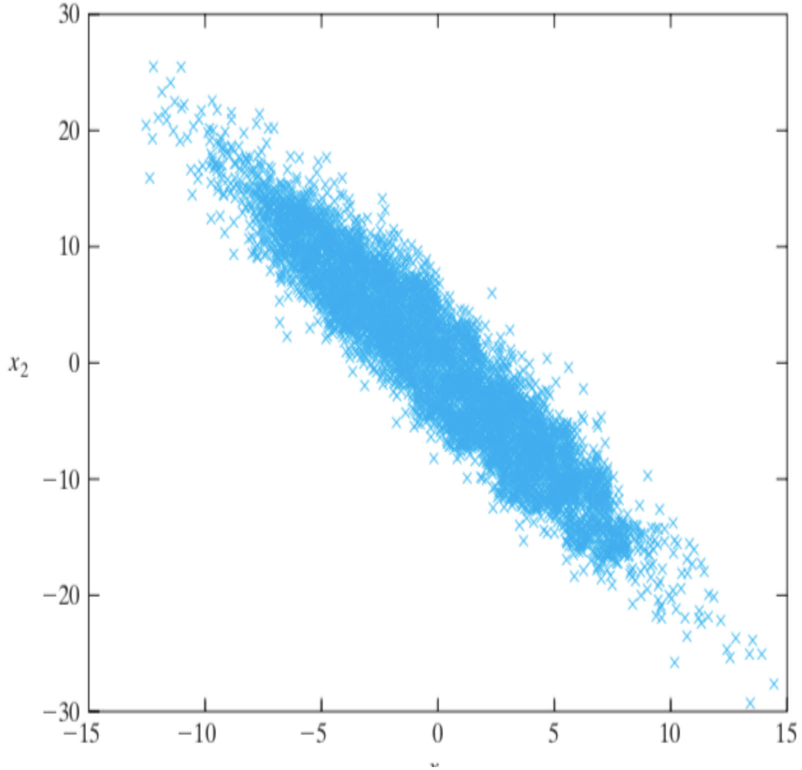


- It is assumed that the source vector $\mathbf{s}$ has zero mean, which, in turn, implies that the observation vector $\mathbf{x}$ has zero mean too. If not, then the mean vector $\mathbb{E}(\mathbf{x})$ is subtracted from $\mathbf{x}$ to make it assume a zero-mean value.

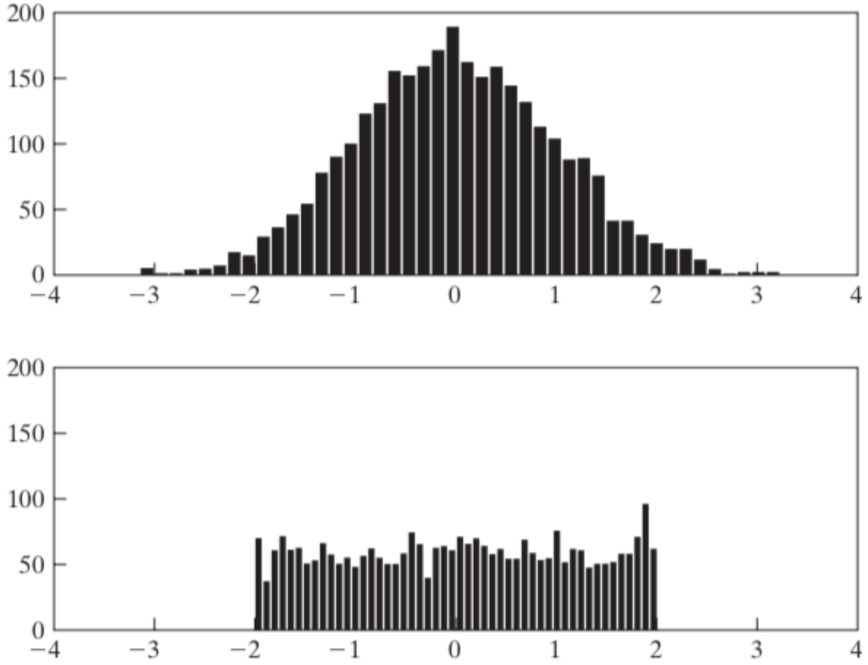
- Whitening: It is also assumed that the observation vector $\mathbf{x}$ has been "whitened", which means that its individual components are uncorrelated, but not necessarily independent. Whitening is achieved by linearly transforming the observation vector so that the correlation matrix $\mathbb{E}(\mathbf{x} \cdot \mathbf{x}^T)$ is equal to the identity matrix.

Example

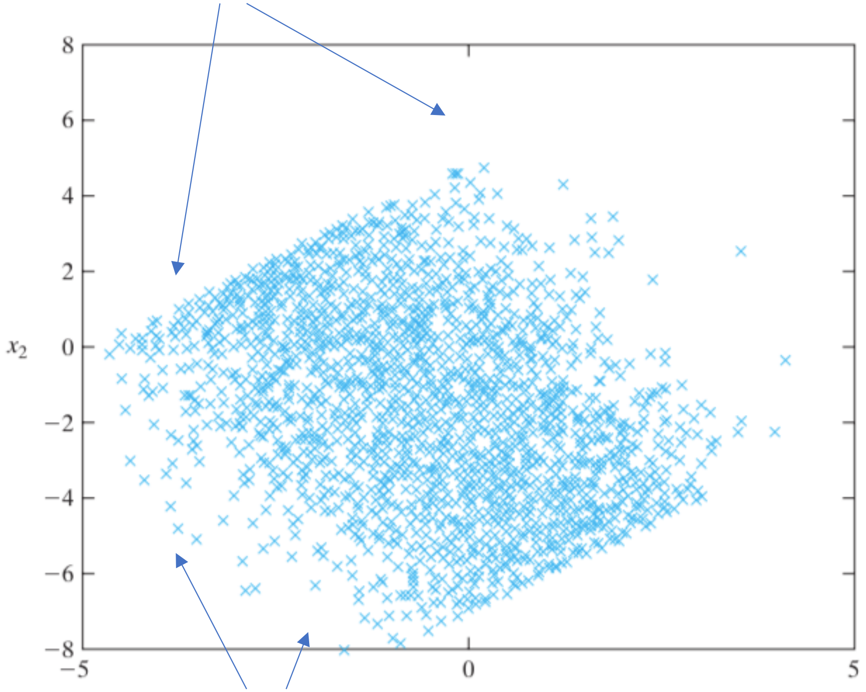

250
200
150
100
50
0
−15
−10
−5
0
5
10
15
250
200
150
100
50
0
−15
−10
−5
0
5
10
15


30
20
10
0
−10
−20
−30
x_2
−15
−10
−5
0
5
10
15

Example



slope is +1



slope is -2

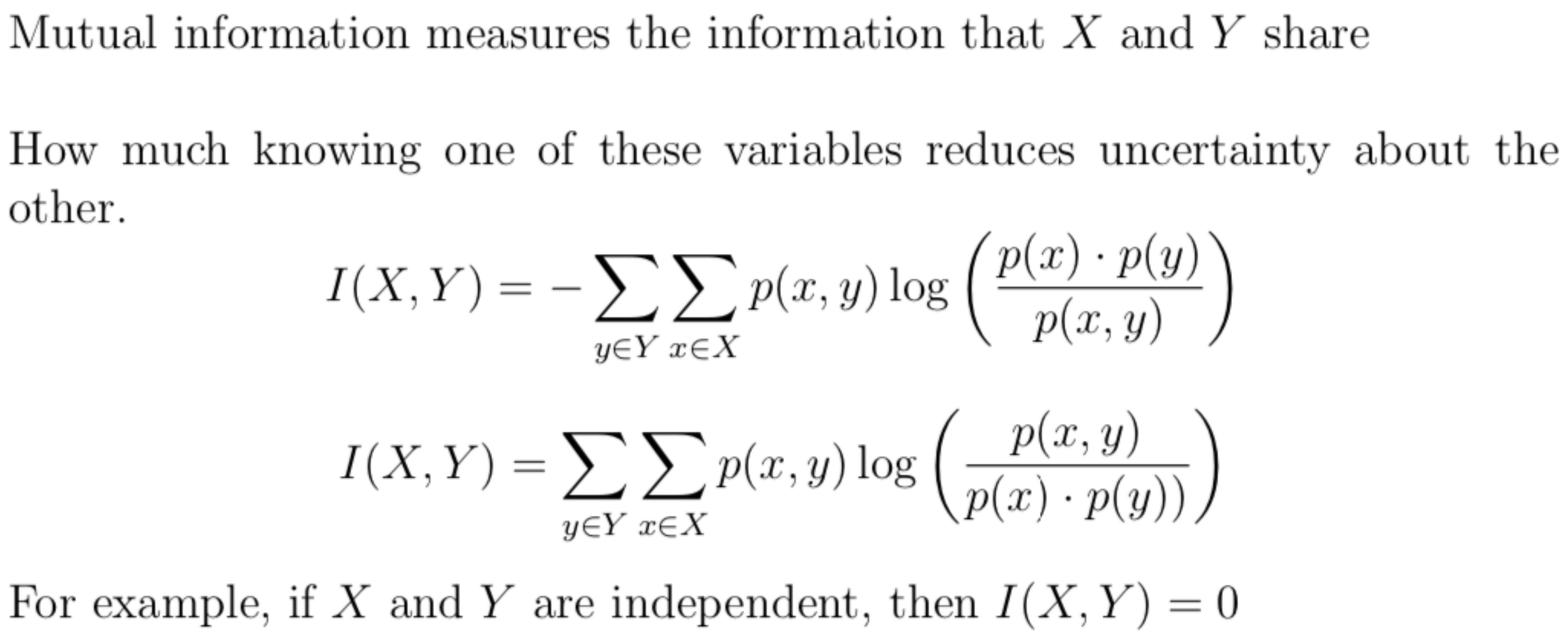
Non-Gaussianity of Sources: A Necessary Requirement for ICA (Except Possibly for One Source)

How should the information content in the observation vector X manifest itself for the separability of source signals to be feasible?

- The source signals $s_1, s_2, \cdots, s_m$ must be non-Gaussian.
- At the very most, only a single source s_k is permitted to have a Gaussian distribution.

Algorithms

- ICA Algorithms rooted in minimization of mutual information



Mutual information measures the information that X and Y share

How much knowing one of these variables reduces uncertainty about the other.

$$I(X,Y) = -\sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x)\cdot p(y)}{p(x,y)}\right)$$

$$I(X,Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)\cdot p(y))}\right)$$

For example, if X and Y are independent, then $I(X,Y) = 0$

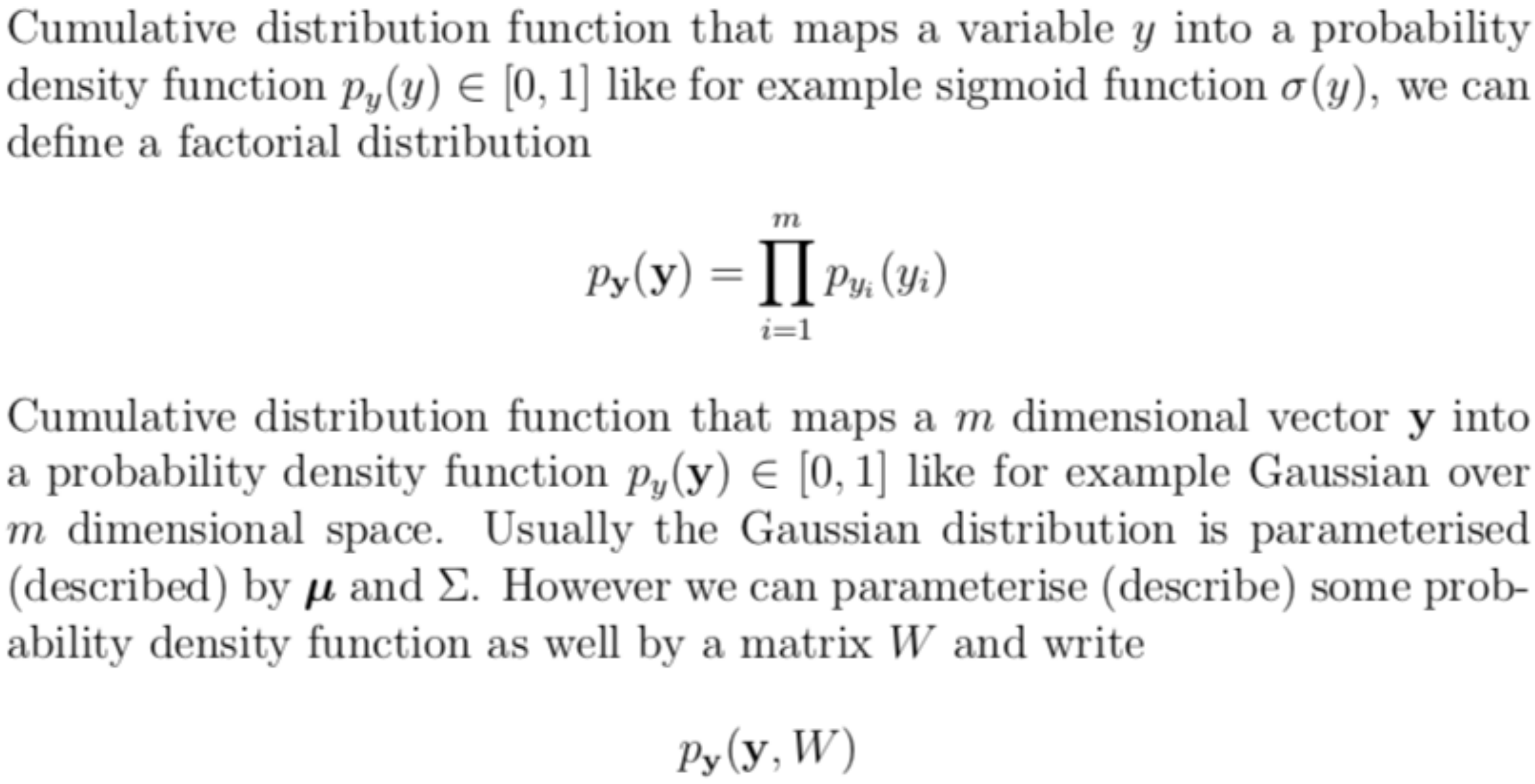
The Mutual information $I(X,Y)$ between X and Y is equal to the Kullback-Leibler divergence between the joint probability density function $p(x,y)$ and the product of the probability functions $p(x)$ and $p(y)$

$$KL(p_{x,y}||p_x, p_y) = I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x) \cdot p(y))} \right)$$

$$KL(p_{x,y}||p_x, p_y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log p(x,y) - \sum_{y \in Y} \sum_{x \in X} p(x,y) \log(p(x) \cdot p(y))$$

$$KL(p_{x,y}||p_x, p_y) = -H(X,Y) - \sum_{y \in Y} \sum_{x \in X} p(x,y) \log(p(x) \cdot p(y))$$

Cumulative Distribution Function



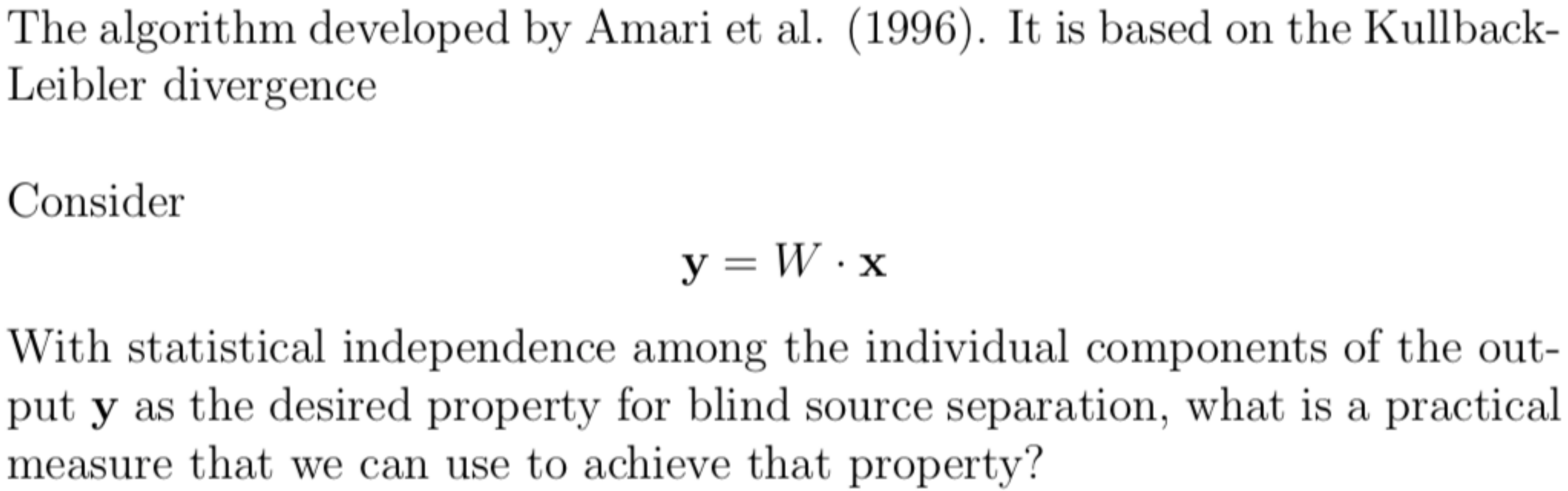
Cumulative distribution function that maps a variable y into a probability density function $p_y(y) \in [0,1]$ like for example sigmoid function $\sigma(y)$, we can define a factorial distribution

$$p_{\mathbf{y}}(\mathbf{y}) = \prod_{i=1}^{m} p_{y_i}(y_i)$$

Cumulative distribution function that maps a m dimensional vector $\mathbf{y}$ into a probability density function $p_y(\mathbf{y}) \in [0,1]$ like for example Gaussian over m dimensional space. Usually the Gaussian distribution is parameterised (described) by $\boldsymbol{\mu}$ and Σ. However we can parameterise (describe) some probability density function as well by a matrix W and write

$$p_{\mathbf{y}}(\mathbf{y}, W)$$

Natural Gradient Learning for ICA



The algorithm developed by Amari et al. (1996). It is based on the Kullback-Leibler divergence

Consider

$$\mathbf{y} = W \cdot \mathbf{x}$$

With statistical independence among the individual components of the output $\mathbf{y}$ as the desired property for blind source separation, what is a practical measure that we can use to achieve that property?

With

$$p_{\mathbf{y}}(\mathbf{y}, W)$$

denote the probability density function of the output $\mathbf{y}$, parameterised by the demixing matrix W

The factorial distribution is defined by

$$p_{\mathbf{y}}(\mathbf{y}) = \prod_{i=1}^{m} p_{y_i}(y_i)$$

and is not parameterised.

Given a random vector $\mathbf{y}$ representing a linear combination of m independent source signals:

The transformation of the observation vector $\mathbf{x}$ into a new random vector $\mathbf{y}$ should be carried out in such a way that

The Kullback-Leibler sdivergence between the parameterized probability density function $p_{\mathbf{y}}(\mathbf{y}, W)$ and the corresponding factorial distribution $p_{\mathbf{y}}(\mathbf{y})$ is minimized with respect to the unknown parameter matrix W.

Relative Entropy

$$R(W) = \sum_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{y}, W) \cdot \log \left(\frac{p_{\mathbf{y}}(\mathbf{y}, W)}{p_{\mathbf{y}}(\mathbf{y})} \right)$$

$$R(W) = \sum_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{y}, W) \cdot \log \left(\frac{p_{\mathbf{y}}(\mathbf{y}, W)}{\prod_{i=1}^{m} p_{y_i}(y_i)} \right)$$

$$R(W) = -H(\mathbf{y}) + \sum_{i=1}^{m} H(y_i)$$

$$R(W) = -H(W \cdot \mathbf{x}) + \sum_{i=1}^{m} H(y_i)$$

$$R(W) = -H(\mathbf{x}) - \log|det(W)| + \sum_{i=1}^{m} H(y_i)$$

$$R(W) = -H(\mathbf{x}) - \log|det(W)| - \sum_{i=1}^{m} \mathbb{E}(\log p_{y_i}(y_i))$$

We redefine a contrast function for stochastic gradient decent with W:

- We ignore with stochastic gradient the operator $\mathbb{E}$ and
- We ignore independent component $H(\mathbf{x})$ since it is independent on W

$$\rho(W) = -\log|det(W)| - \sum_{i=1}^{m} \log p_{y_i}(y_i)$$

$$\rho(W) = -\log|det(W)| - \sum_{i=1}^{m} \log p_{y_i}(y_i)$$

$$\nabla\rho(W) = -\frac{\partial}{\partial W}\log|det(W)| - \frac{\partial}{\partial W}\sum_{i=1}^{m} \log p_{y_i}(y_i)$$

$$\nabla\rho(W) = -(W^T)^{-1} - \frac{\partial}{\partial W}\sum_{i=1}^{m} \log p_{y_i}(y_i)$$

Since

$$y_i = \mathbf{w}_i^T \cdot \mathbf{x}$$

and

$$\frac{\partial \log(p_{y_i}(\mathbf{w}_i^T \cdot \mathbf{x}))}{\partial \mathbf{w}_i} = \frac{\partial y_i}{\partial \mathbf{w}_i} \cdot \frac{\partial}{\partial y_i} \cdot \log(p_{y_i}(y_i)$$

$$\frac{\partial y_i}{\partial \mathbf{w}_i} \cdot \frac{\partial}{\partial y_i} \cdot \log(p_{y_i}(y_i) = \frac{\partial y_i}{\partial \mathbf{w}_i} \cdot \frac{p(y_i)'}{p(y_i)} = \mathbf{x} \cdot \frac{p(y_i)'}{p(y_i)}$$

We define

$$\frac{p(y_i)'}{p(y_i)} \cdot \mathbf{x} = -\phi_i(y_i) \cdot \mathbf{x}$$

with

$$\Phi(\mathbf{y}) = \begin{pmatrix} \phi_1(y_1) \\ \phi_1(y_2) \\ \vdots \\ \phi_m(y_m) \end{pmatrix}$$

we get

$$\nabla \rho(W) = -(W^T)^{-1} + \Phi(\mathbf{y}) \cdot \mathbf{x}^T$$

And we get

$$\Delta W = -\eta \nabla \rho(W)$$

$$\Delta W = \eta((W^T)^{-1} - \phi(\mathbf{y}) \cdot \mathbf{x}^T)$$

and

$$\mathbf{y}^T = \mathbf{x}^T \cdot W^T$$

$$\Delta W = \eta(I - \phi(\mathbf{y}) \cdot \mathbf{x}^T \cdot W^T)(W^T)^{-1}$$

$$\Delta W = \eta(I - \phi(\mathbf{y}) \cdot \mathbf{y}^T)(W^T)^{-1}$$

Learning Rule

So the learning rule is then

$$\mathbf{y}(n) = W(n) \cdot \mathbf{x}(n)$$

$$W(n+1) = W(n) + \eta(n) \cdot (I - \phi(\mathbf{y}(n)) \cdot \mathbf{y}^T(n)) \cdot ((W^T)(n))^{-1}$$

with

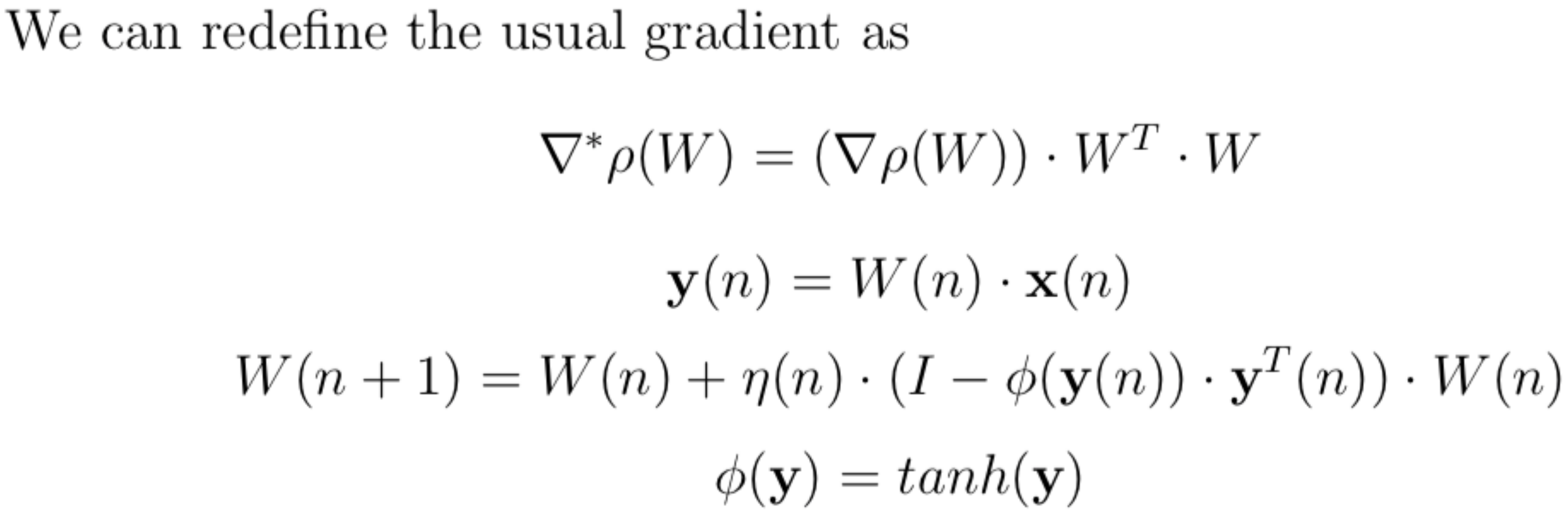
$$I - \phi(\mathbf{y}(n)) \cdot \mathbf{y}^T(n)$$

being the correction term.

According to the ICA robustness theorem for each component of the vector

$$\phi(\mathbf{y}) = tanh(\mathbf{y})$$

Learning Rule



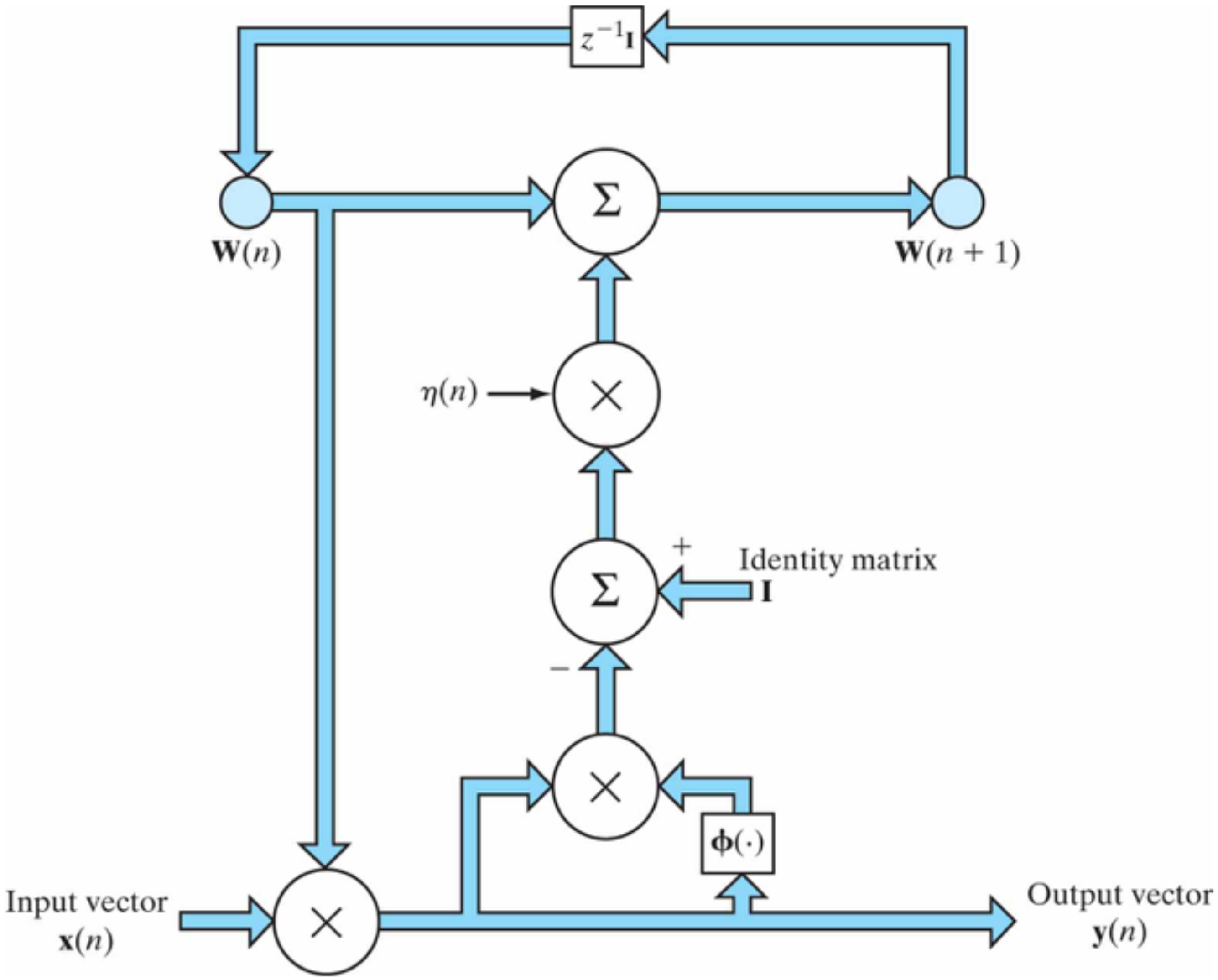
We can redefine the usual gradient as

$$\nabla^* \rho(W) = (\nabla \rho(W)) \cdot W^T \cdot W$$

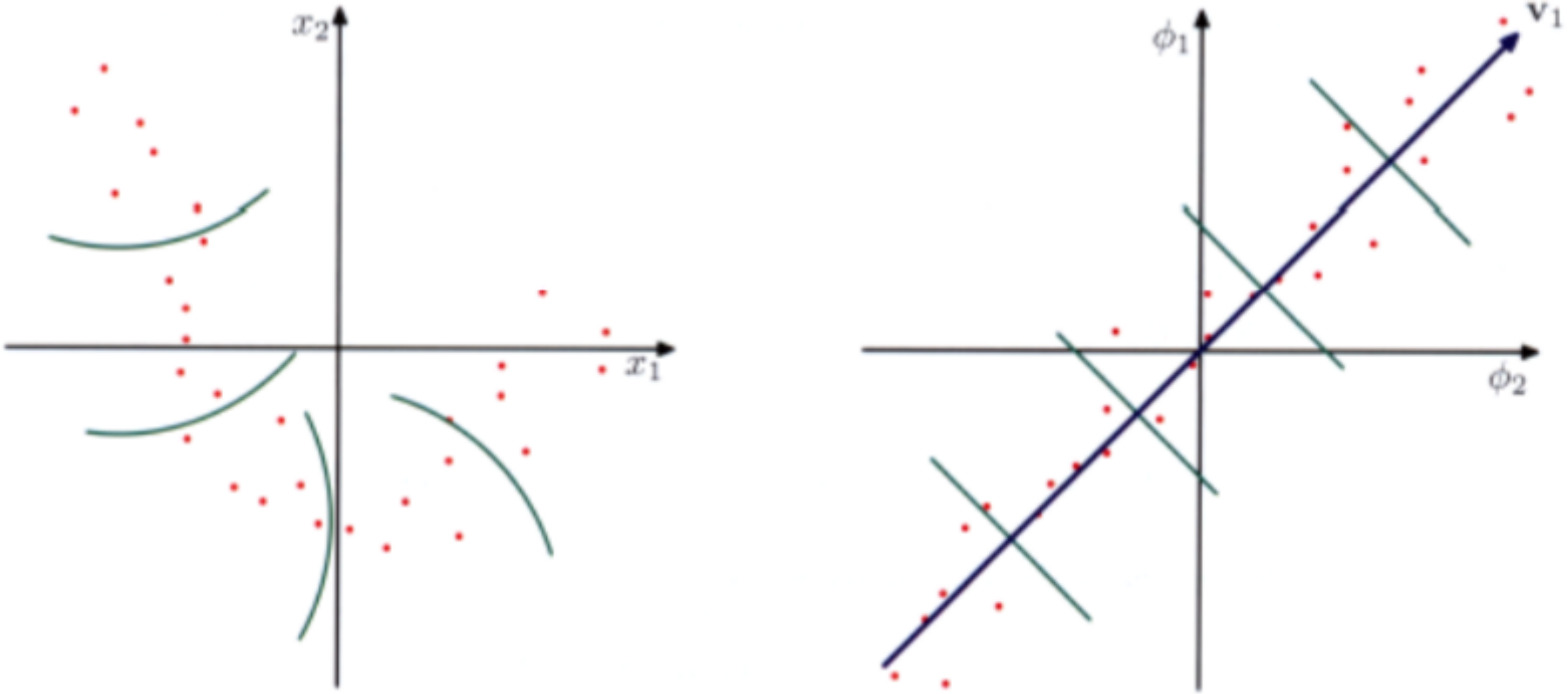
$$\mathbf{y}(n) = W(n) \cdot \mathbf{x}(n)$$

$$W(n+1) = W(n) + \eta(n) \cdot (I - \phi(\mathbf{y}(n)) \cdot \mathbf{y}^T(n)) \cdot W(n)$$

$$\phi(\mathbf{y}) = tanh(\mathbf{y})$$

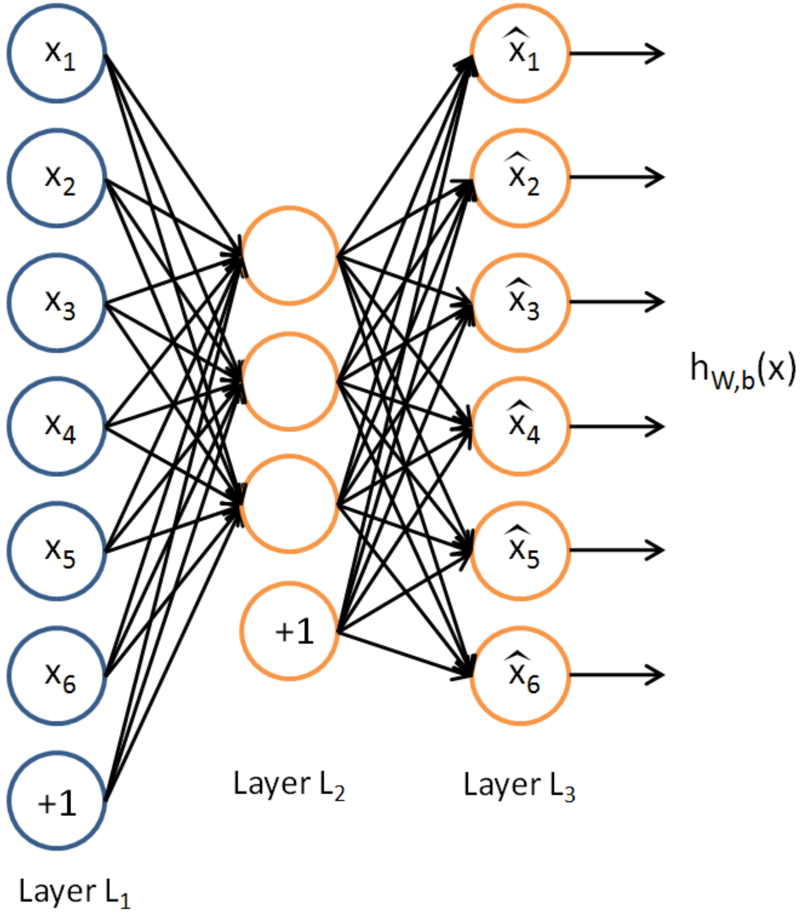

$z^{-1}\mathbf{I}$
$\mathbf{W}(n)$
Σ
$\mathbf{W}(n+1)$
$\eta(n)$
$\times$
$+$
Identity matrix
$\mathbf{I}$
Σ
$-$
$\times$
$\boldsymbol{\phi}(\cdot)$
Input vector
$\mathbf{x}(n)$
$\times$
Output vector
$\mathbf{y}(n)$

Kernel PCA

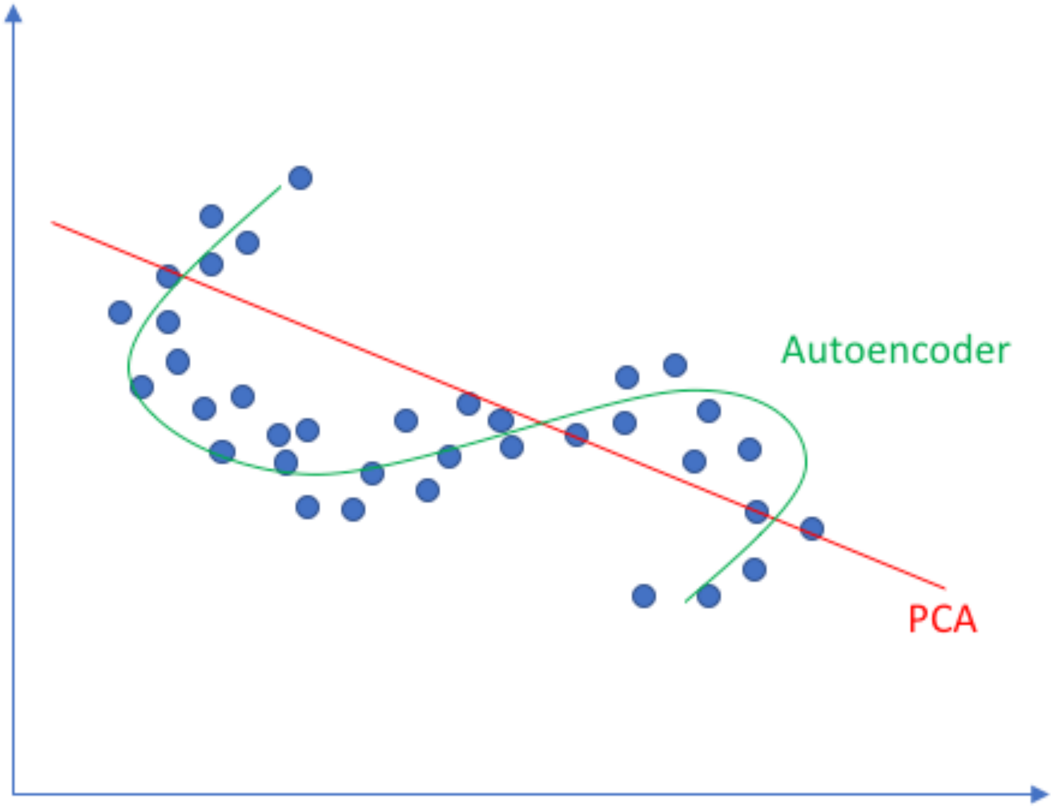


Other Methods for Dimension Reduction

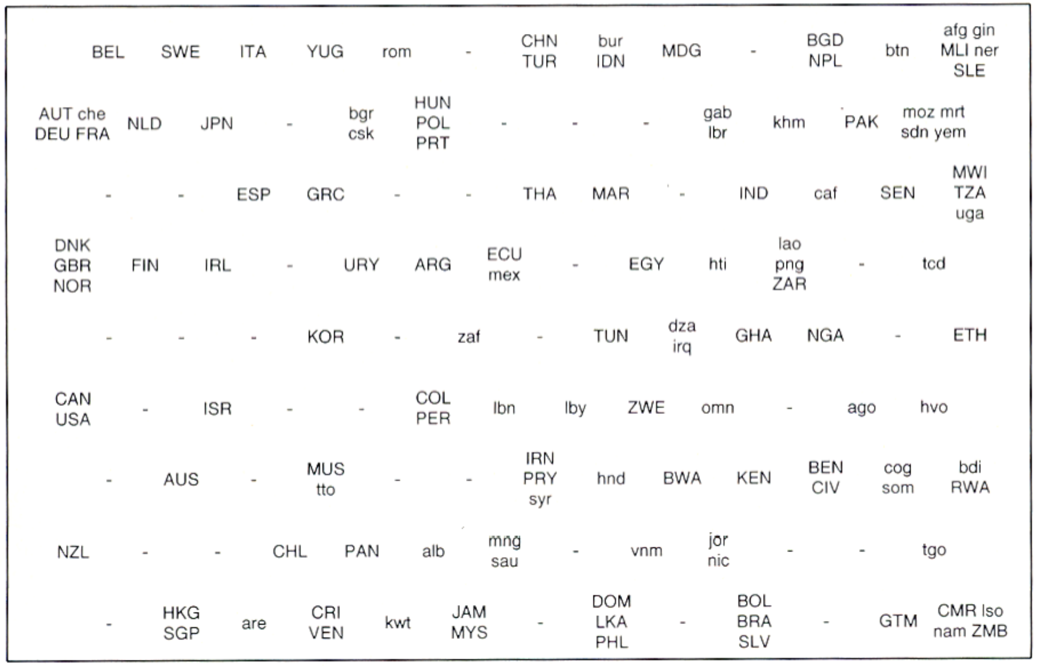
- Unsupervised Learning
  - Data: no labels!
  - Goal: Learn the structure of the data
- Traditionally, autoencoders were used for dimensionality reduction or feature learning.



Linear vs nonlinear dimensionality reduction

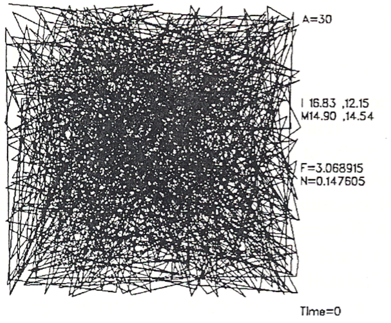


Kohonen Maps

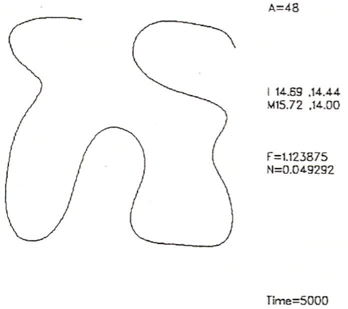


BEL	SWE	ITA	YUG	rom	-	CHN TUR	bur IDN	MDG	-	BGD NPL	btn	afg gin MLI ner SLE
AUT che DEU FRA	NLD	JPN	-	bgr csk	HUN POL PRT	-	-	-	gab lbr	khm	PAK	moz mrt sdn yem
-	-	ESP	GRC	-	-	THA	MAR	-	IND	caf	SEN	MWI TZA uga
DNK GBR NOR	FIN	IRL	-	URY	ARG	ECU mex	-	EGY	hti	lao png ZAR	-	tcd
-	-	-	KOR	-	zaf	-	TUN	dza irq	GHA	NGA	-	ETH
CAN USA	-	ISR	-	-	COL PER	lbn	lby	ZWE	omn	-	ago	hvo
-	AUS	-	MUS tto	-	-	IRN PRY syr	hnd	BWA	KEN	BEN CIV	cog som	bdi RWA
NZL	-	-	CHL	PAN	alb	mng sau	-	vnm	jor nic	-	-	tgo
-	HKG SGP	are	CRI VEN	kwt	JAM MYS	-	DOM LKA PHL	-	BOL BRA SLV	-	GTM	CMR lso nam ZMB

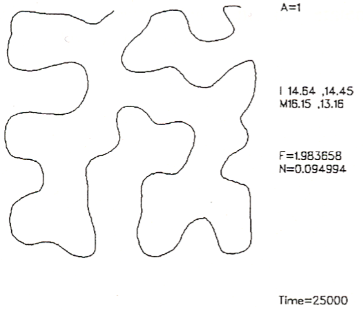
1)



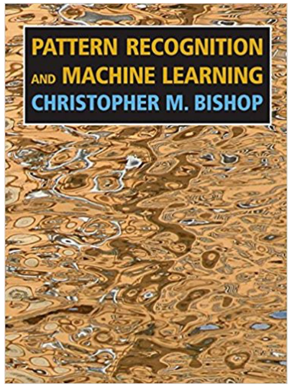
2)



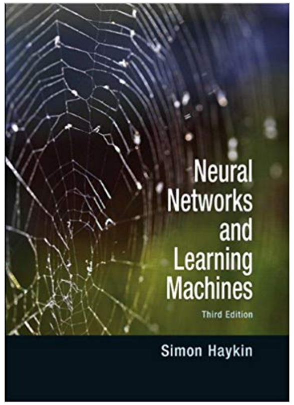
3)



Literature

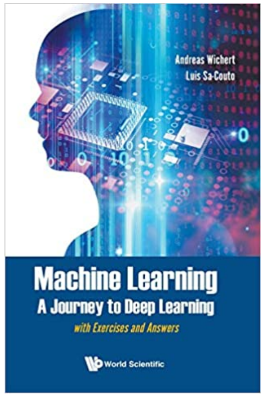


- Christopher M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer 2006
  - Chapter 12

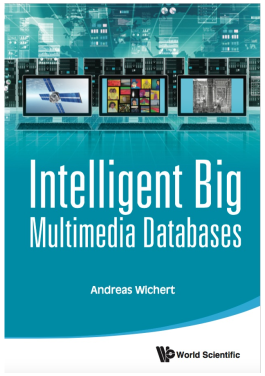


- Simon O. Haykin, Neural Networks and Learning Machine, (3rd Edition), Pearson 2008
  - Chapter 10

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*