Lecture 3:

Linear Methods of Machine Learning:

2) Independent component analysis (ICA)

1. Independent Component Analysis (ICA) - introduction
2. ICA - Cocktail party problem
3. ICA - task definition
4. ICA versus PCA
5. ICA as maximization of nongaussianity
6. ICA as Likelihood maximization
7. ICA as minimization of Mutual Information

Slide from Lecture 1

All original (p-dimensional) features (components of high-dimensional vectors) are actually be governed by a few q, $q < p$, simple variables ("hidden causes", "latent variables", etc.)

Panoramic image $X \in R^p$ ($p = 2^{12} \div 2^{18}$) captured by mobile robot at position $\theta \in R^3$ is determined by latent variable θ: $X = \varphi(\boldsymbol{\theta})$

- all captured panoramic images (features) $\mathbf{M} = \{X\}$ captured by mobile robot at all possible positions $\boldsymbol{\Theta} = \{\theta\}$ is 3-dimensional surface $\mathbf{M} = \{X = \varphi(\boldsymbol{\theta}): \theta \in \boldsymbol{\Theta} \subset R^3\} \subset R^p$ in high dimensional space R^p
- any 3-dimensional one-to-one parameterization $Z = F(X) \in R^3$ of the surface $\mathbf{M}$ gives the new reduced features

Feature space is q-dimensional surface in R^p 'parameterized' by latent variables

Original multiple features $\{X\}$ are indirect measurements of small number of an underlying sources which typically cannot be directly measured

Independent component analysis (ICA)

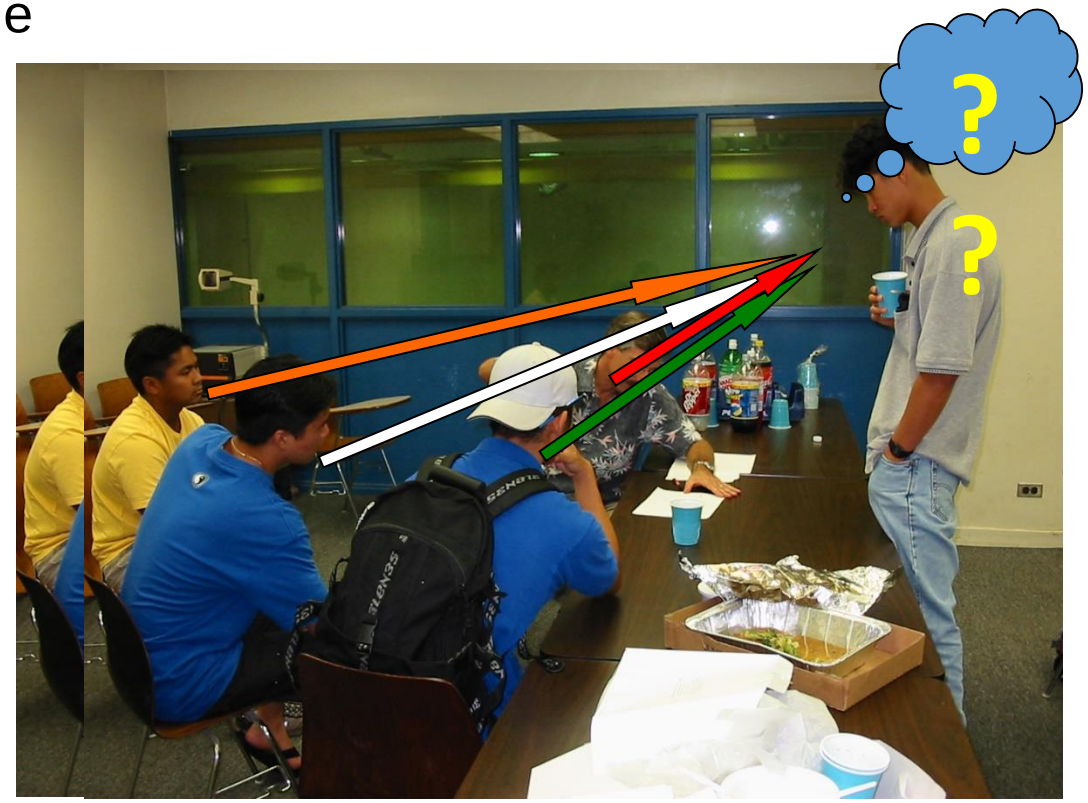
is a method for finding underlying **statistically independent** latent variables (factors, hidden causes, components) **(Jeanny Hérault and Bernard Ans, 1984)**

Most popular example: the **Cocktail party problem**

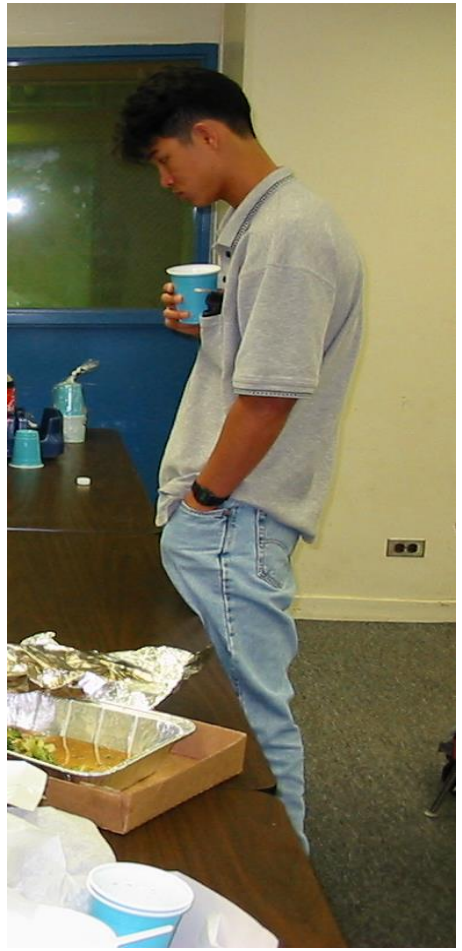
called also **Blind Source Separation** (BSS)

Cocktail party problem

Suppose you are in a crowded room with many people



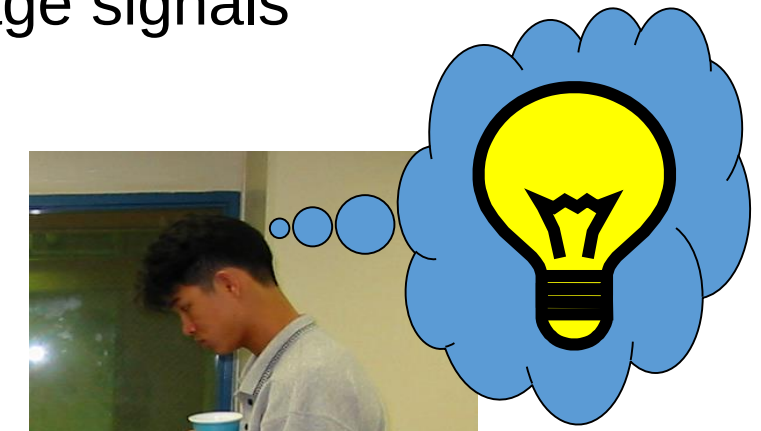
How do you understand what any one person is saying?



Humans can separate multiple signals **with only two ears/sensors**

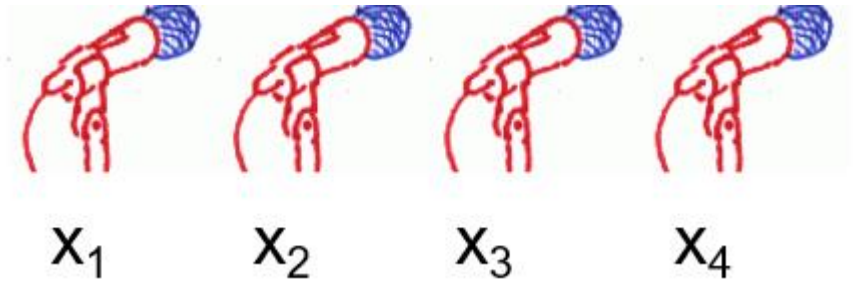
Computer can do the same but **with using many ears/sensors**

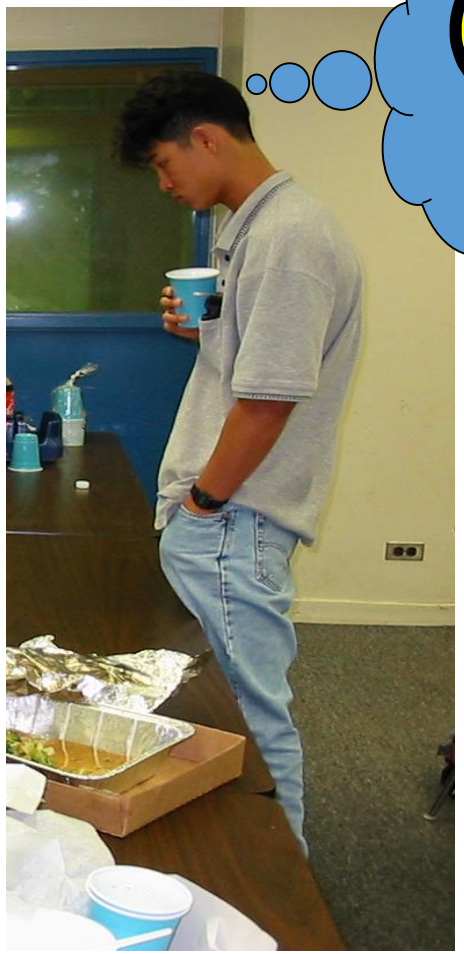
at least, as many ears/sensors as message signals



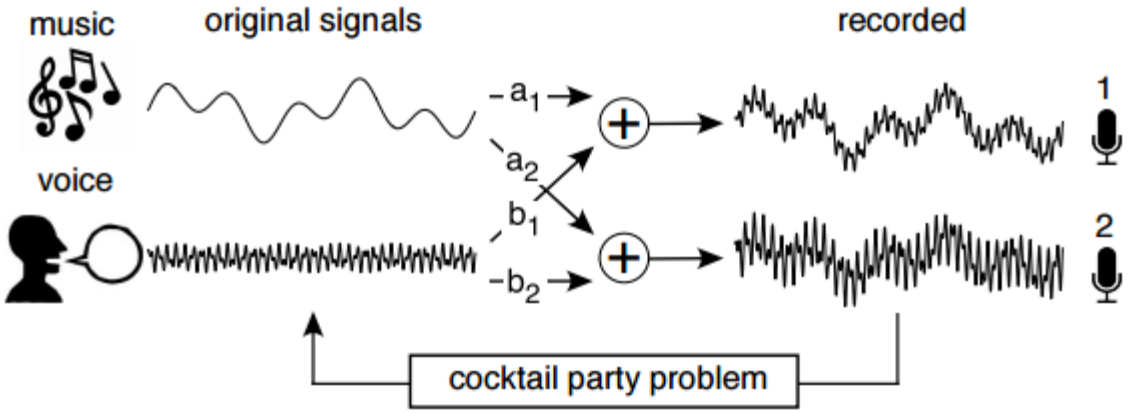
‘Human-computer’ has four ears:

a few mictrophones should be used

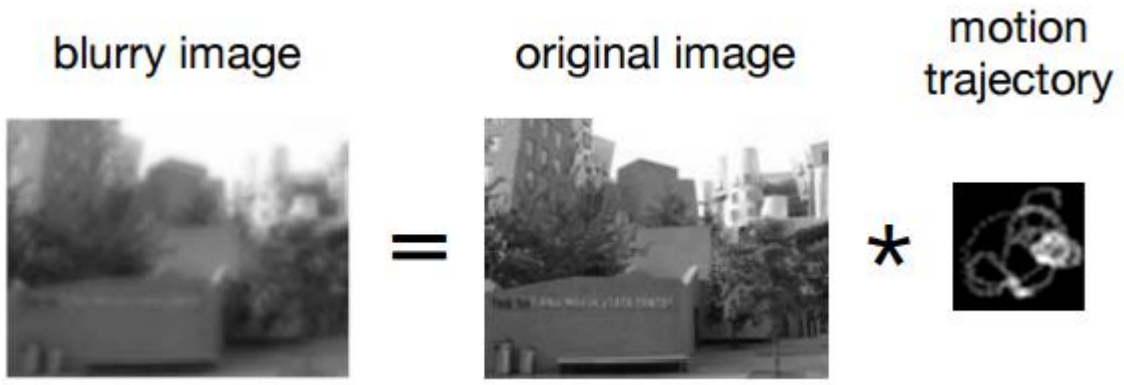




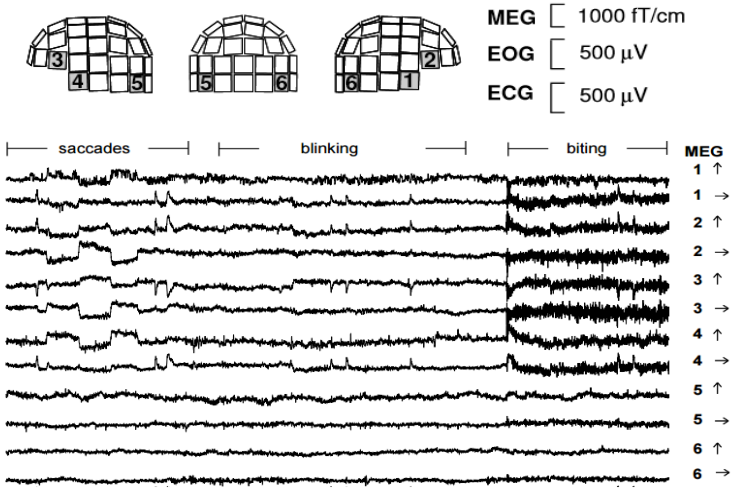
- **Data** (observed signals): mixed signals from **independent sources** obtained by **multiple sensors**
- **Analysis goal**: to find (to distinguish) the independent signals



The goal is to recover the original sources (i.e. music and voice) solely using the microphone recordings

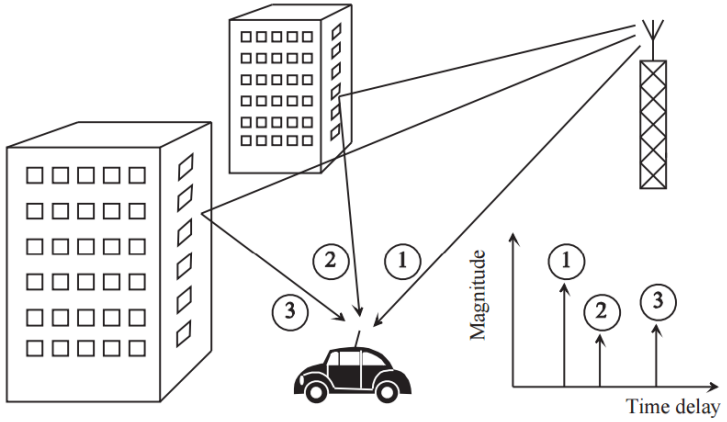


Removing blur from an image due to camera motion: identifying the original image and the motion path from a single blurry image



MagnetoEncephaloGrams (MEG) signals:

- from the frontal, temporal and occipital areas (multiple sensors)
- data contains several types of artifacts (signal sources) - ocular and muscle activity, the cardiac cycle, environmental magnetic disturbances, etc.



Telecommunications:
multipath propagation in urban environment

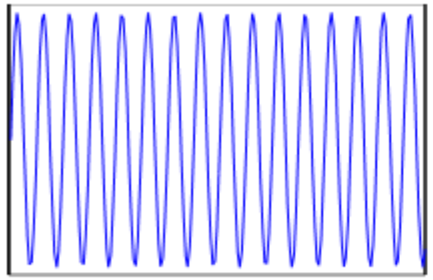
- Series of brain images with multiple voxels caused by different neuronal processes occurring within the brain
- Seismographic data from multiple seismometers

Formal task definition

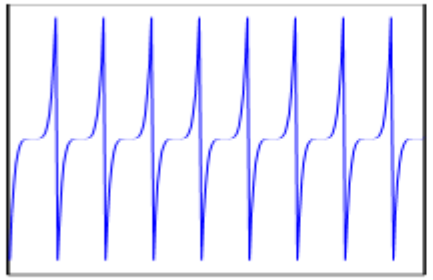
Hidden sources: q 'hidden' sources (latent variables, independent components):

s_j – signal from j^{th} source, $j = 1, 2, \ldots, q$

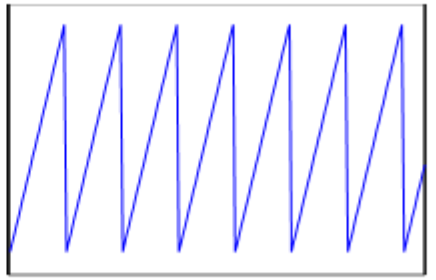
Example: $\{s_j = s_j(t)\}$ – time-dependent processes



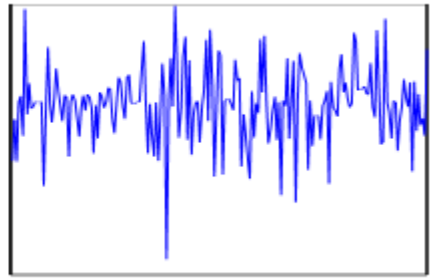
$s_1 = s_1(t)$



$s_2 = s_2(t)$

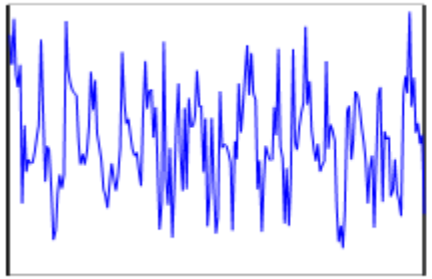


$s_3 = s_3(t)$

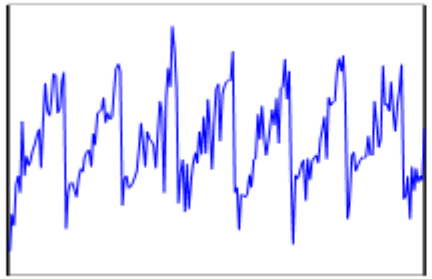


$s_4 = s_4(t)$

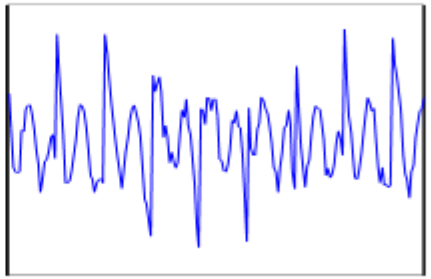
Data: m sensors: x_i – data (signal) from i^{th} sensor, i = 1, 2, … ,m



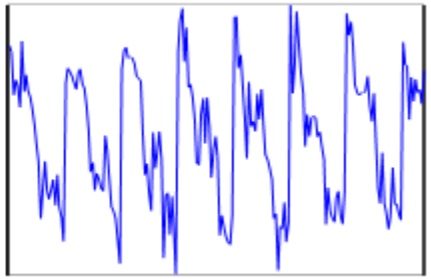
$x_1 = x_1(t)$



$x_2 = x_2(t)$



$x_3 = x_3(t)$



$x_4 = x_4(t)$

Data model - only **linear mixtures** of the source signals are observed:

$$X_i = \sum_{j=1}^{q} a_{ij} s_j, i = 1, 2, \dots, m$$

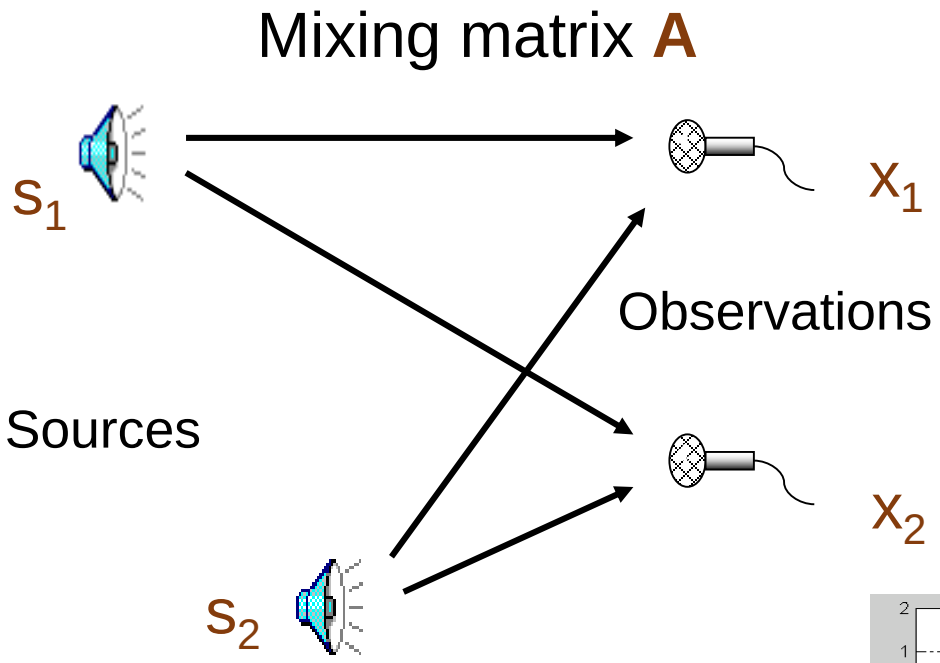
$$\mathbf{X} = \mathbf{A} \times \mathbf{S}$$

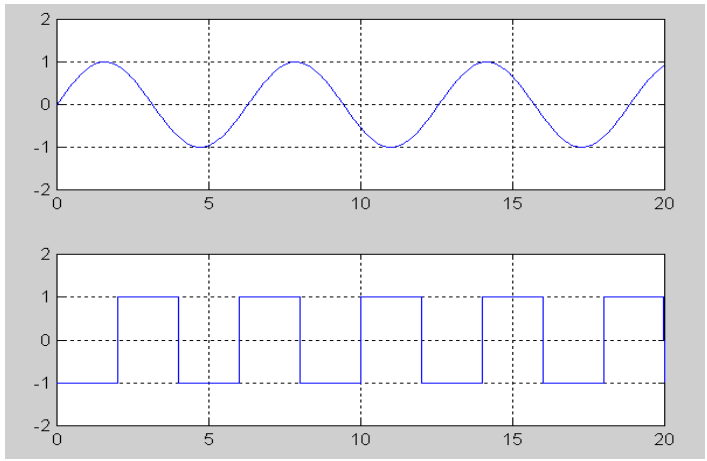
$\mathbf{S} = \begin{pmatrix} s_1 \\ \cdots \\ s_q \end{pmatrix} \in R^q$ - source signal vector

$\mathbf{X} = \begin{pmatrix} x_1 \\ \cdots \\ x_m \end{pmatrix} \in R^m$ - data (observed signal, multiplexed signal) vector

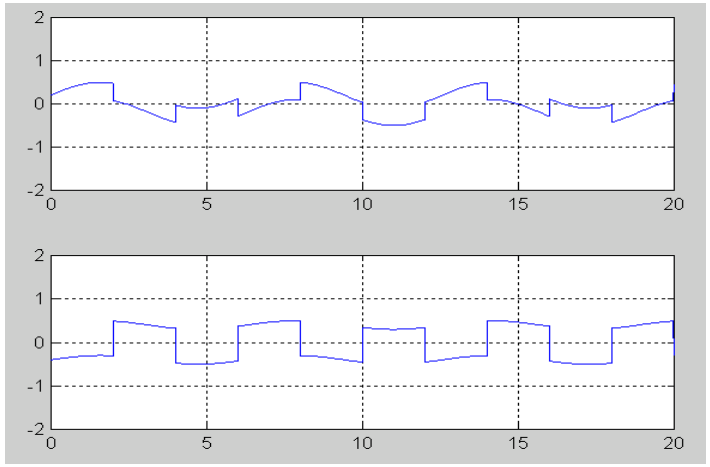
$\mathbf{A} = \|a_{ij}\|$ - $m \times q$ mixing matrix

Cocktail party problem

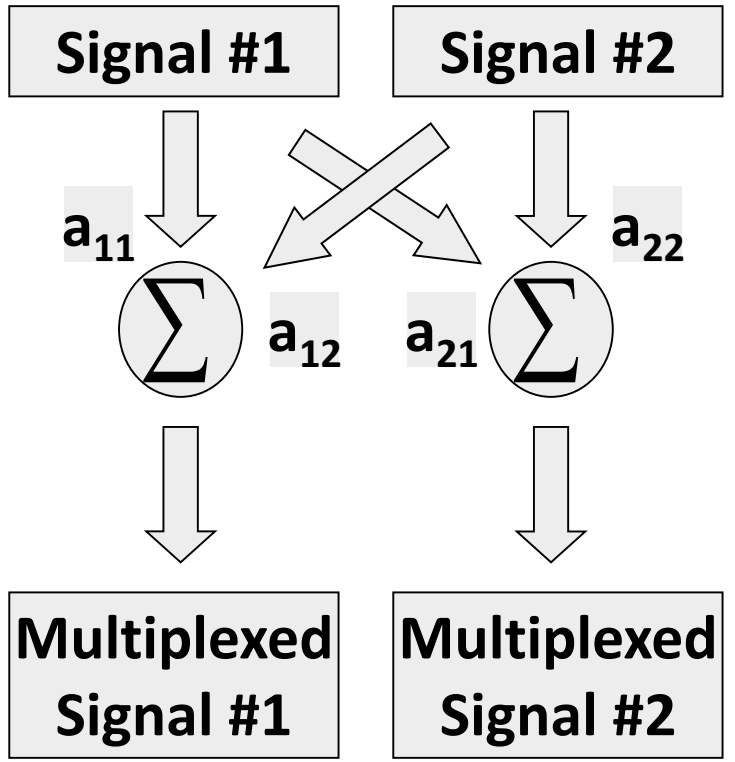




Two Independent sources



Mixture at two Mics



$\{a_{ij}\}$ depend on the distances of the microphones from the speakers

x = As

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

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

ICA problem: to recover

- unknown **independent** ‘source signals’ $\{s_j\}$
- unknown mixing matrix A

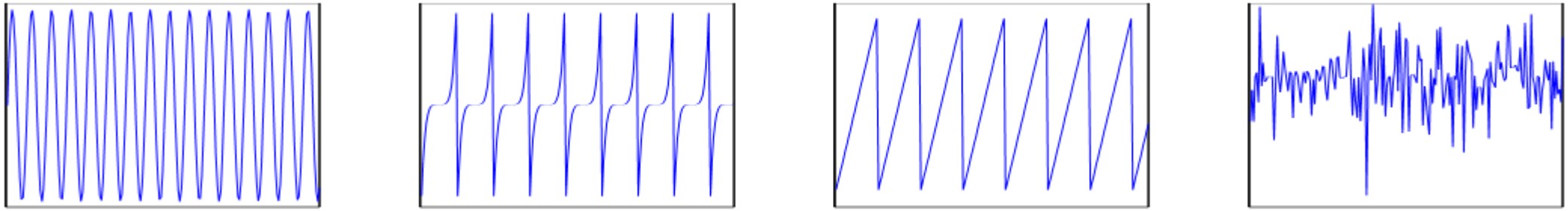
only from mixed data $\{x_i\}$ (unsupervised approach)

Desired **linear** solution – separated (estimated) signals: $\hat{\mathbf{S}} = \mathbf{W} \times \mathbf{X}$

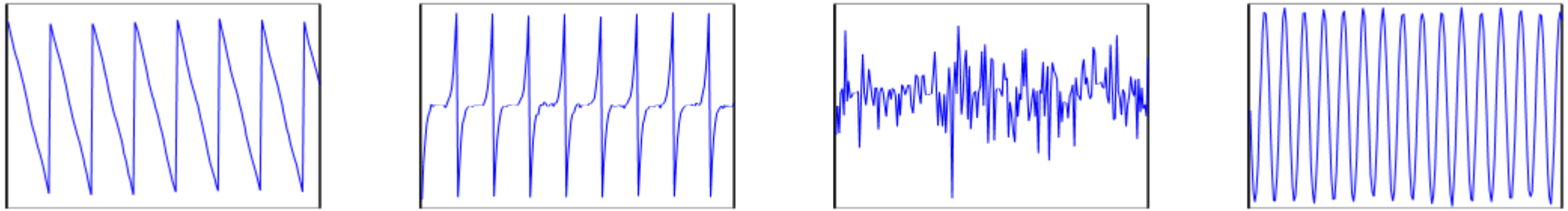
$\mathbf{W}$ - $q \times m$ **unmixing** matrix

ICA allows to recover source signals:

Source signals



ICA-recovered source signals



ICA has two related interpretations:

- **dimensional reduction:** if each source can be identified, a practitioner might choose to selectively delete or retain a single source (e.g. a person's voice, above)

- **filtering operation:** some aspect of the data is selectively removed or retained = is equivalent to projecting out some aspect (or dimension) of the data = a prescription for dimensional reduction

ICA was initially developed to deal with problems closely related to the **Coctail party problem**

Now **ICA** has many other applications too:

- Blind source separation
- Image denoising
- Medical signal processing – fMRI, ECG, EEG
- Modelling of the hippocampus and visual cortex
- Feature extraction, face recognition
- Compression, redundancy reduction
- Watermarking
- Clustering
- Time series analysis
- Topic extraction
- Scientific Data Mining

Desired solution – separated (estimated) signals: $\hat{\mathbf{S}} = \mathbf{W} \times \mathbf{X}$

$\mathbf{W}$ - $q \times m$ **unmixing** matrix

Mixing matrix $\mathbf{A}$ is known: $\mathbf{W}$ is solution of linear equation $\mathbf{X} = \mathbf{A} \times \mathbf{S}$:

$q = m$ and $q \times q$ mixing matrix $\mathbf{A}$ is invertible (nonsingular: $\mathrm{Rank}(\mathbf{A}) = q$): $\mathbf{W} = \mathbf{A}^{-1}$

$q < m$ and $\mathrm{Rank}(\mathbf{A}) = q$: $\mathbf{W} = \mathbf{A}^{+}$ - the solution

$\mathbf{A} = \mathbf{U} \times \Lambda \times \mathbf{V}^{\mathbf{T}}$ - SVD

$\mathbf{X} = (\mathbf{U} \times \Lambda \times \mathbf{V}^{\mathbf{T}}) \times \mathbf{S}$

$\mathbf{S} = (\mathbf{V} \times \Lambda^{-1} \times \mathbf{U}^{\mathrm{T}}) \times \mathbf{X}$ Λ^{-1}: non-zero diagonal elements are replaced by inverse

$\mathbf{A}^{+} = \mathbf{V} \times \Lambda^{-1} \times \mathbf{U}^{\mathrm{T}}$ - (left) pseudoinverse Moore-Penrose matrix to $m \times q$ matrix $\mathbf{A}$

Mixing matrix A is unknown:

- what is possible to estimate in the ICA model
- when it is possible (in which cases)
- how to estimate if it is possible

What is possible to estimate: ICA limitations

1) ICA can‘t determine the order of the Indepeendent componentes

Data model: $\mathbf{X} = \mathbf{A} \times \mathbf{S}$

B - arbitrary $q \times q$ permutation matrix

$\mathbf{B}^{-1}$ – inverse permutation matrix: $\mathbf{B}^{-1} \times \mathbf{B} = I_q$

q = 3:

$$\mathbf{B} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \qquad \mathbf{B}^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

$$\mathbf{X} = \mathbf{A} \times \mathbf{S} = \mathbf{A} \times (\mathbf{B}^{-1} \times \mathbf{B}) \times \mathbf{S} = (\mathbf{A} \times \mathbf{B}^{-1}) \times (\mathbf{B} \times \mathbf{S}) = \mathbf{A}^* \times (\mathbf{B} \times \mathbf{S})$$

$(\mathbf{B}^{-1} \times \mathbf{B}) = I_q$

$$= \mathbf{A} \times \mathbf{B}^{-1}$$

$$\mathbf{B} \times \mathbf{S} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} s_2 \\ s_3 \\ s_1 \end{pmatrix}$$

$$\mathbf{X} = \mathbf{A} \times \begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} = \mathbf{A}^* \times \begin{pmatrix} s_2 \\ s_3 \\ s_1 \end{pmatrix}$$

So, we can call any Independent component as the first one

2) ICA can‘t determine the variances (energies) of the Indepeendent componentes

Data model: $\mathbf{X} = \mathbf{A} \times \mathbf{S}$

B - arbitrary $q \times q$ diagonal matrix

$\mathbf{B}^{-1}$ – inverse diagonal matrix

$q = 3$:

$$\mathbf{B} = \begin{pmatrix} b_1 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & b_3 \end{pmatrix} \qquad \mathbf{B}^{-1} = \begin{pmatrix} b_1^{-1} 0 & 0 & 0 \\ 0 & b_2^{-1} & 0 \\ 0 & 0 & b_3^{-1} \end{pmatrix}$$

$$\mathbf{X} = \mathbf{A} \times \mathbf{S} = \mathbf{A} \times \underbrace{(\mathbf{B}^{-1} \times \mathbf{B})}_{I_q} \times \mathbf{S} = (\mathbf{A} \times \mathbf{B}^{-1}) \times (\mathbf{B} \times \mathbf{S}) = \mathbf{A}^* \times (\mathbf{B} \times \mathbf{S})$$

$$= \mathbf{A} \times \mathbf{B}^{-1}$$

$$\mathbf{B} \times \mathbf{S} = \begin{pmatrix} b_1 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & b_3 \end{pmatrix} \times \begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} b_1 \times s_1 \\ b_2 \times s_2 \\ b_3 \times s_3 \end{pmatrix}$$

$$\mathbf{X} = \mathbf{A} \times \begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} = \mathbf{A}^* \times \begin{pmatrix} b_1 \times s_1 \\ b_2 \times s_2 \\ b_3 \times s_3 \end{pmatrix}$$

So, we can find Independent component up to scaling

The source signals can principally be only recovered up to permutation and scaling

Mixing matrix **A** is unknown:

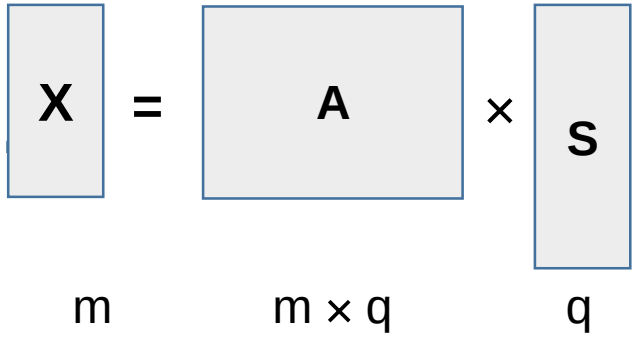
- what is possible to estimate in the ICA model
- **when it is possible (in which cases)**
- how to estimate if it is possible

When mixing matrix and source signals can be identified (estimated):

Requirement 1: the number q of separated (source) signals cannot be larger than the number m of observed variables (inputs)

Understandable!

$$\mathbf{X} = \mathbf{A} \times \mathbf{S}$$



Under $m < q$, even with known matrix **A**, we can not find the vector **S** from the vector **X** **by the only way**

Further, for simplicity, we will assume sometimes that:

- $m = q$
- $q \times q$ mixing matrix **A** is invertible

Requirement 2: source signals $\{s_i\}$ are **mutually statistically independent**

Without any requirements - trivial solution: $\hat{\mathbf{S}} = \mathbf{X}$, $\mathbf{A}$ and $\mathbf{W}$ - unit $q \times q$ matrices

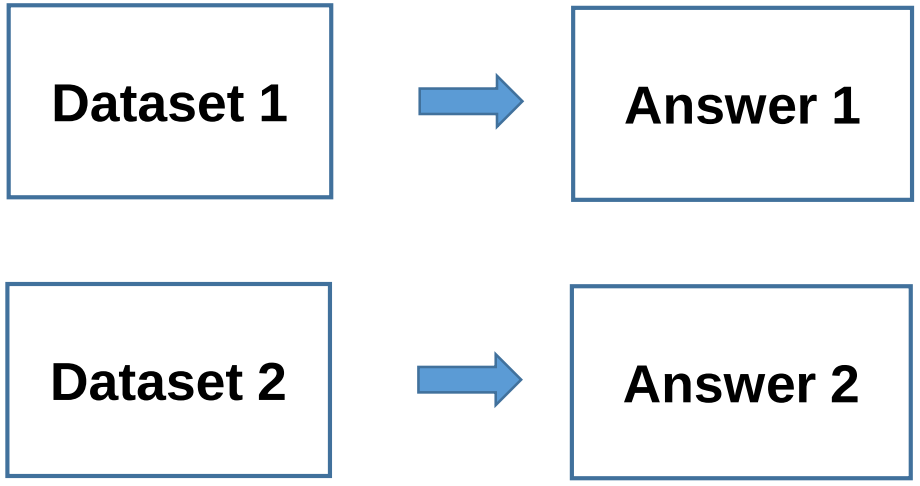
Independent sources:

- a rather natural ‘default’ assumption when we do not want to postulate any specific dependencies between the components
- a **physical interpretation** of independence is also sometimes possible: if the components are created by physically separate and non-interacting entities

Strict mathematical definition requires the concept of a Mathematical data generative model

Data analysis process

Applying the same question and the same procedure to different datasets:



reasonable (acceptable, admissible, satisfactory, agreeable, etc.)

meaningless (nonsensical, senseless), useless, incorrect

We need some assumptions about data (data model) to predict a quality of answer:

which quality we can expect if the data satisfy the assumptions

We need some assumptions about data (data model) to predict a quality of answer

Inverse task: given data model and selected quality criteria, to construct ‘the best procedure’

Physical data model: follows from ‘physical aspects’ which describes data generation process and determines data properties

Math data model: which can generate (simulate) data using certain math procedures (**generative model**) similar to the data obtained from real sources

Generative model:

- **Geometrical model:** Data space - data support - ‘a place where data lie’
- **Extracting model:** how data are extracted (are selected) from the Data space

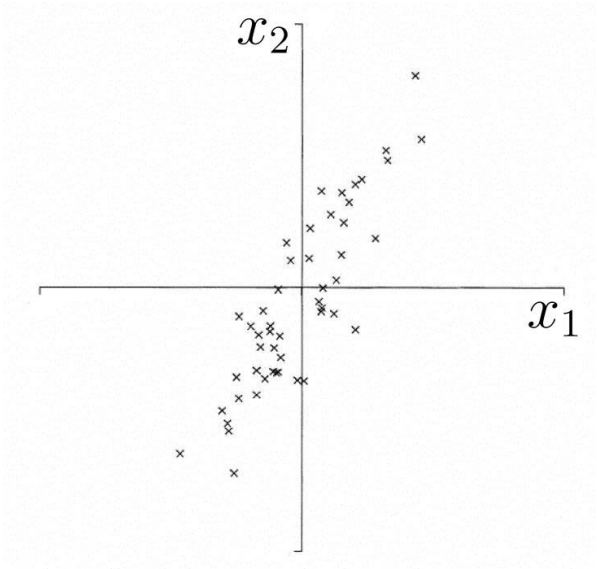
Probabilistic extracting model: data (training data and new, out-of-sample, data) are selected **randomly** from Data space independently of each other according to the certain **unknown probability distribution**

Probability distribution: determines statistical data properties

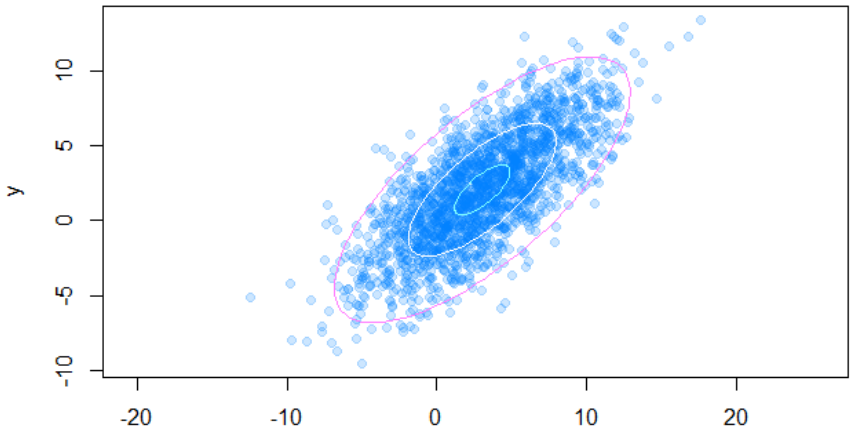
Statistical approach: given training dataset,

- to select appropriate **Probability distribution** which ‘well describes’ the training data
- selected Probability distribution is considered as **Statistical data model** and used for constructing ‘best’ data analysis procedures

Example: training data looks like a single cloud



A sample from two-dimensional Gaussian (normal) distribution looks similar



Thus, we can assume that we deal with random two-dimensional Gaussian (normal) random variables

Independent Component Analysis uses p**robabilistic data model** (assuming data are random) and, thus, p**robabilistic language** for strict formulating the requirements to data and describing the principles underlying the algorithms **is used**

Source signals $\{s_i\}$ are **mutually statistically independent** means that their joint **density function**

$$p(\mathbf{s}) = p(s_1, s_2, \dots, s_q)$$

is factorizable:

$$\boxed{p(s_1, s_2, \dots, s_q) = p(s_1) \times p(s_2) \times \dots \times p(s_q)}$$

Independence implies: conditional distribution of any source signal (say, s_1) under given all remaining signals $\{s_2, s_3, \dots, s_q\}$ doesn't depend on them:

$$\boxed{p(s_1 \mid s_2, \dots, s_q) = p(s_1)}$$

Thus, ICA searches for the rotation W of the observed data such that estimated source signal $\{\hat{s}_j\}$ are mutually statistically independent too: $p(\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_q) = p(\hat{s}_1) \times p(\hat{s}_2) \times \ldots \times p(\hat{s}_q)$

This implies, for example, that $p(\hat{s}_1 \mid \hat{s}_2, \ldots, \hat{s}_q) = p(\hat{s}_1)$

Probability theory: independency of random variables implies noncorrelatedness: $\mathrm{cov}(s_i, s_j) = 0$

Thus, the estimated source signal $\{\hat{s}_j\}$ should be at least uncorrelated

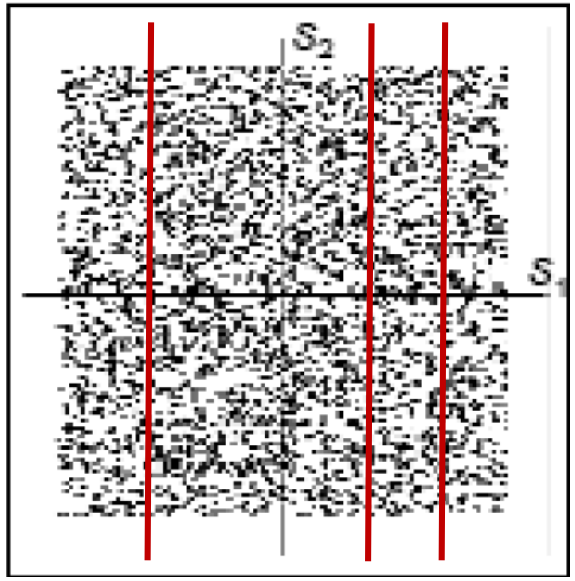
PCA transforms correlated multiplexed signals $\{x_i\}$ to uncorrelated quantities.

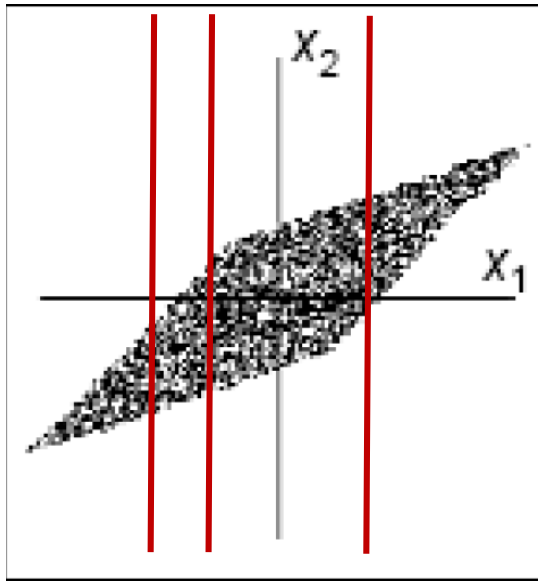
Can PCA give the solution to the ICA task? **Answer - NO**

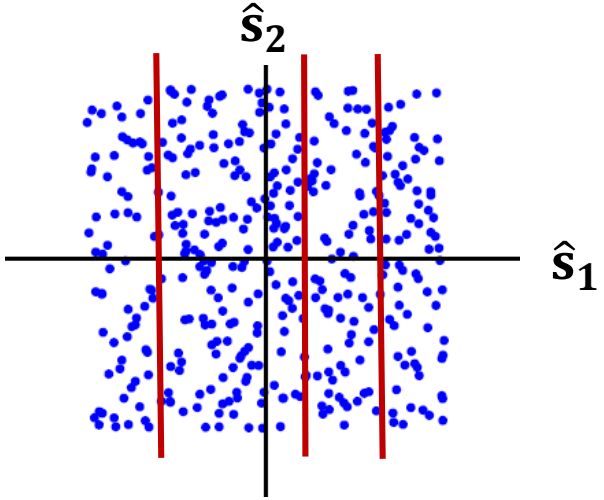
Illustrative example: two components S_1 and S_2 with **uniform distributions**

Mixing matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ -1 & 2 \end{pmatrix}$$

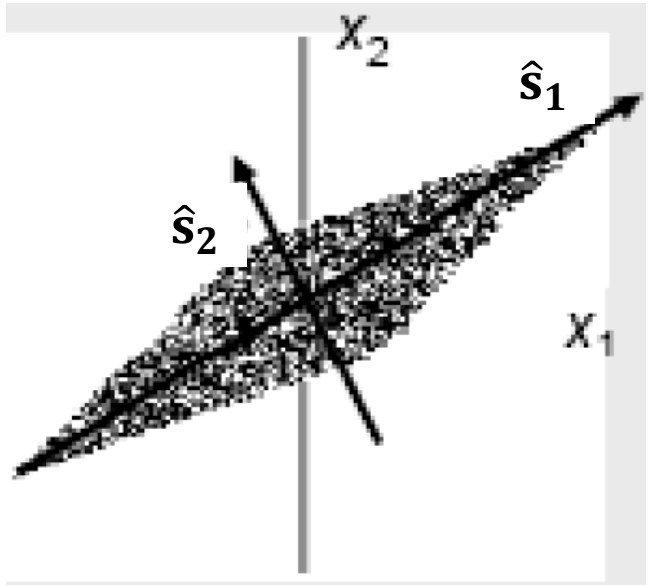




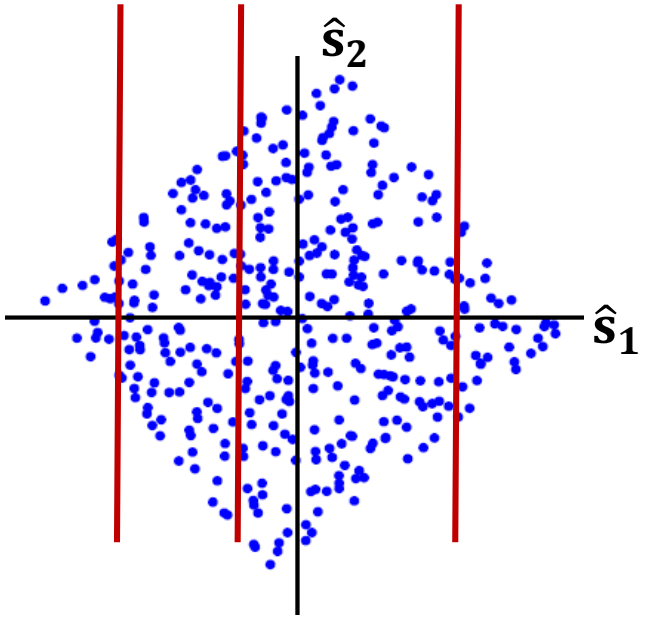


ICA

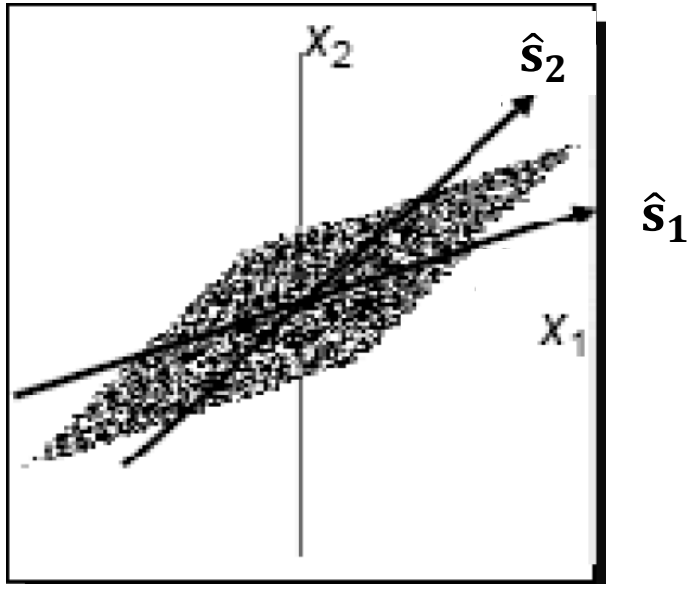
PCA



PCA



ICA



Probability theory: independency of random variables implies their noncorrelatedness

But the converse is not true in general

Example: $\xi \sim N(0, 1)$

$\eta_1 = \xi$ $\quad$ $\mathbf{M}\eta_1 = 0$

$\eta_2 = \xi^2 - 1$ $\quad$ $\mathbf{M}\eta_2 = 0$

η_1 and η_2 are uncorrelated:

$\mathrm{cov}(\eta_1, \eta_2) = \mathbf{M}(\xi \times (\xi^2 - 1)) = \mathbf{M}(\xi^3 - \xi) = 0$

These random varibles are connected by non-random functional dependency

Probability theory: in Gaussian case, a noncorrelatedness implies an independency

Maybe, the PCA will solve **the ICA problem for Gaussian signals**? **Answer - NO**

A surprising and unexpected fact: **no method** can solve **the ICA problem for Gaussian signals**

The **fundamental restriction** in ICA is that the independent components **must be nongaussian** (perhaps, **with the exception of only one** component) for **ICA to be possible**

Why Gaussian variables **are forbidden** (at most one of the sources can be Gaussian)?

Intuitively:

- **a correlatedness** is determined by the **second moments only**
- **an independency is more stronger property** than correlatedness which **is essential for** estimation of **the ICA model** and determined by ‘higher-order information’ about random variables (their higher-order moments)
- the **Gaussian distributions** are “**too simple**”: their higher order moments are completely determined by first two moments

A more strict explanation

two Gaussian source signals s_1 and s_2 → observed multiplexed signals x_1 and x_2 are Gaussian too

Centering

$\mathbf{a} = \mathbf{MX}$ and $\mathbf{X}' = \mathbf{X} - \mathbf{a}$ → $\mathbf{MX}' = 0$

If $\hat{\mathbf{S}}' = \mathbf{W} \times \mathbf{X}'$ - the solution to the centered ICA problem →

$\hat{\mathbf{S}} = \hat{\mathbf{S}}' + \mathbf{W} \times \mathbf{a}$ - the solution to the original ICA problem

Whitening

$\mathbf{cov(X)} = \mathbf{E} \times \mathbf{D} \times \mathbf{E}^T$ – eigenvalue decomposition of positive semidefinite covariance matrix:

E - orthogonal, **D** - diagonal

$\mathbf{V} = \mathbf{E} \times \mathbf{D}^{-1/2} \times \mathbf{E}^T$ - whitening matrix

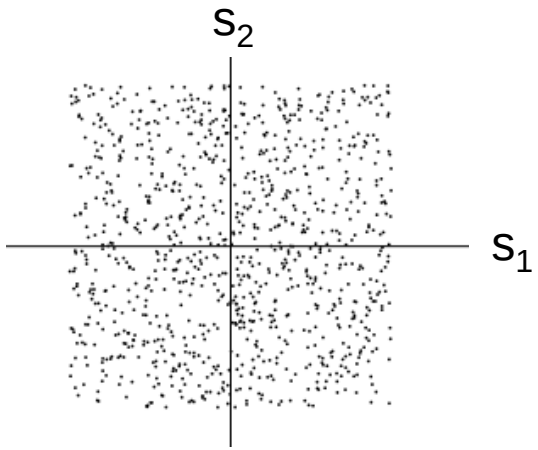
$\mathbf{X}_w = \mathbf{V} \times \mathbf{X}$ →

$\mathbf{cov}(\mathbf{X}_w) = \mathbf{V} \times \mathbf{cov}(\mathbf{X}) \times \mathbf{V}^T = (\mathbf{E} \times \mathbf{D}^{-1/2} \times \mathbf{E}^T) \times (\mathbf{E} \times \mathbf{D} \times \mathbf{E}^T) \times (\mathbf{E} \times \mathbf{D}^{-1/2} \times \mathbf{E}^T)^T = \ldots = I_2$ - unit matrix

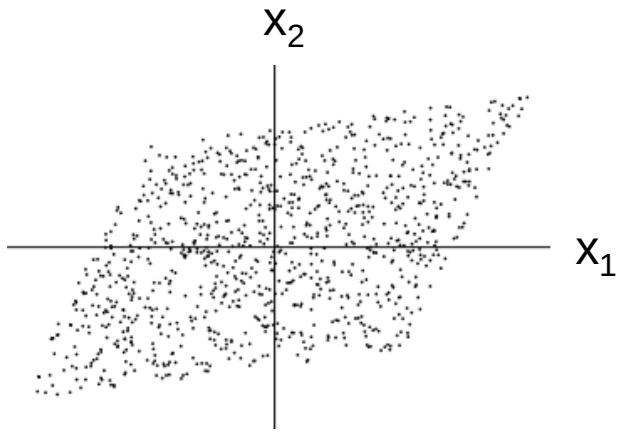
If $\hat{\mathbf{S}}' = \mathbf{W}' \times \mathbf{X}_w$ - the solution to the whitened ICA problem →

$\hat{\mathbf{S}} = \mathbf{W} \times \mathbf{X}$, where $\mathbf{W} = \mathbf{W}' \times \mathbf{V}$, - the solution to the original ICA problem

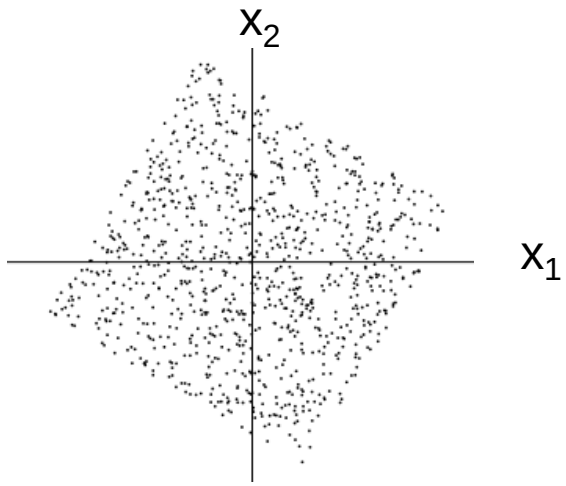
Without loss of generality, we can assume that $\mathbf{X} \sim N(0, I_2)$



Dataset consisting of the independent uniformly distributed signals s_1 and s_2



Dataset consisting of the observed mixtures x_1 and x_2



Dataset consisting of the **whitened** observed mixtures x_1 and x_2

Source signals $\hat{\mathbf{S}}$ can be found up to scaling

$\rightarrow$ we can find Gaussian vector $\mathbf{S}$ with independent components as vector with unit covariance matrix I_2

Model $\mathbf{X} = \mathbf{A} \times \mathbf{S}$, $\mathbf{X} \sim N(0, I_2)$ and $\mathbf{S} \sim N(0, I_2)$, implies that mixing matrix $\mathbf{A}$ is orthogonal

In Gaussian model, mixing matrix $\mathbf{A}$ and, thus, source vector $\mathbf{S}$, **cannot be identified** (estimated) – any orthogonal matrix $\mathbf{A}$ transform Gaussian source vector $\mathbf{S} \sim N(0, I_2)$ with independent components to the random vector $\mathbf{X}$ **with the same distribution** $N(0, I_2)$

ICA is essentially impossible for the Gaussian sources

If **just one** of the independent components is Gaussian, the **ICA model can still be identified**

Fortunately, signals measured by physical sensors are usually quite nongaussian

- R1: number of source signals is equal to number of observed variables (at least not less)
- R2: source signals $\{s_i\}$ are **mutually statistically independent**
- R3: all source signals $\{s_i\}$, with the exception of at most one, must be non-Gaussian

Proven result: under requirements R1 – R3, mixing matrix and components can be identified

ICA model in which some of the components are Gaussian, some non-Gaussian:

- we can estimate all the non-Gaussian components
- the Gaussian components cannot be separated from each other: these components will be arbitrary linear combinations of some constructed Gaussian components
- in the case of just one Gaussian component, we can estimate the model (the single Gaussian component does not have any other Gaussian components that it could be mixed with)

Mixing matrix **A** is unknown:

- what is possible to estimate in the ICA model
- when it is possible (in which cases)
- **how to estimate if it is possible**

ICA preprocessing steps: centering and whitening

Centering

$\mathbf{a} = \mathbf{MX}$ and $\mathbf{X}' = \mathbf{X} - \mathbf{a}$ $\rightarrow$ $\mathbf{MX}' = 0$

$\hat{\mathbf{S}}' = \mathbf{W} \times \mathbf{X}'$ - solution to the centered ICA $\rightarrow$ $\hat{\mathbf{S}} = \hat{\mathbf{S}}' + \mathbf{W} \times \mathbf{a}$ - solution to the original ICA

Without loss of generality, we can assume that $\mathbf{MX} = 0$

When centered observed variables are used, $\mathbf{MS} = \mathbf{M}(\mathbf{W} \times \mathbf{X}) = 0$

ICA model: $\mathbf{X} = \mathbf{A} \times \mathbf{S}$

$\mathbf{S} \in R^q$ - source signal, $\mathbf{X} \in R^m$ - data, $\mathbf{A} = \|a_{ij}\|$ - $m \times q$ mixing matrix ($m \geq q$, $\mathrm{Rank}(\mathbf{A}) = q$)

$\mathbf{cov}(\mathbf{X}) = \mathbf{E} \times \mathbf{D} \times \mathbf{E}^T$ – Eigenvalue decomposition of positive semidefinite covariance $m \times m$ matrix:

$\mathbf{E}$ – orthogonal $m \times m$ matrix, $\mathbf{D}$ – diagonal $m \times m$ matrix at most q nonzero diagonal elements

Whitening

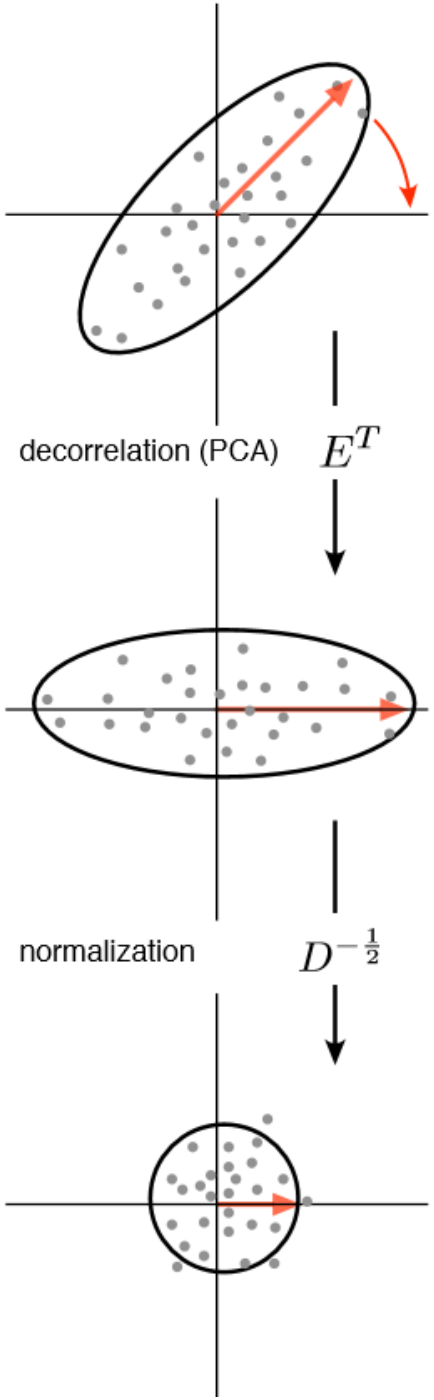
$\mathbf{V} = \mathbf{E}_q \times \mathbf{D}_q^{-1/2} \times \mathbf{E}_q^T$ - whitening $m \times m$ matrix

$\mathbf{E}_q$ - orthogonal $m \times q$ matrix consisting of first q columns of $m \times m$ matrix $\mathbf{E}$

$\mathbf{D}_q$ - diagonal $q \times q$ matrix with positive eigenvalues of matrix $\mathbf{cov}(\mathbf{X})$: $\mathrm{Rank}(\mathbf{cov}(\mathbf{X})) = q$

In fact, 'PCA technique' has been applied to the observed data!

$\mathbf{X}_w = \mathbf{V} \times \mathbf{X}$ - whitened data vector $\rightarrow$ $\mathbf{cov}(\mathbf{X}_w) = I_q$ - unit matrix



If $\hat{S}' = W' \times X_w$ - the solution to the whitened ICA problem →

$\hat{S} = W \times X$, $W = W' \times V$, - the solution to the original ICA problem

Without loss of generality, we can assume that $cov(X_w) = I_q$

Source signals $\hat{\mathbf{S}}$ can be found up to scaling

$\rightarrow$ we can find $\mathbf{S}$ with independent components and **unit variances**

$$\mathrm{Cov}(\mathbf{S}) = \mathbf{M}(\mathbf{S} \times \mathbf{S}^T) = I_q$$

Model $\mathbf{X} = \mathbf{A} \times \mathbf{S}$ $\rightarrow$ $\mathbf{X}_w = \mathbf{V} \times \mathbf{A} \times \mathbf{S}$ $\rightarrow$ $\mathbf{X}_w = \mathbf{A}_w \times \mathbf{S}$, where $\mathbf{A}_w = \mathbf{V} \times \mathbf{A}$

$$\mathbf{cov}(\mathbf{X}_w) = \mathrm{cov}(\mathbf{A}_w \times \mathbf{S}) = \mathbf{A}_w \times \mathrm{cov}(\mathbf{S}) \times (\mathbf{A}_w)^T = \mathbf{A}_w \times (\mathbf{A}_w)^T$$

$\mathbf{cov}(\mathbf{X}_w) = I_q$ $\rightarrow$ $\mathbf{A}_w \times (\mathbf{A}_w)^T = I_q$ $\rightarrow$ $\mathbf{A}_w$ - orthogonal $m \times q$ matrix

We can restrict our search for the mixing matrix to the space of **orthogonal** $m \times q$ matrices:

Case $m = q$:

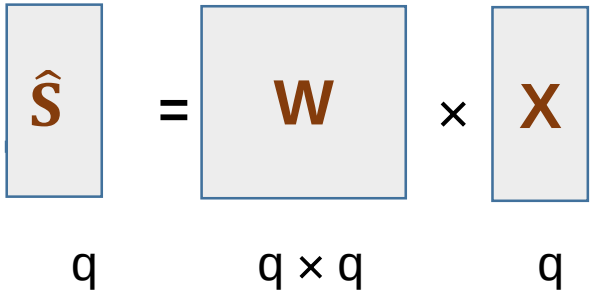
- original mixing matrix A depends on q^2 parameters which must be estimated
- orthogonal mixing matrix A_w depends on $q(q - 1)/2$ parameters only

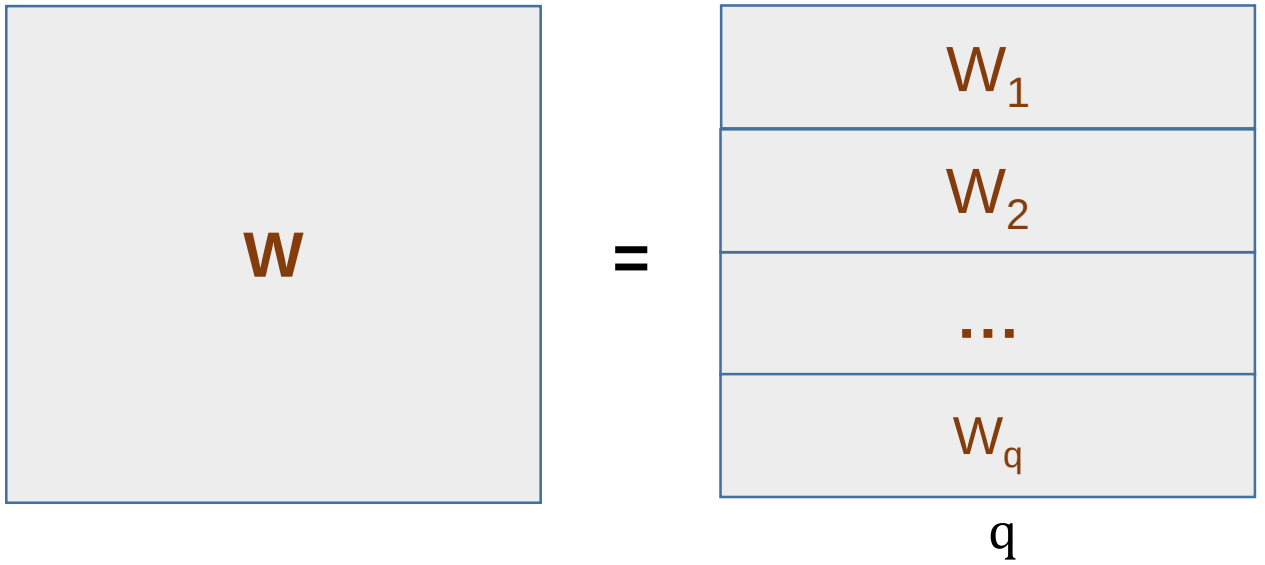
whitening solves half of the problem of ICA by reducing the complexity of the problem

Fundamental approaches to the solution of the ICA problem:

- maximization the nongaussianity of the marginals
- maximum likelihood approach
- minimization of mutual information between components

$\hat{S} = W \times X$





$W_1, W_2, \ldots, W_q$ - rows of **W**

Recovered signals: $\hat{s}_i = W_i \times X$

Assumption: source signals $\{s_i\}$ are **random**, **mutually statistically independent**, and all signals, with the exception of at most one, are **non-Gaussian**

Data vector $\mathbf{X}$ is random → we assume (after centering and whitening) that

$$\mathbf{MX} = 0 \quad \rightarrow \quad \mathbf{cov}(\mathbf{X}) = I_q \text{ - unit matrix}$$

$\mathbf{MX} = 0 \rightarrow \mathbf{MS} = 0$

Source signals can be found up to scaling → we assume that $\mathbf{cov}(\mathbf{S}) = I_q$

$\mathbf{cov}(\mathbf{X}) = I_q$, $\mathbf{cov}(\mathbf{S}) = I_q$ → mixing matrix $\mathbf{A}$ is orthogonal

Source signals can be found up to permutation and scaling → we will search:

- independent recovered signals $\hat{\mathbf{S}} = \mathbf{W} \times \mathbf{X}$ satisfying relations $\mathbf{M}\hat{\mathbf{S}} = 0$, $\mathbf{cov}(\hat{\mathbf{S}}) = I_q$
- orthogonal unmixing matrix $\mathbf{W}$ → the rows $W_1, W_2, \ldots, W_q$ of $\mathbf{W}$ - **orthonormal** vectors

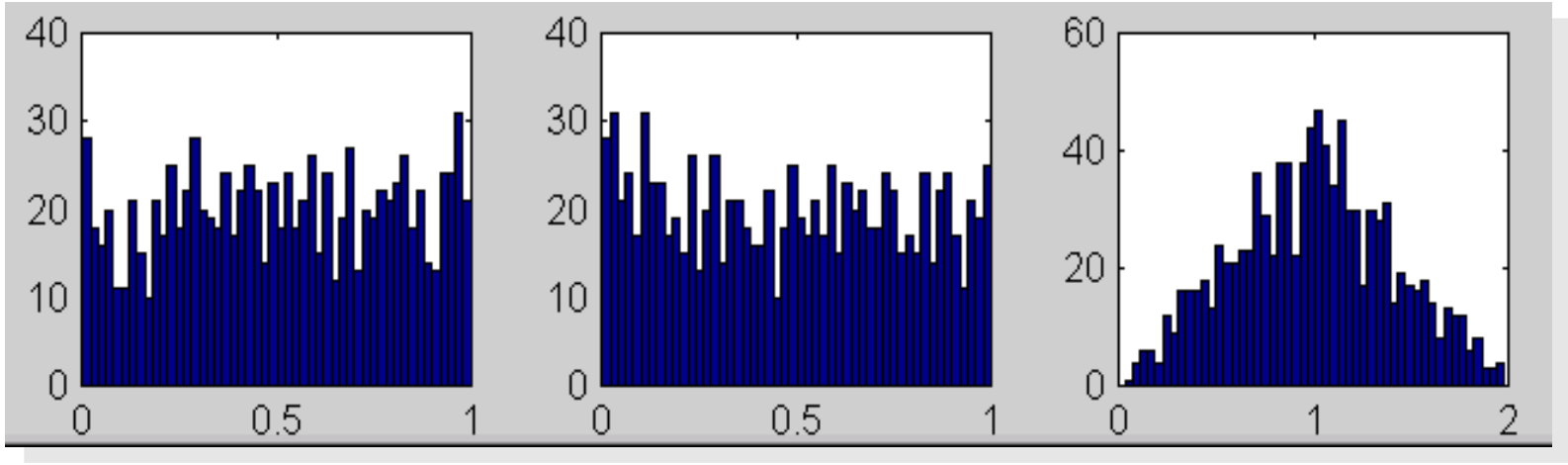
Fundamental approaches to the solution of the ICA problem:

1) Maximum nongaussianity principle: find maximally '**nongaussian**' (with respect to **chosen measure of nongaussianity**) **independent** signals $\{\hat{s}_i = W_i \times \mathbf{X}\}$

Independency as maximum nongaussianity: motivation

P**robability theory (Central Limit Theorem)** - *informally***:** The distribution of a sum of independent random variables tends toward a Gaussian distribution

A sum of a few independent random variables **is more gaussian** than individual random variables



density of s_1 density of s_2 density of $(s_1 + s_2)$

$$\hat{S} = W \times X$$

W_k - k-th row of desired unmixing matrix **W** →

$\hat{s}_k = W_k \times \mathbf{X}$ - k-th component of the estimated signal $\hat{S}$

$\hat{s}_k = W_k \times \mathbf{X} = W_k \times \mathbf{A} \times \mathbf{S}$ - a weighted linear combination of independent signals $\{s_i\}$

$\hat{s}_k$ is 'more Gaussian' than source signals $\{s_i\}$ **AND becomes least gaussian when** equals to one of $\{s_i\}$

Starting from the first row, we could take W_1 as a vector which **maximizes 'non-gaussianity'** of $W_1 \times \mathbf{X}_w$

After choosing $\hat{s}_1 = W_1 \times \mathbf{X}$, we will search for next estimated signal $\hat{s}_2 = W_2 \times \mathbf{X}_w$ (W_2 - 2nd row of desired matrix **W**) by similar way taking into account that $\hat{s}_2$ and $\hat{s}_1$ are independent variables

→ they must be uncorrelated

and so on

$f(\xi)$ - a chosen measure of non-Gaussianity of random variable ξ, $\mathbf{M}\xi = 0$ and $\mathbf{M}\xi^2 = 1$

$b \in R^q$ - arbitrary row of desired unmixing matrix $\mathbf{W}$ which determines estimated signal $\hat{s} = b \times \mathbf{X}$

$F(b) = \mathbf{M}f(b \times \mathbf{X})$ - objective function defined by chosen measure of non-Gaussianity which should be maximized under constraint $b^T \times b = 1$ with taking into account that $\mathbf{cov}(\mathbf{X}_w) = I_q$

$X_1, X_2, \ldots, X_T$ - training dataset consisting of centering and whitened observed signals

$F_n(b)$ - some estimator of the objective function $F(b)$ based on training dataset

For example: $F_n(b) = \frac{1}{n}\sum_{t=1}^{T} F\left(b^T \times X_t\right)$

ICA: $F_n(b) \to \max$ under constraint $b^T \times b = 1$ with taking into account that $\frac{1}{n}\sum_{t=1}^{T}\left(X_t \times X_t^T\right) = I_q$

We look for the rows of matrix **W** one after another, starting from the first row W_1 by maximizing an objective function $F(W_1)$ under constraint $W_1^T \times W_1 = 1$

After choosing $\hat{s}_1 = W_1 \times \mathbf{X}$, we will search for next estimated signal $\hat{s}_2 = W_2 \times \mathbf{X}$ by maximizing $F(W_2)$ under constraints $W_2^T \times W_2 = 1$

and under additional constraint $W_2 \times W_1^T = 0$ which follows from independence of $\hat{s}_2$ and $\hat{s}_1$:

$$\mathbf{Cov}(\hat{s}_2, \hat{s}_1) = \mathbf{M}(\hat{s}_2 \times \hat{s}_1) = \mathbf{M}\{(W_2 \times \mathbf{X}) \times (W_1 \times \mathbf{X})\} = W_2 \times \mathbf{M}(\mathbf{X} \times X^T) \times W_1^T = W_2 \times W_1^T = 0$$

and so on

ICA: used quantitative measures of non-Gaussianity

ξ - random variable, $\mathbf{M}\xi = 0$, $\mathbf{M}\xi^2 = 1$

1) **Kurtotis:** $\mathbf{Kurt}(\xi) = \mathbf{M}\xi^4 - 3(\mathbf{M}\xi^2)^2$, $\mathbf{Kurt}(\xi_{Gauss}) = 0$ for Gauss random variable ξ_{Gauss}

$f(\xi) = |\mathrm{Kurt}(\xi)|$ - a measure of non-Gaussianity

$$F(b) = f(b \times \mathbf{X}) = \mathbf{M}(b \times \mathbf{X})^4 - 3(\mathbf{M}(b \times \mathbf{X})^2)^2 = \frac{1}{T}\sum_{t=1}^{T}(b \times X_t)^4 - 3 \times \left(\frac{1}{T}\sum_{t=1}^{T}(b \times X_t)^2\right)^2$$

$X_1, X_2, \ldots, X_T$ - training dataset consisting of centering and whitened observed signals

$$\frac{1}{T}\sum_{t=1}^{T}(b \times X_t)^2 = \frac{1}{T}\sum_{t=1}^{T}\left\{(b \times X_t) \times \left(X_t^T \times b^T\right)\right\} = b \times \frac{1}{T}\sum_{t=1}^{T}\left\{X_t \times X_t^T\right\} \times b^T = b \times b^T = 1$$

$\mathbf{cov}(\mathbf{X}) = I_q$

$$F(b) = \frac{1}{T}\sum_{t=1}^{T}(b \times X_t)^4 - 3 \rightarrow \max \quad \text{under constraint } b \times b^T = 1$$

$$F(b) = \left|\mathrm{Kurt}(b \times \mathbf{X})\right| = \frac{1}{T}\sum_{t=1}^{T}(b \times X_t)^4 - 3 \rightarrow \max$$ under constraint $b \times b^T = 1$

Gradient descent method is usually used

Let $w = b^T$ - desired vector-column of unmixing matrix **W** and $F(b) = F(w)$

$$\nabla F(w) = \frac{\partial}{\partial w}\left|\mathrm{Kurt}(w^T \times \mathbf{X})\right| = 4 \times \mathrm{sign}(\mathrm{Kurt}(w^T \times \mathbf{X})) \times \{\mathbf{M}(\mathbf{X} \times (w^T \times \mathbf{X})^3) - 3w \times \|w\|^2\}$$

$= 1$

Iterations in the gradient descent method: $w \leftarrow w + \Delta w$, $\Delta w \propto \nabla F(w)$

- since $\|w\|^2 = 1$, iterations $w \leftarrow w + \Delta w$ must be complemented by projecting on the unit sphere after every step: $w \leftarrow w/\|w\|$
- the term $-3w$ in the gradient $\nabla F(w)$ can be omitted: it changes the norm of w in the gradient algorithm which will be normalized

Iterations in the gradient descent method:

- $w \leftarrow w + \Delta w$,
- $w \leftarrow w/\|w\|$

$$\Delta w \propto \text{sign}(\text{Kurt}(w^T \times \mathbf{X})) \times \mathbf{M}(\mathbf{X} \times (w^T \times \mathbf{X})^3)$$

$$\mathbf{M}(\mathbf{X} \times (w^T \times \mathbf{X})^3) = \frac{1}{T}\sum_{t=1}^{T} X_t \times \left(w^T \times X_t\right)^3$$

A fast fixed-point algorithm using kurtosis- FastICA algorithm (kurtosis version)

Math: at a stable point of the gradient algorithm, the gradient $\nabla F(w)$ must point in the direction of w

- the gradient $\nabla F(w)$ must be equal to w multiplied by some scalar constant - only in such a case, adding the gradient to w does not change its direction

Iterations in the FastICA method:

- $w \leftarrow \mathbf{M}(\mathbf{X} \times (w^T \times \mathbf{X})^3) - 3w$
- $w \leftarrow w/\|w\|$

Kurtosis-based measure of non-Gaussianity has some drawbacks in practice, when kurtosis value has to be estimated from a measured sample: **kurtosis can be very sensitive to outliers**

Kurtosis is not a robust measure of nongaussianity

2) Negentropy: $p_\xi(y)$ - density of random variable ξ

Entropy (measure of uncertainty): $H(\xi) = - \mathbf{M}(\log_2 p_\xi(\xi)) = - \int p_\xi(y) \times \log_2 p_\xi(y) dy$

Probability theory: which random variable ξ has maximal entropy?

The answer depends on available information about ξ (its density p_ξ)

Let available information has a form: for given functions $h_i(\xi)$, $i = 1, 2, \dots, m$:

$$\mathbf{M} h_i(\xi) = \int p_\xi(y) \times h_i(y) dy = k_i$$

Example: $h_1(\xi) = \xi$, $k_1 = 0$; $h_2(\xi) = \xi^2$, $k_2 = 1$

Under these constraints, a density with maximum entropy is

$$p_0(y) = A \times \exp\{\sum_{i=1}^{m} \alpha_i \times h_i(y)\}$$

Example: $p_0(y) = A \times \exp\{\alpha_1 \times y + \alpha_2 \times y^2\}$ - Gaussian density

Negentropy: $J(\xi) = H(\xi_{Gauss}) - H(\xi) \geq 0$

$H(\xi_{Gauss})$ - entropy of Gaussian random variable ξ_{Gauss} ~ such that $Var(\xi_{Gauss}) = Var(\xi)$

$J(\xi_{Gauss}) = 0 \quad \rightarrow \quad f(\xi) = J(\xi)$ - a measure of distance from Gaussianity

$\xi = b \times \mathbf{X} \quad \rightarrow \quad \mathbf{cov}(\xi) = \mathbf{M}(b \times \mathbf{X})^2 = 1 \quad \rightarrow$

$\xi_{Gauss} \sim N(0, 1) \quad \rightarrow \quad H(\xi_{Gauss}) = \frac{1}{2}(1 + \log_2(2\pi))$

$J(b \times \mathbf{X}) = \frac{1}{2}\{1 + \log_2(2\pi)\} - H(b \times \mathbf{X}) = \frac{1}{2}\{1 + \log_2(2\pi)\} + \mathbf{M}\log_2(p_s(b \times \mathbf{X}))$

$= \frac{1}{2}\{1 + \log_2(2\pi)\} + \frac{1}{T}\sum_{t=1}^{T}\{\log_2(p_s(b \times X_t))\}$ p_s - density function of random signal s

$F(b) = \frac{1}{T}\sum_{t=1}^{T}\{\log_2(p_s(b \times X_t))\} \rightarrow \max$ under constraint $b \times b^T = 1$

The trouble: computationally very difficult, density function p_s of random signal s is unknown and should be estimated → simpler approximations of the negentropy/entropy are used

Approximation 1 based on Gram-Charlier expansion for density estimation

Probability theory: $p_\xi(y)$ - unknown density of random variable ξ

$\varphi(y) = (2\pi)^{-1/2} \times \exp(-y^2/2)$ - density of N(0, 1)

$H_3(y)$, $H_4(y)$ - Chebyshev-Hermite polynomials of order 3 and 4

$$p_\xi(y) \approx p^*(y) = \varphi(y) \times \left\{1 + k_3(\xi) \times \frac{H_3(y)}{3!} + k_4(\xi) \times \frac{H_4(y)}{4!}\right\}$$

$K_3(\xi) = M\xi^3$, $K_4(\xi) = M\xi^4 - 3$

Negentropy approximation:

$$H^*(\xi) = -\int p^*(y) \times \log_2 p^*(y) dy \approx H(\xi)$$

$$J^*(\xi) = \frac{1}{12}(\mathbf{M}\xi^3)^2 + \frac{1}{48}\left(\mathbf{Kurt}(\xi)\right)^2 \approx J(\xi)$$

$\mathbf{M}\xi^3 = 0$ for random variables with symmetric distributions → $J^*(\xi) = \frac{1}{48}\left(\mathbf{Kurt}(\xi)\right)^2$

→ maximization of $f(\xi) = J^*(\xi)$ is equivalent to maximization of $|\mathbf{Kurt}(\xi)|$

$$p_\xi(y) \approx p^*(y) = \varphi(y)\times\left\{1 + k_3(\xi) \times \frac{H_3(y)}{3!} + k_4(\xi) \times \frac{H_4(y)}{4!}\right\}$$

Approximation 2 replaces polynomial functions $H_3(y)$ and $H_4(y)$ by other functions and is based on 'maximum entropy' density estimation

Assumption: $M\xi = 0$, $M\xi^2 = 1$ and we have observed (in practice, estimated) a number of expectations $\mathbf{M}h_i(\xi)$ of functions $\{h_i(\xi)\}$:

$$\mathbf{M}h_i(\xi) = \int p_\xi(y)\times h_i(y)dy = k_i, \qquad i = 1, 2, \dots, m,$$

which:

- form orthonormal system: $\int \varphi(y)\times h_i(y)\times h_j(y)dy = \delta_{ij}$ (Kronecker symbol)
- orthogonal to all polynomials of second degree: $\int \varphi(y)\times h_i(y)\times y^k dy = 0, \qquad k = 0, 1, 2$

We can choose arbitrary not-orthonormal functions $\{h_i(\xi)\}$ and then use standard Gram-Schmidt orthonormalization technique

Approximative maximum entropy density: $p_\xi(y) \approx p^{**}(y) = \varphi(y)\times\{1 + \sum_{i=1}^{m} k_i \times h_i(y)\}$

with nonquadratic functions $\{h_i(y)\}$

$$H^{**}(\xi) = -\int p^{**}(y)\times\log_2 p^{**}(y)dy \approx H(\xi)$$

Negentropy approximation: $J^{**}(\xi) = \sum_{i=1}^{m} k_i \times \{\mathbf{M}h_i(\xi) - \mathbf{M}h_i(\xi_{Gauss(0,1)})\}^2 \approx J(\xi)$

$\{k_i\}$ - positive constants, $\xi_{Gauss(0,1)} \sim N(0, 1)$

How many and how to choose the functions $\{h_i(\xi)\}$?

$m = 1$: $J^{**}(\xi) = \{\mathbf{M}h(\xi) - \mathbf{M}h(\xi_{Gauss(0,1)})\}^2$ h - arbitrary nonquadratic function

ξ - symmetric and $h(y) = y^4$ - kurtosis based approximation

$m = 2$: $J^{**}(\xi) = k_1 \times \{Mh_1(\xi) - Mh_1(\xi_{Gauss(0,1)})\}^2 + k_2 \times \{Mh_2(\xi) - Mh_2(\xi_{Gauss(0,1)})\}^2$

The following choices of $\{h_i(y)\}$ have proved very useful ($m = 2$)

$h_1(y) = \frac{1}{\alpha_1}\log\cosh(\alpha_1 \times y)$, $1 \leq \alpha_1 \leq 2$; $h_2(y) = -\exp(-y^2/2)$

Gradient descent algorithm - ($m = 1$)

$F(w) = \{\mathbf{M}h(w^T \times \mathbf{X})) - \mathbf{M}h(\xi_{Gauss(0,1)})\}^2$

$\nabla F(w) = \mathbf{M}(\mathbf{X} \times g(w^T \times \mathbf{X}))$

$h(y) = \frac{1}{\alpha_1}\log\cosh(\alpha_1 \times y)$

$g(y) = h'(y) = \tanh(\alpha_1 \times y)$

Iterations in the gradient descent method:

- $w \leftarrow w + \Delta w$, $\quad \Delta w \propto \mathbf{M}(\mathbf{X} \times g(w^T \times \mathbf{X}))$
- $w \leftarrow w/\|w\|$

$$\mathbf{M}(\mathbf{X} \times g(w^T \times \mathbf{X})) = \frac{1}{T}\sum_{t=1}^{T} X_t \times g\left(w^T \times X_t\right)$$

A fast fixed-point algorithm - FastICA ($m = 1$)

Iterations in the FastICA method:

- $w \leftarrow \mathbf{M}(\mathbf{X} \times h(w^T \times \mathbf{X})) - \mathbf{M}(g(w^T \times \mathbf{X})) \times w$
- $w \leftarrow w/\|w\|$

Fundamental approaches to the solution of the ICA problem:

2) Likelihood maximization

Probability theory: $\mathbf{X}$ and $\mathbf{S}$ - random vectors with densities $p_{\mathbf{X}}(x)$ and $p_{\mathbf{S}}(s)$, respectively

$\mathbf{X} = \mathbf{A} \times \mathbf{S} \quad \rightarrow \quad p_{\mathbf{X}}(\mathbf{x}) = |\mathrm{Det}(\mathbf{B})| \times p_{\mathbf{S}}(\mathbf{s}), \quad \mathbf{B} = \mathbf{A}^{-1}$

$\mathbf{S} = \mathbf{B} \times \mathbf{X} \qquad \mathbf{B} = (b_1, b_2, \dots, b_q)^T$

$s_1, s_2, \dots, s_q$ - independent signals: $p_{\mathbf{S}}(s) = p_{\mathbf{S}}(s_1, s_2, \dots, s_q) = p_1(s_1) \times p_2(s_2) \times \dots \times p_q(s_q)$

$s_i = b_i^T \times \mathbf{X} \qquad p_{\mathbf{X}}(\mathbf{x}) = |\mathrm{Det}(\mathbf{B})| \times p_1(b_1^T \times \mathbf{X}) \times p_2(b_2^T \times \mathbf{X}) \times \dots \times p_q(b_q^T \times \mathbf{X})$

$X_1, X_2, \dots, X_T$ - training dataset

Likelihood: $L(\mathbf{B}, p_s) = \left\{\prod_{t=1}^{T}\prod_{i=1}^{q} p_i(b_i^T x_t)\right\} \times |\mathrm{Det}(\mathbf{B})|^T \to \max$

If density p_s of independent signals are known from some prior knowledge, likelihood depends only on unknown matrix **B**

$\mathbf{W} = \arg\max L(\mathbf{B}, p_s)$ - desired unmixing matrix

Under unknown density p_s, it is approximated by some estimator p^* and $L(\mathbf{B}, p^*)$ depends only on **B**:

$\mathbf{W} = \arg\max L(\mathbf{B}, p^*)$ - desired unmixing matrix

Maximum likelihood approach:

- a choice of ‘appropriate’ densities $p^*(s)$
- maximization $L(\mathbf{B}, p^*)$ over $\mathbf{B}$

Math: $\{p_i\}$ - chosen densities, $p^*(s_1, s_2, \dots, s_q) = p_1(s_1) \times p_2(s_2) \times \dots \times p_q(s_q)$.

Maximum likelihood estimator $\mathbf{W} = \arg\max L(\mathbf{B}, p^*)$ is locally consistent, if the assumed densities meet the conditions

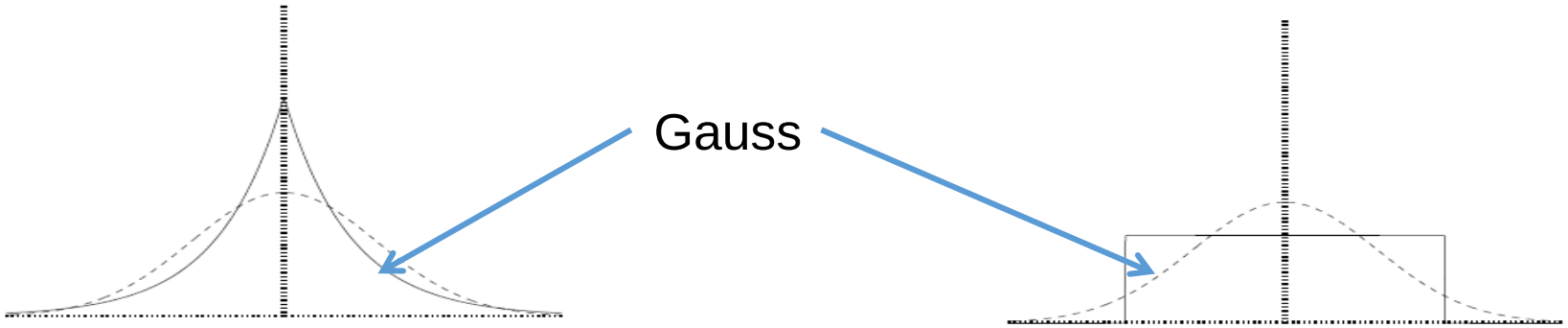
$$\textbf{Index } \Psi(p_i) = \mathbf{M}\{s_i \times g_i(s_i) - (g_i(s_i))'\} > 0, \qquad g_i(s) = (\log p_i(s))'$$

for all i.

Example:

$\log(p_1(s)) = \alpha_1 - 2\log\cosh(s)$ $\log(p_2(s)) = \alpha_2 - [s^2/2 - \log\cosh(s)]$

$p_1(s)$ - supergaussian density with positive kurtosis - with ‘heavy’ tails (the density is relatively large at zero and at large values of the variable, while being small for intermediate values



$p_2(s)$ - subgaussian density with negative kurtosis - have typically a “flat” density which is rather constant near zero, and very small for larger values of the variable

$\log(p_1(s)) = \alpha_1 - 2\log\cosh(s)$

$\log(p_2(s)) = \alpha_2 - [s^2/2 - \log\cosh(s)]$

$\Psi(p_1) = 2\mathbf{M}\{-\tanh(s) \times s + (1 - \tanh^2 s)\}$

$\Psi(p_2) = \mathbf{M}\{\tanh(s) \times s - (1 - \tanh^2 s)\}$

$\Psi(p_i) = -2\Psi(p_2)$

Maximization $L(\mathbf{B}, p) = \left\{\prod_{t=1}^{T}\prod_{i=1}^{q} p(b_i^T \times X_t)\right\} \times |\mathrm{Det}(\mathbf{B})|^T$ over $\mathbf{B}$ $\qquad p_i(s) = p(s)$

$L(\mathbf{B}, p) \rightarrow \frac{1}{T}\log L(\mathbf{B}, p) = \sum_{t=1}^{T}\sum_{i=1}^{q} \log p(b_i^T \times X_t) + \log|\mathrm{Det}\mathbf{B}|$

$$\nabla = \frac{1}{T}\frac{\partial}{\partial \mathbf{B}}\log L(\mathbf{B}, p) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{g}(\mathbf{B} \times X_t) \times X_t^T + [\mathbf{B}^T]^{-1}$$

$$\mathbf{g}(\mathbf{B}\times X_t) = \begin{pmatrix} g(b_1^T \times X_t) \\ \cdots \\ g(b_q^T \times X_t) \end{pmatrix} \qquad g(s) = (\log p(s))'$$

Iterations in the gradient descent maximization:

$$\mathbf{B} \leftarrow \mathbf{B} + \Delta\mathbf{B}, \qquad \Delta\mathbf{B} \propto \frac{1}{T}\sum_{t=1}^{T} \mathbf{g}(\mathbf{B} \times X_t) \times X_t^T + [\mathbf{B}^T]^{-1}$$

There exists fast fixed-point algorithm for likelihood maximization - FastICA

Fundamental approaches to the solution of the ICA problem:

3) Minimization of Mutual Information

Probability theory: $\mathbf{S}$ - random vector

Mutual information $I(s_1, s_2, \dots, s_q)$ between $s_1, s_2, \dots, s_q$ is:

$$I(s_1, s_2, \dots, s_q) = \sum_{i=1}^{q} H(s_i) - H(s_1, s_2, \dots, s_q)$$

$$p_{\mathbf{s}}(s) = p_{\mathbf{s}}(s_1, s_2, \dots, s_q) \rightarrow$$

$$H(s_1, s_2, \dots, s_q) = -\int p_{\mathbf{s}}(y_1, y_2, \dots, y_q) \times \log_2 p_{\mathbf{s}}(y_1, y_2, \dots, y_q) dy_1 dy_2 \dots dy_q$$

$$p_i(s_i) \rightarrow H(s_i) = -\int p_i(y) \times \log_2 p_i(y) dy$$

$s_1, s_2, \dots, s_q$ - independent signals: $p_{\mathbf{S}}(s) = p_{\mathbf{S}}(s_1, s_2, \dots, s_q) = p_1(s_1) \times p_2(s_2) \times \dots \times p_q(s_q)$

$$H(s_1, s_2, \dots, s_q) = \sum_{i=1}^{q} H(s_i) \quad \rightarrow \quad I(s_1, s_2, \dots, s_q) = 0$$

Probability theory: Kullback-Leibler divergence.

p_1, p_2 - densities $\rightarrow$ $d(p_1, p_2) = \int p_1(y) \log \frac{p_1(y)}{p_2(y)} dy \geq 0$

- kind of a distance between the two probability densities, because $d(p_1, p_2)$ is always nonnegative,

and zero if and only if $p_1(y) = p_2(y)$

$p_1 = p_{\mathbf{S}}(s_1, s_2, \dots, s_q)$, $p_2 = p_1(s_1) \times p_2(s_2) \times \dots \times p_q(s_q)$

$\rightarrow$ $d(p_1, p_2) = I(s_1, s_2, \dots, s_q)$

Mutual information $I(s_1, s_2, \dots, s_q)$ has minimal value when $s_1, s_2, \dots, s_q$ - independent

$\mathbf{S} = \mathbf{B} \times \mathbf{X}$ $\rightarrow$ $p_{\mathbf{X}}(\mathbf{x}) = |\mathrm{Det}(\mathbf{B})| \times p_{\mathbf{S}}(\mathbf{s})$

$H(\mathbf{S}) = H(X) + \log |\mathrm{Det}(\mathbf{B})|$

$I(s_1, s_2, \dots, s_q) = \sum_{i=1}^{q} H(s_i) - H(\mathbf{X}) - \log |\mathrm{Det}(\mathbf{B})|$

B - orthogonal matrix $\rightarrow$ $\log |\mathrm{Det}(\mathbf{B})| = 0$

$s_i = b_i^T \times \mathbf{X}$: $\sum_{i=1}^{q} H\left(b_i^T \times \mathbf{X}\right) = -\mathbf{M} \sum_{i=1}^{q} \log_2 p\left(b_i^T \times \mathbf{X}\right) \rightarrow \min$

$F(b) = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{q} \log_2 p\left(b_i^T \times X_t\right) \rightarrow \max$ under constraint $b \times b^T = 1$

- Negentropy maximization

Mutual information minimization - Negentropy maximization