Pattern Recognition (PR)

Prof. Dr.-Ing. Andreas Maier
Pattern Recognition Lab (CS 5), Friedrich-Alexander-Universität Erlangen-Nürnberg
Winter Term 2020/21

This is a printable version of the slides of the lecture

Pattern Recognition (PR)
Winter term 2020/21
Friedrich-Alexander University of Erlangen-Nuremberg.

Erlangen, January 8, 2021
Prof. Dr.-Ing. Andreas Maier

Independent Component Analysis

Cocktail-Party Problem

Example

Imagine the following situation:

- You have two microphones in a room at different locations.
- The microphones record time signals $x_1(t), x_2(t)$.
- Each recorded signal is a weighted sum of two speakers $s_1(t), s_2(t)$:

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

Parameters a_{ij} depend on the distance of the microphones to the speakers.

Cocktail-Party Problem

Example

For simplicity, we assume just a very simple mixing model without any time delays or other factors.

Observations:

- If we knew the a_{ij}, the problem of reconstructing s_i is to solve the linear equations by classical methods.
- But: We do not know the a_{ij}! Thus, the problem is considerably more difficult!

Cocktail-Party Problem

Example

Original sound sources:


Samples at the cocktail-party:

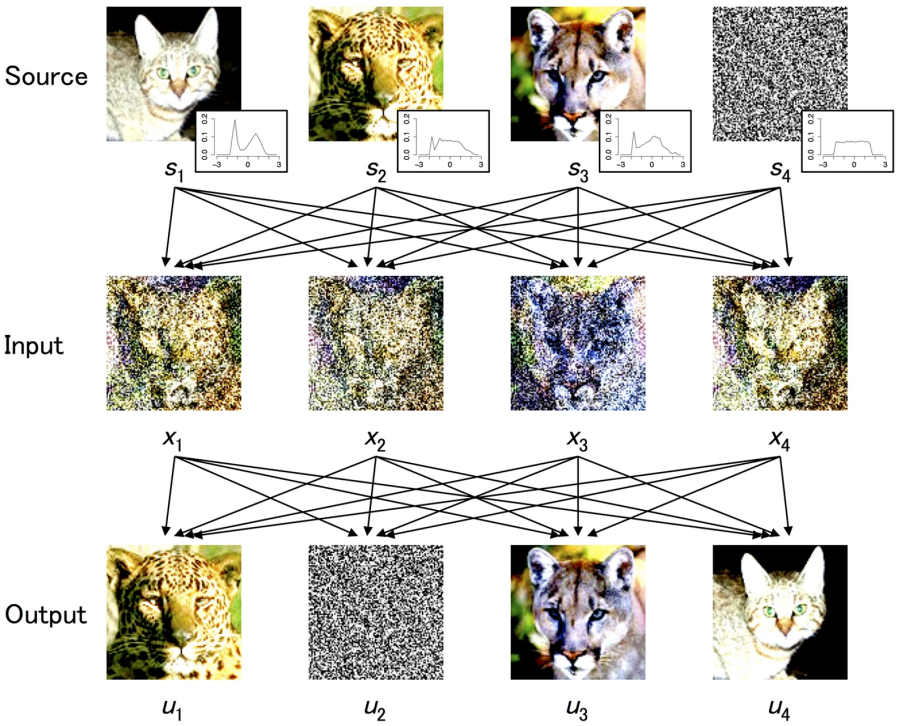

Reconstructed sound sources:


Cocktail-Party Problem

The principle for solving the cocktail-party problem has a lot of other interesting applications:

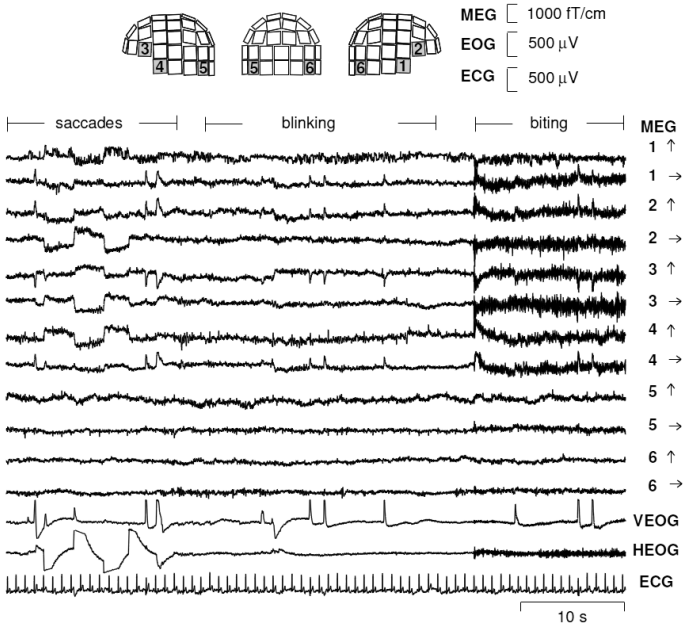
- speech signal recovery: telecommunications
- recovery of images from mixed signals: MRI, fMRI
- electrical recordings of brain activity:
  - EEG (electroencephalography)
  - MEG (magnetoencephalography)
- feature extraction
- multispectral image analysis

Separate Natural Images(cont.)



Isomura, T., Toyoizumi, T. A Local Learning Rule for Independent Component Analysis. Sci Rep 6, 28073 (2016)

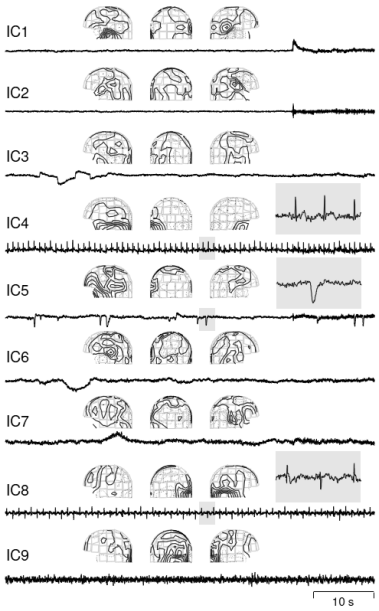
MEG: Recovery of Brain Activity



R. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari, E. Oja. "Independent component analysis for identification of artifacts in magnetoencephalographic recordings, Advances in Neural Information Processing Systems"

Fig.: Principle of MEG acquisition.

MEG: Recovery of Brain Activity (cont.)



R. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari, E. Oja. "Independent component analysis for identification of artifacts in magnetoencephalographic recordings, Advances in Neural Information Processing Systems"

Fig.: Recovered signals.

Common Framework

Idea:

- Use some information about the statistical properties of the signals $s_i(t)$ to estimate a_{ij}.

Surprisingly, it turns out that the only statistical assumption that we have to make is that the $s_i(t)$ are *statistically independent* at each time point t.

Formulation in a unified mathematical framework (Hérault and Jutten, 1984-1991):

ICA – Independent Component Analysis

Latent Variables and Factor Analysis

Statistical latent variables model:

- Rewrite the time series into n linear mixture observations $x_1, \ldots, x_n$
- Each mixture x_i as well as each component s_j are random variables

$$x_i = \sum_{j=1}^{m} a_{ij} s_j \,, \qquad i = 1, \ldots, n$$

In matrix notation:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$$

where

- $\boldsymbol{A}$ is a constant *mixing* matrix
- s_j are latent random variables (independent components)
- both $\boldsymbol{A}$ and s_j have to be estimated based on observations x_i

First Approach: Decorrelation

Assuming $\bar{x} = 0$, we already know a latent variable representation (see chapter *Discriminant Analysis I*).

From:

$$E\{\boldsymbol{x}\boldsymbol{x}^T\} = \boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$$

and

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{U}^T = \left(\boldsymbol{U}\boldsymbol{D}^{-\frac{1}{2}}\right) \cdot \boldsymbol{I} \cdot \left(\boldsymbol{U}\boldsymbol{D}^{-\frac{1}{2}}\right)^T$$

we compute a mapping:

$$\tilde{x} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{x}$$

First Approach: Decorrelation

For zero-mean vectors, the mapping:

$$\tilde{x} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{x}$$

is called *Whitening Transform*.

It has some interesting properties:

- The mapped random variables $\tilde{x}_i$ are uncorrelated.
- $\tilde{x}$ has unit variance:

$$\begin{aligned} E\{\tilde{x}\tilde{x}^T\} &= E\left\{\left(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{x}\right)\left(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{x}\right)^T\right\} = E\left\{\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{U}\boldsymbol{D}^{-\frac{1}{2}}\right\} \\ &= \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{\Sigma}\boldsymbol{U}\boldsymbol{D}^{-\frac{1}{2}} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{U}^T \cdot \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T \cdot \boldsymbol{U}\boldsymbol{D}^{-\frac{1}{2}} \\ &= \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{D}\boldsymbol{D}^{-\frac{1}{2}} = \boldsymbol{I} \end{aligned}$$

First Approach: Decorrelation

We could interpret the mapped random variable $\tilde{x}$ as an estimate of the latent variable model:

$$\boldsymbol{s} = \tilde{x}$$

But this would give poor results.

Problem: The whitening transform is not unique!

Consider for example an arbitrary orthogonal matrix $\boldsymbol{R} \in \mathbb{R}^{n \times n}$:

$$\begin{aligned} \hat{x} &= \boldsymbol{R}\tilde{x} = \boldsymbol{R}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{x} \\ E\{\hat{x}\hat{x}^T\} &= E\left\{\boldsymbol{R}\left(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{x}\right)\left(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{x}\right)^T\boldsymbol{R}^T\right\} \\ &= \boldsymbol{R}\boldsymbol{I}\boldsymbol{R}^T \\ &= \boldsymbol{I} \end{aligned}$$

Pattern Recognition Lab

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Next Time in Pattern Recognition



Second Approach: Independence

Observations:

- Lack of correlation determines the second-degree cross-moments (covariances) of a multi-variate distribution.
- Statistical independence is stronger, as it determines all of the cross-moments.

Given 2 statistically independent random variables y_1, y_2 and 2 functions h_1, h_2:

$$\begin{aligned} E\{h_1(y_1)\,h_2(y_2)\} &= \iint h_1(y_1)\,h_2(y_2)\,p(y_1,y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 \\ &= \iint h_1(y_1)\,p(y_1)\,h_2(y_2)\,p(y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 \\ &= \int h_1(y_1)\,p(y_1)\,\mathrm{d}y_1 \int h_2(y_2)\,p(y_2)\,\mathrm{d}y_2 \\ &= E\{h_1(y_1)\}\,E\{h_2(y_2)\} \end{aligned}$$

Second Approach: Independence

These extra moment conditions allow us to identify the elements of $\boldsymbol{A}$ uniquely.

Case of Gaussian distribution:

- Gaussian distribution is determined by its second moments alone.
- Any Gaussian independent components can be determined only up to a rotation

Therefore, we assume that the s_i are independent and non-Gaussian.

Whitening Transform

The whitening transform is usually done before ICA as a pre-processing step:

- Mixing matrix $\boldsymbol{A}$ has n^2 degrees of freedom.
- Whitening transforms the mixing matrix $\boldsymbol{A}$ into $\tilde{A}$:

$$\tilde{x} = \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{U}^T \boldsymbol{A}\boldsymbol{s} = \tilde{A}\boldsymbol{s}$$

- The new mixing matrix is orthogonal:

$$E\{\tilde{x}\tilde{x}^T\} = \tilde{A}\, E\{\boldsymbol{s}\boldsymbol{s}^T\}\, \tilde{A}^T = \boldsymbol{I}$$

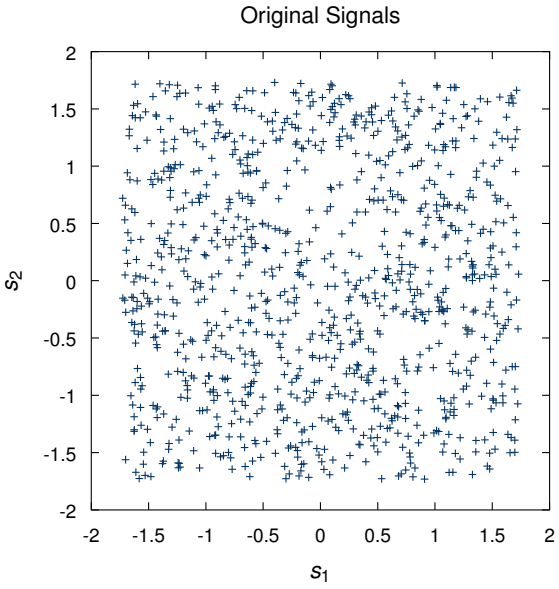
- $\tilde{A}$ is orthogonal and has $n(n-1)/2$ degrees of freedom

Thus, applying the whitening transform solves roughly half of the problem.

Illustration of ICA

Consider two independent components with the following uniform distributions:

$$p(s_i) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{, if } |s_i| \leq \sqrt{3} \\ 0 & \text{, otherwise} \end{cases}$$



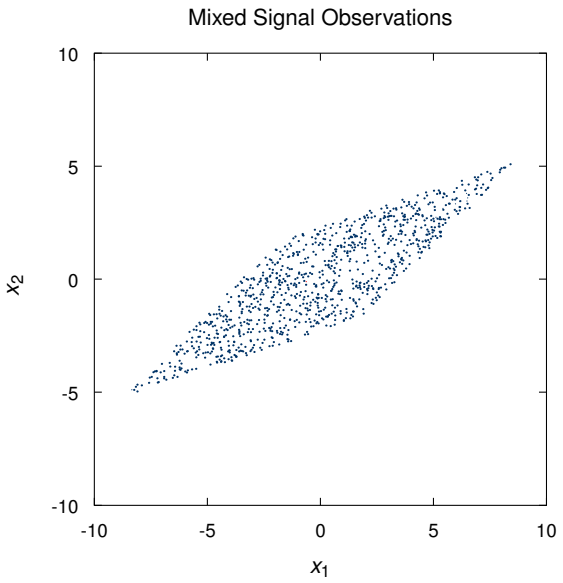
Properties of the joint pdf:

- signal components are independent
- joint pdf is uniform on square
- zero mean
- variance is equal to one

Illustration of ICA (cont.)

The two independent components are mixed with the matrix:

$$\boldsymbol{A} = \begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix}, \text{ which results in } \boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}\,.$$



Properties of the mixed signal:

- joint pdf of mixed signals is uniform on a parallelogram

More important:

- x_1, x_2 are not independent any more

Illustration of ICA (cont.)

Intuitive way of estimating $\boldsymbol{A}$:

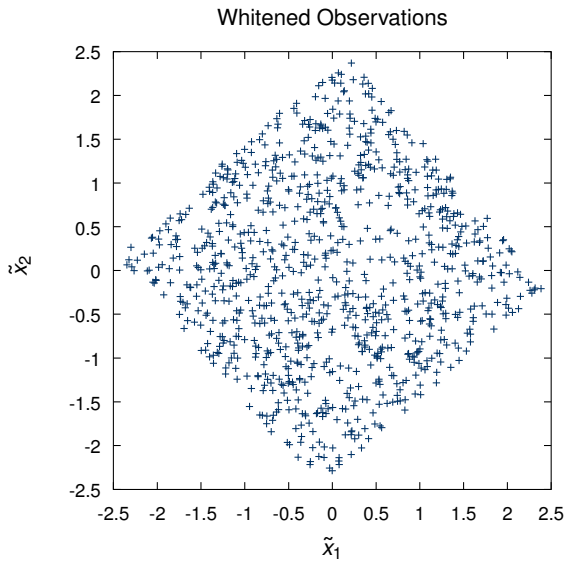
- Edges of the parallelogram are in the directions of the columns of $\boldsymbol{A}$.
- In principle, we could first estimate the joint pdf of x_1, x_2.
- If we locate the edges of the joint pdf, we can estimate $\boldsymbol{A}$.

But:

- Computationally quite expensive
- This principle works only with exactly uniform distributions.

Illustration of ICA (cont.)

Effect of the whitening transform applied to the data:



Properties of the whitened observations:

- Joint pdf of $\tilde{x}$ is uniform on a square.
- Components are determined except for rotation.
- Problem of recovering $\tilde{A}$ is much simpler.

Basic Properties

Assumptions for the ICA model:

- We assume that the s_j are mutually independent.
- The s_j have to be non-Gaussian in order to determine them from the x_i.
- For simplicity, we assume that $\boldsymbol{A}$ is square.

Ambiguities of the ICA model:

- The s_j are defined only up to a multiplicative constant.
- The s_j are not ordered.

Ambiguities of ICA

Writing the ICA model in terms of the colums of $\boldsymbol{A}$:

$$\boldsymbol{x} = \sum_{i=1}^{n} \boldsymbol{a}_i s_i$$

- Any scalar multiplier for s_i can be eliminated by scaling $\boldsymbol{a}_i$ appropriately.
- The matrix $\boldsymbol{A}$ can be adapted to restrict the s_i to have unit variance.
- This still leaves the ambiguity of the sign:
  multiplying s_i by ± 1 does not affect the model.
- This ambiguity is usually insignificant in most applications.

Ambiguities of ICA (cont.)

Ambiguity of the ordering:

$$\boldsymbol{x} = \sum_{i=1}^{n} \boldsymbol{a}_i s_i$$

- As both $\boldsymbol{s}$ and $\boldsymbol{A}$ are unknown we can change the order of the summation.
- Formalization using a permutation matrix $\boldsymbol{P}$:

$$\boldsymbol{x} = \underbrace{\boldsymbol{A}\boldsymbol{P}}_{\boldsymbol{A}^*} \underbrace{\boldsymbol{P}^{-1}\boldsymbol{s}}_{\boldsymbol{s}^*}$$

- $\boldsymbol{A}^*$ is just a new mixing matrix to be solved.

Basic Principle of ICA

So far, if we know $\boldsymbol{A}$, we could compute its inverse $\boldsymbol{A}^{-1}$ to obtain the independent components. Consider a linear combination of x_i with a weight vector $\boldsymbol{w}$:

$$y = \boldsymbol{w}^T \boldsymbol{x}$$

Clearly, y equals one of the independent components if $\boldsymbol{w}$ is one row of $\boldsymbol{A}^{-1}$.

Basic Principle of ICA (cont.)

Change in variables:

$$\boldsymbol{z} = \boldsymbol{A}^T \boldsymbol{w}$$

Applied to the linear combination:

$$y = \boldsymbol{w}^T \boldsymbol{x} = \boldsymbol{w}^T \boldsymbol{A} \boldsymbol{s} = \boldsymbol{z}^T \boldsymbol{s}$$

Result from the Central Limit Theorem:

- The sum of a number of independent random variables tends toward a normal distribution.
- $\boldsymbol{z}^T \boldsymbol{s}$ is *more Gaussian* than any of the s_i
- $\boldsymbol{z}^T \boldsymbol{s}$ is *least Gaussian* when it equals one of the s_i

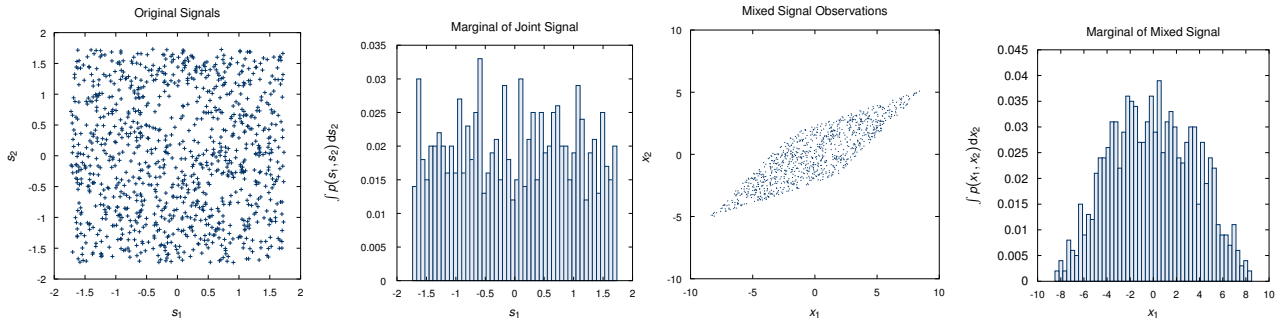
Basic Principle of ICA (cont.)

Key principle of ICA:

Maximizing the non-Gaussianity of $\boldsymbol{w}^T\boldsymbol{x}$ results in the independent components!

Basic Principle of ICA (cont.)

Marginal distributions of the joint and the mixed signal:



Basic Principle of ICA (cont.)

Reasons for the importance of the non-Gaussianity:

- In case of Gaussian random variables, the ICA model can only be estimated up to an orthogonal transformation.

  Note:
  If just one of the components is Gaussian, ICA still works.

- The Gaussian is the *most random* distribution within the family of pdfs with given mean and variance.
- Therefore, it is the least informative pdf with respect to the underlying data.

Distributions that have the *least resemblance* to the Gaussian reveal more structure associated with the data.

Importance of Non-Gaussianity

The *randomness* can be measured using the concept of entropy from Shannon's information theory:

- Entropy is a measure of the uncertainty of an event, or the randomness of a measure.

Differential Entropy

The *differential entropy* $H(X)$ of a continuous random variable X with density $p(x)$ is defined as

$$H(p) = -\int p(x) \log p(x) \, \mathrm{d}x$$

Importance of Non-Gaussianity (cont.)

Theorem

The Gaussian maximizes the entropy over all distributions with the same mean and the same covariance.

Proof:

- Let x be the random variable, $p(x)$ the pdf that has the highest randomness.
- Rewrite moments M_i equations using a set of polynomials $\{r_i(x)\}$:

$$\int p(x) r_i(x)\,\mathrm{d}x = M_i \text{ , where } M_i \text{ are called moments.}$$

- Using $r_0(x) \equiv 1$, $M_0 = 1$ constrains $p(x)$ to be a pdf.

Importance of Non-Gaussianity (cont.)

Lagrangian functional for the maximum entropy problem:

$$\underset{p(x)}{\operatorname{argmin}} J \equiv \underset{p(x)}{\operatorname{argmin}} \int p(x) \log p(x)\,\mathrm{d}x - \sum_{i=0}^{N} \lambda_i \left(\int p(x) r_i(x)\,\mathrm{d}x - M_i \right)$$

Taking the functional derivative with respect to $p(x)$ (Gâteaux derivative) and setting it to zero:

$$\frac{\delta J}{\delta p} = \log p(x) + 1 - \sum_{i=0}^{N} \lambda_i r_i(x) \overset{!}{=} 0$$

yields the family of exponential distributions:

$$p(x) = \exp\left(-1 + \sum_{i=0}^{N} \lambda_i r_i(x) \right)$$

Importance of Non-Gaussianity (cont.)

Result for using first and second moments for mean μ and variance σ^2:

$$p(x) = e^{-(1-\lambda_0-\lambda_1 x-\lambda_2(x-\mu)^2)}$$

Plug the form into the constraints:

$$\begin{aligned}
\int e^{-(1-\lambda_0-\lambda_1 x-\lambda_2(x-\mu)^2)}\,\mathrm{d}x &= 1 \\
\int x e^{-(1-\lambda_0-\lambda_1 x-\lambda_2(x-\mu)^2)}\,\mathrm{d}x &= \mu \\
\int (x-\mu)^2 e^{-(1-\lambda_0-\lambda_1 x-\lambda_2(x-\mu)^2)}\,\mathrm{d}x &= \sigma^2
\end{aligned}$$

Importance of Non-Gaussianity (cont.)

Integrate analytically (non-trivial) and solve for Lagrangian multipliers:

$$\begin{aligned}
\lambda_0 &= 1 - \frac{1}{2}\log(2\pi\sigma^2) \\
\lambda_1 &= 0 \\
\lambda_2 &= -\frac{1}{2\sigma^2}
\end{aligned}$$

Insert the results into the form of $p(x)$:

$$\begin{aligned}
p(x) &= e^{-(1-\lambda_0-\lambda_1 x-\lambda_2(x-\mu)^2)} \\
&= e^{-\frac{1}{2}\log(2\pi\sigma^2)} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\
&= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\end{aligned}$$

□

ICA Estimation Algorithm

Basic ICA Estimation Algorithm:

Apply centering transform
Apply whitening transform
$i \leftarrow 1$
Take a random vector $\boldsymbol{w}_i$
Maximize non-Gaussianity of $\boldsymbol{w}_i^T \boldsymbol{x}$ subject to $\|\boldsymbol{w}_i\| = 1$ $\boldsymbol{w}_j^T \boldsymbol{w}_i = 0 \ , \ j < i$
$i \leftarrow i + 1$
$i > n$ (n: number of independent components)
Use weight matrix: $\boldsymbol{W} = \left(\boldsymbol{w}_1^T, \boldsymbol{w}_2^T, \ldots, \boldsymbol{w}_n^T\right)$ to compute $\boldsymbol{s}$
Output: independent components $\boldsymbol{s}$

ICA Estimation Algorithm (cont.)

Notes:

- Estimation by maximizing non-Gaussianity of independent components.
- There exist equivalent algorithms for solving the ICA:
  - Gradient descent methods
  - Fast ICA
- Relation to Projection Pursuit approach (Friedman and Tukey, 1974):
  - Projection Pursuit is a method for visualization and exploratory data analysis.
  - Attempts to show clustering structure by finding *interesting* projections.
  - Interestingness is usually measured by non-Gaussianity.

Pattern Recognition Lab

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Next Time in Pattern Recognition



Measures of Non-Gaussianity

- So far, we have seen that the key principle in estimating independent components is the non-Gaussianity.
- In order to optimize the independent components, we need a quantitative measure of non-Gaussianity.

Consider the random variable y and assume that it has zero mean and unit variance (enforced by pre-processing).

We will consider three measures of non-Gaussianity:

- Kurtosis
- Negentropy
- Mutual Information

Kurtosis

Definition

The *Kurtosis* of y is defined as:

$$\text{kurt}(y) = E\{y^4\} - 3\left(E\{y^2\}\right)^2$$

Since y has unit variance, the equation simplifies to:

$$\text{kurt}(y) = E\{y^4\} - 3$$

Linearity properties for independent random variables y_1, y_2:

$$\begin{aligned} \text{kurt}(y_1 + y_2) &= \text{kurt}(y_1) + \text{kurt}(y_2) \\ \text{kurt}(\alpha y) &= \alpha^4 \,\text{kurt}(y)\,, \quad \alpha \in \mathbb{R} \end{aligned}$$

Kurtosis (cont.)

Kurtosis for a Gaussian distribution:

- The n-th central moment of a Gaussian distribution $p(y) = \mathcal{N}(y|\mu, \sigma^2)$ with mean μ and variance σ^2 is:

$$E\{(y-\mu)^n\} = \begin{cases} (n-1)!! \cdot \sigma^n & , \; n \text{ even} \\ 0 & , \; n \text{ odd} \end{cases}$$

  Note: $(n)!!$ denotes the double factorial, i. e. the product of every odd number from 1 to n.

- Thus, for a zero mean, unit variance random variable y that is normally distributed:

$$\text{kurt}(y) = 0$$

Kurtosis (cont.)

Properties of Kurtosis:

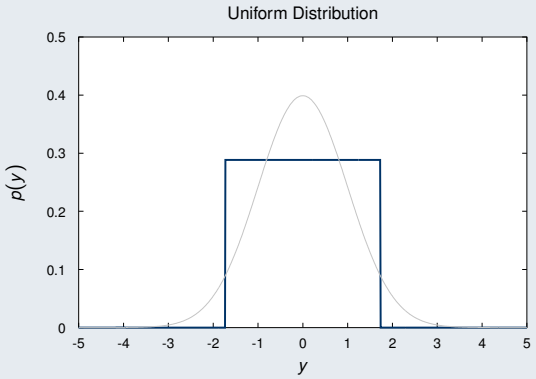
- Kurtosis is zero for a Gaussian random variable.
- For most (but not all) non-Gaussian random variables, Kurtosis is nonzero.
- Kurtosis can be positive or negative.
- Typically, non-Gaussianity is measured as:
  - $|\text{kurt}(y)|$ or
  - $\text{kurt}(y)^2$

Kurtosis (cont.)

Example

Subgaussian pdf:

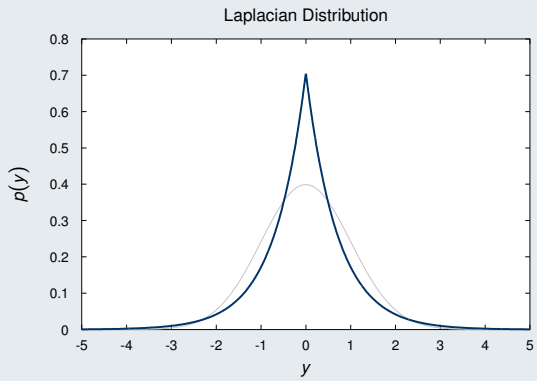
$$p(y) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{, if } |y| \leq \sqrt{3} \\ 0 & \text{, otherwise} \end{cases}$$



$\text{kurt}(y) < 0$

Supergaussian pdf:

$$p(y) = \frac{1}{\sqrt{2}} \exp(-\sqrt{2}|y|)$$



$\text{kurt}(y) > 0$

Kurtosis (cont.)

- Consider 2-D case using the linear combination:

$$\begin{aligned} y &= \boldsymbol{w}^T\boldsymbol{x} = \boldsymbol{w}^T\boldsymbol{A}\boldsymbol{s} = \boldsymbol{z}^T\boldsymbol{s} = z_1 s_1 + z_2 s_2 \\ \mathrm{kurt}(y) &= \mathrm{kurt}(z_1 s_1) + \mathrm{kurt}(z_2 s_2) = z_1^4\,\mathrm{kurt}(s_1) + z_2^4\,\mathrm{kurt}(s_2) \end{aligned}$$

- As y has unit variance, concerning also s_1, s_2:

$$E\{y^2\} = z_1^2 + z_2^2 = 1$$

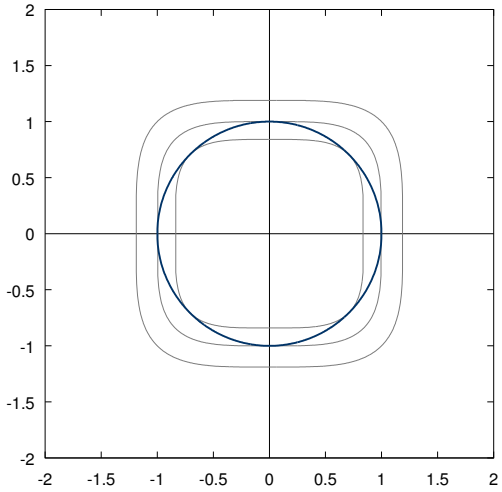
  which constrains $\boldsymbol{z}$ to the unit circle in the 2-D plane.

- Thus, we have to find the maximum of the following function on the unit circle w. r. t. $\boldsymbol{z}$:

$$|\mathrm{kurt}(y)| = |z_1^4\,\mathrm{kurt}(s_1) + z_2^4\,\mathrm{kurt}(s_2)|$$

Kurtosis

Optimization landscape for Kurtosis in 2-D plane:



- Thick curve is the unit circle
- Thin curves are isocontours of the objective function
- The maxima are located at sparse values of $\boldsymbol{z}$, i. e. when:

$$y = \pm s_i$$

Kurtosis (cont.)

Maximizing the non-Gaussianity of a vector $\boldsymbol{w}$ in practice:

- Start with some initial vector $\boldsymbol{w}$.
- Use a gradient descent method to optimize:

$$\underset{\boldsymbol{w}}{\operatorname{argmax}} \left|\operatorname{kurt}(y)\right| = \underset{\boldsymbol{w}}{\operatorname{argmax}} \left|\operatorname{kurt}(\boldsymbol{w}^T \boldsymbol{x})\right|$$

- Plug this optimization into the ICA estimation algorithm (see above).

Kurtosis (cont.)

Kurtosis as a function of the direction of the projection:

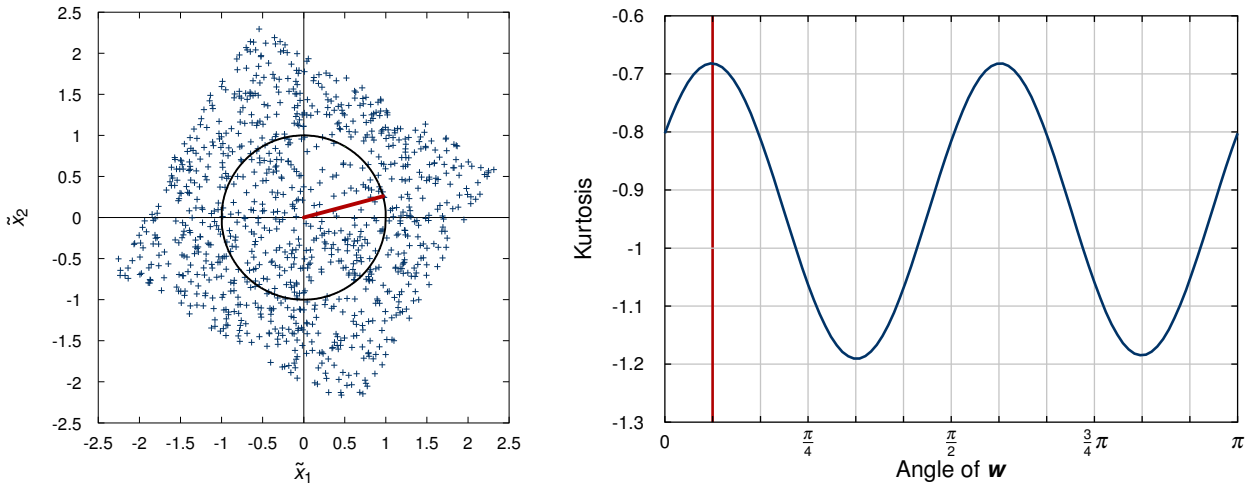


Fig.: Example for negative Kurtosis.

Kurtosis (cont.)

Kurtosis as a function of the direction of the projection:

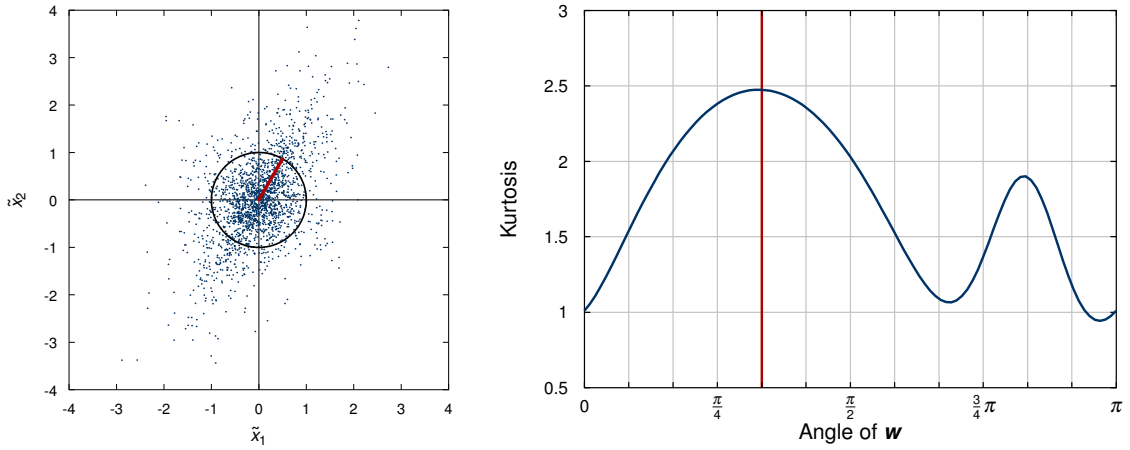


Fig.: Example for positive Kurtosis.

Kurtosis (cont.)

Despite being a valid measure for non-Gaussianity, Kurtosis has some drawbacks in practice when it is computed on a set of measurement samples:

- Kurtosis can be very sensitive to outliers due to higher order statistical moments.
- Not optimal for supergaussian variables even without outliers.
- It is not a robust measure of non-Gaussianity.

Negentropy

Observations:

- A Gaussian variable has the largest entropy among all random variables of equal variance.
- Entropy is small for distributions that are “spiky”.
- Negentropy can be used as a measure for non-Gaussianity:
  it is zero for Gaussian random variables and always non-negative.

Negentropy

The *Negentropy* $J(\boldsymbol{y})$ of a random variable $\boldsymbol{y}$ is defined as follows:

$$J(\boldsymbol{y}) = H(\boldsymbol{y}_{\text{Gauss}}) - H(\boldsymbol{y})$$

where $\boldsymbol{y}_{\text{Gauss}}$ is a Gaussian random variable with the same covariance as $\boldsymbol{y}$.

Negentropy (cont.)

Properties of Negentropy:

- Measure is well justified by statistical theory.
- In theory, negentropy is an optimal statistical estimator of non-Gaussianity.
- Computing the negentropy from a measured sample set requires the estimation of the pdf.
- The (non-parametric) estimation of a pdf from samples is non-trivial and often computationally expensive.

Approximations for negentropy exist that are both robust and computationally more efficient than the direct pdf-based approach.

Mutual Information

Information-theoretic approach:

- Negentropy measures the difference in terms of information value to Gaussian random variables.
- Instead, we could measure the statistical dependency between the random variables directly.
- *Mutual Information* is a concept to measure the entire dependency structure of random variables (and not only the covariance)

Mutual Information

Mutual Information

The *Mutual Information* between n scalar random variables $\boldsymbol{y} = y_1, \ldots, y_n$ is defined as:

$$\begin{aligned} \mathrm{MI}(\boldsymbol{y}) &= \sum_{i=1}^{n} H(y_i) - H(\boldsymbol{y}) \\ &= \mathrm{D}_{\mathrm{KL}}\left(p(\boldsymbol{y}) \| \prod_{i=1}^{n} p(y_i) \right) \end{aligned}$$

where H denotes the (differential) entropy and D_{KL} the Kullback-Leibler divergence.

Mutual Information (cont.)

Interpretation:

- Entropy can be regarded as a measure for code length.
- $H(y_i)$ is a measure for the code length necessary to encode y_i.
- $H(\boldsymbol{y})$ can be regarded as the code length necessary to encode the entire vector $\boldsymbol{y}$.

- In this context, MI shows the reduction in code length obtained when encoding $\boldsymbol{y}$ instead of the components y_i separately.

Mutual Information (cont.)

Properties of MI:

- For an invertible linear transform $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$:

$$\mathrm{MI}(\boldsymbol{y}) = \sum_{i=1}^{n} H(y_i) - H(\boldsymbol{x}) - \log\left|\det \boldsymbol{W}\right|$$

- For uncorrelated y_i of unit variance, MI and negentropy differ only by a constant and a sign:

$$\mathrm{MI}(\boldsymbol{y}) = C - \sum_{i=1}^{n} J(y_i)$$

Therefore, instead of maximizing the negentropy we can minimize the Mutual Information to compute the direction of highest non-Gaussianity.

Lessons Learned

- ICA is a simple model based on a linear non-Gaussian latent variables
- Non-Gaussianity as key principle
- Estimation by maximizing non-Gaussianity of independent components
- Equivalence between Kurtosis, Negentropy and Mutual Information

Pattern Recognition Lab

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Next Time in

Pattern Recognition



Further Readings

Examples and various content have been taken mainly from the overview paper:

- A. Hyvärinen, E. Oja:
  Independent Component Analysis: Algorithms and Applications,
  Neural Networks, 13(4-5):411-430, 2000.

Further reading:

- T. Hastie, R. Tibshirani, J. Friedman:
  The Elements of Statistical Learning,
  2nd Edition, Springer, 2009.
- T. M. Cover, J. A. Thomas:
  Elements of Information Theory,
  2nd Edition, John Wiley & Sons, 2006.

Comprehensive Questions

- What is the latent variables model behind independent component analysis?
- Why are whitening transformed observations not the independent components?
- Why is non-Gaussianity so important for the independent components?
- How can non-Gaussianity be measured?