group of machine gmum learning research

Independent Component Analysis

Przemysław Spurek

Table of Contents

Table of Contents

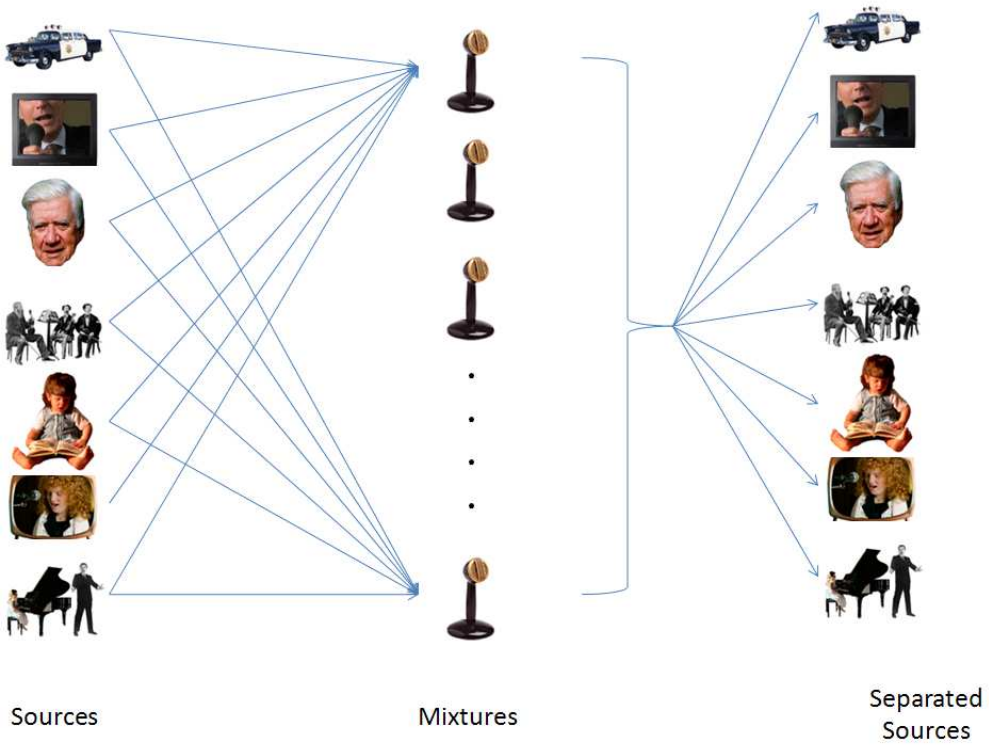
Cocktail-party problem

Imagine that you are in a room where two people are speaking simultaneously. We could express this as a linear equation:

$$\begin{cases} x_1(t) = a_{11}s_1 + a12s_2 \\ x_2(t) = a_{21}s_1 + a22s_2 \end{cases}$$

It would be very useful if you could now estimate the two original speech signals s1(t) and s2(t), using only the recorded signals x1(t) and x2(t).

- `https://github.com/przem85/ICA_presentation/blob/master/ICA_signals.ipynb`



https://onionesquereality.wordpress.com/tag/cocktail-party-problem/

Definition of ICA

Assume that we observe n linear mixtures $x_1, \ldots, x_n$ of n independent components

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \ldots + a_{jn}s_n, \text{ for all } j.$$

- In the ICA model, we assume that each mixture x_j as well as each independent component s_k is a random variable.
- Without loss of generality, we can assume that both the mixture variables and the independent components have zero mean.

Vector-matrix notation

- Let us denote by $\mathbf{x}$ the random vector whose elements are the mixtures $x_1, \ldots, x_n$,
- Let us denote by $\mathbf{s}$ the random vector with elements $s_1, \ldots, s_n$.
- Let us denote by $\mathbf{A}$ the matrix with elements a_{ij}.

The above mixing model is written as

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

Sometimes we need the columns of matrix $\mathbf{A}$; denoting them by $\mathbf{a}_j$ the model can also be written as

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

Vector-matrix notation

After estimating the matrix $\mathbf{A}$, we can compute its inverse, say $\mathbf{W}$, and obtain the independent component simply by:

$$\mathbf{s} = \mathbf{W}\mathbf{x}.$$

Ambiguities of ICA

- We cannot determine the variances (energies) of the independent components. The reason is that, both **s** and **A** being unknown, any scalar multiplier in one of the sources s_i could always be cancelled by dividing the corresponding column $\mathbf{a}_i$ of **A** by the same scalar. The most natural way to do this is to assume that each has unit variance: $E\{s_i^2\} = 1$.
- Note that this still leaves the ambiguity of the sign: we could multiply the an independent component by -1 without affecting the model.
- We cannot determine the order of the independent components.

Table of Contents

Illustration of ICA

- https://github.com/przem85/ICA_presentation/blob/master/ICA_basis.ipynb

Definition and fundamental properties

To define the concept of independence, consider two scalar-valued random variables y_1 and y_2.

Basically, the variables y_1 and y_2 are said to be independent if information on the value of y_1 does not give any information on the value of y_2, and vice versa.

Definition and fundamental properties

Technically, independence can be defined by the probability densities. Let us denote by p(y1,y2) the joint probability density function (pdf) of y_1 and y_2. Let us further denote by $p_1(y_1)$ the marginal pdf of y_1, i.e. the pdf of y_1 when it is considered alone:

$$p_1(y_1) = \int p(y_1, y_2) dy_2,$$

and similarly for y2.

Definition and fundamental properties

Then we define that y1 and y2 are independent if and only if the joint pdf is factorizable in the following way:

$$p(y_1, y_2) = p_1(y_1)p_2(y_2).$$

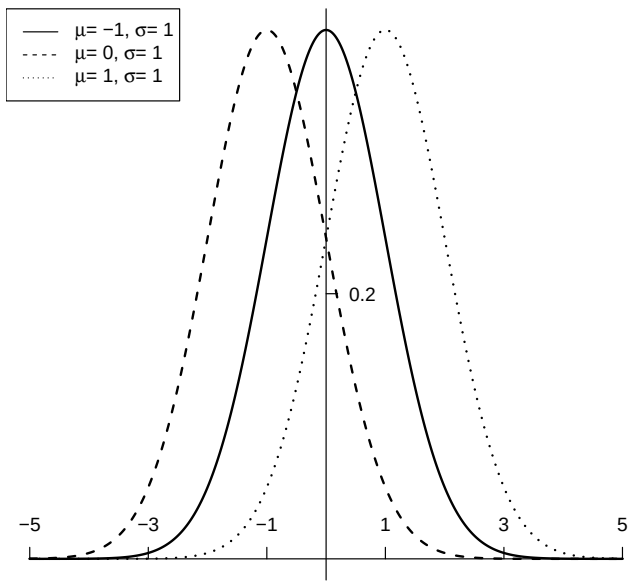
Table of Contents

Why Gaussian variables are forbidden

The fundamental restriction in ICA is that the independent components must be nongaussian for ICA to be possible.

Normal distribution

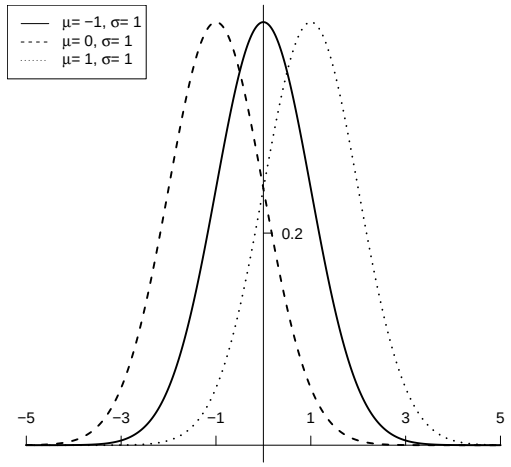
$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$



Normal distribution

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

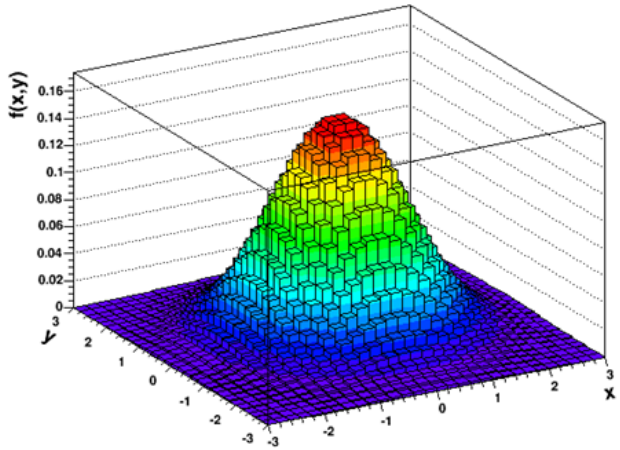
$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)\right), \quad x \in \mathbb{R}$$

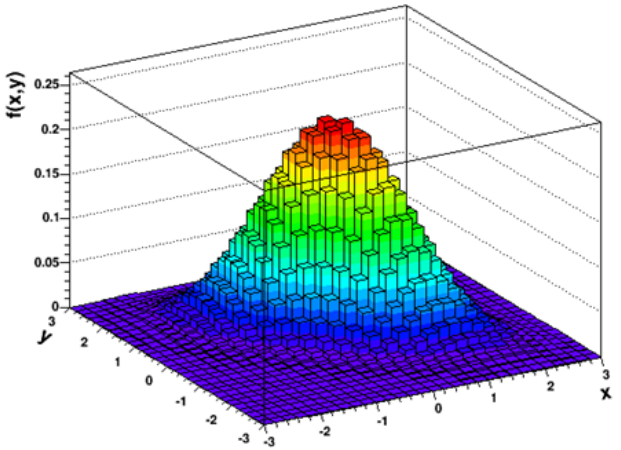


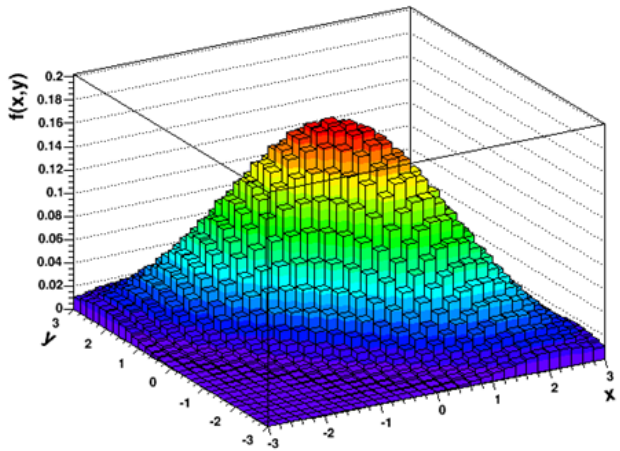
The N-dimensional normal distribution

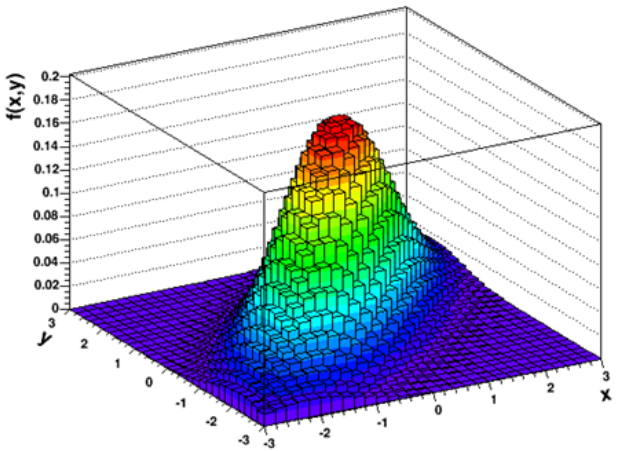
The multivariate normal distribution is said to be "non-degenerate" when the covariance matrix Σ of the multivariate normal distribution is symmetric and positive definite. In this case the distribution has density

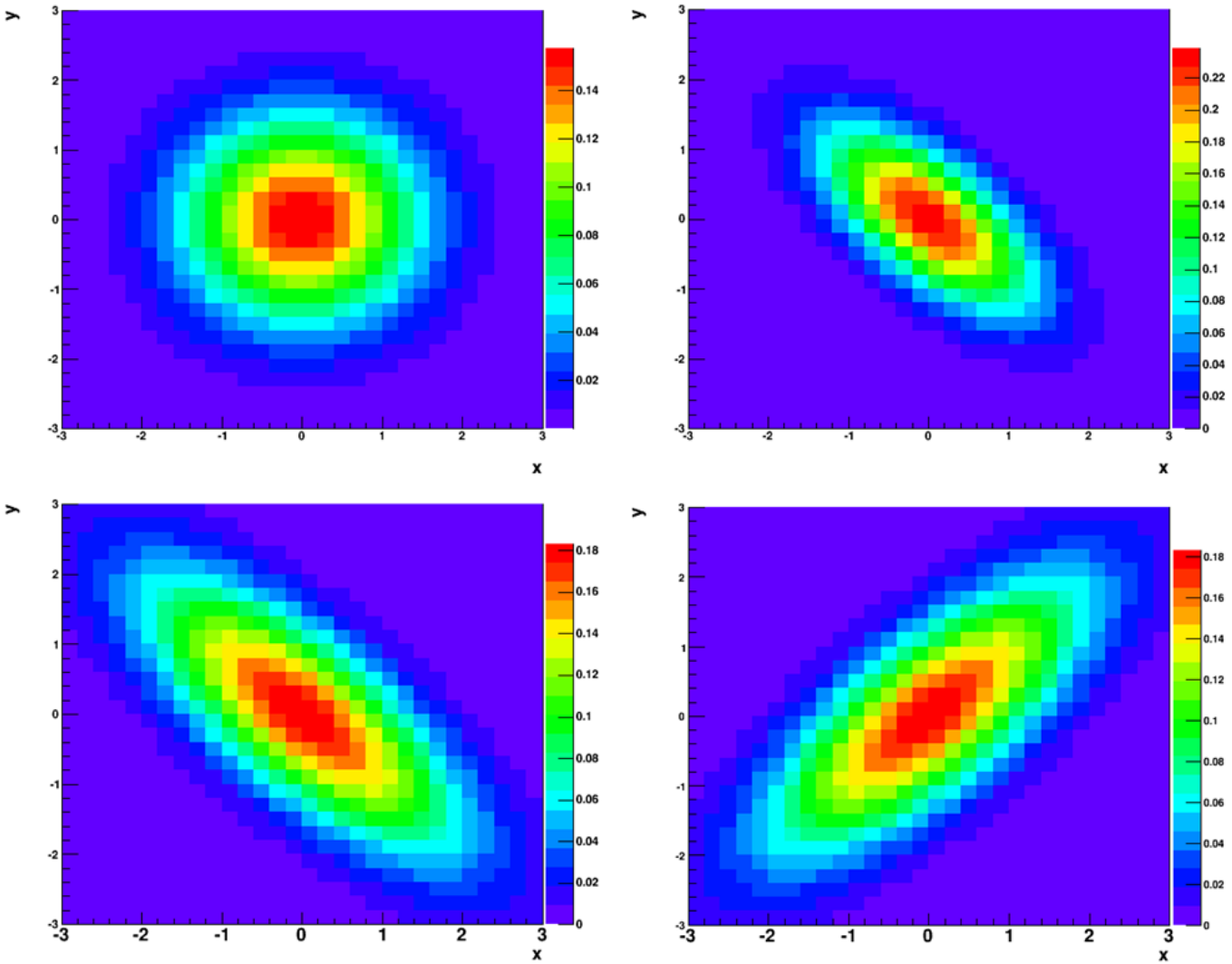
$$f(x_1, \ldots, x_n, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$



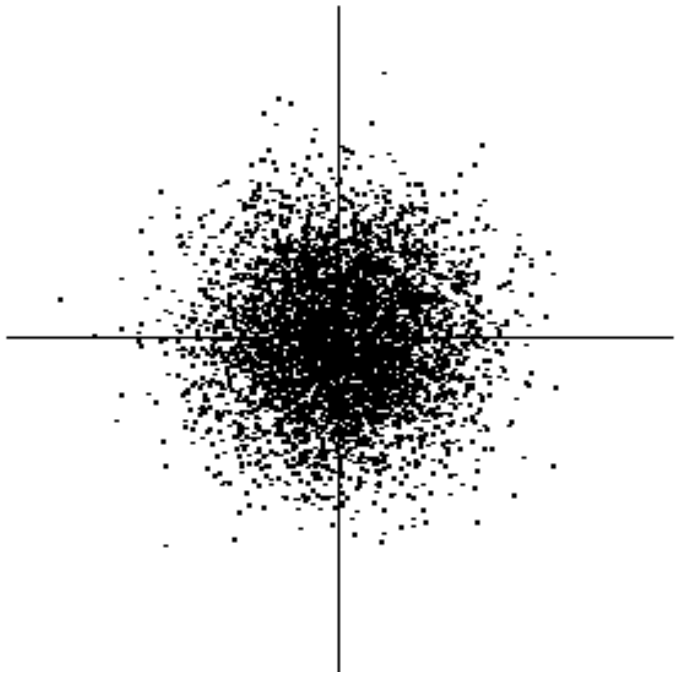








The Figure shows that the density is completely symmetric. Therefore, it does not contain any information on the directions of the columns of the mixing matrix **A**. This is why **A** cannot be estimated.



More rigorously, one can prove that the distribution of any orthogonal transformation of the gaussian (x_1, x_2) has exactly the same distribution as (x_1, x_2), and that x_1 and x_2 are independent. Thus, in the case of gaussian variables, we can only estimate the ICA model up to an orthogonal transformation. In other words, the matrix $\mathbf{A}$ is not identifiable for gaussian independent components.

- `https://github.com/przem85/ICA_presentation/blob/master/ICA_nongaussian_1.ipynb`
- `https://github.com/przem85/ICA_presentation/blob/master/ICA_nongaussian_2.ipynb`

Table of Contents

“Nongaussian is independent”

http://fourier.eng.hmc.edu/e161/lectures/ica/node3.html

Kurtosis

The classical measure of nongaussianity is kurtosis or the fourth-order cumulant. The kurtosis of y is classically defined by

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

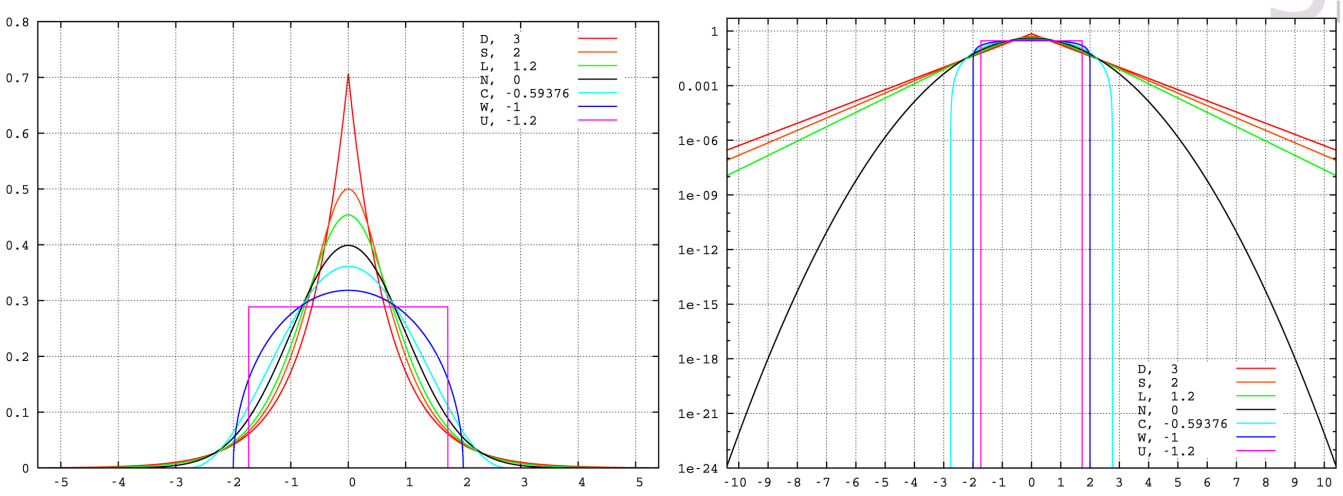
Kurtosis

For a gaussian y, the fourth moment equals $3(E\{y^2\})^2$.
Thus, kurtosis is zero for a gaussian random variable. For most (but not quite all) nongaussian random variables, kurtosis is nonzero.

Kurtosis

- Kurtosis can be both positive or negative.
- Random variables that have a negative kurtosis are called subgaussian, and those with positive kurtosis are called supergaussian.
- Supergaussian random variables have typically a "spiky" pdf with heavy tails, i.e. the pdf is relatively large at zero and at large values of the variable, while being small for intermediate values.

Kurtosis



Example 1

A typical example is the Laplace distribution, whose pdf (normalized to unit variance) is given by

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

Kurtosis

Typically nongaussianity is measured by the absolute value of kurtosis. The square of kurtosis can also be used. These are zero for a gaussian variable, and greater than zero for most nongaussian random variables. There are nongaussian random variables that have zero kurtosis, but they can be considered as very rare.
`http://cis.legacy.ics.tkk.fi/aapo/papers/IJCNN99_tutorialweb/node13.html`

Entropy

Entropy H is defined for a discrete random variable Y as

$$H(Y) = -\sum_i P(Y = a_i) \log P(Y = a_i).$$

where the a_i are the possible values of Y.

Entropy

This very well-known definition can be generalized for continuous-valued random variables and vectors, in which case it is often called differential entropy. The differential entropy H of a random vector $\mathbf{y}$ with density $f(\mathbf{y})$ is defined as:

$$H(\mathbf{y}) = -\int f(\mathbf{y}) \log f(\mathbf{y}) \mathrm{d}\mathbf{y}.$$

Negentropy

To obtain a measure of nongaussianity that is zero for a gaussian variable and always nonnegative, one often uses a slightly modified version of the definition of differential entropy, called negentropy. Negentropy J is defined as follows

$$J(\mathbf{y}) = H(\mathbf{y}_{gauss}) - H(\mathbf{y})$$

where $\mathbf{y}_{gauss}$ is a Gaussian random variable of the same covariance matrix as $\mathbf{y}$. Due to the above-mentioned properties, negentropy is always non-negative, and it is zero if and only if $\mathbf{y}$ has a Gaussian distribution.
`http://cis.legacy.ics.tkk.fi/aapo/papers/IJCNN99_tutorialweb/node15.html`

The likelihood

A very popular approach for estimating the ICA model is maximum likelihood estimation.
Suppose that the original density F of the data $X \subset \mathbb{R}^D$ is independent in the coordinate system A. Then in the coordinates given by A the density F factors as

$$F(x) = \frac{1}{\det(A)} f_1(\alpha_1) \cdot \ldots \cdot f_D(\alpha_D) \text{ where } x = \alpha_1 a_1 + \ldots + \alpha_D a_D.$$

and f_i are the marginal densities, where α_i denote the coefficients of x in the base $A = [a_1, \ldots, a_D]$. Observe that

$$x = A^{-1}\alpha \text{ where } \alpha = [\alpha_1, \ldots, \alpha_D]^T.$$

This means that by the likelihood the probability of drawing the data X by the factors defined on the RHS of the above equation is maximal.

Inverse reasoning

Thus to find the coordinate system A for which the data is independent, we can apply the inverse approach, and try to maximize with respect to the coordinate system the probability of drawing the data from the marginal densities.
It is possible to formulate directly the likelihood in the noise-free ICA model and then estimate the model by a maximum likelihood method. Denoting by $\mathbf{W} = (\mathbf{w}_1, ..., \mathbf{w}_n)^T$ the matrix $\mathbf{A}^{-1}$, the log-likelihood takes the form:

$$L = \sum_{t=1}^{T} \sum_{i=1}^{n} \log f_i(\mathbf{w}_i^T \mathbf{x}(t)) + T \log |\det \mathbf{W}|$$

where the f_i are the density functions of the s_i (here assumed to be known), and the $\mathbf{x}(t), t = 1, ..., T$ are the realizations of $\mathbf{x}$.

Choice of the density family for marginal densities

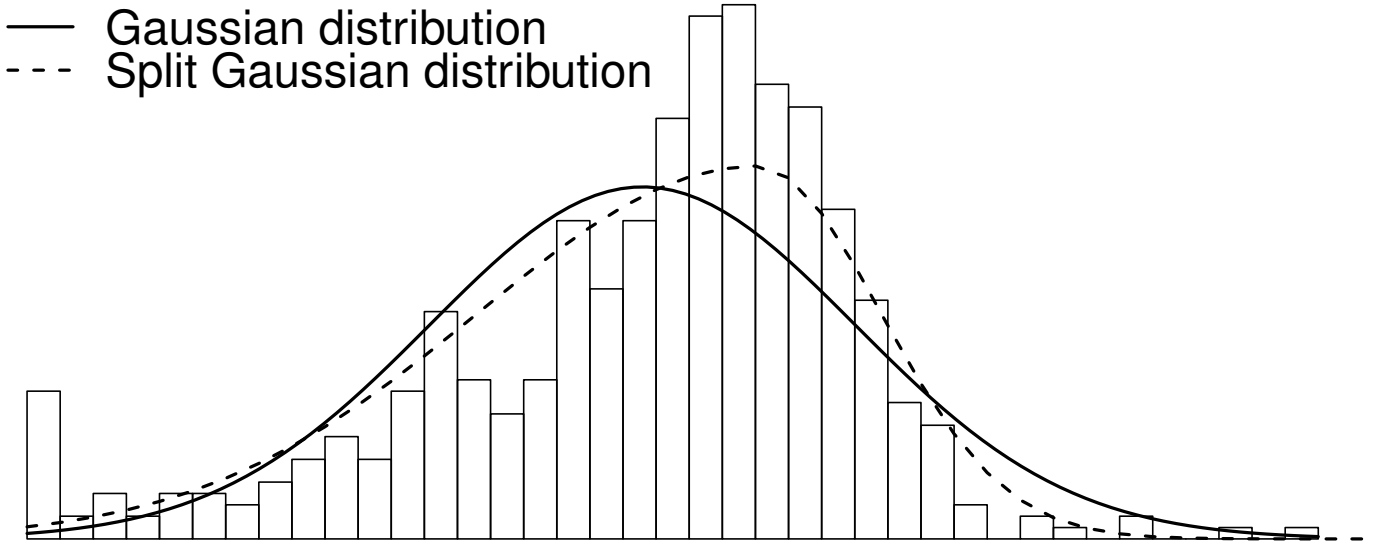
There appears a problem how to estimate the marginal densities (observe, that in practice we only have the data). Luckily, we are in one dimensional space, which allows to work efficiently with density estimation. One can use an arbitrary density family which is larger then gaussians. Most authors use sub or super gaussians, which differentiate from gaussians basing on the size of the tails. In our opinion it is not the best choice, as the data is typically bounded and the reliable estimation of tails is not easy.

Split gaussians

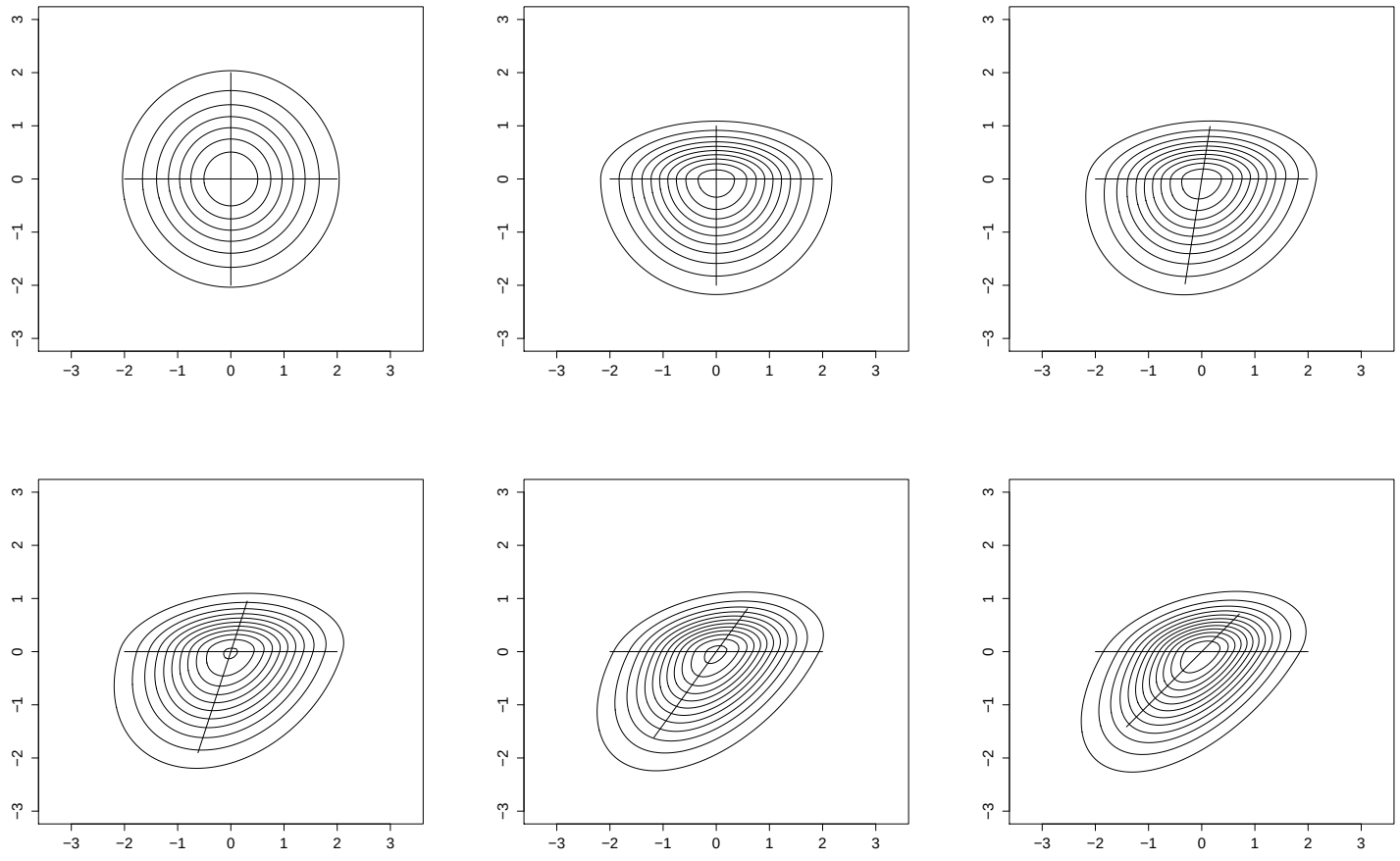
In [ICA based on asymmetry P Spurek, J Tabor, P Rola, M Ociepka Pattern Recognition 67, 230-244] we have decided to use the asymmetry of the data as the way of differentiating it from normality. From the large family of possible choices we have decided to take Split-Gaussians:

$$SN(x; m, \sigma^2, \tau^2) = \begin{cases} c \cdot \exp[-\frac{1}{2\sigma^2}(x-m)^2], & \text{where } x \leq m \\ c \cdot \exp[-\frac{1}{2\tau^2\sigma^2}(x-m)^2], & \text{where } x > m \end{cases}$$

where $c = \sqrt{\frac{2}{\pi}}\sigma^{-1}(1+\tau)^{-1}$.
They are easy to deal with as they come as the "gluing" of two gaussians in their common modal value. We can fit this distribution to the data easily – given the choice of m we can calculate the optimal σ and τ.



Comparison between fitting Gaussian distribution and Split Gaussian distribution.



- http://ww2.ii.uj.edu.pl/~spurek/publications/ica.pdf
- https://arxiv.org/pdf/1802.05550.pdf
- https://github.com/przem85/ICA_presentation/blob/master/ICA_sound_1.ipynb
- https://github.com/przem85/ICA_presentation/blob/master/ICA_sound_2.ipynb
- https://github.com/przem85/ICA_presentation/blob/master/ICA_img.ipynb
- https://github.com/przem85/ICA_presentation/blob/master/ICA_EEG.ipynb

Table of Contents

Non-linear ICA

Model random vector x as a function (not necessarily linear as in standard ICA) of statistically independent sources s:

$$x = f(s)$$

$$x = [x_1, ..., x_n]^T$$

$$s = [s_1, ..., s_n]^T$$

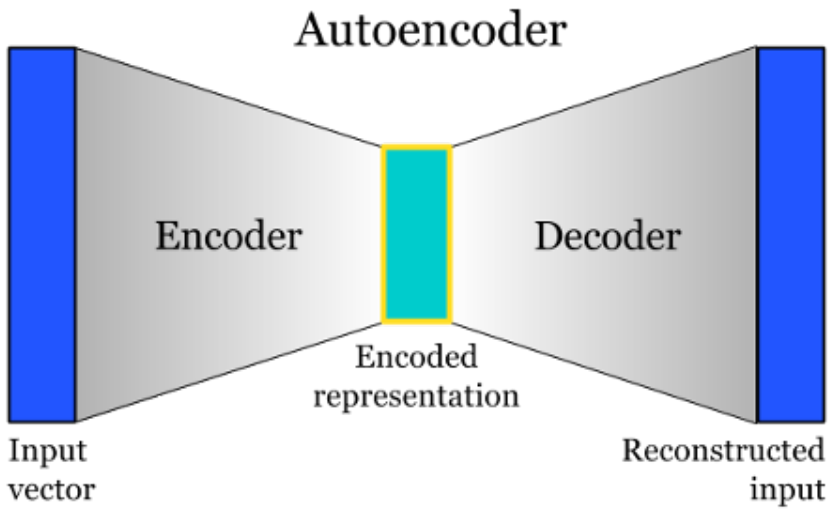
Problem: Observing only x, can we retrieve s? i.e. we want to find an unmixing function g such that the joint probability $g(x)$ factorizes.

- The fundamental problem in solving NICA is that the solution is in principle non-identifiable.
- Without any constraints on the space of the mixing functions, there exists an infinite number of solutions.
- To illustrate, consider that there is an infinite number of possible nonlinear decompositions of a random vector into independent components, and those decompositions are not related to each other in any trivial way.

AutoEncoder

Generalization of PCA, idea based on compression of dataset $X = (x_i) \subset \mathbb{R}^N$ to a linear space Z of smaller dimension D (*latent space*).
We have an encoder $\mathbb{R}^N \ni x \to \mathcal{E}x \in Z$ and decoder $Z \ni z \to \mathcal{D}z \in \mathbb{R}^N$.
We want to find such encoder and decoder which minimize reconstruction error:

$$MSE(X; \mathcal{E}, \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \mathcal{D}(\mathcal{E}x_i)\|^2.$$



(N)ICA with an Autoencoder

Given instances of random vector x as the input for the network our goal is to find a encoding $z = \phi(x)$ such that the latent variable z is independent. So we worry about the standard reconstruction loss **and** the independence loss measured on z:

$$loss(\theta) = \beta loss_{rec}(\theta) + loss_{ind}(\theta, z)$$

Question: How to measure independence?

(N)ICA with an Autoencoder

How to measure independence?

- Weighted independence index `https://arxiv.org/pdf/2001.04147.pdf`
- GAN `https://arxiv.org/pdf/1710.05050.pdf`
- Cramer Wold distance `https://arxiv.org/pdf/1903.00201.pdf`

NICE

- NICE: g_θ modeled by neural network, such that the Jacobian is trivial, and g_θ is invertible.
- `https://arxiv.org/pdf/1410.8516.pdf`
- `https://lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html`