APPLIED MACHINE LEARNING

Independent Component Analysis (ICA)

Algorithm

Adapted from Hyvarinen @ 2000

ICA: Algorithm

Consider a N-dimensional observation vector $x \in \mathbb{R}^N$

Assume that x was generated by a linear variable model

$$x = As$$

A: is an unknown $N \times N$ mixing matrix

$s \in \mathbb{R}^N$ are unknown latent random variables,
referred to as the sources.

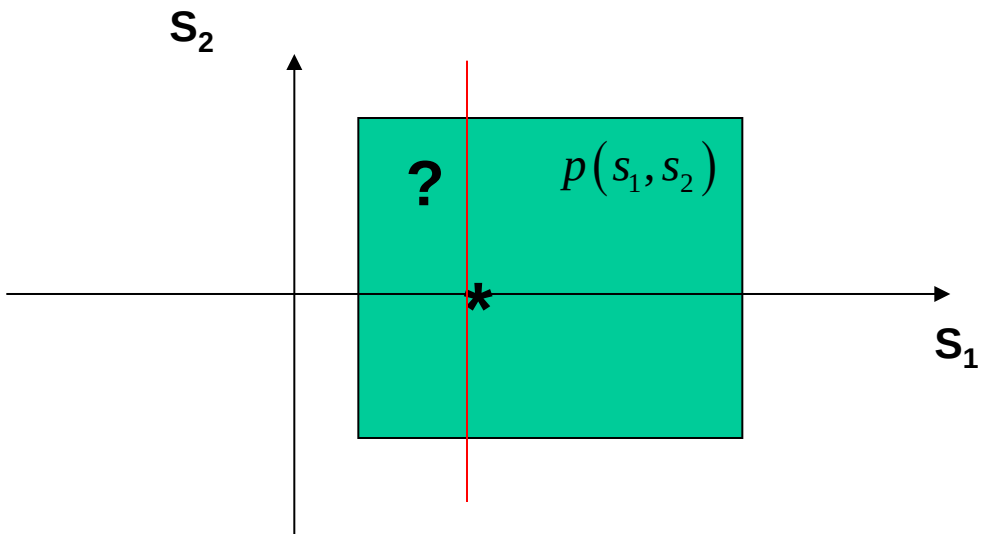
ICA estimates both the mixing matrix A and the sources s knowing only x.

1st ambiguity of ICA: x is product of A and s

ICA: Hypotheses

Two hypotheses: 1: $s := \{s_1, ..., s_N\}$ are statistically independent.

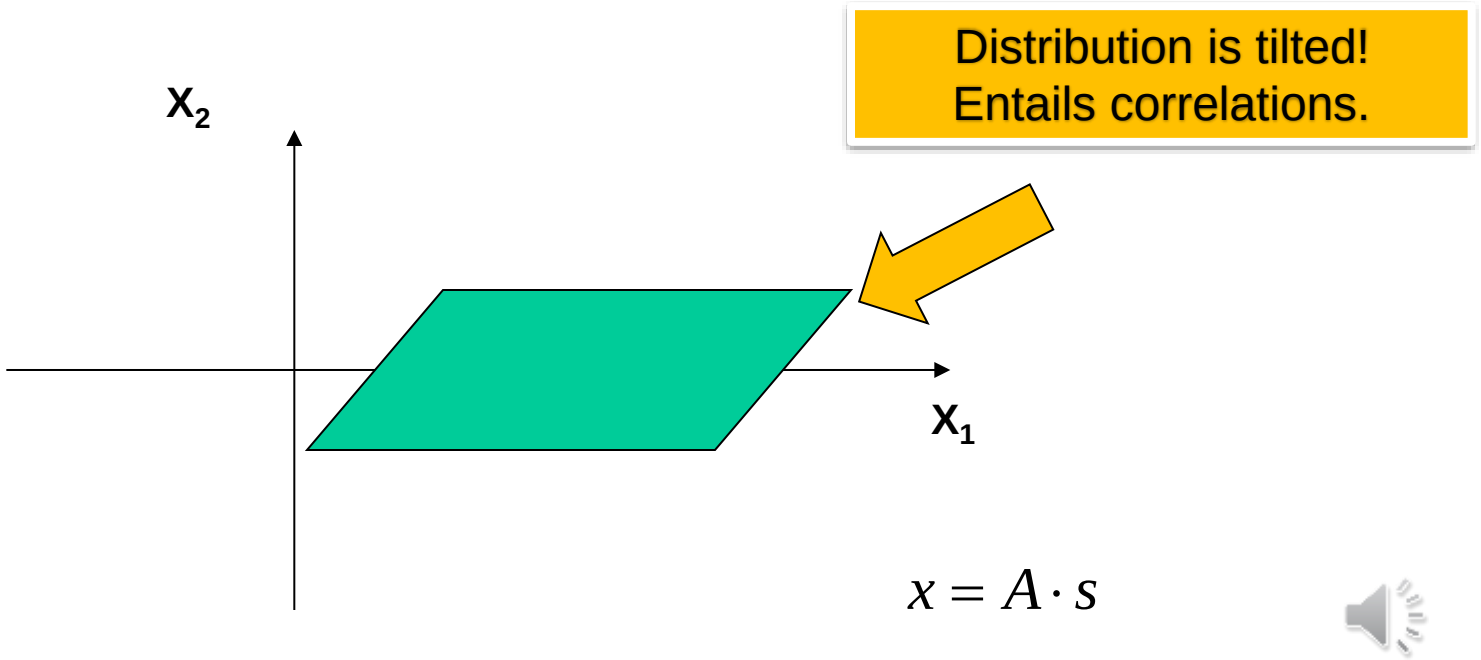
2: The distribution of s is likely to be non-Gaussian.



Here is an example of two independent sources s_1 and s_2, whose distribution is non-Gaussian (here uniform distribution).

ICA: Statistical dependency

Observables X_1 and X_2, result from a linear combination of the independent sources. The distribution of the mixed observables is not statistically independent.



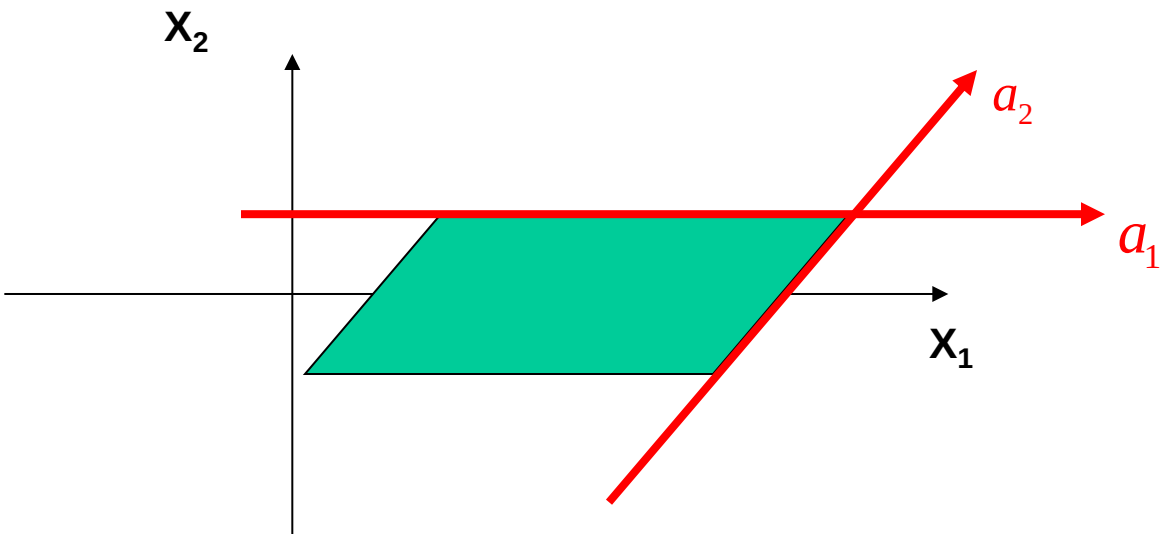
Approach: Use the correlation to find A, then compute $s = A^{-1}x$

ICA: Finding mixing matrix - intuition

Idea:
Finds the columns of the matrix *A*.
The vectors embeds correlation across variables

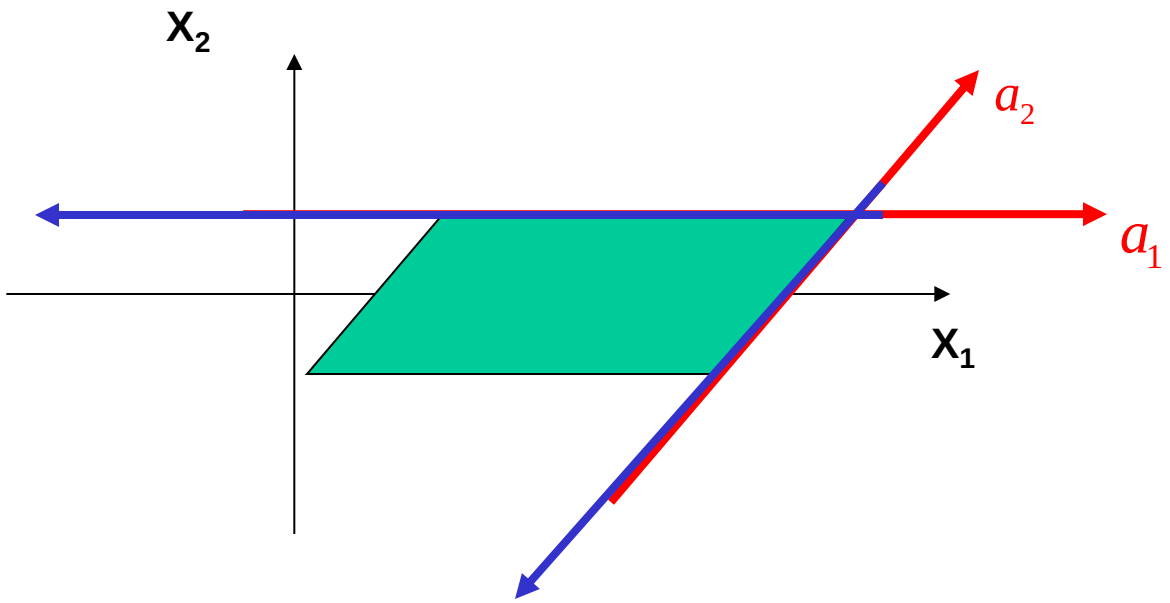
$$x = A \cdot s$$

$$A = \begin{bmatrix} a_1 & a_2 \end{bmatrix}$$



Once projected onto these vectors, the distribution is statistically independent

ICA: Ambiguity



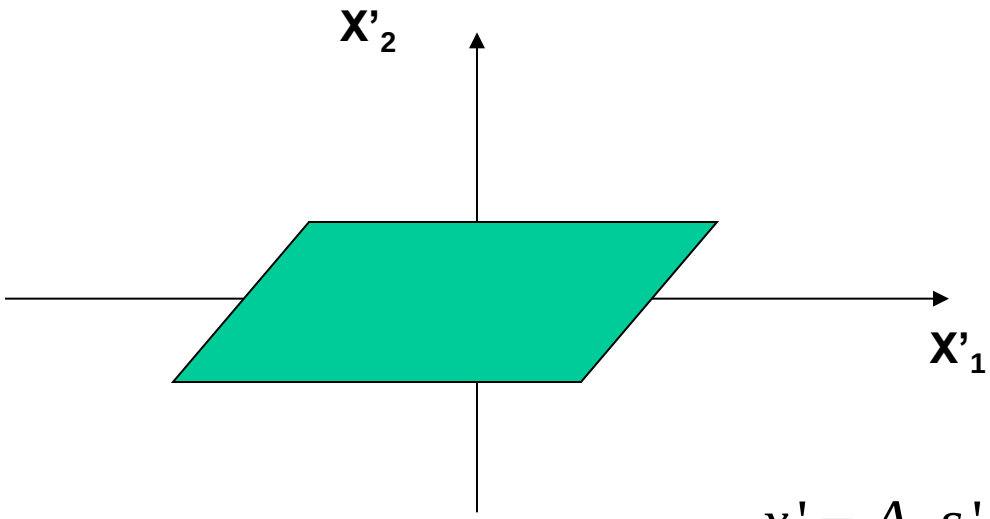
2nd ambiguity of ICA: cannot determine the sign of the vectors

ICA: Preprocessing - Centering

ICA first pre-processing step starts

by centering the distribution such that: $x' = x - E\{x\}$

$$\Rightarrow \quad E\{x'\} = 0$$



$$x' = A \cdot s'$$

$$s = s' + \mu_S, \quad \mu_S = A^{-1} \cdot E\{x\}$$

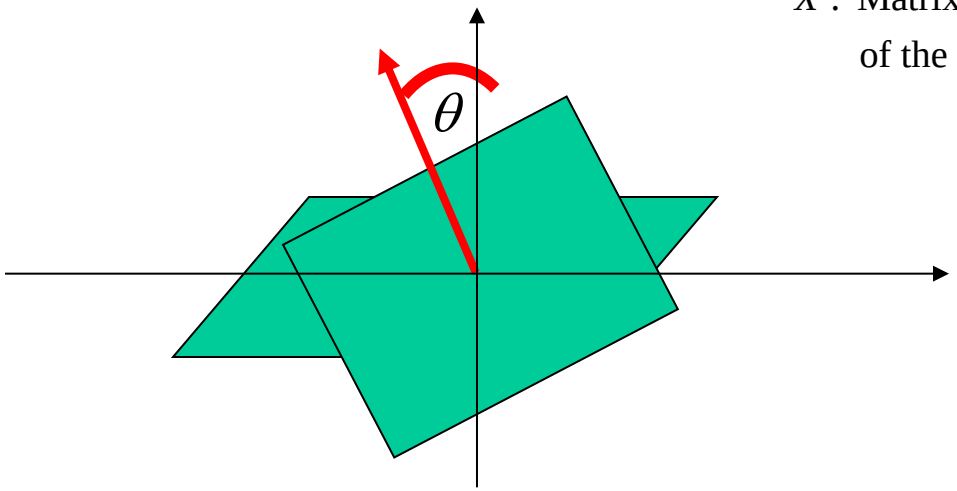
ICA: Preprocessing - Whitening

The joint distribution $p(x)$ of two variates x_1, x_2 is said to be white if the variates are uncorrelated and their variance is equal to unity.

Uncorrelated: $E\{x_1, x_2\} = E\{x_1\} E\{x_2\}$

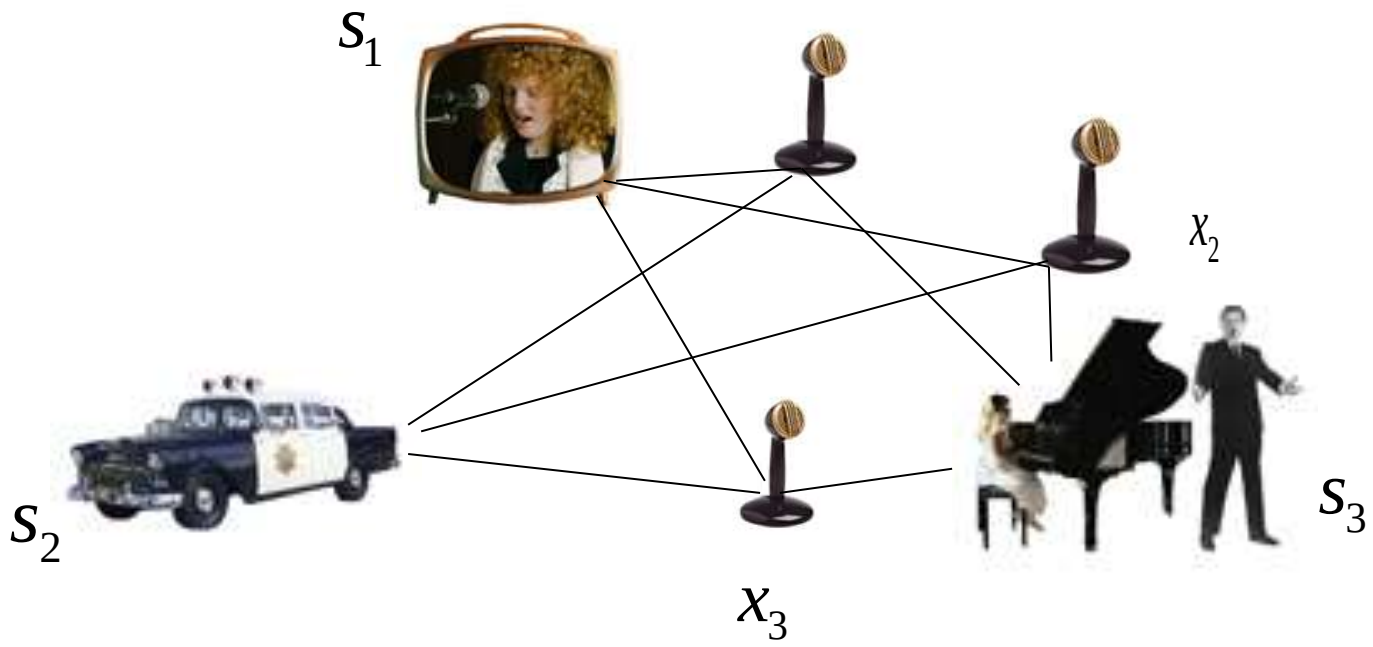
White: $E\{XX^T\} = I$

X : Matrix composed of all instances of the observations x



Advantage: Need only to compute the angle of rotation θ

N-dimensional observation vector $x \in \mathbb{R}^N$ and sources $s \in \mathbb{R}^N$, $N = 3$.



But at each time step of the recording, we obtain a new observation of both mixtures and sources.

If we do T measurements, the total dimension of the dataset is X: $N \times T$

ICA: Preprocessing - Whitening

Determine a matrix V, s.t.:

$X'' = VX'$

$\mathrm{E}\{(X''X''^{T})\} = I$

X : Matrix composed of all instances of the observations x

Use the eigenvalue decomposition

$E\{XX^T\} = EDE^T$

$D = diag\{\lambda_1, ..., \lambda_N\}, \quad E = \{e^1, ..., e^N\}$

$V = D^{-\frac{1}{2}}E^T$

$\mathrm{E}\{X''X''^{T}\} = V\mathrm{E}\{XX^T\}V^T = D^{-\frac{1}{2}}E^T\mathrm{E}\{XX^T\}ED^{-\frac{1}{2}}$

$\mathrm{E}\{X''X''^{T}\} = D^{-\frac{1}{2}}E^TEDE^TED^{-\frac{1}{2}} = I$

Whitening is essentially decorrelation followed by scaling.

Whitening is bringing us closer to unmix the observables and find the sources.
However, it only guarantees to find projections that decorrelate the data.
It does not guarantee to find projections such that the dataset becomes statistically independent.

ICA: Determining the sources

The distribution of s is non-Gaussian.

Find sources how joint distribution optimizes
a measure of non-gaussianity
Negentropy or Kurtosis

ICA: Optimizing – Fast ICA

Estimating the negentropy is difficult as it requires to estimate the distribution of the data

→ Minimizes Kurtosis instead

$$kurtosis(x) = E\left\{x^4\right\};\ kurtosis(s) = E\left\{s^4\right\} - 3\left(E\left\{s^2\right\}\right)^2$$

Iterative process: candidate source $y = s$?

Look for an optimum of $J(y) \propto \left[E\left\{\left(w^T x\right)^4\right\} - 3E\left(\left(w^T x\right)^2\right)\right]^2$

Since data is white and uncorrelated $E\left\{xx^T\right\} = I \Rightarrow E\left\{y^2\right\} = 1$

Search optimum for: $J(y) = E\left\{\left(w^T x\right)^4\right\} - 3E\left(\|w\|^4\right)$

Fast – ICA optimization

For any non-linear transformation G, we would have:

$$E\{G\{s_1 \cdot s_2\}\} = E\{G\{s_1\}\} \cdot E\{G\{s_2\}\}$$

Set to minimize: $J(y) = E\left\{G\left(w^T y\right)^4\right\} - 3E\left(\|w\|^4\right)$

It is important to choose well the G functions.

- Non-quadratic functions, that do not grow too fast.
- Asymmetric derivative

Good candidates are:

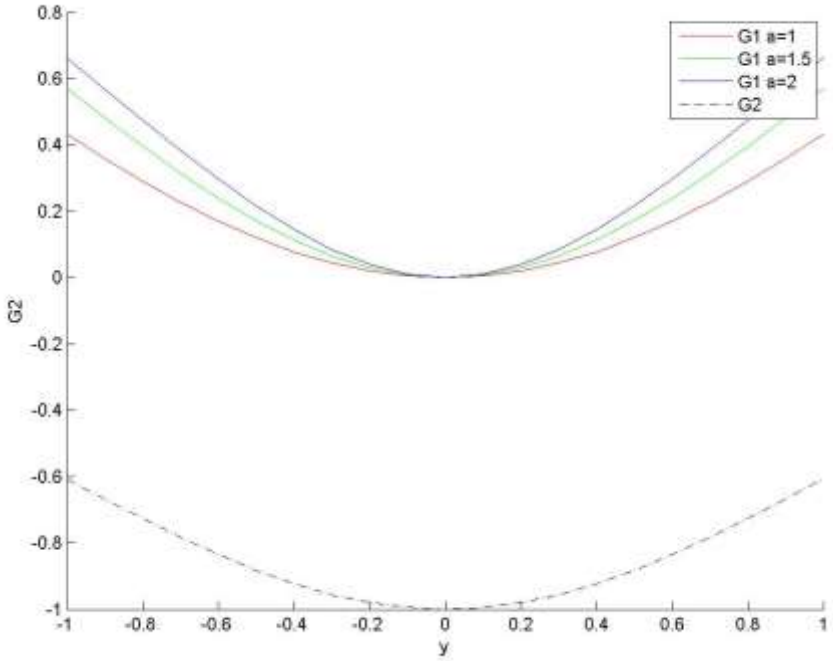
$$G_1(y) = \frac{1}{a}\log(\cosh(a \cdot y)), \quad 1 \le a \le 2$$

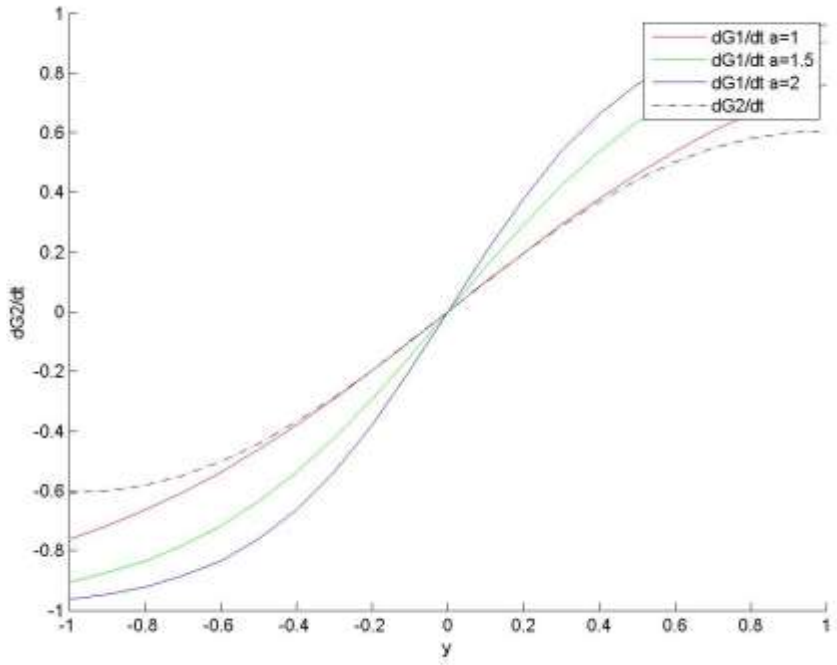
$$G_2(y) = -e^{-\frac{1}{2}y^2}$$

$$\frac{dG_1(y)}{dy} = g_1(y) = \tanh(a \cdot y) \quad 1 \le a \le 2$$

$$\frac{dG_2(y)}{dy} = g_2(y) = y \cdot e^{-\frac{1}{2}y^2}$$

Fast – ICA optimization





Good candidates are :

$$G_1(y) = \frac{1}{a}\log(\cosh(a \cdot y)), \qquad 1 \le a \le 2$$

$$G_2(y) = -e^{-\frac{1}{2}y^2}$$

$$\frac{dG_1(y)}{dy} = g_1(y) = \tanh(a \cdot y) \qquad 1 \le a \le 2$$

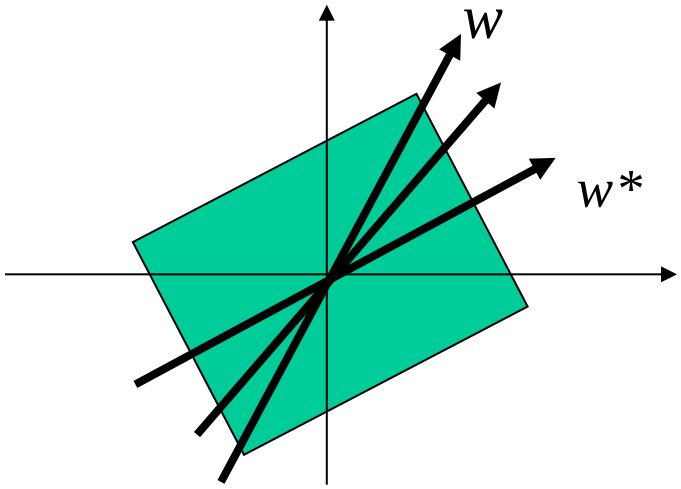
$$\frac{dG_2(y)}{dy} = g_2(y) = y \cdot e^{-\frac{1}{2}y^2}$$

Fast-ICA: Determining the sources iteratively

Idea: find each source one after the other one (iterative process)

Let w be a candidate projection, such that:

$$y = w^T x = w^T As = zs, \quad z = w^T A$$



y is more gaussian than s

(central limit theorem), unless $y = s$.

Search w, to seek optimum: $\frac{dJ\left(y = w^T x\right)}{dw} = 0$

=> iterate until one finds a w that maximizes the non-gaussianity of y.

Fast – ICA optimization

Let *x''* be the centered and whitened projection of the data

1. Choose an initial (e.g. random) weight vector *w*.

2. Compute the quantity $w^+ = E\left\{xG\left(w^T x\right)\right\} - E\left\{g\left(w^T x\right)\right\} w$ (derivative of J)

3. Proceed to a normalization of the weight vector:

$$w = \frac{w^+}{\left\|w^+\right\|}$$

$$E\left\{yy^T\right\} = 1 \Leftrightarrow w^T E\left\{xx^T\right\} w = w^T w = I$$

4. If the weight has not converged, that is if: $w^T\left(t-1\right) \cdot w\left(t\right) \neq 1$ go back to step 2.

ICA: iteration for several sources

To estimate several independent components, we must prevent the different vectors from converging to the same maximum.
→ *decorrelate* the vectors at each iteration.

1. Estimate the independent components one by one $w_1, \ldots, w_p$
2. Run the one-unit fixed-point algorithm for w_{p+1}
3. after every iteration step subtract from w_{p+1} the other``projections" $w_{p+1}^T w_j w_j, \quad j = 1, \ldots p$

1. Let $\mathbf{w}_{p+1} = \mathbf{w}_{p+1} - \sum_{j=1}^{p} \mathbf{w}_{p+1}^T \mathbf{w}_j \mathbf{w}_j$
2. Let $\mathbf{w}_{p+1} = \mathbf{w}_{p+1} / \sqrt{\mathbf{w}_{p+1}^T \mathbf{w}_{p+1}}$

Guaranteed to converge to one of the independent components
Hyvarinen & Oja (1997)

ICA Estimation and Limitations

ICA Estimation

Centering (mean=0)

Whitening (variance = 1)

Fast-ICA for one component

Fast-ICA for several components + decorrelation at each time step

Limitations of ICA

Reconstructed sources may be inverted (sign of vector unknown)

Source vectors cannot be ordered according to importance (unlike PCA)