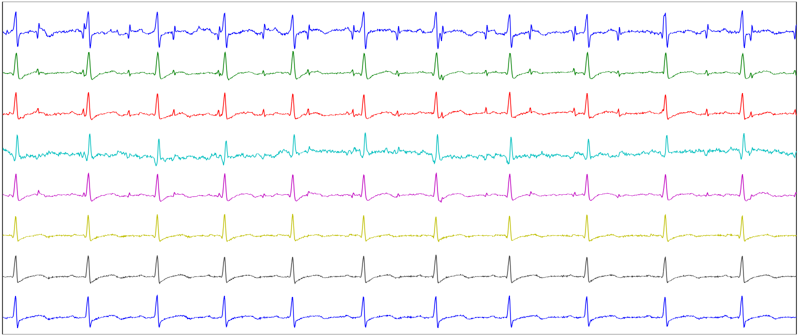
Principal Component Analysis (PCA) & Independent Component Analysis (ICA)

Alexandre Gramfort

Télécom ParisTech
alexandre.gramfort@telecom-paristech.fr

based on slides by Cédric Févotte and Jean-François Cardoso

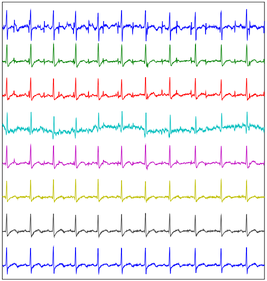
Motivating example: ECG data



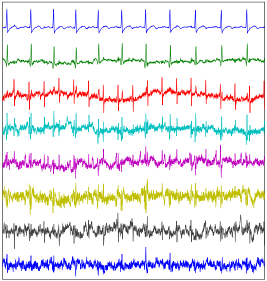
8 electrodes located on the thorax and the abdomen of a pregant woman.

Data Credits: Daisy database

Principal Component Analysis on ECG data

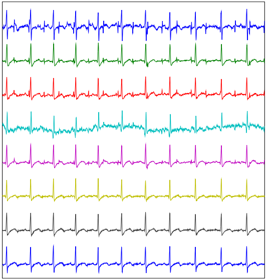


$$
= \begin{bmatrix}
95 & -51 & -33 & 8 & 2 & 0 & 2 & 2 \\
87 & 53 & 16 & 6 & -2 & -5 & 4 & 4 \\
94 & 17 & 14 & 0 & 11 & 6 & 1 & 1 \\
-9 & 105 & -28 & -5 & 1 & 0 & 0 & 0 \\
97 & 26 & 0 & 5 & 0 & -1 & -9 & -9 \\
89 & 7 & 1 & 0 & -11 & 8 & 0 & 0 \\
97 & -6 & 1 & -4 & -1 & -4 & 1 & 1 \\
92 & -32 & -3 & -15 & 0 & -1 & 0 & 0
\end{bmatrix} \times
$$

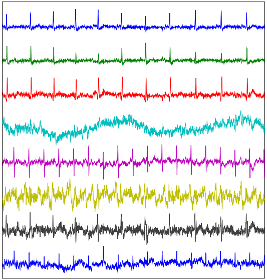


PCA works by assuming that the sources are orthogonal: $\frac{1}{N}\sum_n y_i(n)y_j(n) = 0$ for all $i \neq j$ (uncorrelated sources)

Independent Component Analysis on ECG data



$$= \begin{bmatrix} 98 & 60 & -71 & 5 & -34 & 17 & 3 & 3 \\ 93 & 114 & 37 & 0 & 10 & 5 & 6 & 6 \\ 101 & 99 & -4 & 0 & 13 & 6 & 6 & 6 \\ -17 & 47 & 79 & 33 & -19 & -13 & -1 & -1 \\ 106 & 102 & 7 & 0 & -10 & 4 & 10 & 10 \\ 110 & 71 & -8 & 0 & -1 & 1 & 0 & 0 \\ 107 & 82 & -22 & 0 & 0 & 7 & 17 & 17 \\ 112 & 42 & -43 & 0 & -1 & 11 & 20 & 20 \end{bmatrix} \times$$



ICA assumes that the sources are statistically independent

Only independence is at work but it must go beyond decorrelation

Objectives

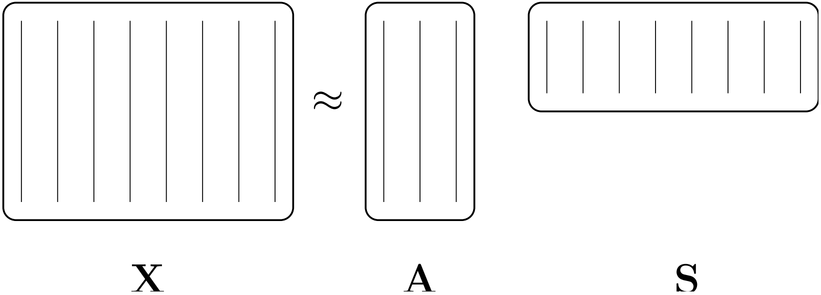
In this course we search for **unsupervised decompositions** of data such that

$\mathbf{x}_n$	$\approx$	$\mathbf{A}$	$\mathbf{s}_n$
data vector		"explanatory variables"	"regressors"
		"basis", "dictionary"	"expansion coefficients"
		"patterns"	"activation coefficients"

and $\mathbf{A}$ is learnt from a set of data vectors $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_N]$.

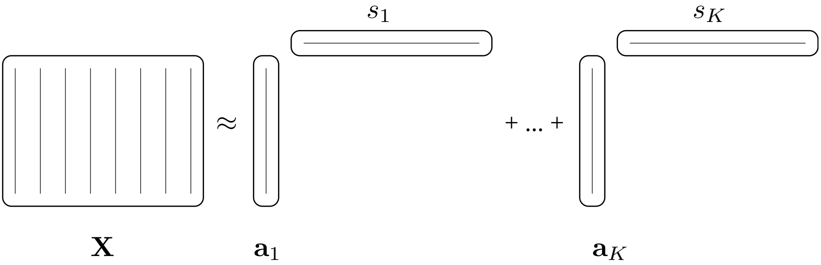
- $\mathbf{x}_n$ is a vector of size F
- $\mathbf{s}_n$ is a vector of size K
- $\mathbf{A}$ is a matrix of size $F \times K$, with usually $F \geq K$.

Example : dimensionality reduction



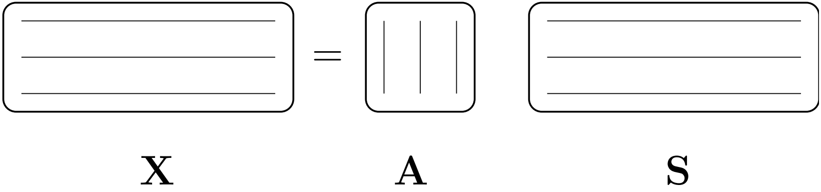
The $F \times N$ data matrix is approximated by $K(F+N)$ coefficients.

Example : dimensionality reduction (ctd)



The factorization is akin to a rank-K approximation.

Example : source separation



The rows of **X** are mixed signals, **A** is a mixing matrix and the sources form the rows of **S**.

Questions

In matrix form, we search for the following factorization

$$\mathbf{X} \approx \mathbf{A}\mathbf{S}$$

- What should the "$\approx$" entail ?
- What constraints should be imposed on $\mathbf{A}$ and/or $\mathbf{S}$?

1. Principal Component Analysis (PCA)
   - Concept
   - Sphering
2. Independent Component Analysis (ICA)
   - Concept
   - Nongaussian is independent
   - FastICA algorithms

Principal Component Analysis (PCA)

Concept

- The data is assumed real-valued ($\mathbf{x}_n \in \mathbb{R}^F$) and centered ($E\{\mathbf{x}_n\} = 0$)
- PCA returns a dictionary $\mathbf{A}_{PCA} \in \mathbb{R}^{F \times K}$ such that

$$\mathbf{x}_n \approx \hat{\mathbf{x}}_n = \sum_k < \mathbf{a}_k, \mathbf{x}_n > \mathbf{a}_k = \mathbf{A}(\mathbf{A}^T \mathbf{x}_n)$$

and such that the least squares error is minimized

$$\mathbf{A}_{PCA} \quad = \quad \min_{\mathbf{A}} \frac{1}{N} \sum_n \|\mathbf{x}_n - \hat{\mathbf{x}}_n\|_2^2 = \frac{1}{N} \|\mathbf{X} - \mathbf{A}\mathbf{A}^T\mathbf{X}\|_F^2$$

Solution

The solution can be shown to be of the form

$$\mathbf{A}_{PCA} = \mathbf{E}_{1:K}\mathbf{U}$$

where $\mathbf{E}_{1:K}$ denotes the K dominant eigenvectors of $\mathbf{C}_\mathbf{x} = \mathrm{E}\{\mathbf{x}\mathbf{x}^T\} \approx \frac{1}{N}\sum_n \mathbf{x}_n\mathbf{x}_n^T$ such that

$$\mathbf{C}_\mathbf{x}\mathbf{e}_k = d_k\mathbf{e}_k$$

and where $\mathbf{U}$ is any unitary matrix of size $K \times K$.

Compression

The residual least square error of the decomposition can be shown to be

$$\frac{1}{N}\sum_{n}\|\mathbf{x}_n - \hat{\mathbf{x}}_n\|_2^2 = \sum_{i=K+1}^{F} d_i$$

where $\{d_i\}_i$ are the eigenvalues of $\mathbf{C}_{\mathbf{x}}$, sorted in order of decreasing value.

PCA can be used for compression : the original $F \times N$ data matrix $\mathbf{X}$ can be approximated by $FK + KN$ coefficients of the matrices $\mathbf{A}_{PCA}$ and $\mathbf{S}_{PCA} = \mathbf{A}_{PCA}^T\mathbf{X}$.

Uncorrelatedness

When $\mathbf{U} = \mathbf{I}$, the expansion coefficients in the PCA model are uncorrelated; indeed, we have

$$\begin{aligned} \mathbf{C_s} &= \mathrm{E}\{(\mathbf{A}^T\mathbf{x})(\mathbf{A}^T\mathbf{x})^T\} \\ &= \mathbf{A}^T\mathbf{C_x}\mathbf{A} \\ &= \mathrm{diag}\left([d_1, \ldots, d_K]\right) \\ &\overset{\text{def}}{=} \mathbf{D}_K \end{aligned}$$

Sphering (aka whitening)

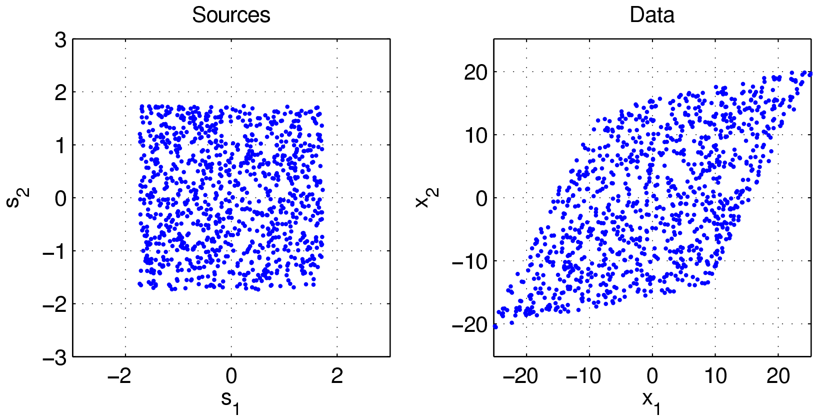
Besides decorrelation, the variance of the entries of $\mathbf{s}$ can be normalized to 1. This achieved for $\mathbf{S}_{SPH} = \mathbf{A}_{SPH}^T\mathbf{X}$ where

$$\mathbf{A}_{SPH} = \mathbf{E}_{1:K}\mathbf{D}_K^{-\frac{1}{2}}$$

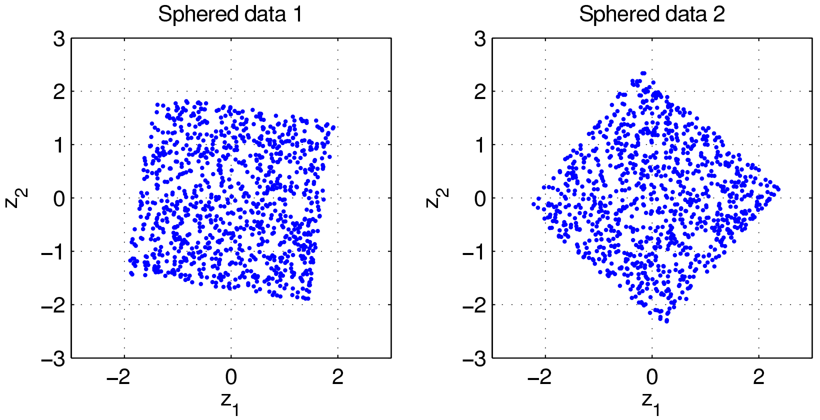
The sphering matrix $\mathbf{A}_{SPH}$ is not unique. Indeed, for any unitary matrix $\mathbf{U}$ of size $K \times K$, the matrix $(\mathbf{A}_{SPH}\mathbf{U})$ is also a sphering matrix, as we may write

$$\begin{aligned} \mathrm{E}\{(\mathbf{A}_{SPH}\mathbf{U})^T\mathbf{x}\mathbf{x}^T(\mathbf{A}_{SPH}\mathbf{U})\} &= \mathbf{U}^T\mathbf{A}_{SPH}^T\mathbf{C}_\mathbf{x}\mathbf{A}_{SPH}\mathbf{U} \\ &= \mathbf{U}^T\mathbf{U} \\ &= \mathbf{I} \end{aligned}$$

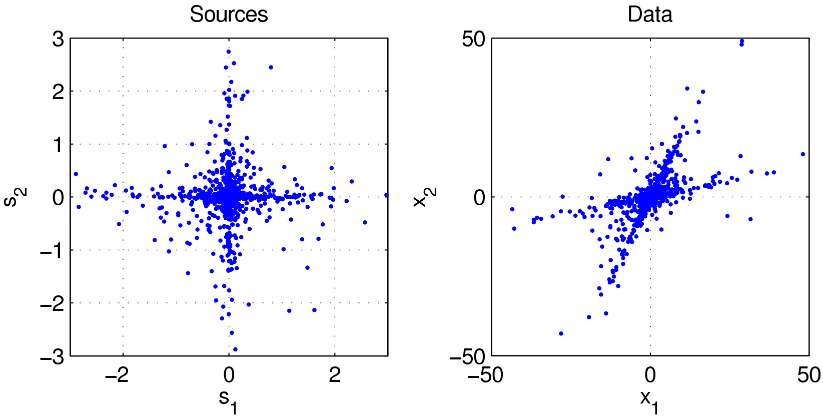
Example : uniform coefficients



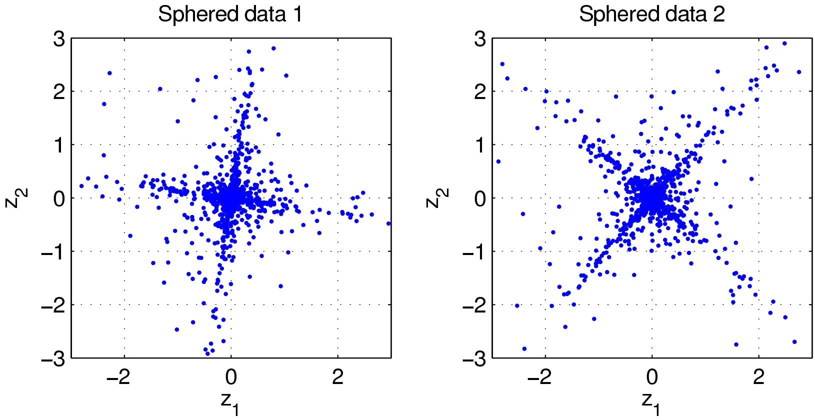
Example : uniform coefficients (ctd)



Example : sparse coefficients



Example : sparse coefficients (ctd)



Independent Component Analysis (ICA)

Concept

Sphering returns coefficients $\mathbf{z} = \mathbf{A}_{SPH}^T \mathbf{x}$ that are uncorrelated and with unit variance, i.e., $\mathrm{E}\{\mathbf{z}\mathbf{z}^T\} = \mathbf{I}$.

Sphering is not unique as any rotation $\mathbf{Uz}$ is also white. Hence, one may choose the arbitrary rotation $\mathbf{U}$ so that $\mathbf{Uz}$ satisfies an additional criterion.

ICA aims at finding $\mathbf{U}_{ICA}$ so that the components of

$$\mathbf{s}_{ICA} = \mathbf{U}_{ICA}\mathbf{z} = \mathbf{U}_{ICA}\mathbf{A}_{SPH}^T \mathbf{x}$$

are sphered and **mutually independent**.

Concept (ctd)

Assume for simplicity that $F = K$. In other words, ICA decomposes the data as

$$\mathbf{x} = \mathbf{A}_{ICA}\,\mathbf{s}$$

such that the entries of $\mathbf{s}$ are mutually independent :

$$p(\mathbf{s}) = \prod_k p(s_k)$$

Given what precedes, ICA can be achieved in two steps :

1) Sphere the observations as $\mathbf{z} = \mathbf{A}_{SPH}^T\,\mathbf{x}$,
2) Find $\mathbf{U}_{ICA}$ such that the entries of $\mathbf{U}_{ICA}\,\mathbf{z}$ are mutually independent.

Concept (ctd)

Hence, in practice, given sphered data $\mathbf{z}_n = \mathbf{A}_{SPH}^T \mathbf{x}_n$ we need to

1) Construct a numerical criterion $C(\mathbf{Y})$ measuring the independence of the entries of the random vector $\mathbf{y}$ given realizations $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]$.
2) Solve the following optimization problem

$$\max_{\mathbf{U}} C(\mathbf{UZ})$$

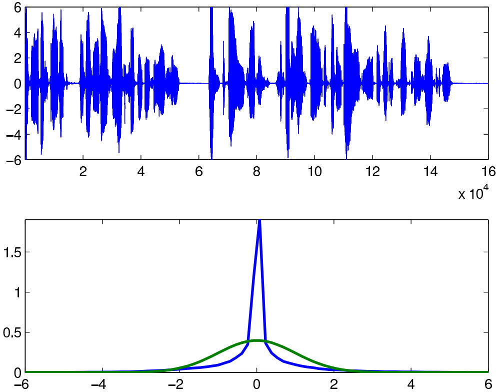
Nongaussian is independent

The Central Limit Theorem tells that the distribution of the sum of independent random variables tends towards a Gaussian distribution (under certain conditions).

Basically, it implies that the sum of two random variables is “more Gaussian” than the original random variables.

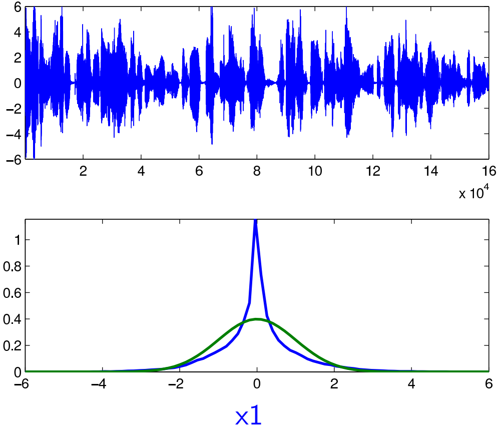
This suggests that the entries of $\mathbf{y} = \mathbf{U}\mathbf{z}$ should be searched as *nongaussian* as possible.

Nongaussian is independent (ctd)

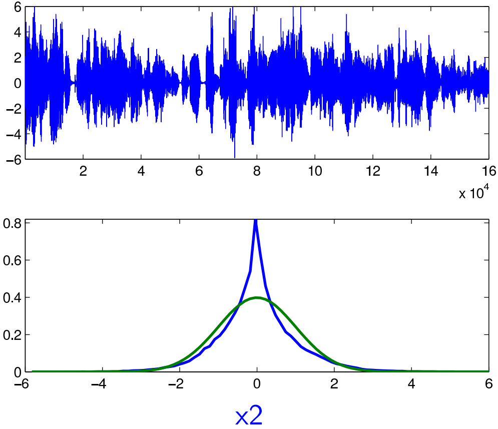


s1

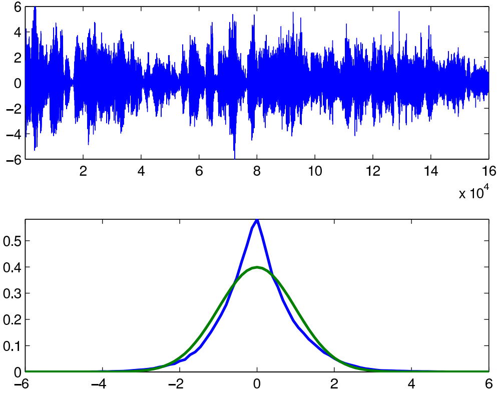
Nongaussian is independent (ctd)



Nongaussian is independent (ctd)

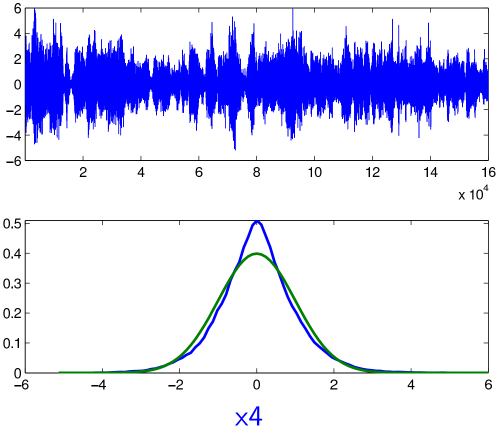


Nongaussian is independent (ctd)

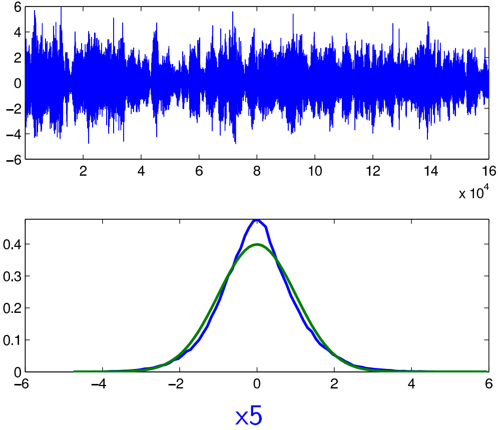


x3

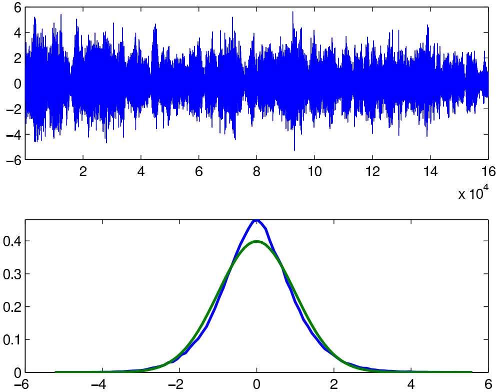
Nongaussian is independent (ctd)



Nongaussian is independent (ctd)

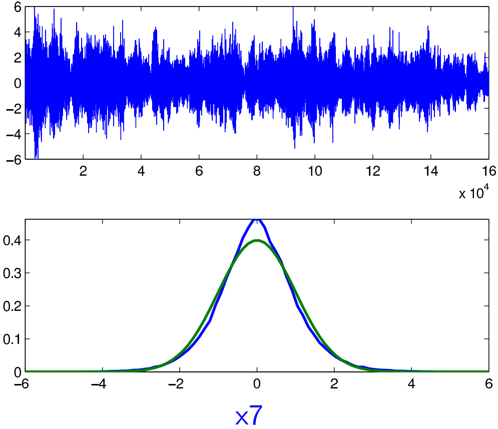


Nongaussian is independent (ctd)



x6

Nongaussian is independent (ctd)



Nongaussian is independent (ctd)

This intuition can be made rigorous via the properties of *mutual information*, defined as

$$\begin{aligned} I\{\mathbf{y}\} &= KL[p(\mathbf{y})|\prod_i p(y_i)] \\ &= \int p(\mathbf{y}) \log \frac{p(\mathbf{y})}{\prod_i p(y_i)} d\mathbf{y} \\ &= \sum_f H\{y_f\} - H\{\mathbf{y}\} \end{aligned}$$

where $H\{\mathbf{y}\} = -\int p(\mathbf{y}) \log p(\mathbf{y}) d\mathbf{y}$ denotes *differential entropy*.

Nongaussian is independent (ctd)

The mutual information of $\mathbf{y} = \mathbf{U}\mathbf{z}$ can be written

$$\begin{aligned} I\{\mathbf{y}\} &= \sum_f H\{y_f\} - H\{\mathbf{z}\} - \log|\det \mathbf{U}| \\ &= \sum_f H\{y_f\} + cst \end{aligned}$$

Hence, because the Gaussian is the distribution with highest entropy (for given variance), minimizing $I\{\mathbf{y}\}$ (i.e., enforcing mutual independence) is equivalent to minimizing $\sum_f H\{y_f\}$ (i.e., enforcing nongaussianity).

Identifiability of ICA

This discussion implies that ICA cannot separate Gaussian sources.

This is because the sum of Gaussian random variables is itself Gaussian.

The ICA model $\mathbf{X} = \mathbf{AS}$ (with $F = K$) is identifiable (up to scale and order ambiguities) when at most one source is Gaussian. *[Comon 94]*

Measures of nongaussianity

A quantitative measure of nongaussianity is the *kurtosis*, defined by

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

The Gaussian distribution has zero kurtosis. Distributions "flatter" than the Gaussian are called *subgaussian* and have kurtosis < 0. Distributions "peakier" than the Gaussian are called *supergaussian* and have kurtosis > 0.

Another common measure of nongaussianity is the *negentropy*, defined by

$$J\{y\} = H\{y_G\} - H\{y\}$$

where y_G denotes a Gaussian variable with same variance than y.

FastICA algorithms

Using the kurtosis as a quantitative measure of nongaussianity, we are left with the following optimization problem

$$\max_{\mathbf{U}} \sum_k |\mathrm{kurt}\{[\mathbf{U}^T \mathbf{z}]_k\}| \quad \text{subject to} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}$$

For simplicity, let's first consider the problem of finding only one maximally nongaussian component, i.e, solve

$$\max_{\mathbf{u}} \; C(\mathbf{u}) = |\mathrm{kurt}\{\mathbf{u}^T \mathbf{z}\}| \quad \text{subject to} \quad \mathbf{u}^T\mathbf{u} = 1$$

FastICA algorithms

For sphered, centered data $\mathbf{z}$, the criterion writes

$$C(\mathbf{u}) = |\mathrm{E}\{(\mathbf{u}^T\mathbf{z})^4\} - 3|$$

Its gradient thus writes...

FastICA algorithms

For sphered, centered data **z**, the criterion writes

$$C(\mathbf{u}) = |\mathrm{E}\{(\mathbf{u}^T\mathbf{z})^4\} - 3|$$

Its gradient thus writes...

$$\nabla_{\mathbf{u}} C(\mathbf{u}) = 4\,\mathrm{sign}(\mathrm{E}\{(\mathbf{u}^T\mathbf{z})^4\} - 3)\,\mathrm{E}\{(\mathbf{u}^T\mathbf{z})^3\mathbf{z}\}$$

FastICA algorithms (ctd)

A suitable projected gradient ascent algorithm writes

```
Initialize $\mathbf{u}^{(0)}$
for $i = 1 : n_{iter}$ do
    $\mathbf{u}^{(i)} \leftarrow \mathbf{u}^{(i-1)} + \alpha^{(i)} \nabla_{\mathbf{u}} C(\mathbf{u}^{(i)})$
    $\mathbf{u}^{(i)} \leftarrow \frac{\mathbf{u}^{(i)}}{\|\mathbf{u}^{(i)}\|}$
end for
```

where $\alpha^{(i)}$ is a (decreasing) sequence of positive step sizes.

FastICA algorithms (ctd)

A faster algorithm, free of tuning parameters, may be obtained by observing that a stationary point of the criterion must point in the direction of the gradient.
Indeed the Lagrangian to the original problem

$$\max_{\mathbf{u}} \; C(\mathbf{u}) = |\mathrm{kurt}\{\mathbf{u}^T \mathbf{z}\}| \quad \text{subject to} \quad \mathbf{u}^T \mathbf{u} = 1$$

writes

$$L(\mathbf{u}, \lambda) = C(\mathbf{u}) + \lambda(1 - \|\mathbf{u}\|^2)$$

so that a stationary point $\mathbf{u}^\star$ must satisfy $\nabla_{\mathbf{u}} C(\mathbf{u}^\star) = 2\lambda \mathbf{u}^\star$.

FastICA algorithms (ctd)

Hence, a fast fixed point algorithm can be obtained as

```
Initialize $\mathbf{u}^{(0)}$
for $i = 1 : n_{iter}$ do
    $\mathbf{u}^{(i)} \leftarrow \frac{\nabla_{\mathbf{u}} C(\mathbf{u}^{(i)})}{\|\nabla_{\mathbf{u}} C(\mathbf{u}^{(i)})\|}$
end for
```

Though based on a heuristic, the convergence of this algorithm to a stationary point of the original constrained problem can be shown.

FastICA algorithms (ctd)

In practice, the expectation appearing in the gradient is replaced by sample averages, i.e,

$$\mathrm{E}\{(\mathbf{u}^T\mathbf{z})^3\mathbf{z}\} \approx \frac{1}{N}\sum_n (\mathbf{u}^T\mathbf{z}_n)^3\mathbf{z}_n$$

The estimation may however be quite sensitive to outliers so that other algorithms based on robust approximations of the negentropy should be used.

However, the optimization concepts hold, and lead to the family of FastICA algorithms.

Fast ICA algorithms (ctd)

The “one-unit” optimization can be generalized to optimization of the whole matrix **U** through orthogonalization.

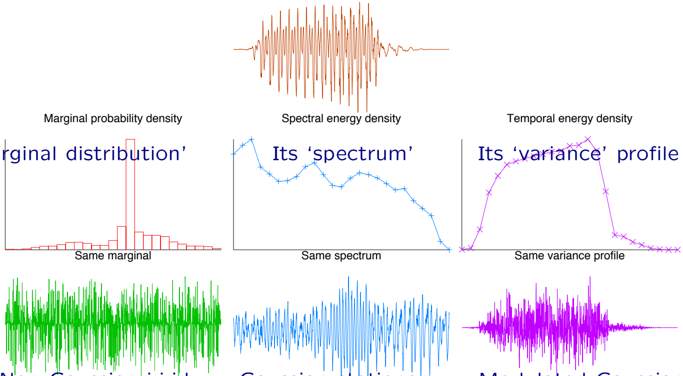
```
Initialize $\mathbf{u}_1^{(0)}, \ldots, \mathbf{u}_K^{(0)}$ (randomly)
for $i = 1 : n_{iter}$ do
    Do one iteration of a one-unit algorithm on every $\mathbf{u}_k$ in parallel
    Orthogonalize the set of vectors $\mathbf{u}_1^{(i)}, \ldots, \mathbf{u}_K^{(i)}$
        $\mathbf{U} \leftarrow (\mathbf{U}\mathbf{U}^T)^{-\frac{1}{2}}\mathbf{U}$
end for
```

Demos

See Jupyter / IPython Notebooks

Beyond Gaussian whiteness

A random (?!) sequence



Its 'marginal distribution'

Its ‘spectrum’

Its ‘variance’ profile

Non Gaussian i.i.id

Gaussian stationary

Modulated Gaussian i.i.d.

Beyond Gaussian whiteness (cnt)

Simple statistical models for the i-th source

- 0) Gaussian i.i.d. with variance σ_i^2.
- 1) Non Gaussian i.i.d. with marginal distribution p_i.
- 2) Gaussian stationary with power spectrum $C_i(f)$.
- 3) Gaussian independent with variance profile $\sigma_i^2(n)$.

Model 0) has a single parameter and is not strong enough for blind separation: its likelihood leads to measuring independence by global decorrelation.
Models 1), 2) and 3) are parameterized by a one-dimensional density: a probability density, a spectral density, a temporal density. They are strong enough for blind separation.

References

- A. Hyvärinen, J. Karhunen and E. Oja. *Independent Component Analysis*. Wiley, 2001.
- A. Hyvärinen and E. Oja. *Independent Component Analysis : Algorithms and Applications*. Neural Networks, 2000. [online]
- The FastICA package for MATLAB. `http://www.cis.hut.fi/projects/ica/fastica/`
- J.-F. Cardoso. *Blind Signal Separation: Statistical Principles*. Proceeding of the IEEE, 1998. [online]
- P. Comon and C. Jutten, editors. Handbook of Blind Source Separation, Independent Component Analysis and Applications. Academic Press, Oxford UK, Burlington USA, 2010.