Méthode de séparation de sources
Modèles et algorithmes

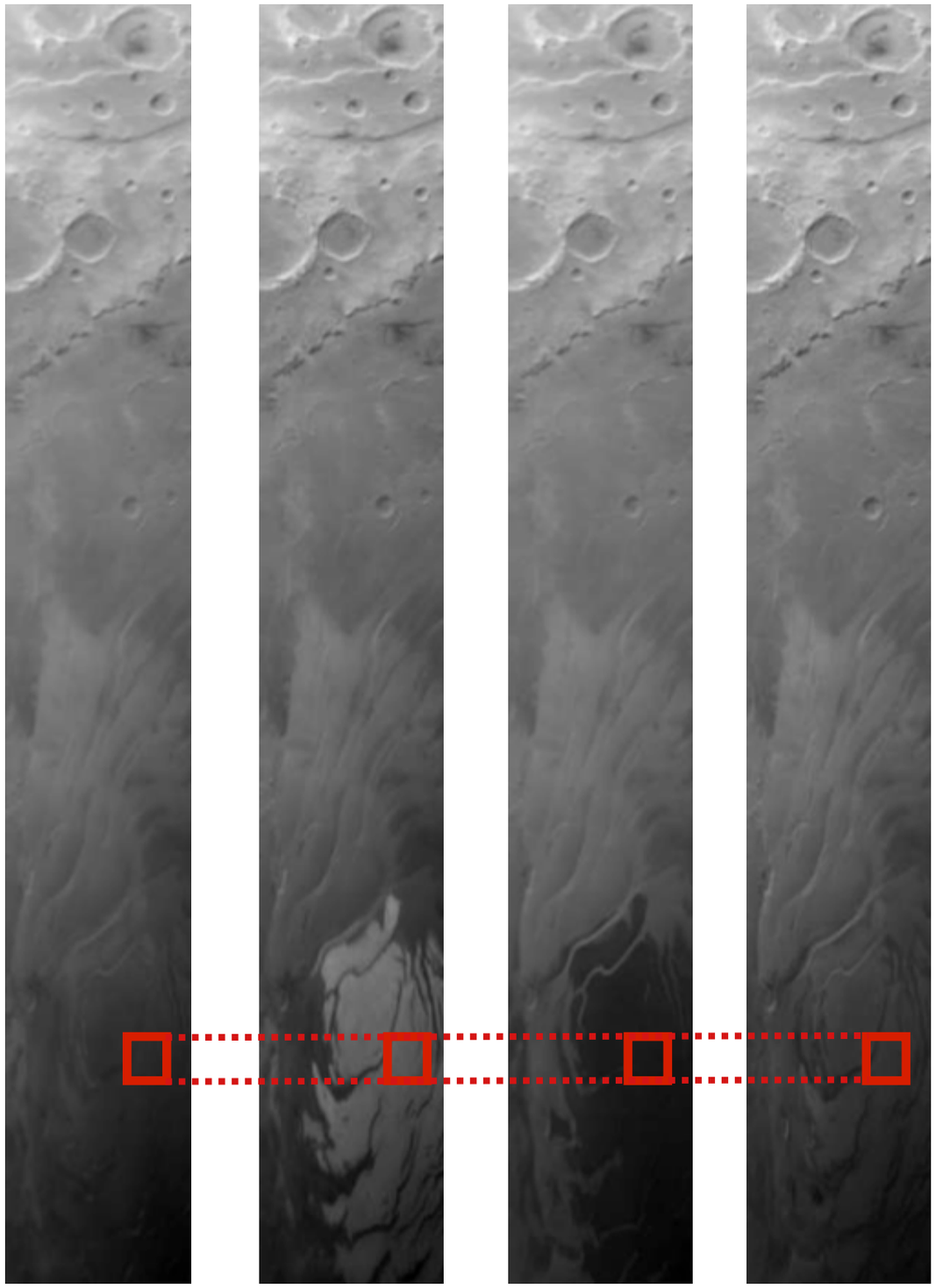
Applications en Astrophysique

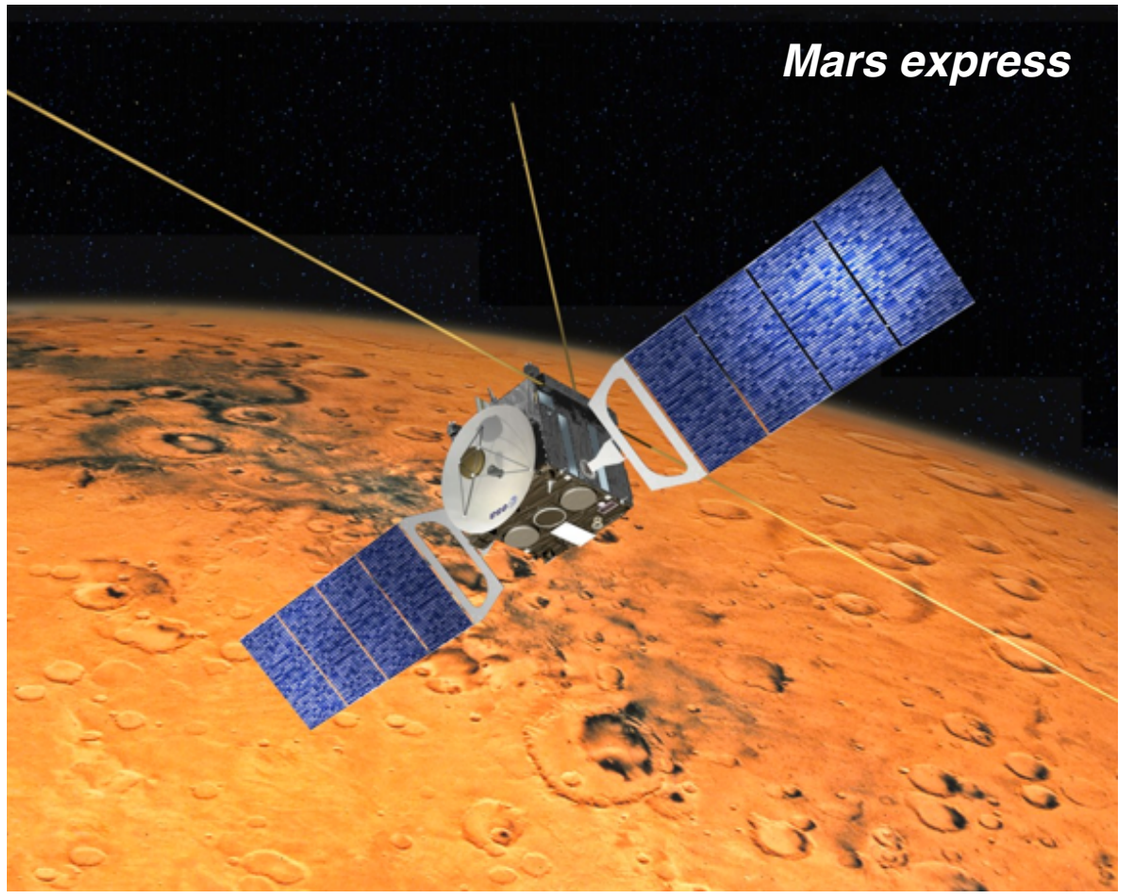
Statistical approaches - take I

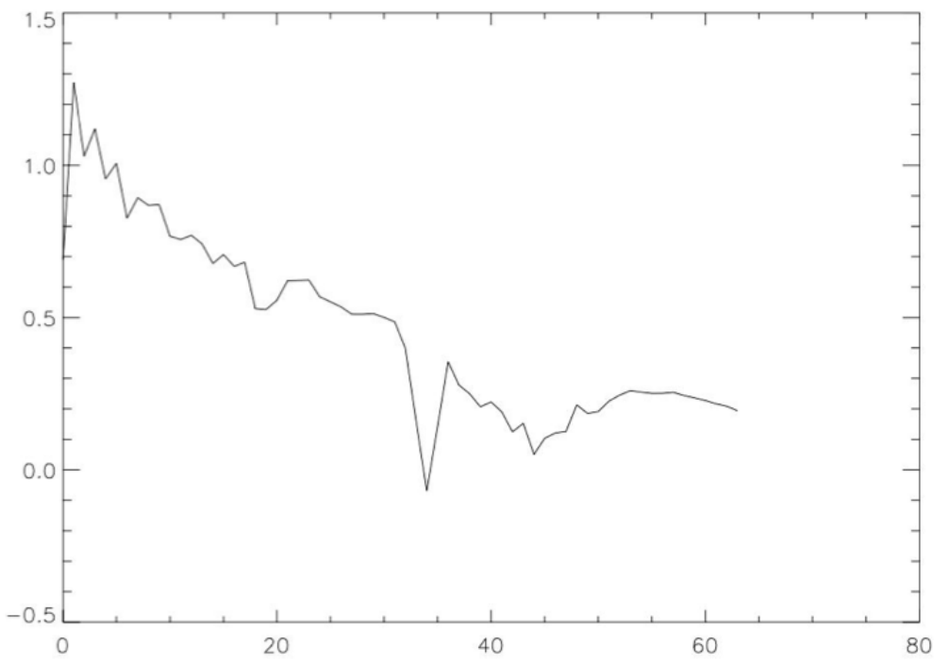
J.Bobin/C. Kervazo

jerome.bobin@cea.fr - christophe.kervazo@telecom-paris.fr

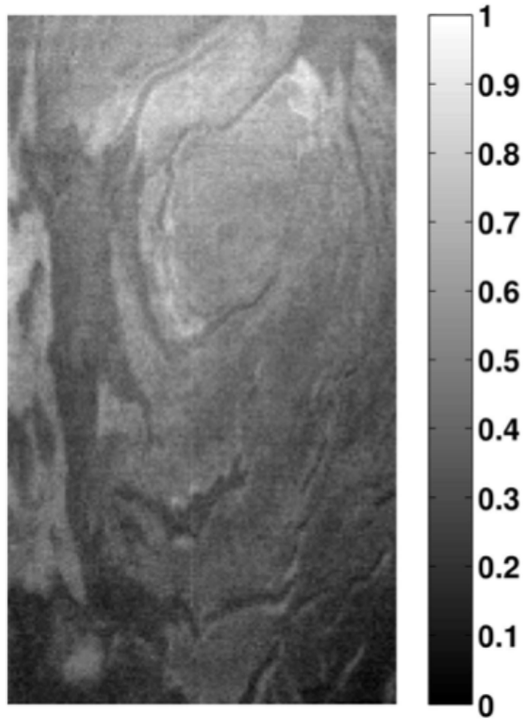
Context

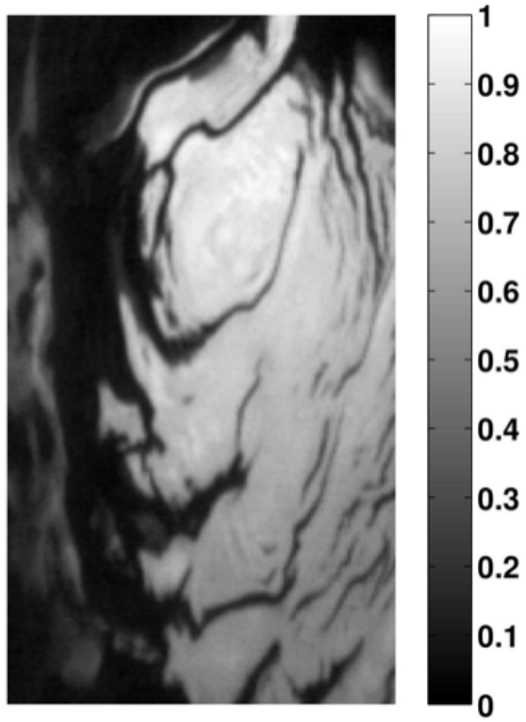


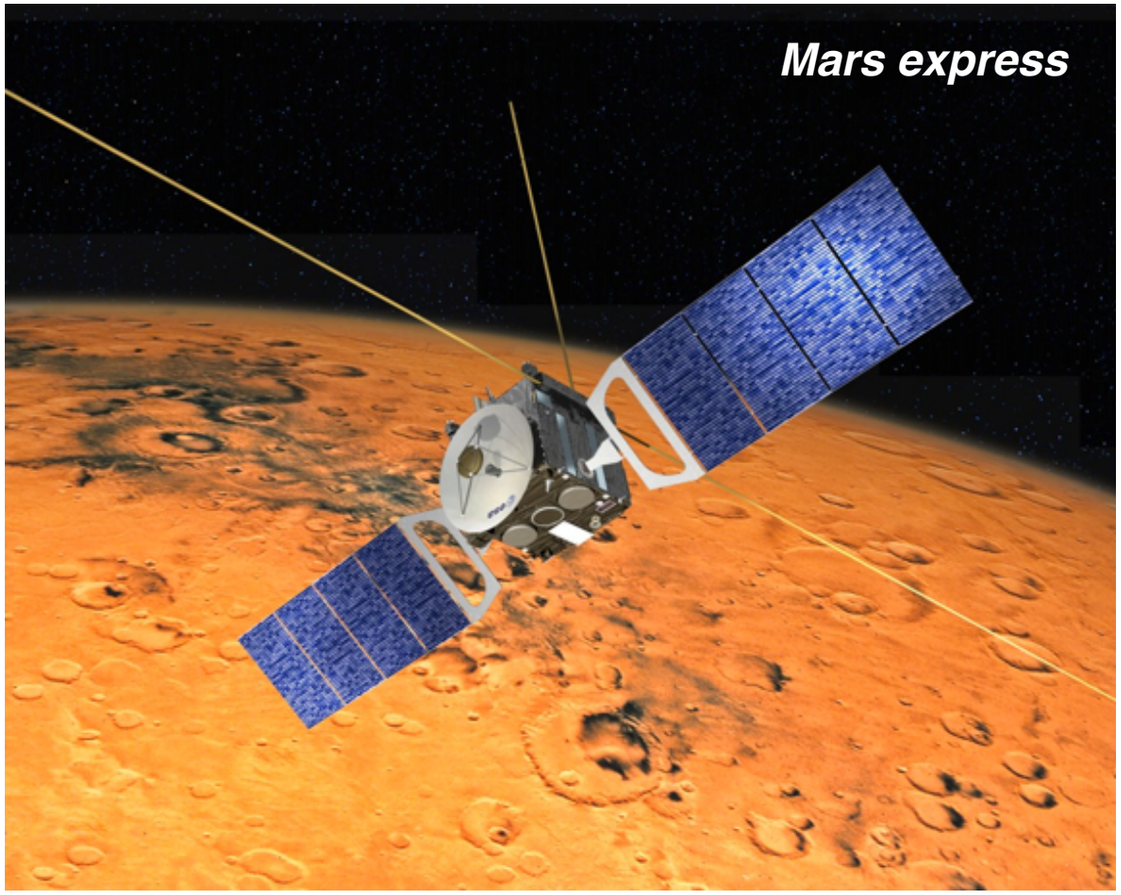


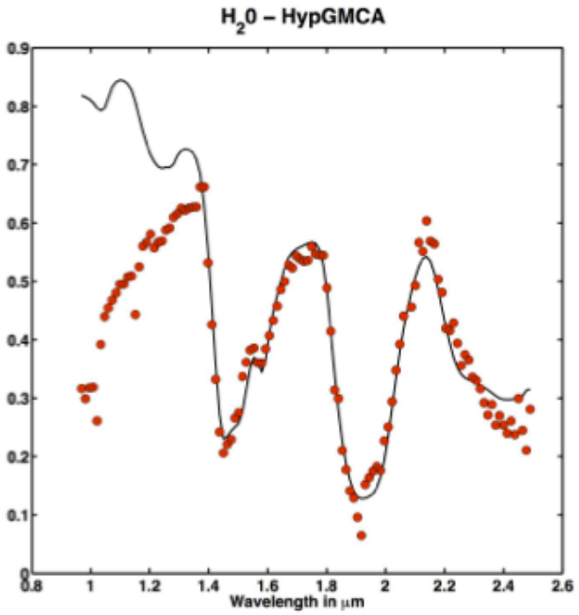


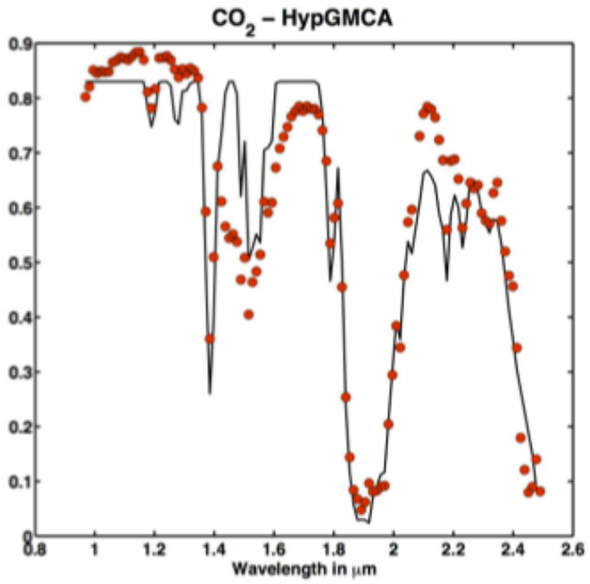
Context







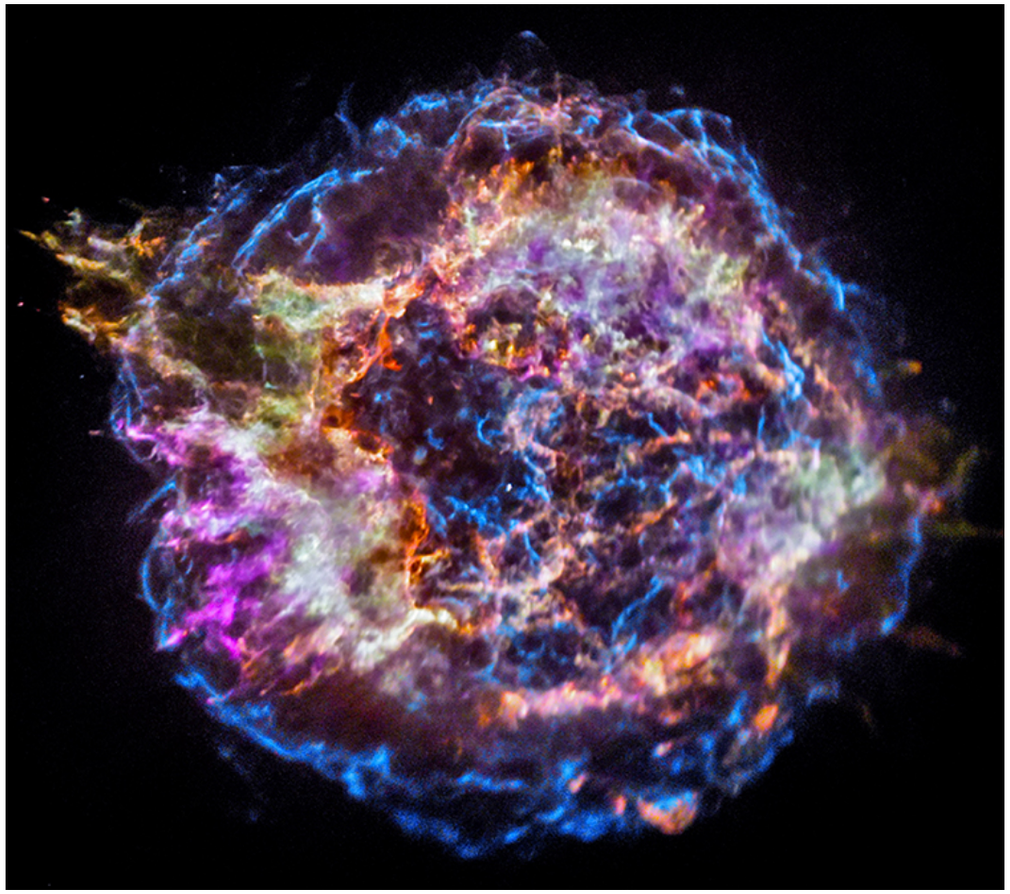


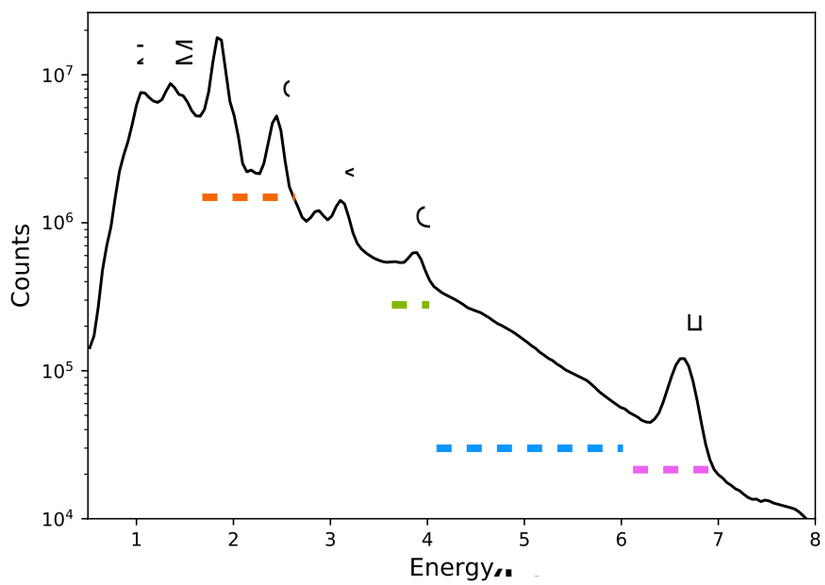


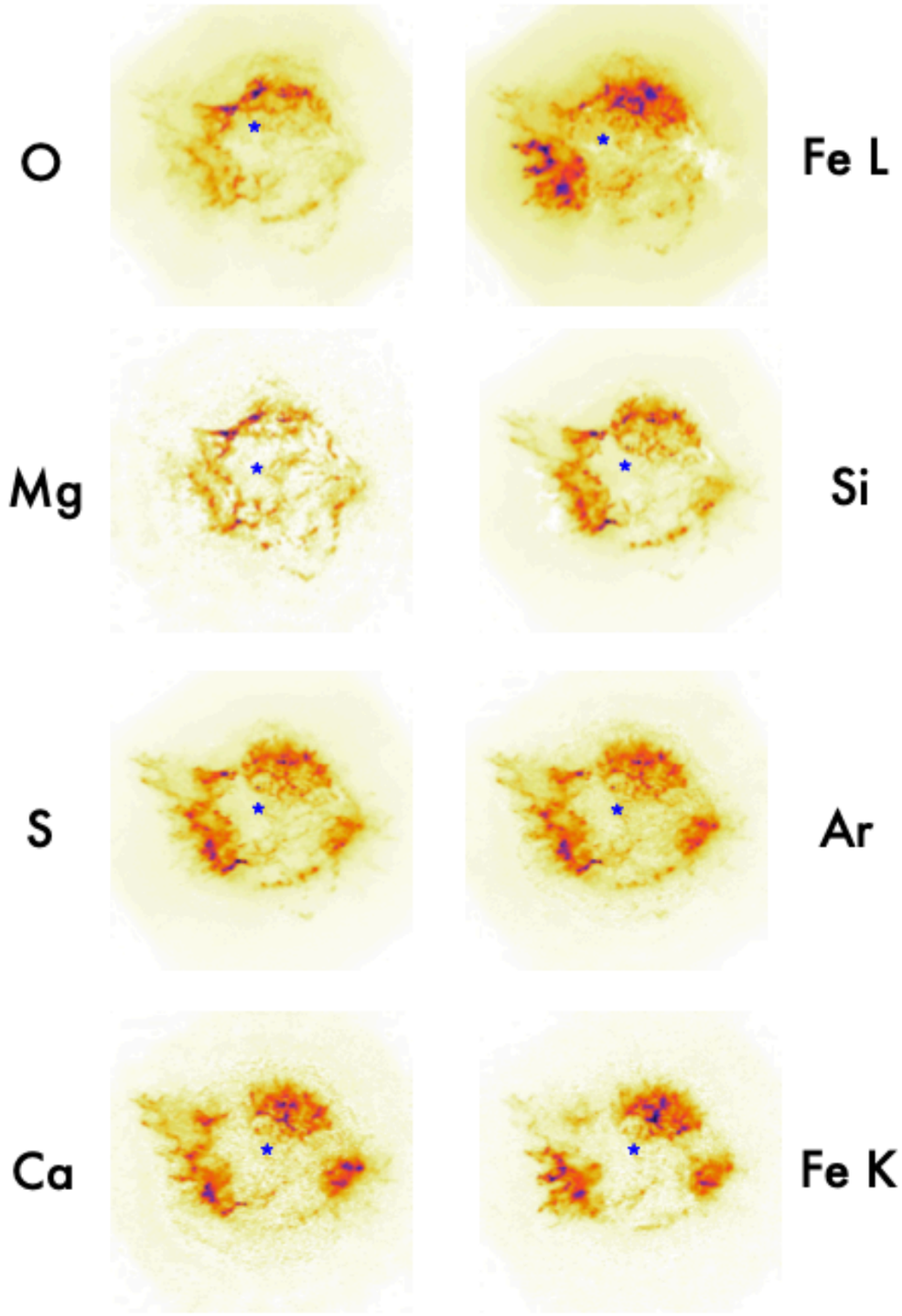
Objective:

Disentangling between the various components (water, carbon dioxyde, dust, minerals, etc ...)

Context





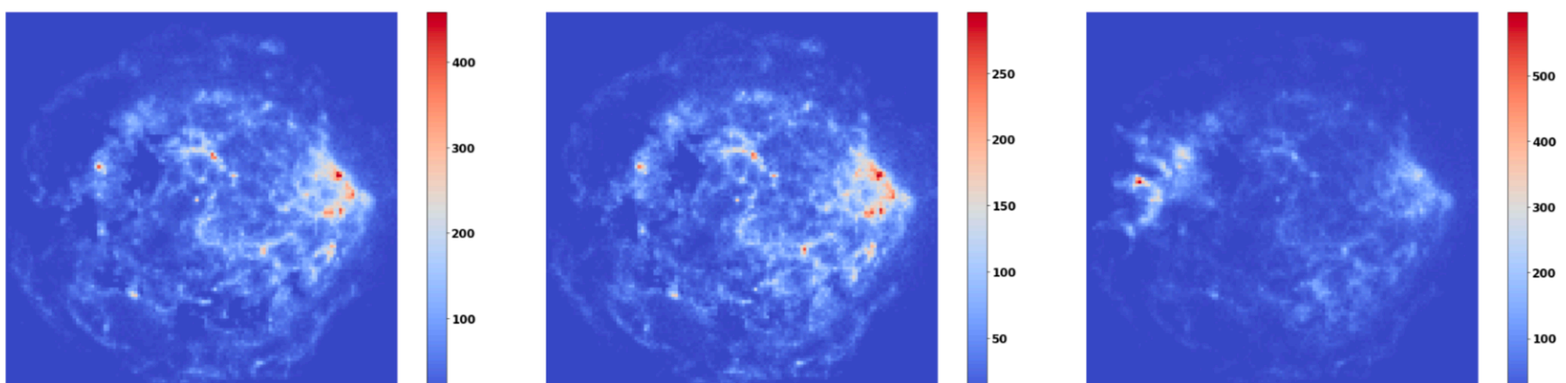


Multispectral X-ray observations with the Chandra telescope

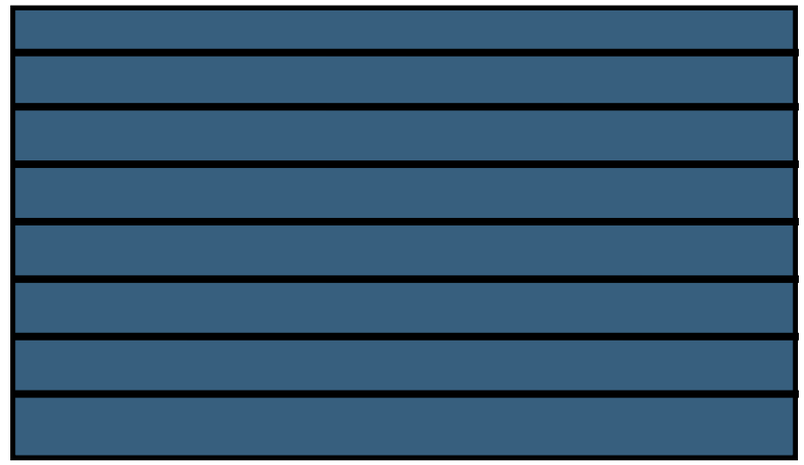
Context

Different scientific fields but …

common problems: mixtures of elementary signals or sources



(a) 5279 eV (b) 5548 eV (c) 5817 eV



Matrix formulation:

Whatever the format of the data,
a single observation is put in a matrix row

$\mathbf{X}$

[Not so] simple mixture model

More formally :

$$\mathbf{X} = \mathbf{AS}$$

$$\forall i = 1,\cdots,m; k = 1,\cdots,t; \quad x_i[k] = \sum_{j=1}^{n} a_{ij} s_j[k]$$

Hypotheses:

- At least as many observations as sources : $m \geq n$

- A is invertible / pseudo-invertible : $\left|\det(A^T A)\right| > 0$

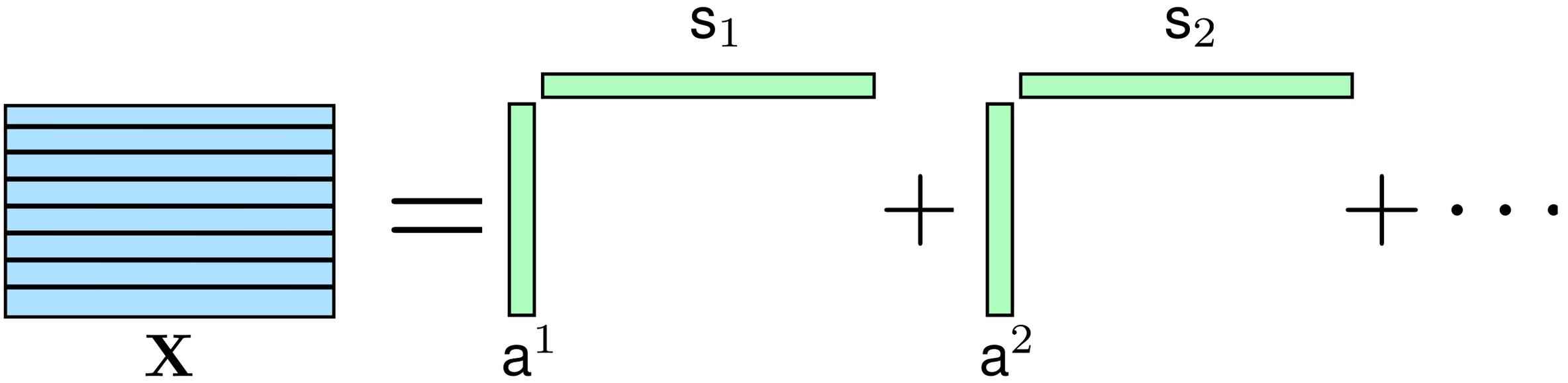
[Not so] simple mixture model

Le modèle de mélange

$$\mathbf{X} = \mathbf{AS}$$

A garder en mémoire :

Chaque ligne de X correspond à une observation
(signal 1D, image vectorisée ... etc.)



[Not so] simple mixture model

Let's take

$$\mathbf{X} = \mathbf{AS}$$

If A and S are fully unknown, are they identifiable ?

What are the limits of their identifiability ?

[Not so] simple mixture model

It's an ill-posed problem :

There exists an infinite number of solutions $\{A', S'\}$ such that :

$$X = AS = A'S'$$

Specifically for any invertible matrix U :

$$A' = AU \text{ et } S' = U^{-1}S'$$

What we can't avoid :

$$U = DP$$

D : diagonal matrix
P : permutation matrix

The sources will always be estimated up to a scale/ permutation factor

What can we do with an ill-posed problem ?

To sum this up :

1 - Blind source separation is an ill-posed problem

2 - spectral diversity (different mixtures) is not enough

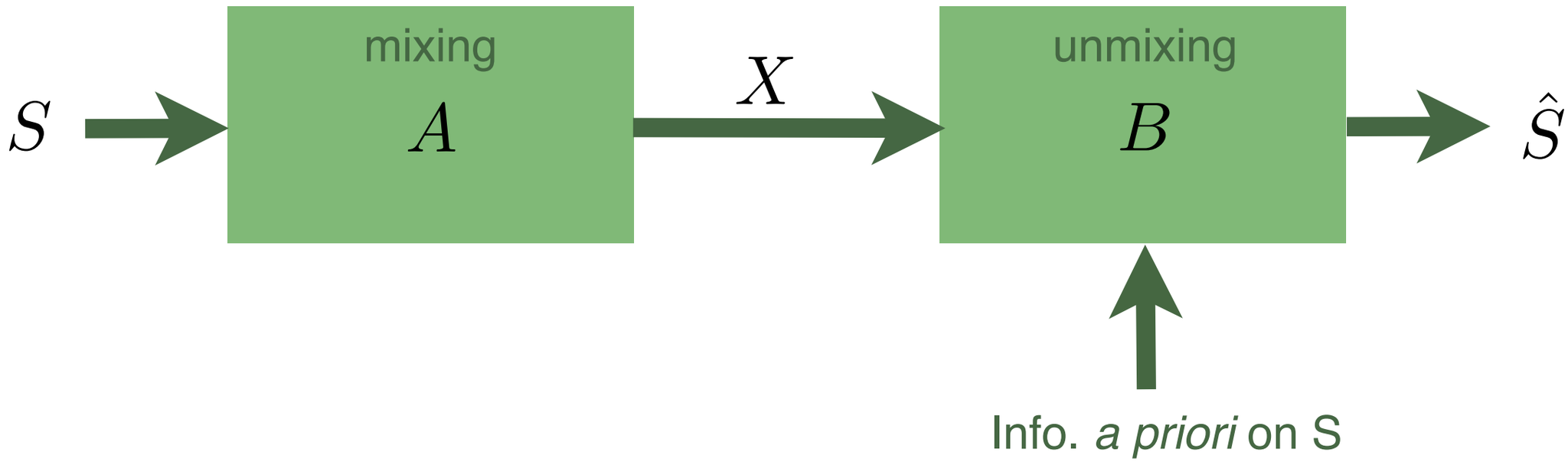
We need to better constrain the sources :

- **Exploiting inter-sources information :**
  decorrelation, independance, morphological diversity ... etc

- **Exploiting intra-sources information :**
  non-stationarity, positivity, bounded support, discrete-valued, sparsity, colour ... etc

What can we do with an ill-posed problem ?

Objective :



Statistical characterization or how can we say that the sources are different ?

- Sources are i.i.d. : exploiting joint statistics (decorrelation, higher order statistics, independance ...etc)

- Sources are not i.i.d. : exploiting their non-stationarity

A first (naive ?) attempt

What we measure**:** $$X = AS$$

Objective - seek the sources **S**, and the mixing matrix **A** so that :

$$R_S \quad \textbf{is diagonal}$$

Principal component analysis :

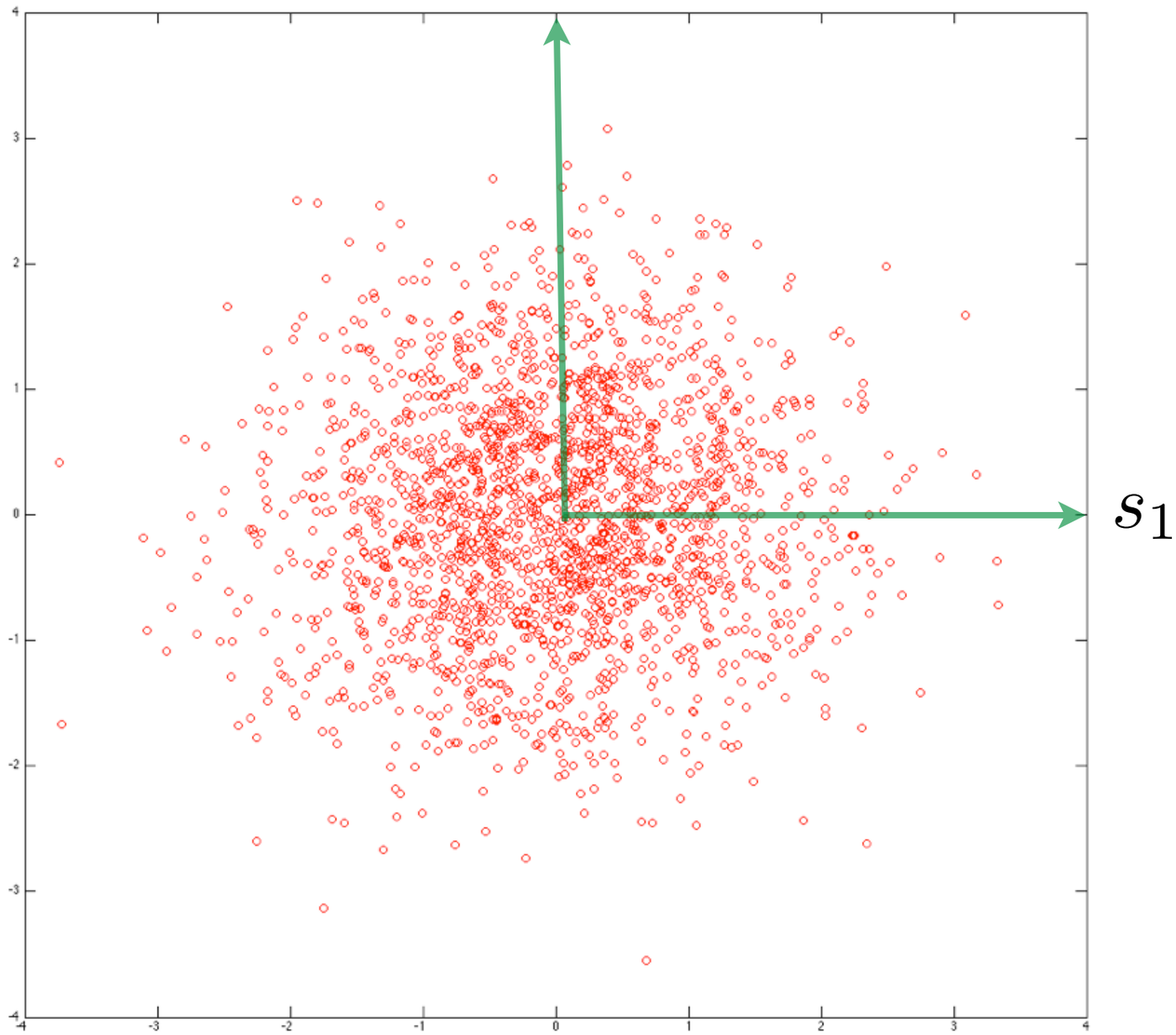
$$R_X = AR_SA^T$$

Assuming **A** is orthogonal, this boils down to a simple eigen values/vectors decomposition of the data covariance matrix R_X

$$S = A^T X$$

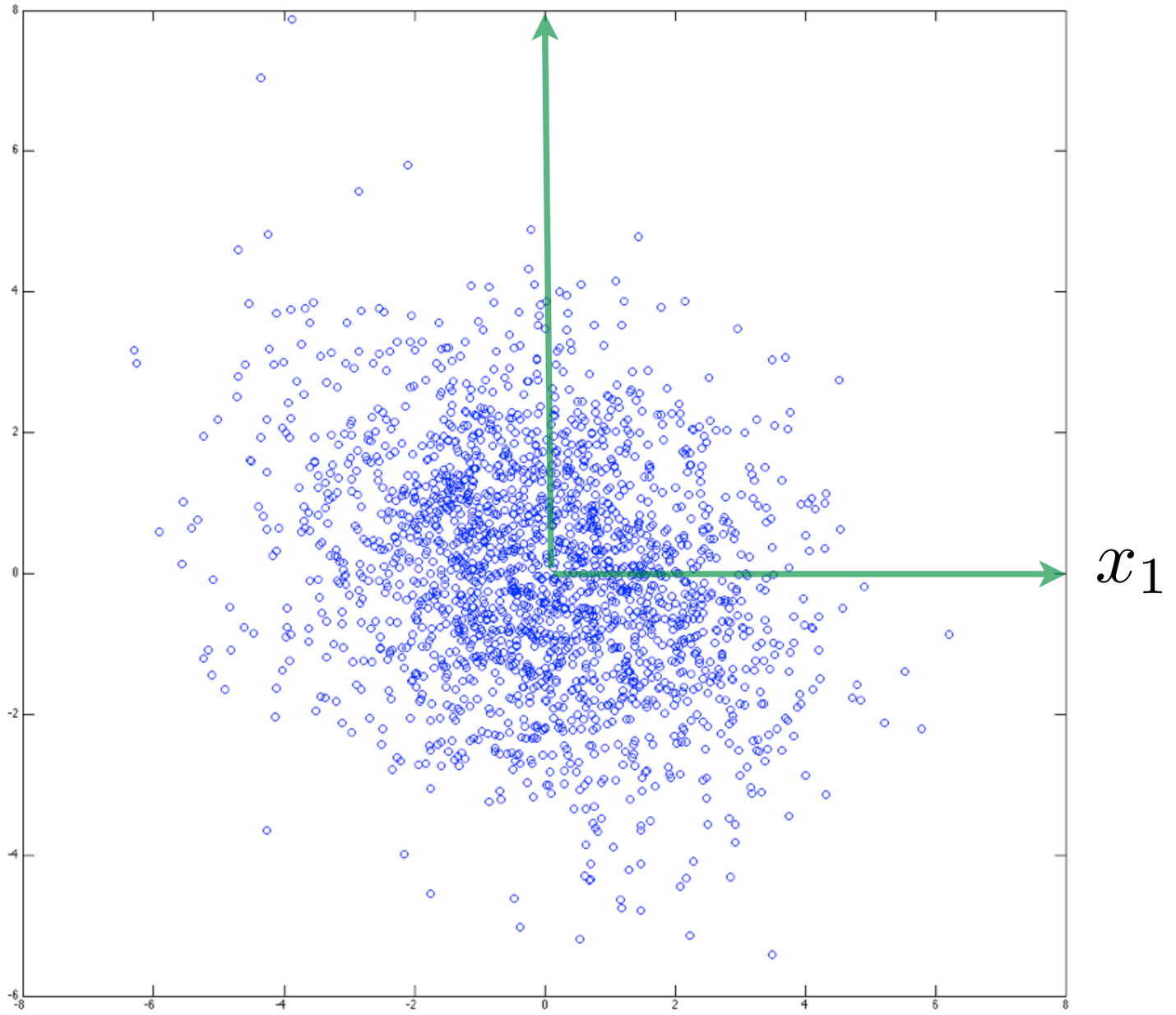
A (useful) geometric interpretation

Setting:



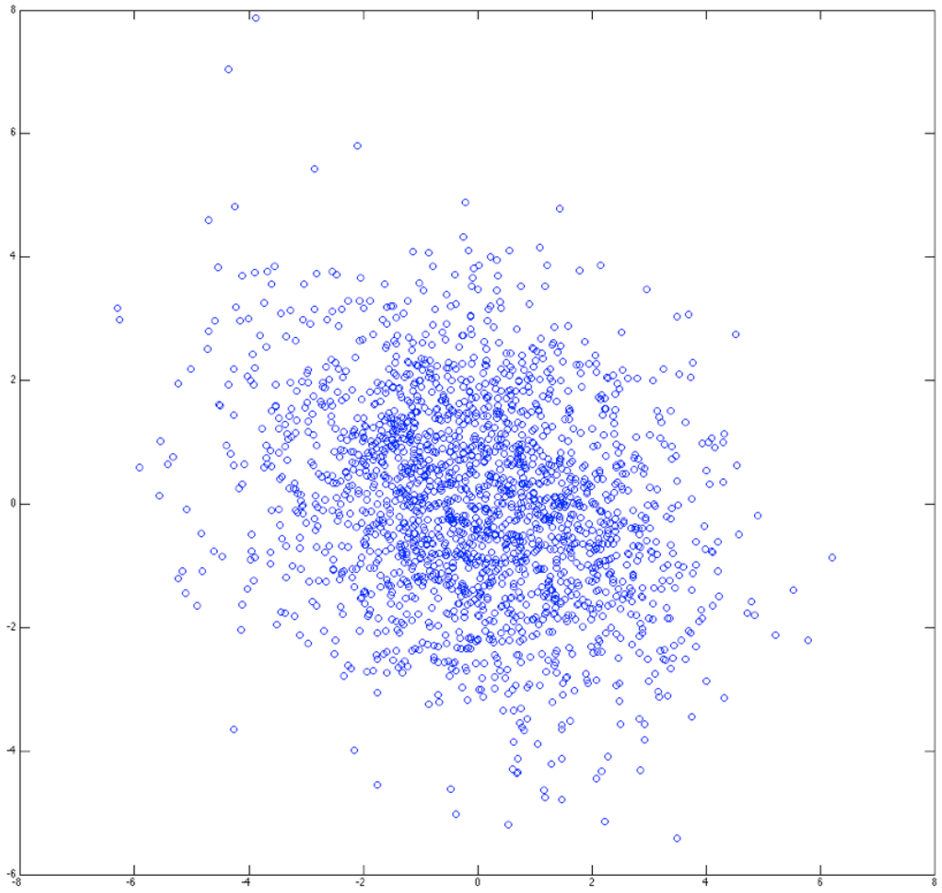
Sampling of $f_S(s_1,s_2)$

A (useful) geometric interpretation

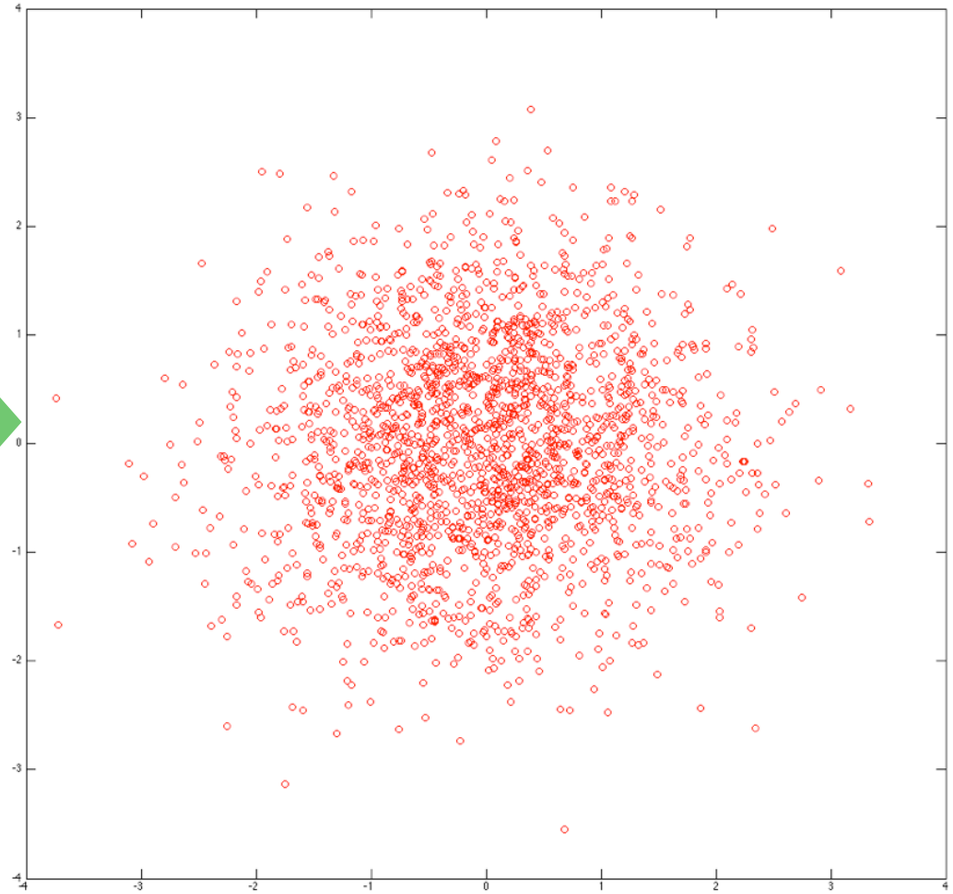


Sampling of $f_X(x_1,x_2)$

A (useful) geometric interpretation



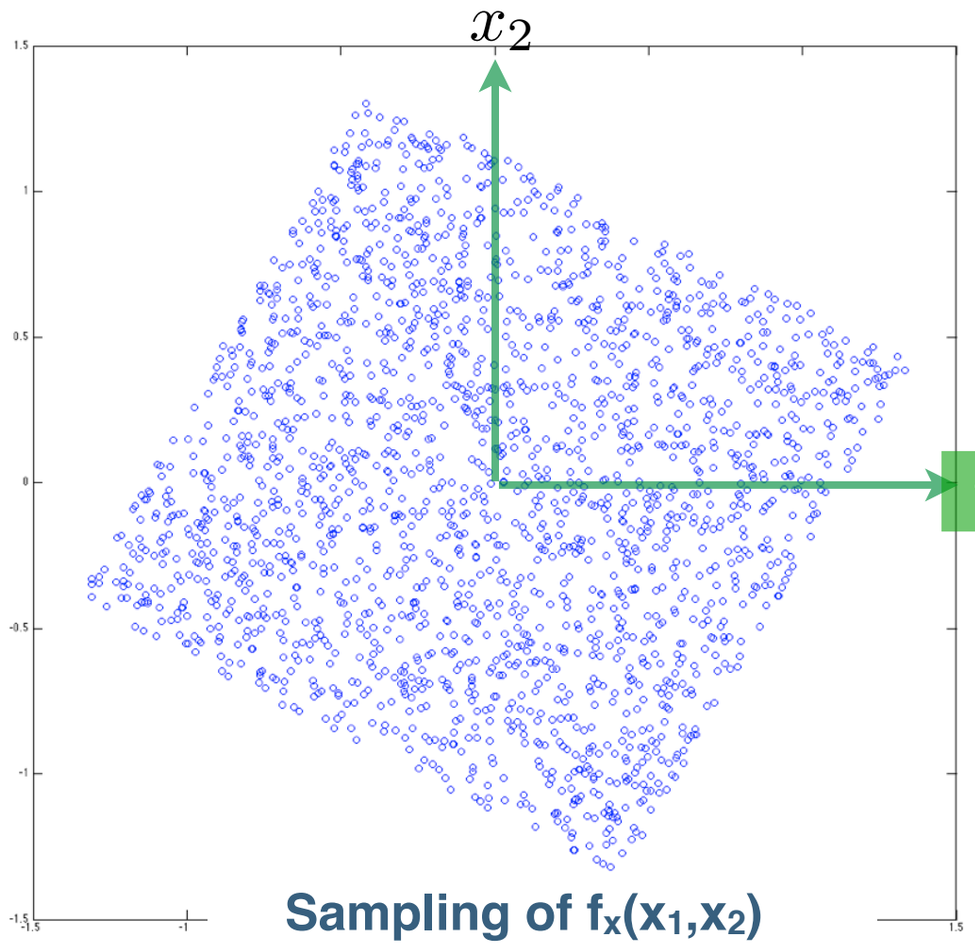
?



The two distribution are uncorrelated

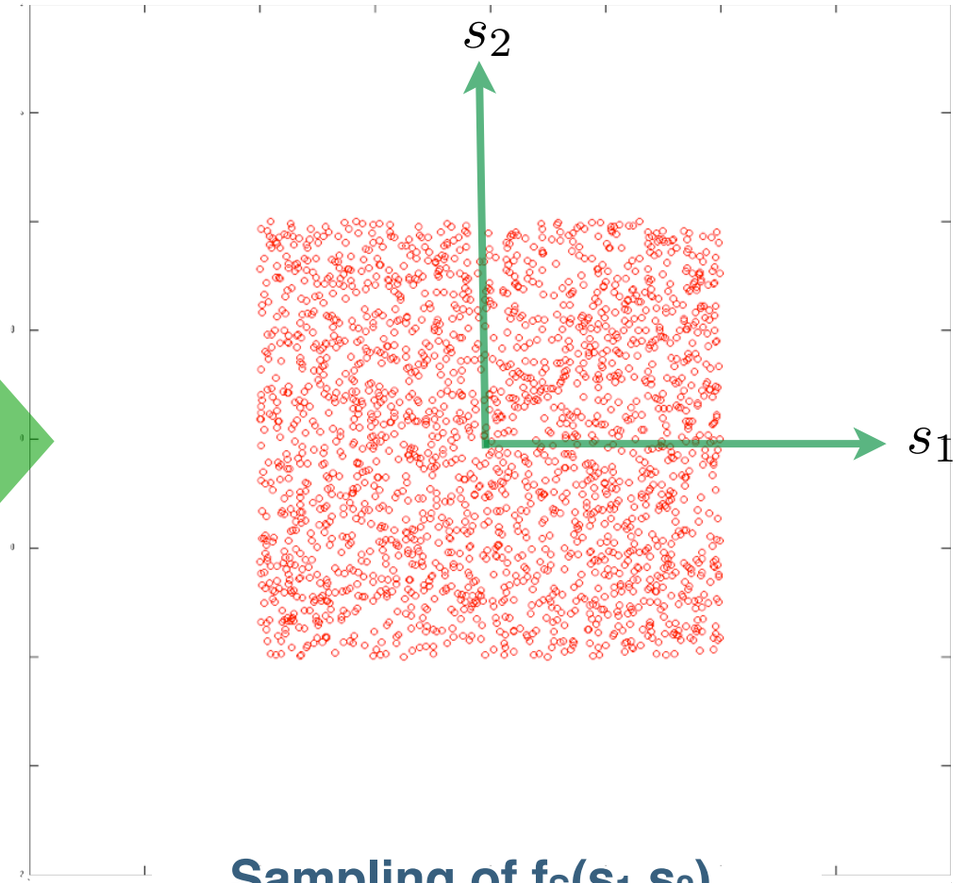
Second-order statistics

Example :



Sampling of $f_x(x_1,x_2)$

?



Sampling of $f_S(s_1,s_2)$

Independent Component Analysis

General principles

Setting things up

Whitening is a pre-processing step that allows to reduce the parameter space, and allows to:

- **Constrain the amplitude of the sources S** (*i.e.* their norm)

- **Limits the estimation procedure to orthogonal matrices**

In practice, the objective is to transform the data so that :

$$\tilde{X} = \Pi X \qquad R_{\tilde{X}} = \mathbb{E}\{\tilde{X}\tilde{X}^T\} = I$$

Setting things up

In practice, the objective is to transform the data so that :

$$\tilde{X} = \Pi X \qquad R_{\tilde{X}} = \mathbb{E}\{\tilde{X}\tilde{X}^T\} = I$$

How can we do that ?

Step 1 - Décomposition in eigenvalues/vectors of R_x :

$$R_X = P\Sigma P^T$$

Step 2 - Whitening the data :

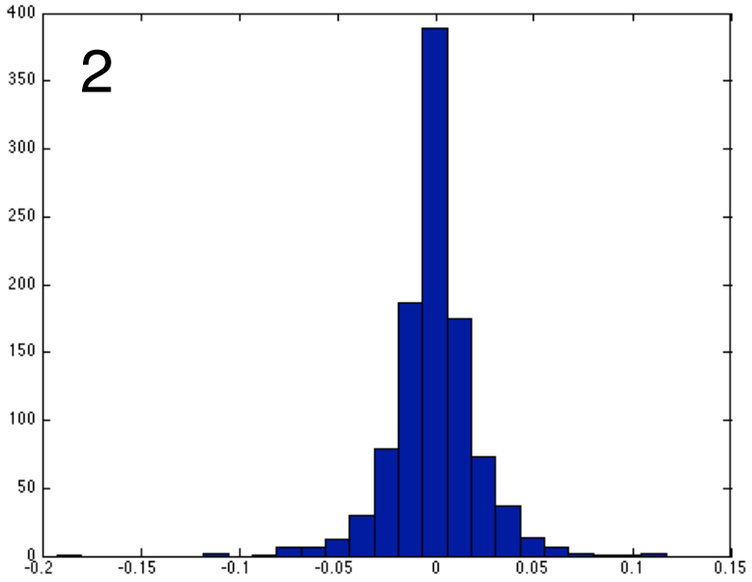
$$\tilde{X} = \Sigma^{-1/2} P^T X$$

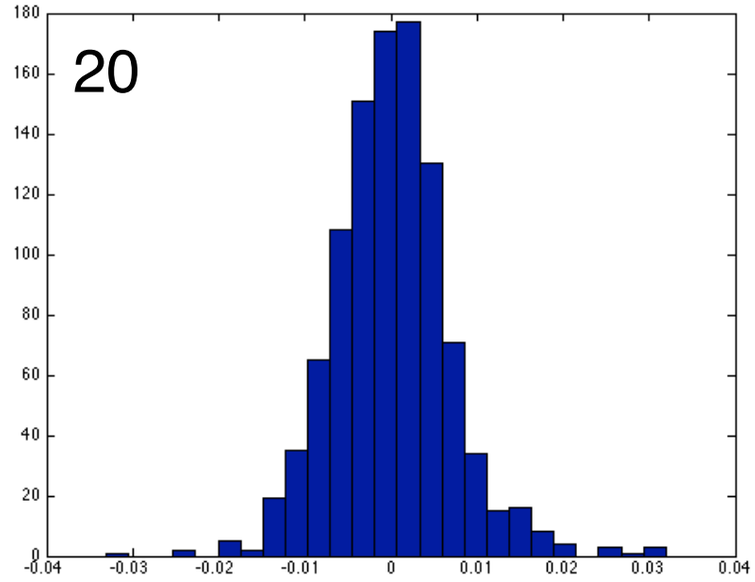
The mixture model then boils down to :

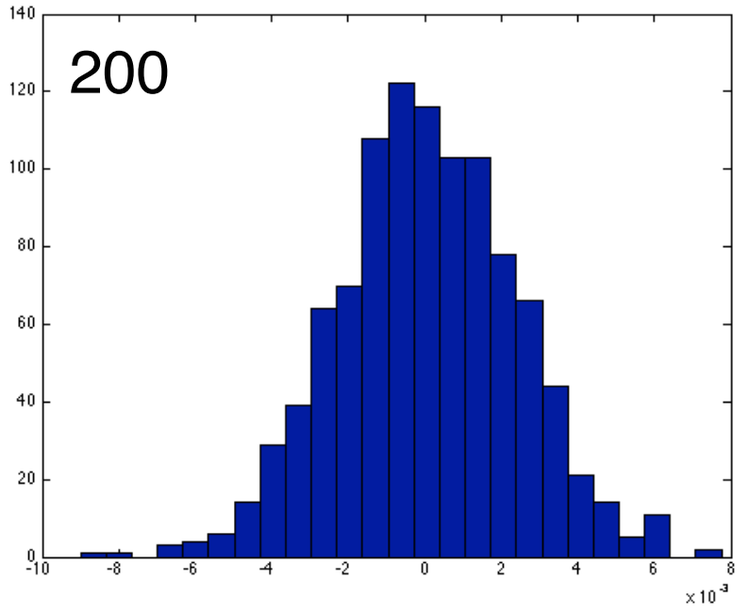
$$\tilde{X} = \Sigma^{-1/2} P^T A S = \tilde{A} S$$

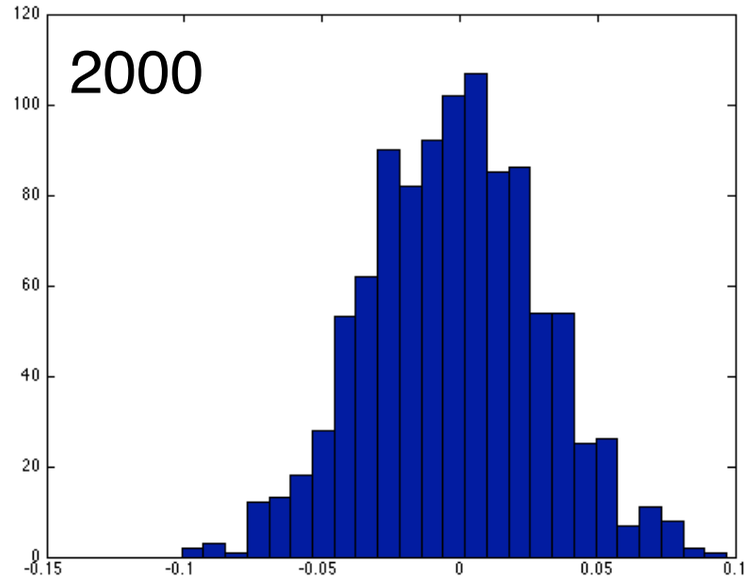
Assuming that $R_S = \mathbb{E}\{SS^T\} = I$ **then** $\tilde{A}\tilde{A}^T = I$

Let's get some intuition









Mixing → **"Gaussianizing"**

Let's get some intuition

Fundamentally, each observation is a mixture of independent R.V. :

$$x_i = \sum_{j=1}^{n} a_{ij} s_j$$

Assuming unitary weights, and i.i.d. sources :

$$x_i = \sum_{j=1}^{n} s_j$$

Thanks the central limit theorem :

$$\forall t; \quad \frac{1}{\sqrt{n}} x_i[t] = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} s_j \longrightarrow x \sim \mathcal{N}(0,1)$$

Mixing ⟷ **"Gaussianizing"**

A theoretical motivation

A theoretical result - the Darmois-Linnik theorem (1953)

Hypotheses : - i.i.d sources
- A is invertible
- **at most one source has a Gaussian distribution**

Results : if $\hat{S} = BX$ independent then $\hat{S}$ is an admissible solution

This motivates the estimation of B so that the sources are independent ...

But … what is statistical independence

Separating out the sources by enforcing their independence amounts to factorizing the joint density probability of the estimated sources :

$$f_{s_1 \cdots s_n}(z_1, \cdots, z_n) = \prod_j f_{s_j}(z_j)$$

That would yield to handle an equality between multi-valued data
which is hard to perform in practice

It is then necessary to resort to an indirect measure of independence !

An information-theoretic approach

Rather than manipulating functional equalities :

$$f_{s_1 \cdots s_n}(z_1, \cdots, z_n) = \prod_j f_{s_j}(z_j)$$

It is much simpler to handle a "divergence" between both terms :

$$\mathcal{D}\left(f_{s_1 \cdots s_n}(z_1, \cdots, z_n), \prod_j f_{s_j}(z_j) \right) = 0$$

In that case, the divergence of choice is the Kullback-Leibler divergence

An information-theoretic approach

Mutual information is defined as follows:

$$I(\mathbf{S}) = \int \cdots \int f_{\mathbf{S}}(\mathbf{z}) \log \frac{f_{\mathbf{S}}(\mathbf{z})}{\prod_j f_{s_j}(z_j)} d\mathbf{z}$$

$$= \mathcal{D}\left(f_{\mathbf{S}}, \prod_j f_{s_j} \right)$$

It turns out to be the KL divergence between the joint probability and the product of the marginal distributions

Properties : - positivity : $I(\mathbf{S}) \geq 0$

- minimum : $I(\mathbf{S}) = 0 \text{ ssi } f_{\mathbf{S}} = \prod_j f_{s_j}$

An information-theoretic approach

To enforce independence between the sources, it suffices to minimize their mutual information.

It is standard to formulate the MI as a function of the entropy function :

$$I(\mathbf{S}) = \sum_j H(s_j) - H(\mathbf{S})$$

Since the mixing matrix is invertible, this yields:

$$I(\mathbf{S}) = \sum_j H(s_j) - H(\mathbf{X}) - \log|\det(B)|$$

Then: $$\min_B I(\mathbf{S}) \Longleftrightarrow \min_B \sum_j H(s_j) - \log|\det(B)|$$

Introducing the score function

Let's seek the minimum of the MI with respect to the unmixing matrix:

$$\frac{dI}{dB} = \sum_j \frac{dH}{dB}(s_j) - \frac{d\log|\det(B)|}{dB}$$

$$= \sum_j \mathbb{E}\left\{ -\frac{d\log f_{s_j}(s_j)}{ds_j}\frac{ds_j}{dB} \right\} - B^{-T}$$

We then define:

$$\phi_j(z) = -\frac{d\log f_{s_j}(z)}{dz}$$

$\phi_j(z)$ **is fully characterized by the probability density of the sources, it is called the score function.**

Introducing the score function

If we look for the minimum of the MI:

$$\begin{aligned}\frac{dI}{dB} &= \sum_j \mathbb{E}\left\{\phi_j(s_j)\frac{ds_j}{dB}\right\} - B^{-T} \\ &= \mathbb{E}\left\{\varphi_S(S)X^T\right\} - B^{-T} \\ &= \left(\mathbb{E}\left\{\varphi_S(S)S^T\right\} - I\right)B^{-T}\end{aligned}$$

This leads to the following optimality condition:

$$\mathbb{E}\left\{\varphi_S(S)S^T\right\} = I$$

InfoMAX

Proposed in 1995 by Bell & Sejnowski *(neural computation, #7)*, the **InfoMAX** algorithm looks for the minimum of the mutual information:

$$\begin{aligned}\hat{B} &= \mathrm{Argmin}_B \; I(S) \\ &= \mathrm{Argmin}_B \; I(BX)\end{aligned}$$

It makes use of first-order information about the MI:

$$\begin{aligned}\frac{dI}{dB} &= \left(\mathbb{E}\left\{\varphi_S(S)S^T\right\} - I\right) B^{-T} \\ &= \mathbb{E}\left\{\varphi_S(BX)X^T\right\} - B^{-T}\end{aligned}$$

It is based on a gradient-descent algorithm:

1) Initialize $B^{(0)}$ randomly

2) Repeat until convergence :

$$B^{(k+1)} = B^{(k)} - \alpha(t)\nabla_B I(B^{(k)}X)$$

↑

"learning rate"

$$B^{(k+1)} = B^{(k)} + \alpha(t)\left(B^{(k)^{-T}} - \varphi_S(B^{(k)}X)X^T\right)$$

Some remarks about InfoMAX :

1) In practice, **expectations are only empirical**, the quality of separation will therefore by dependent on the number of samples

2) What choice for φ_S ?

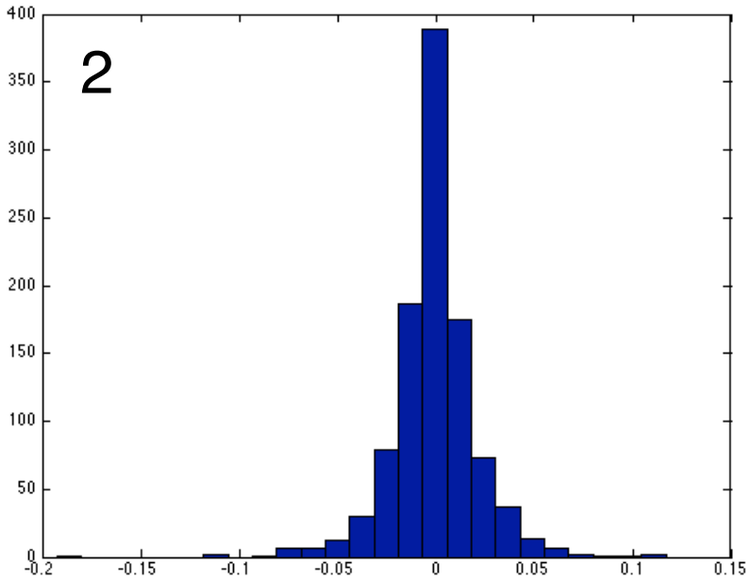
In theory : $$\forall i; \phi_{s_i}(s_i) = -\frac{d \log f_{s_i}}{dz}(s_i)$$

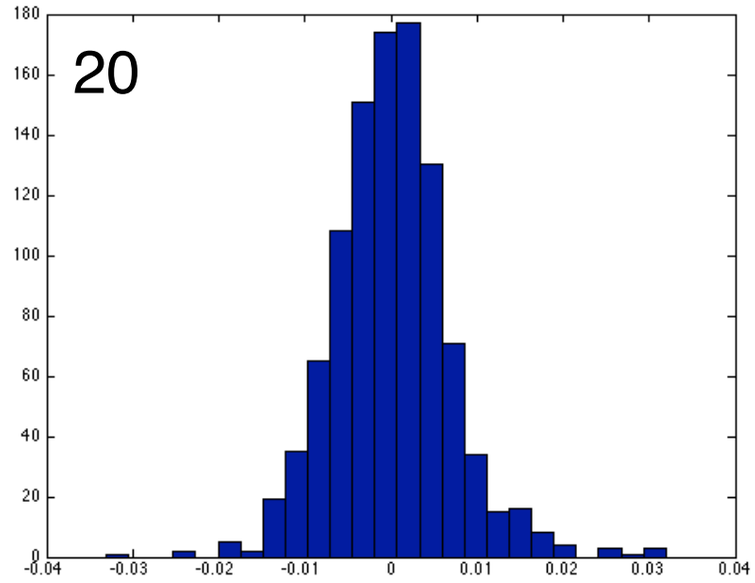
En pratique, it is unknown and is generally approximated with well-chosen non-linear functional such as :

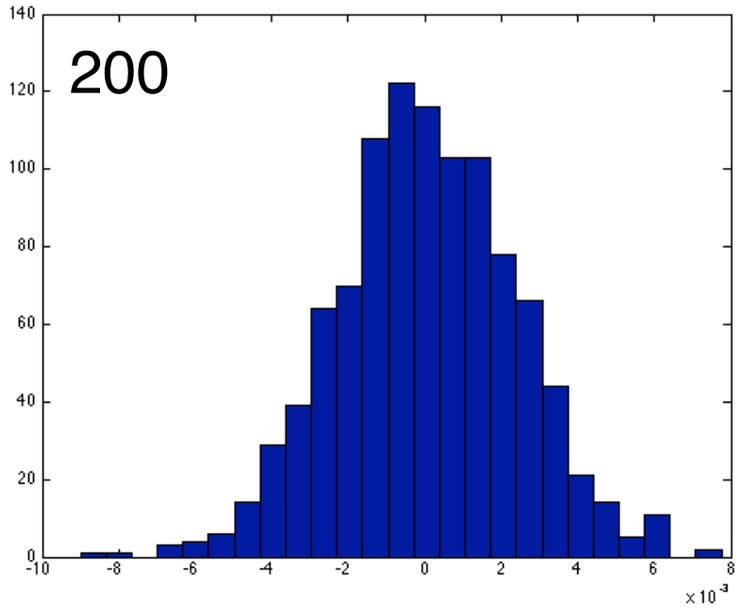
$$g(s_i) = \tanh(s_i)$$ dans le cas super-gaussien

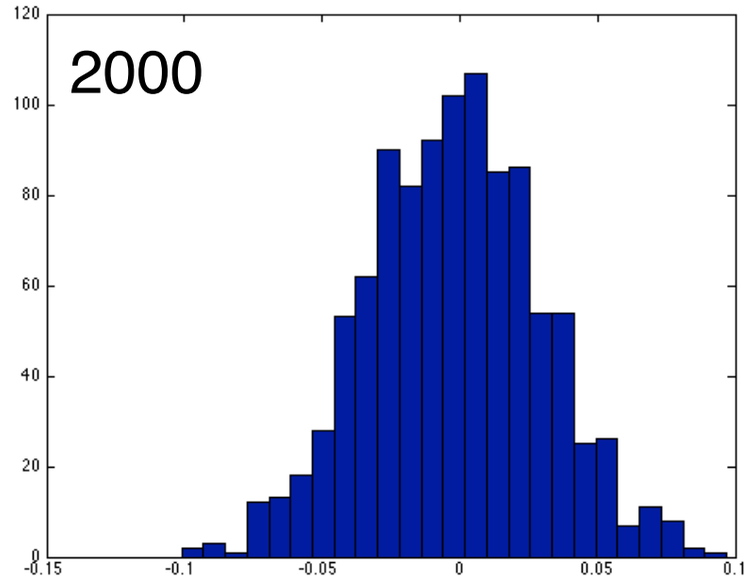
$$g(s_i) = s_i - \tanh(s_i)$$ dans le cas sous-gaussien

Maximising the non-gaussianity of the sources









Mixing → **“Gaussianizing”**

Maximising the non-gaussianity of the sources

IDEA : estimating the unmixing matrix so that the sources are the "least Gaussian possible".

This can be done by maximizing some contraste function that is sensitive to the non-Gaussianity of the source *(Hyvarinen, 97)* **:**

$$\Gamma_G(\mathbf{S}) = \sum_{i=1}^{n} \mathbb{E}\{G_i(s_i)\}$$

Where HOS appear such as the 4-th order statistics :

$$G_i(s_i) = s_i^4 \qquad \textit{sub-gaussian sources}$$

$$G_i(s_i) = -s_i^4 \qquad \textit{super-gaussian sources}$$

Maximising the non-gaussianity of the sources

To that end, the FastICA algorithm *(Hyvarinen,99)* **has been proposed to maximize :**

$$\Gamma_G\left(\mathbf{S}\right) = \sum_{i=1}^{n} \mathbb{E}\{G_i(s_i)\}$$

or equivalenty

$$\Gamma_G\left(\mathbf{B}\right) = \sum_{i=1}^{n} \mathbb{E}\left\{G_i\left(b^{i^T}\mathbf{X}\right)\right\}$$

that is maximized:

$$\hat{\mathbf{B}} = \mathrm{Argmax}_{\mathbf{B}}\ \Gamma_G\left(\mathbf{B}\right)$$

Maximum likelihood estimation

Principle : estimating the unmixing matrix so that the likelihood is maximized

More precisely:

$f_X(u|A)$ **is the likelihood of A**

By a simple change of variables:

$$\log f_X(u|A) = \log f_S(A^{-1}u) + \log|\det(A^{-1})|$$

(A corresponds to the Jacobian of the transformation)

Maximum likelihood estimation

If one looks at its maximum, it leads to:

$$\frac{\partial \log f_X}{\partial B} = \frac{\partial \log f_S}{\partial B} - \frac{d \log |\det(B)|}{dB}$$

$$= \left(\mathbb{E}\left\{ \varphi_S(S) S^T \right\} - I \right)$$

Equivalent to minimizing the MI !!!

Summing things up !

Different approches :

- Minimizing mutual information

- Maximizing the non-Gaussianity of the sources

- Maximizing the likelihood of the mixing matrix

are “equivalent”

$$\mathbb{E}\left\{\varphi_S(S)S^T\right\} = I$$

A unifying information-theoretic framework for independent component analysis **- Lee et al. 1997**

Independent Component Analysis

Practical aspects

A non-convex minimisation problem

InfoMax Algorithm - *Bell & Sejnowski (1995)*

The gradient of the MI with respect to B is given by :

$$\frac{dI}{dB} = \varphi_S(S)X^T - (B^T)^{-1}$$

The MI can therefore be minimized by gradient descent :

$$B^{(k+1)} = B^{(k)} + \alpha(k)\left(B^{(k)^{-T}} - \varphi_S(B^{(k)}X)X^T\right)$$

[Natural] gradient descent

Usual gradient:

$$\ell(u+h) = \ell(u) + \langle \nabla \ell(u), h \rangle + o\left(\|h\|\right)$$

if u belongs to a multiplicative group (e.g. mixing matrixes), it is much more efficient to keep the multiplicative structure in the descent process :

$$u + h \longrightarrow u + wu$$

Relative/natural gradient :

$$\ell(u+wu) = \ell(u) + \langle \nabla^R \ell(u), w \rangle + o\left(\|w\|\right)$$

$$\nabla^R \ell(u) = \nabla \ell(u) u^T$$

[Natural] gradient descent

Usual gradient of the MI :

$$\nabla I(B) = \varphi_S(S)X^T - (B^T)^{-1}$$

Relative gradient of the MI:

$$\nabla^R I(B) = \left(\varphi_S(S)X^T - (B^T)^{-1}\right) B^T$$

$$= \varphi_S(S)S^T - I$$

[Natural] gradient descent

From the multiplicative structure, one has:

$$B^{(k+1)} = (I + w)B^{(k)}$$

with :

$$w = -\mu \nabla^R I(B^{(k)})$$

So that:

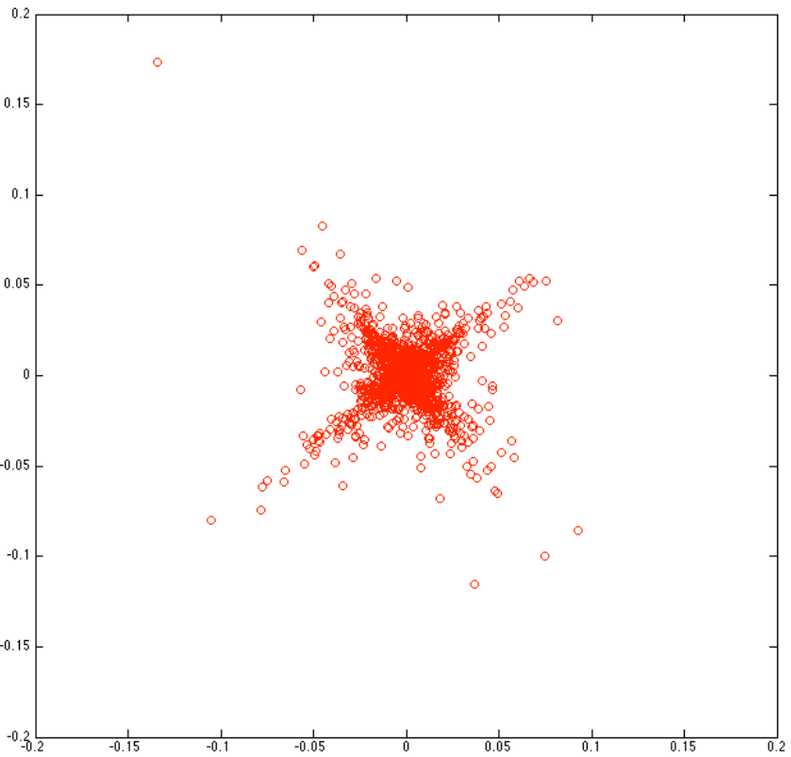
$$B^{(k+1)} = ((1+\mu)I - \mu\varphi_S(S)S^T)B^{(k)}$$

Relative gradient algorithm

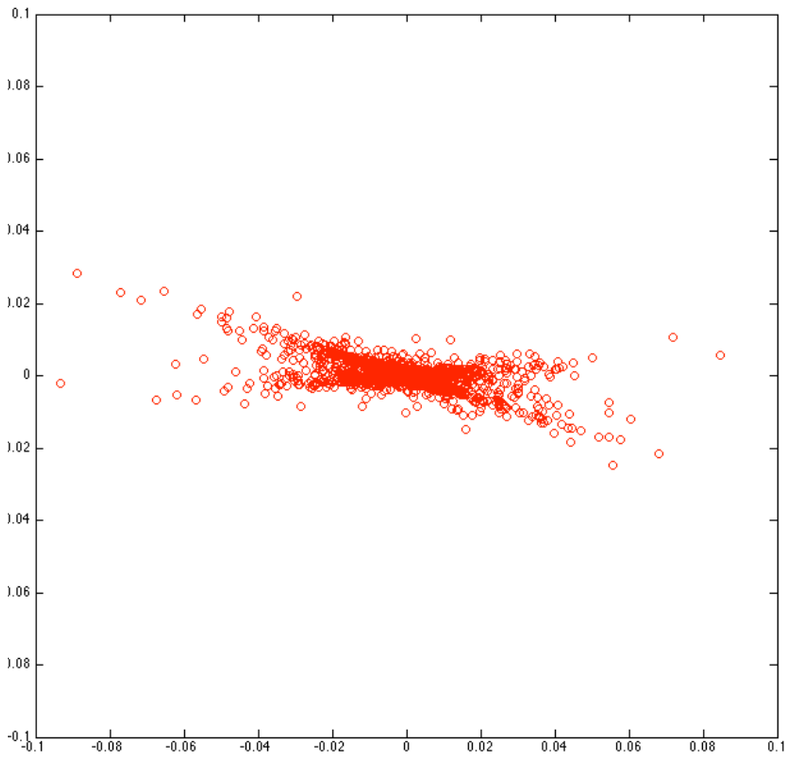
Equivariance

1 - Convergence performances of the relative gradient algorithm are close to the Newton algorithm ***(see Amari, Natural gradient works efficiently in learning,1998)***

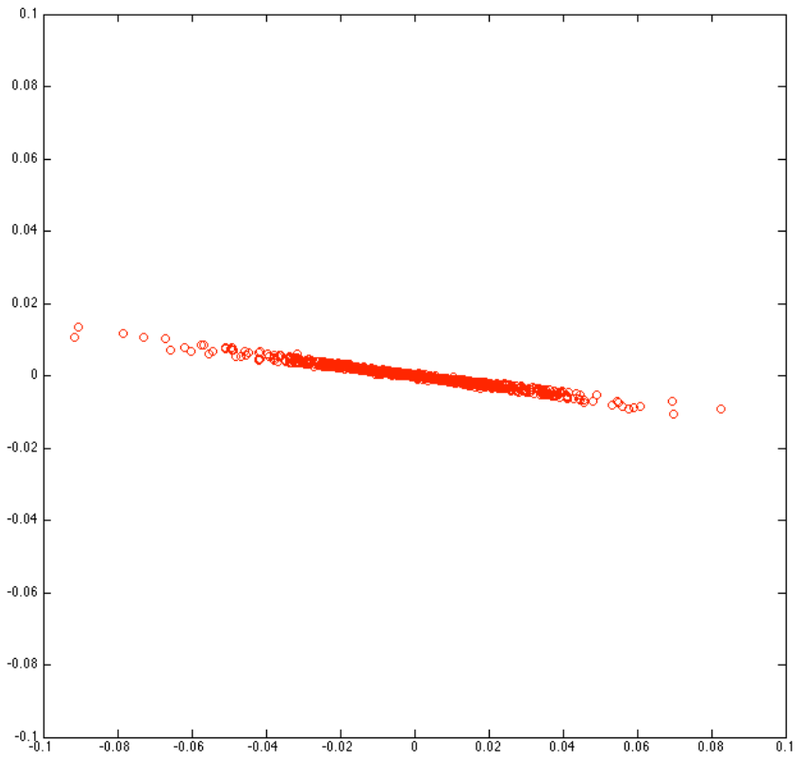
2 - Equivariance property : the convergence performances of this algorithm do not depend on the condition number of A !



A orthogonal



cond(A) = 5

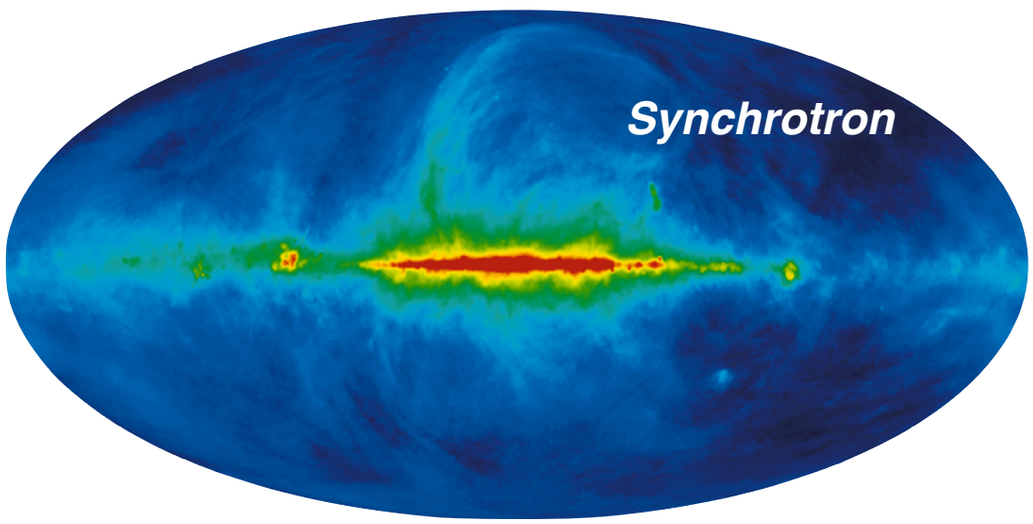


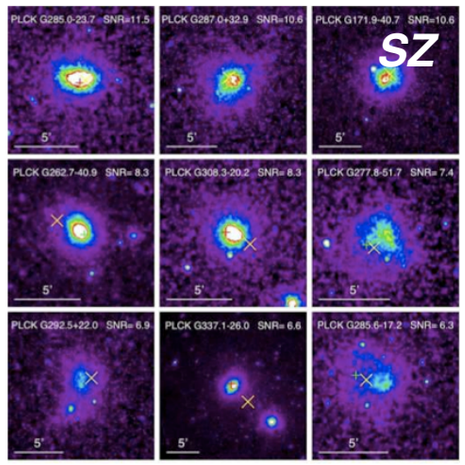
cond(A) = 50

Independent Component Analysis

Beyond basic statistical independence

How to account for the sources' structures





How can we account for the sources inner structures ?

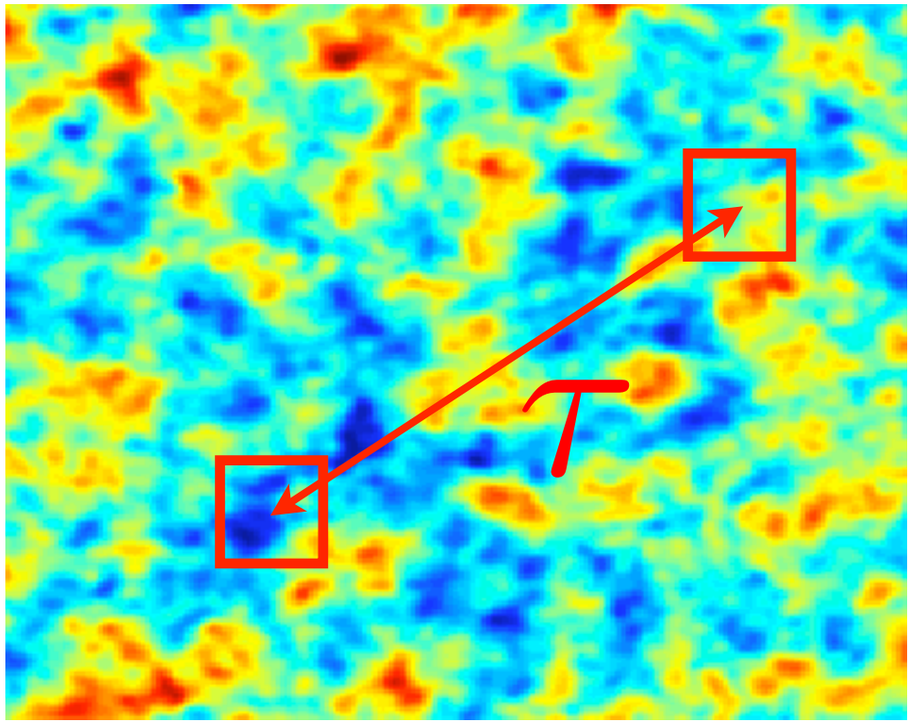
Accounting for correlations

Hypotheses et modeling :

- we assume that each source is stationary (in the wide sense) :

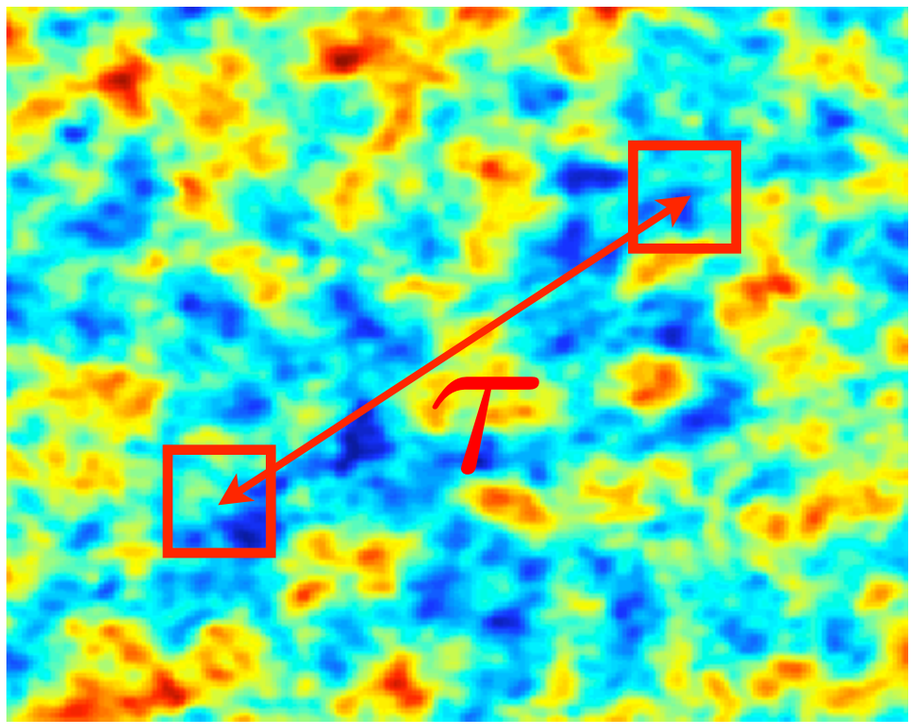
$$\mathbb{E}\{s_i[k]s_i[k']\} = c_i(|k-k'|)$$
$$= c_i(\tau)$$

autocorrelation function

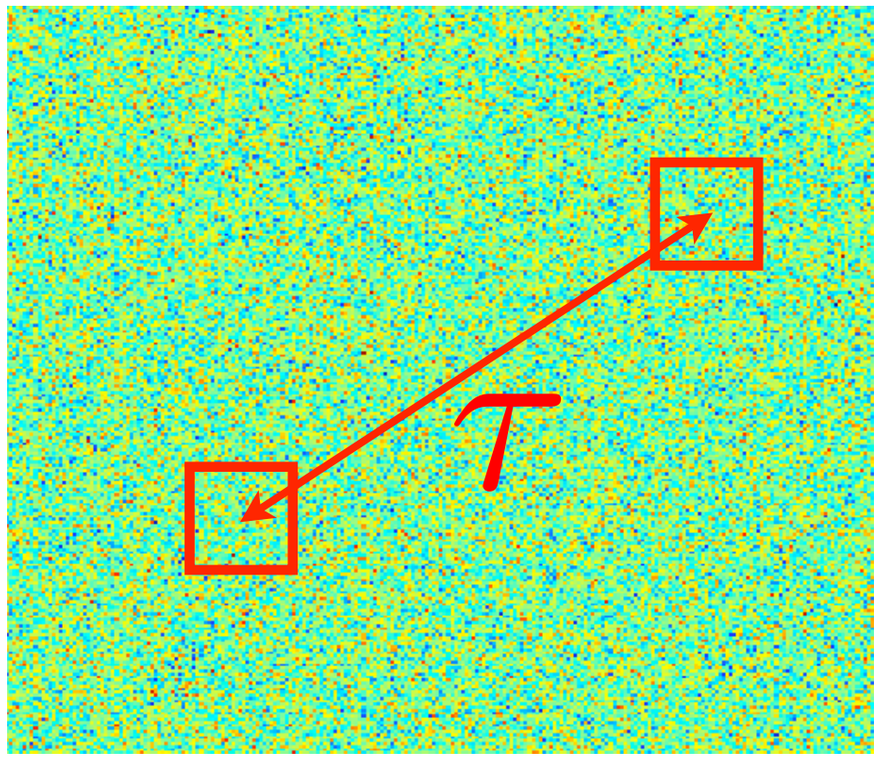


The correlation between two samples only depends on their distance

Accounting for correlations



Source



Source 2

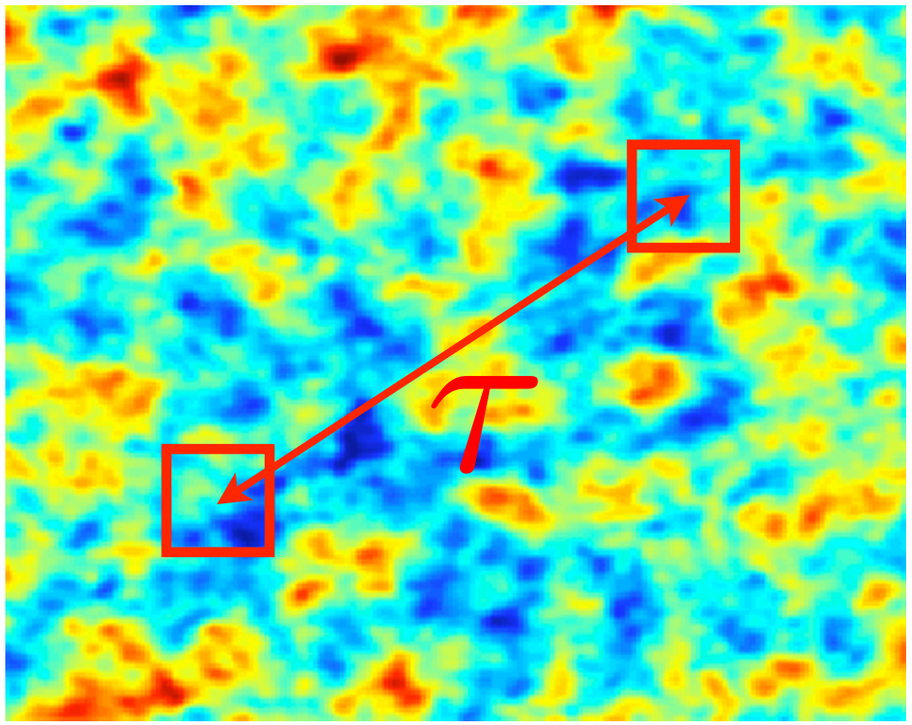
$$\mathbb{E}\{s_1[k]s_2[k']\} = 0$$

Decorrelation between the sources

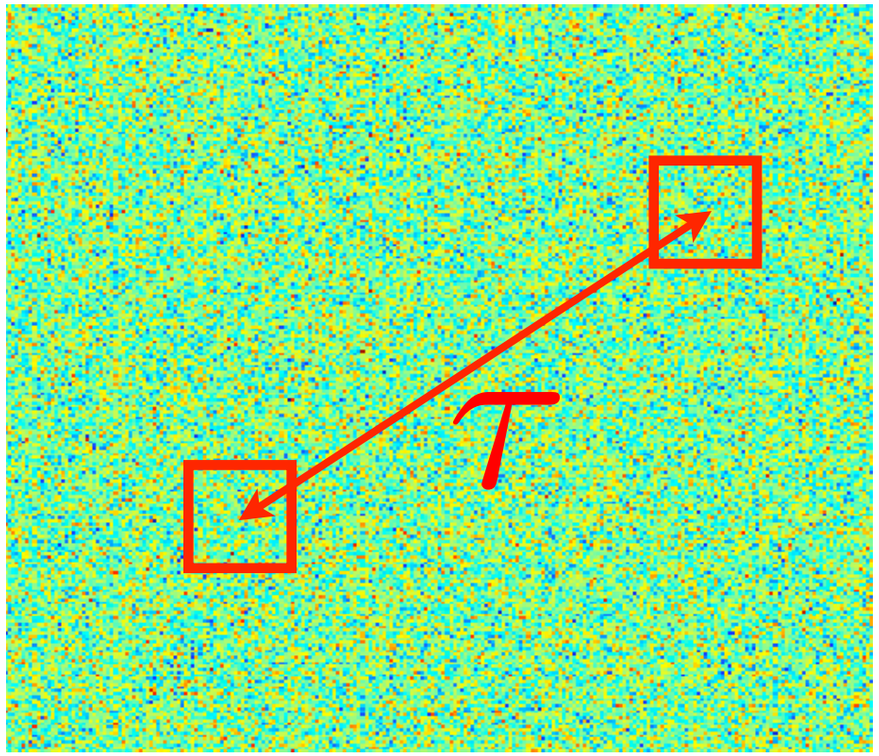
$$\mathbb{E}\{s_1[k]s_1[k']\} = c_1[\tau]$$

$$\mathbb{E}\{s_2[k]s_2[k']\} = c_2[\tau]$$

Accounting for correlations



Source

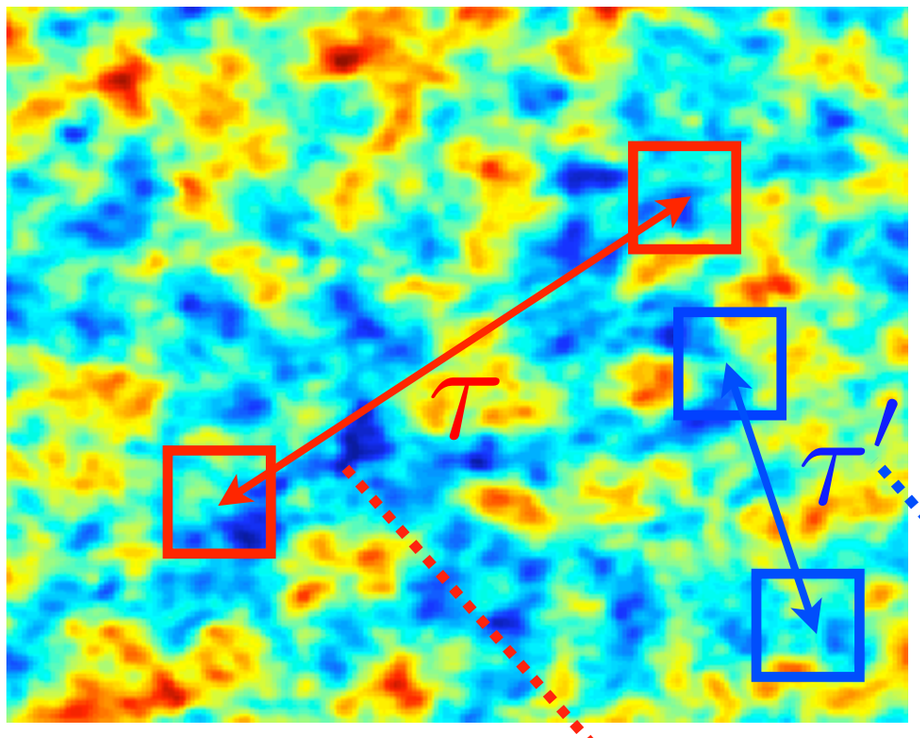


Source 2

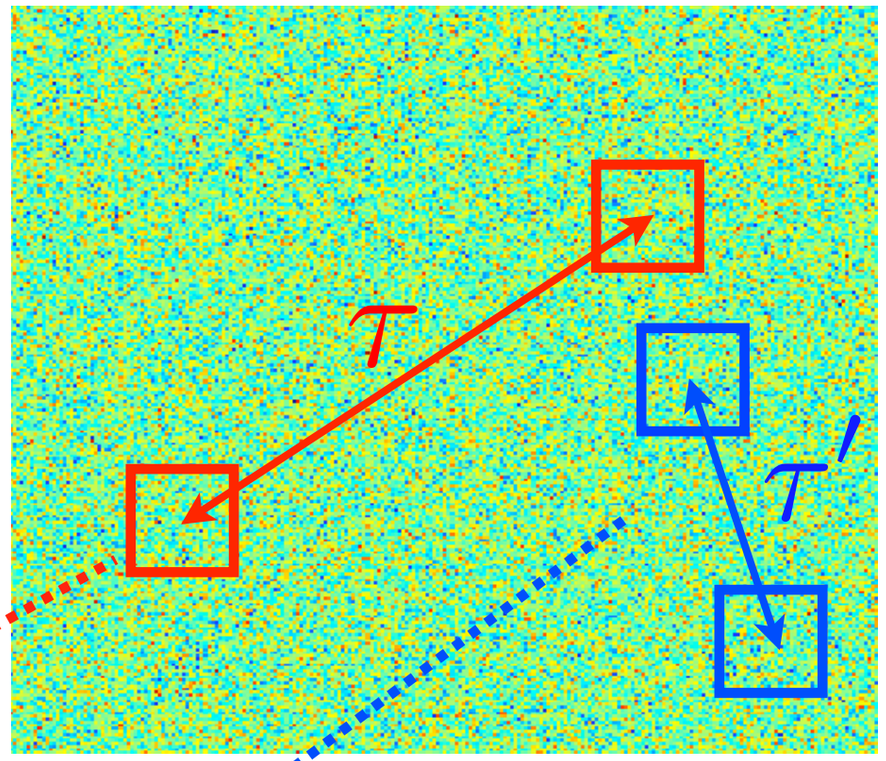
$$\mathbb{E}\{S[k]S[k']\} = \begin{bmatrix} c_1[\tau] & & 0 \\ & \ddots & \\ 0 & & c_n[\tau] \end{bmatrix}$$

How can we separation from this single correlation matrix ?

Accounting for correlations



Source 1



Source 2

$$\mathbb{E}\{S[k]S[k']\} = \begin{bmatrix} c_1[\tau] & & 0 \\ & \ddots & \\ 0 & & c_n[\tau] \end{bmatrix} \qquad \mathbb{E}\{S[k]S[k'']\} = \begin{bmatrix} c_1[\tau'] & & 0 \\ & \ddots & \\ 0 & & c_n[\tau'] \end{bmatrix}$$

These two matrices should be jointly diagonal

Joint-diagonalisation algorithms

Objective :

- We measure several covariance matrices in different frequency bands

- The mixing matrix is then estimated by joint diagonalization:

$$\forall \Omega; \quad \hat{R}_{\tilde{X}}[\Omega] = A \hat{R}_{\tilde{S}}[\Omega] A^T$$

In astrophysics : SMICA - Spectral matching ICA (Delabrouille, 02)

- Allows to account for a priori information about the source power spectrum (CMB)

- which bands, how many ?

- Somewhat limited: assumes that the sources are all stationary Gaussian processes, which is not always a good assumption - w-SMICA (Moudden, 05)

ICA - limitations

Score functions

ICA benefits from highly interesting statistical properties (think of the ML viewpoint):

- **Equivariance of the estimator**

- **Cramér-Rao bound for the mixing matrix estimator**

- **ICA estimators reach the CR bound as long as:**

$$\phi_j(z) = -\frac{d \log f_{s_j}(z)}{dz}$$

However, score functions are rarely known !

Score functions

Practical options are the following:

- Estimation of the source probability density within a parameterized family :

$$f_{s_i}(z) \simeq f(z|\theta_i \in \Theta)$$

- Approximate the score function :

$$g(S) \simeq \varphi_i(S)$$

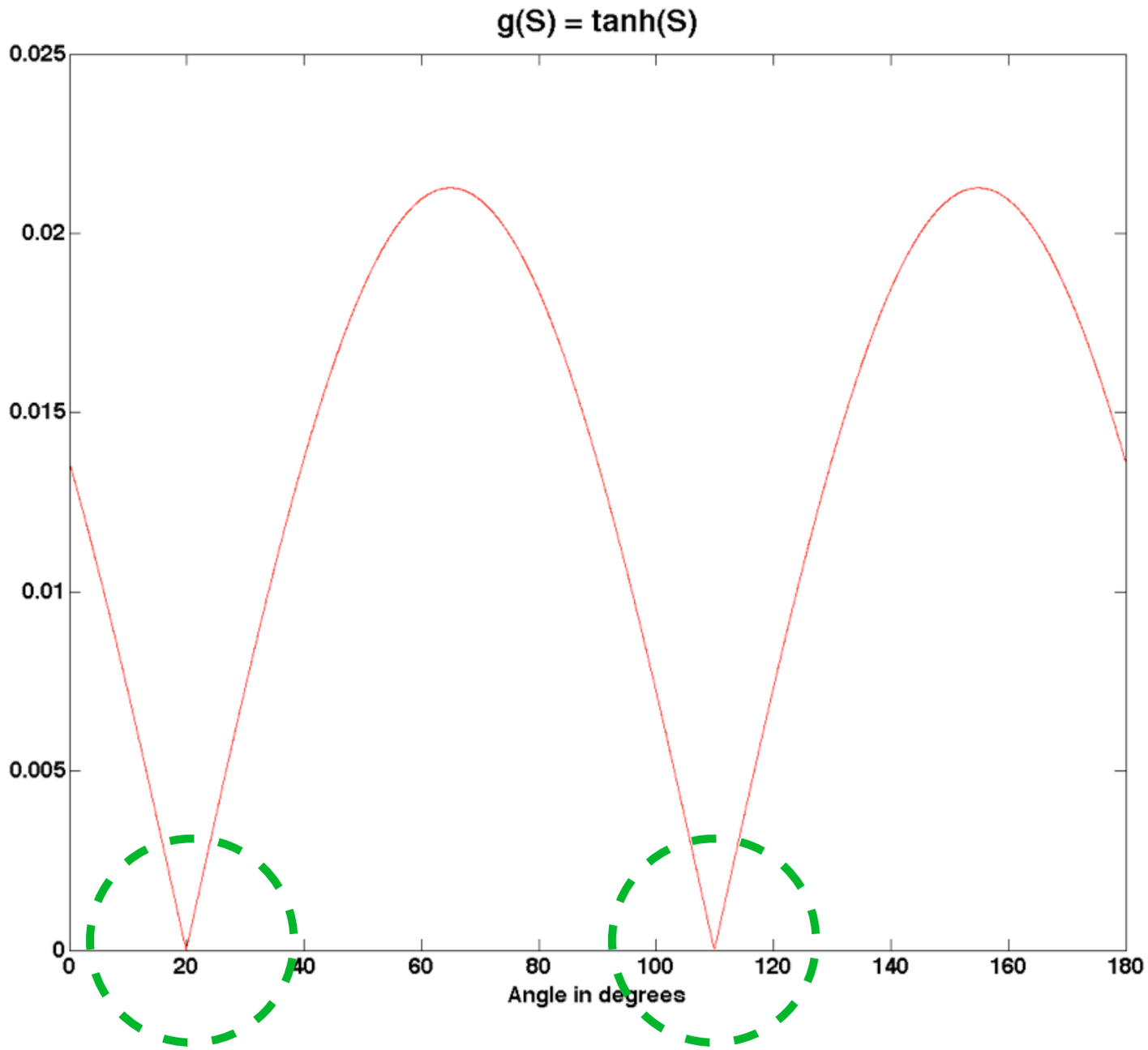
Good in general but …

- No optimality (CR bound)

- Stability is not always guaranteed

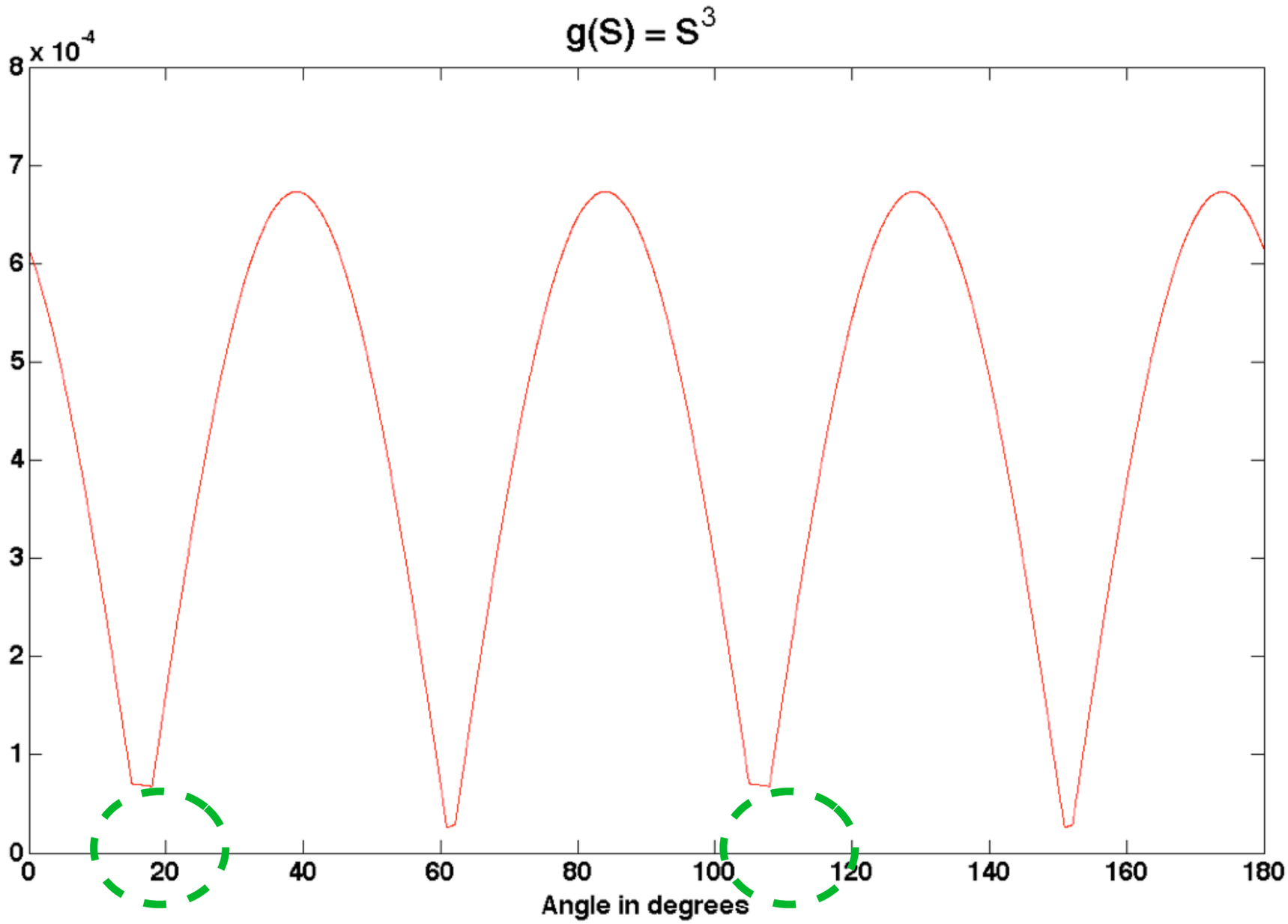
Score functions

FastICA with an approximating function :



Score functions

FastICA with an approximating function :

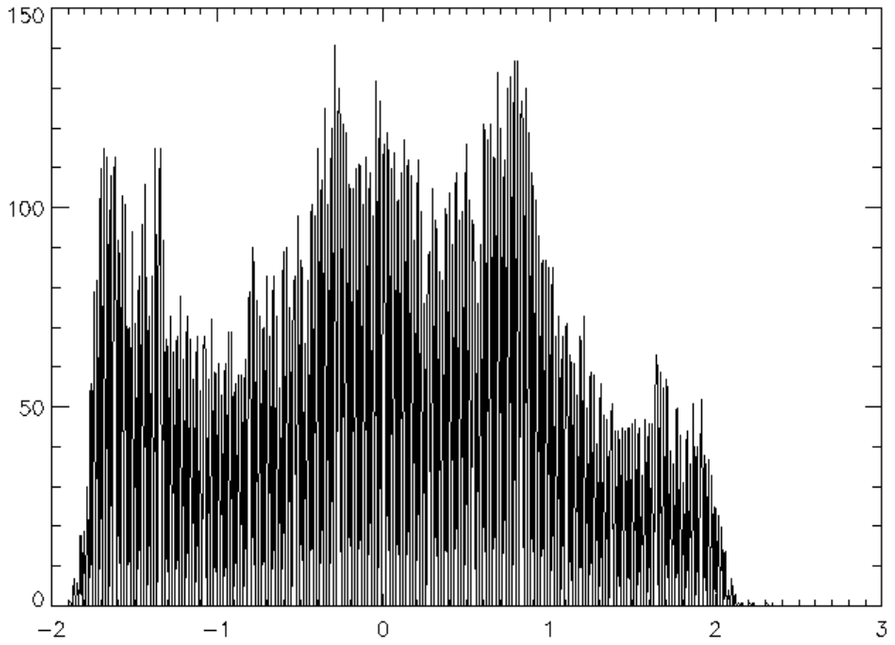


Source modelling

Basic ICA methods assume that the source samples are i.i.d. :

$$\forall t = 1, \cdots, T; \quad s_i[t] \sim f_{s_i}$$

Is that relevant for realistic sources ?



- This modeling is generally too simplistic

- The probability density of realistic sources is much more complex

Noise/model imperfections

Does not generally account for noise :

$$X = AS$$

Noise is standardly modeled by adding an additive perturbation :

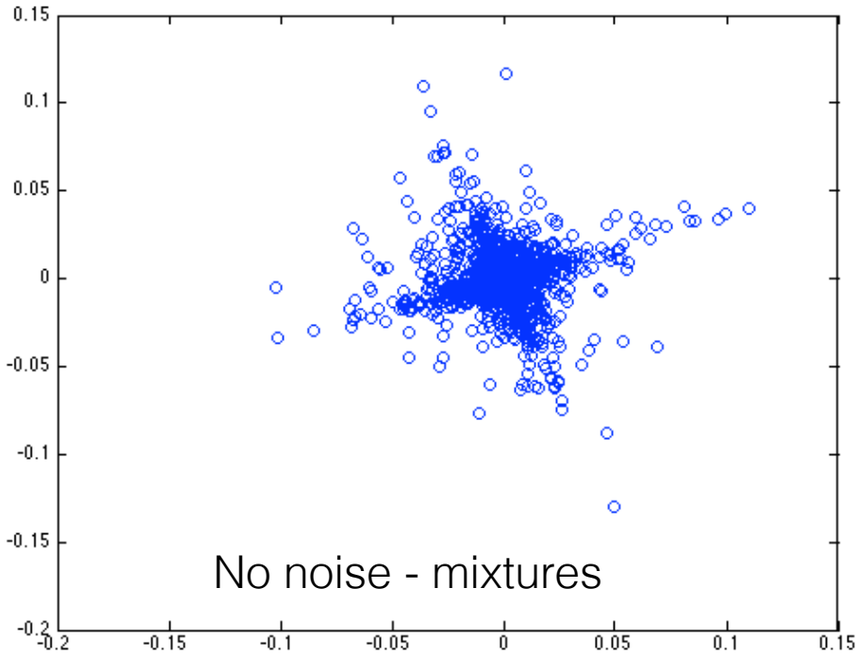
$$X = AS + N$$

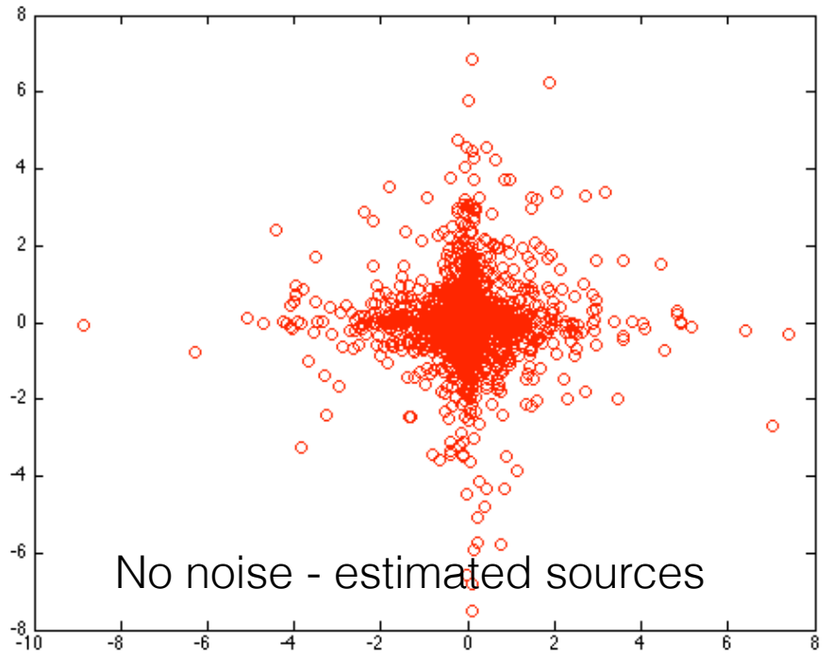
Where N models:

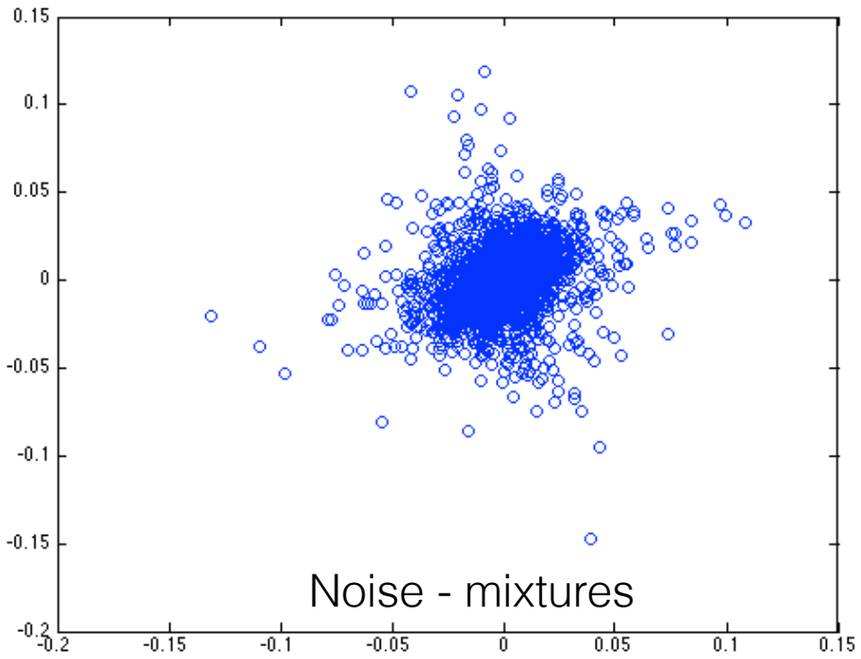
- **Additive instrumental noise**

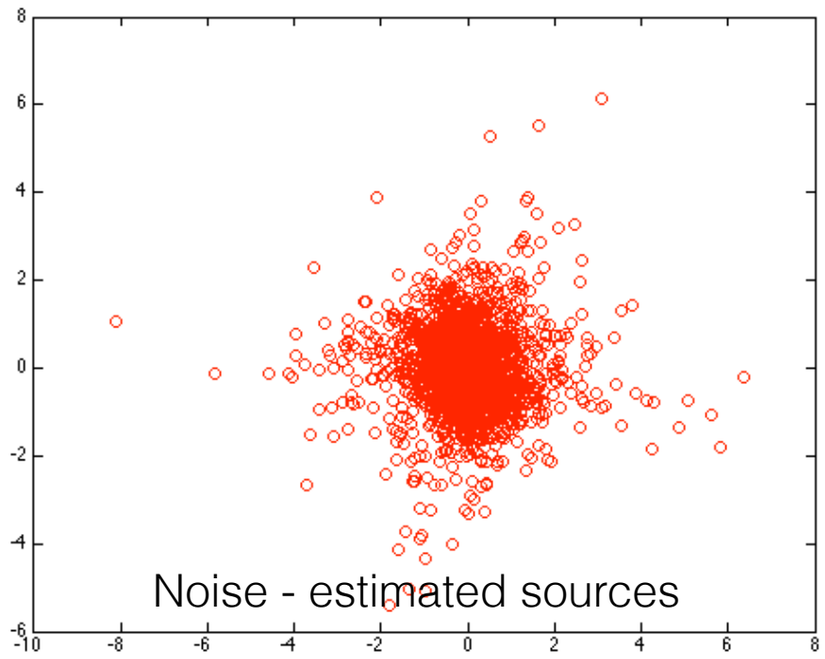
- **Imperfection of the model**

Noise/Model imperfections









Non-convexity of the optimisation problem

The linear mixture model is bilinear:

- It essentially leads to optimization problems that are

non convex

- So far, we have already seen that …solutions are equivalent up to a scaling/permutation

- Functions to be minimized (MI, log-ML, etc ...) have local minima

No guarantee of convergence to an admissible solution
Convergence highly depends on the initial point

Non-convexity of the optimisation problem