Information Geometry of ICA

J.-F. Cardoso, C.N.R.S.
Institut d'Astrophysique de Paris

Colloquium Prairie. February 5, 2020

About models

Climate Models Are Running Red Hot, and Scientists Don't Know Why

The simulators used to forecast warming have suddenly started giving us less time.

By **Eric Roston**

February 3, 2020, 11:00 AM GMT+1

There are dozens of climate models, and for decades they've agreed on what it would take to heat the planet by about 3° Celsius. It's an outcome that would be disastrous–flooded cities, agricultural failures, deadly heat–but there's been a grim steadiness in the consensus among these complicated climate simulations.

www.bloomberg.com/news/features/2020-02-03

What is the carbon footprint of learning with insanely large neural networks ?

Are we blowing up our carbon quota ? Would we like to have one ?

.

Information Geometry of Independent Component Analysis

Outline:

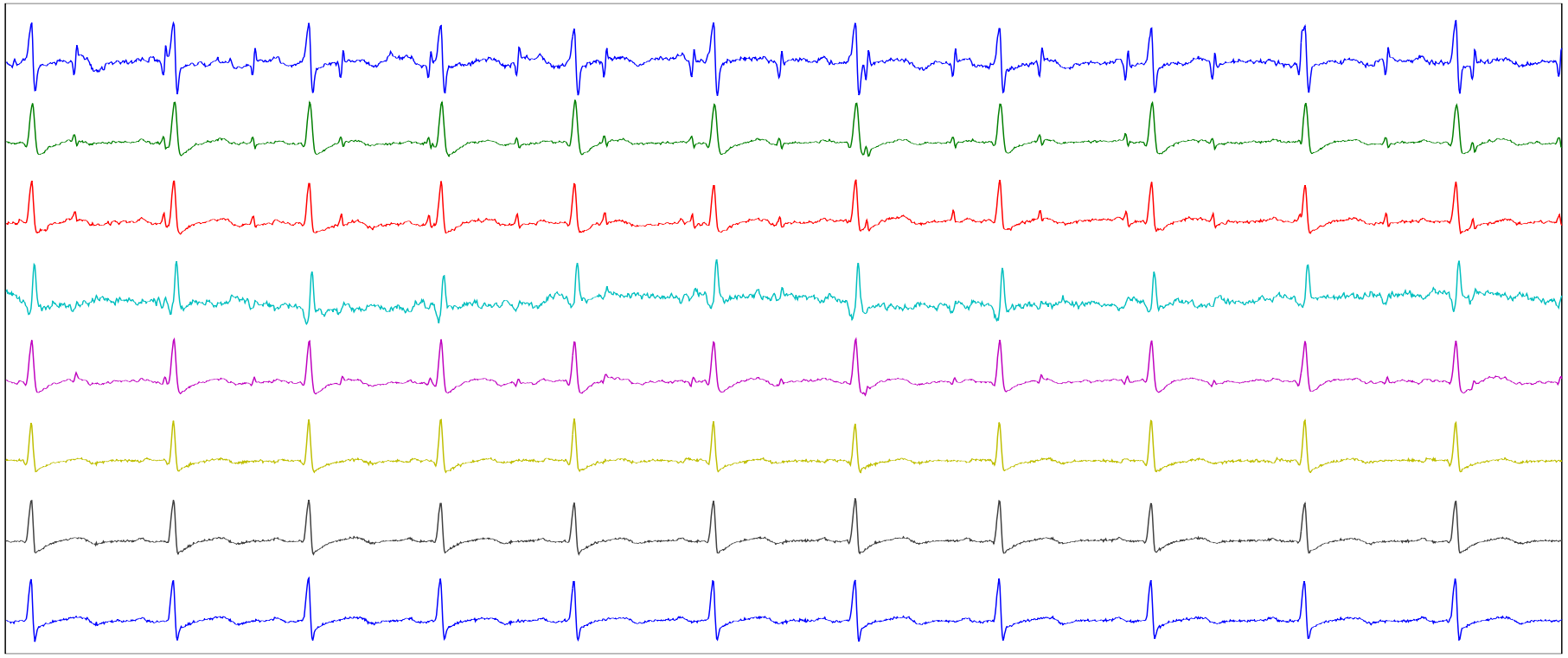
1. Independent Component Analysis

2. Information Geometry

3. Information Geometry of Independent Component Analysis

Independent Component Analysis

Multi-channel time series: a biomedical example



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

Looking for linear decompositions: Data = Mixing matrix $\times$ Sources.

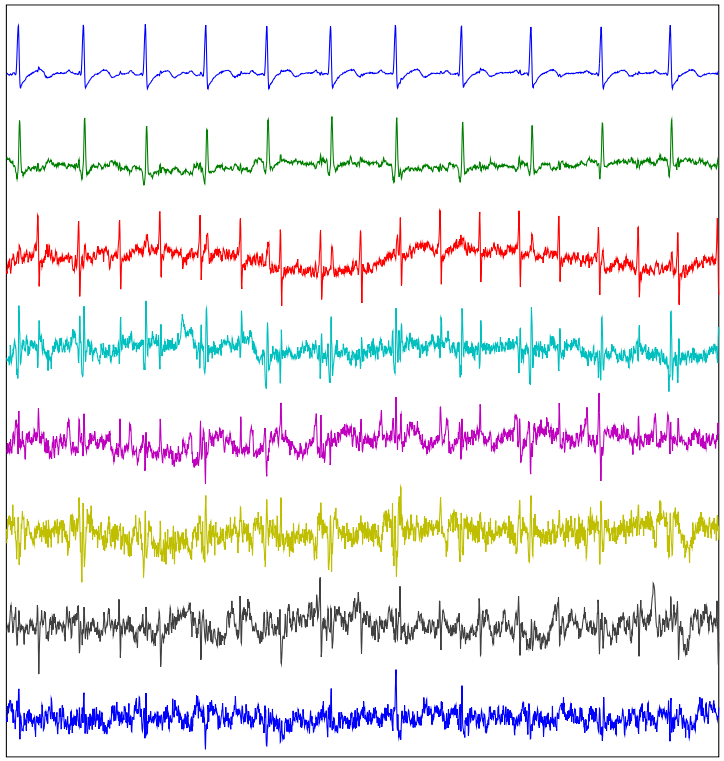
Data Credits: Daisy database

Principal component analysis

=

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

×

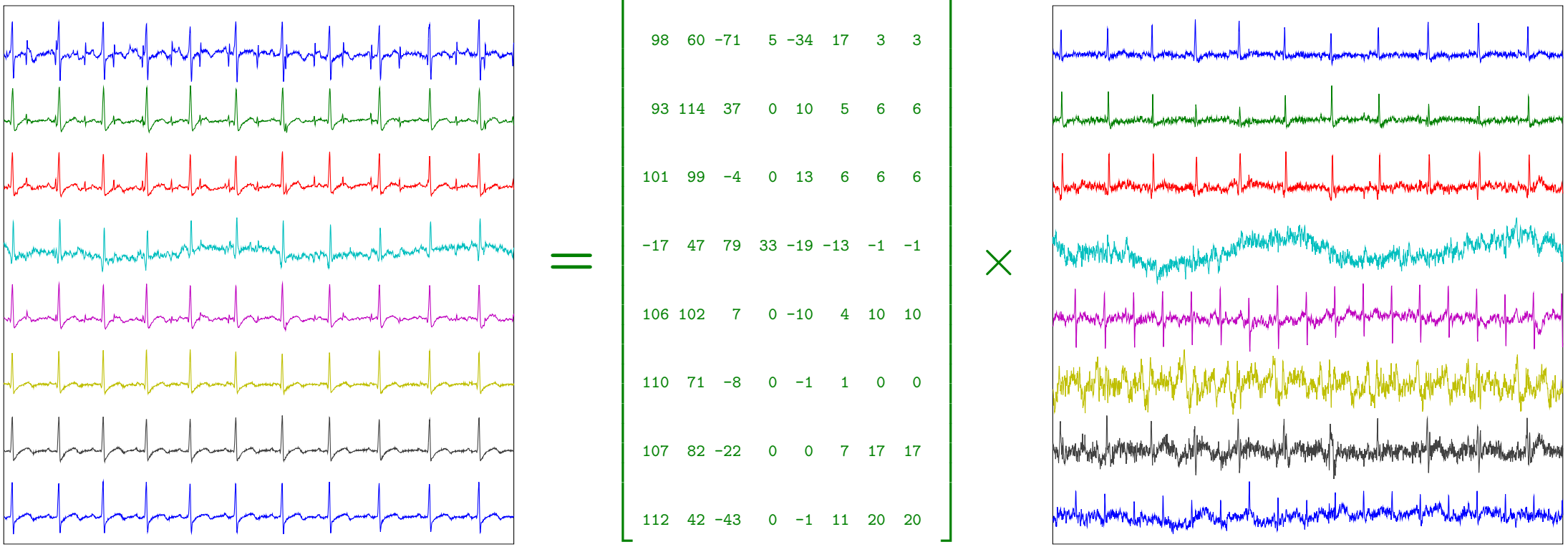


- Orthogonal mixture, uncorrelated components :

$$\frac{1}{T}\sum_t y_i(t)y_j(t) = 0 \text{ for } i \neq j$$

- Decorrelation is weak (always posible), orthogonality is implausible.

Independent component analysis



- Linear decomposition in "the most independent sources".
- Blind: only independence is at work, but we must go beyond decorrelation.
- Independence is statistically very strong but often physically plausible.
- Weak assumptions $\longrightarrow$ wide applicability
- How to do it ? Principles ? Efficiency ? Robustness ?

How to do it ? Add constraints ?

For all pairs $i \neq j$, replace the $n(n-1)/2$ sample decorrelation conditions

$$\widehat{\mathbf{Cov}}(y_i, y_j) = 0$$

by something more constraining/less symmetric like the $n(n-1)$ conditions

$$\widehat{\mathbf{Cov}}(\psi_i(y_i), \phi_j(y_j)) = 0$$

for well chosen (?) non-linear functions $\psi_i()$, $\phi_i()$.

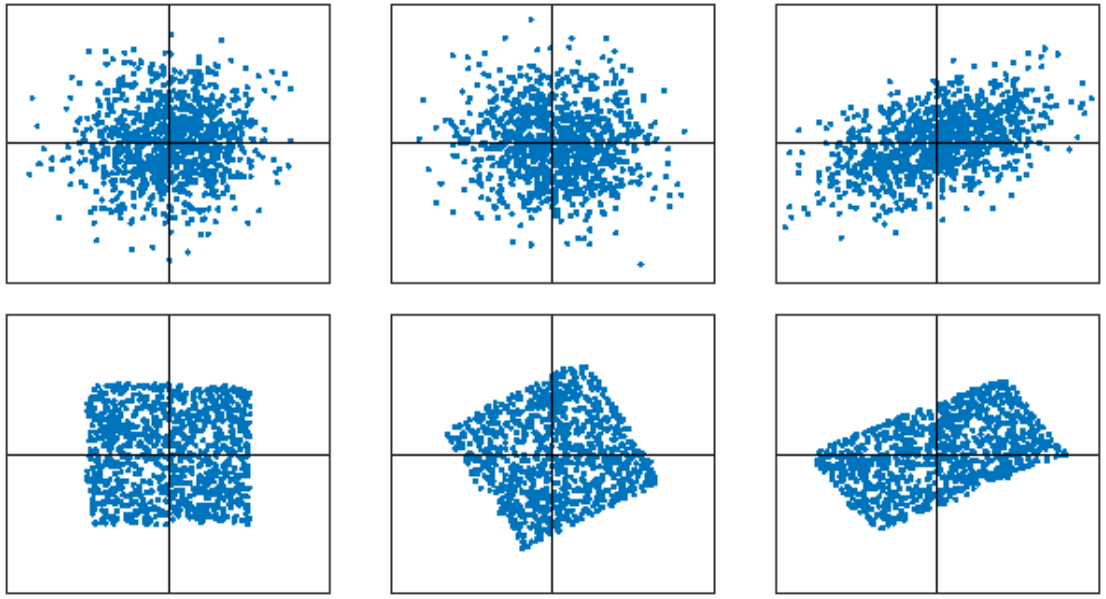
That would mimick the property that, for two independent variables, U and V, and any two functions f and g, one has

$$\mathbf{Cov}(f(U), g(V)) = 0$$

Jutten & Hérault (1991) designed the first adaptive component separation algorithm based on that property and 'some' non linear functions.

Gaussianity makes you blind (or one-eyed)

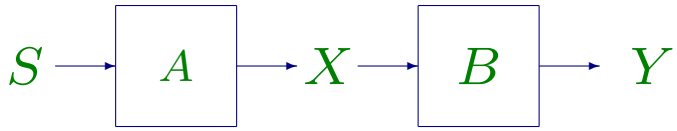
Mixing two Gaussian (top row) or uniform(bottom) random variables with the identity (left), a rotation (center), a generic transform (right):



Rotation is invisible unless the model **and** the data are non Gaussian.

Mixing and unmixing independent non Gaussian variables

We observe $X = AS$, a mixture of independent variables in vector S



- **Q**: If $Y = BX$ has independent entries, do we have $BA = I$ and $Y = S$?

A: Yes, essentially, if at most one Gaussian entry in S [Darmois '53, Comon '92].

That is: Mixing induces dependence; restoring independence unmixes.

- **Q**: If mixing gaussianizes, does degaussianizing unmix ?

A: yes, locally: If B is a separating matrix then (without any CLT)

$$G(Y_i + \alpha Y_j) = G(Y_i) - \frac{1}{2}\,\frac{\alpha^2\sigma_j^2}{\sigma_i^2}\;\cdot\;\kappa_i\;+\;o\left(\frac{\sigma_j^2}{\sigma_i^2}\right)$$

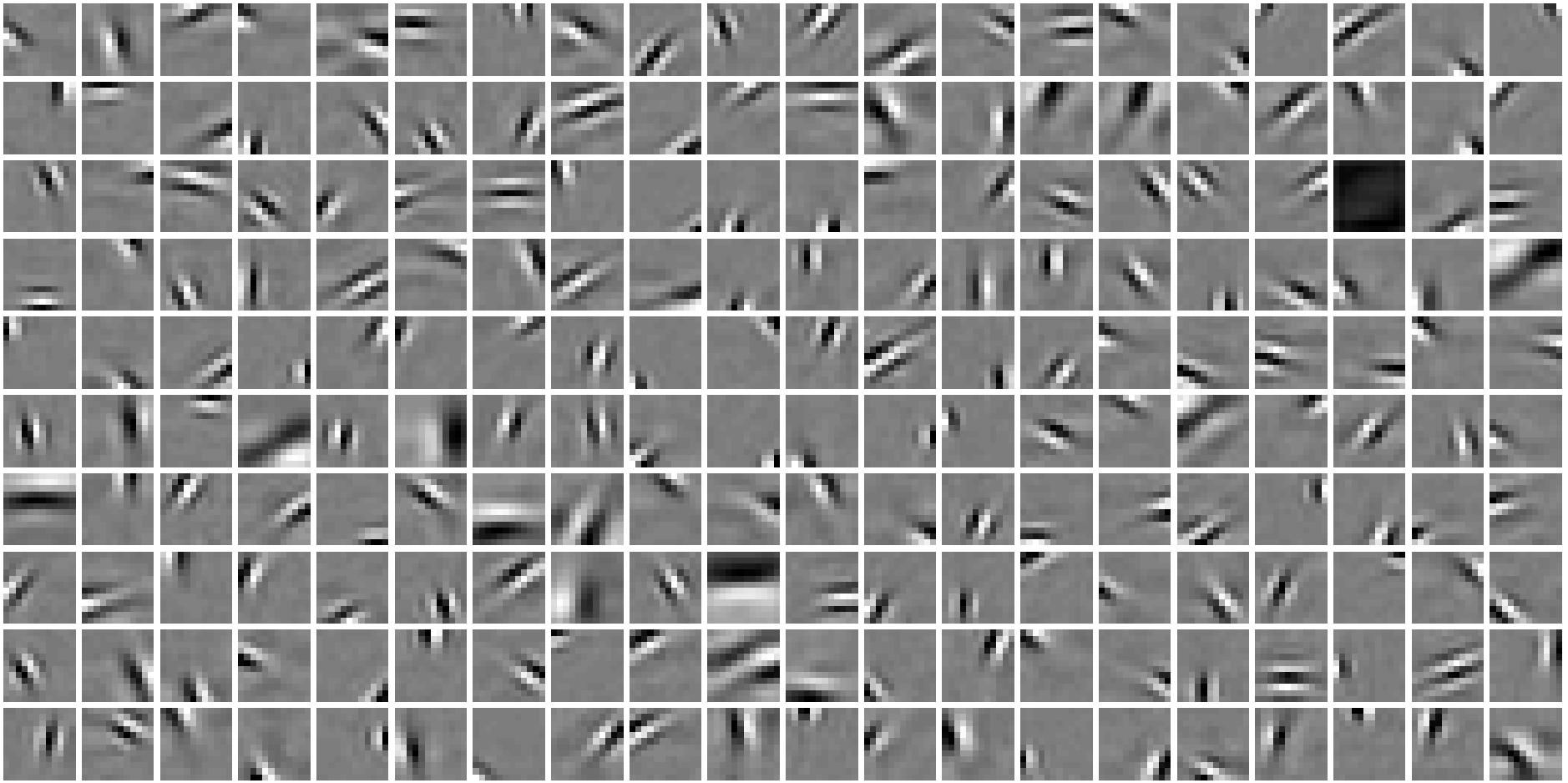
for any 'good' (*cf infra*) measure $G()$ of non-Gaussianity.

The scalar κ_i is positive, and strictly so for non Gaussian Y_i.

.

Independent (?) components of natural scenes

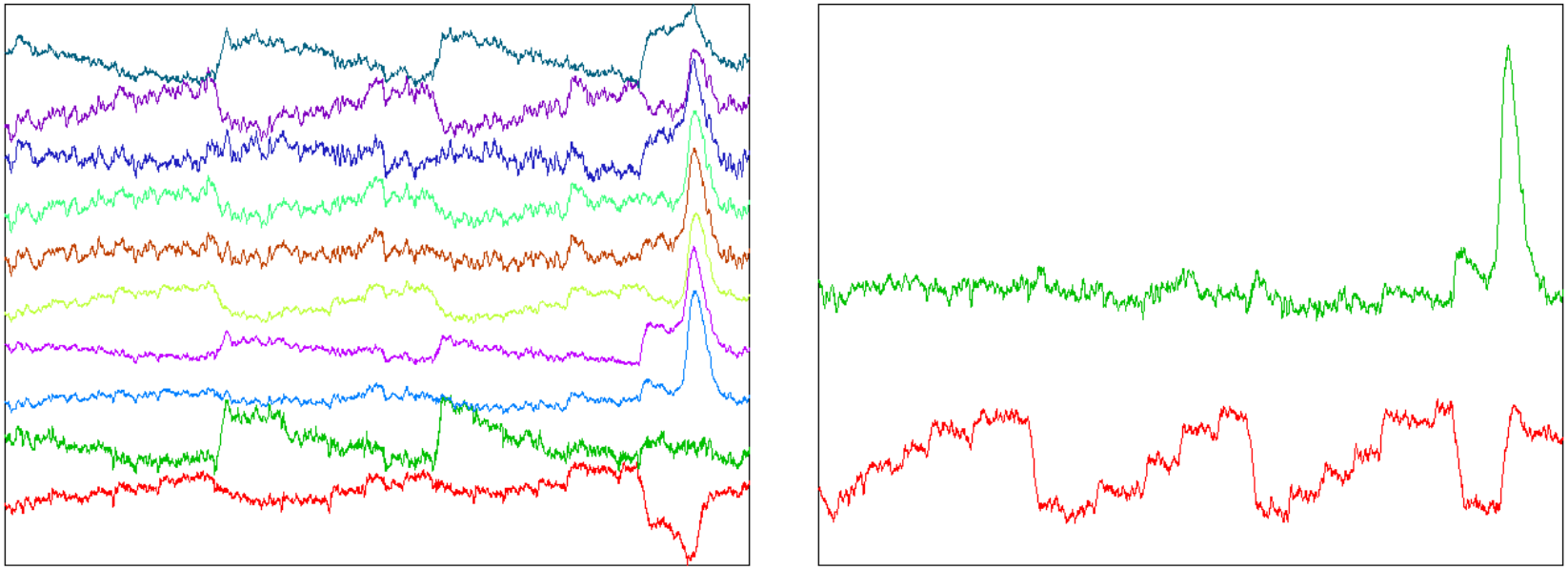
16×16 patches from natural images represented as linear combinations of "atoms" from a visual dictionary learned from data.



Looking for sparse (B. Olshausen) or
for independent (Bell & Sejnowski) components in 'natural images'.

Sparsity is one style of non Gaussianity

Electro-oculogram (Data courtesy I. Gorodnitsky, UCSD.)



Left: 10 electrodes located around the eyes: saccades and blinks.

Right: the linear combinations with extremal kurtosis:

$$\text{kurtosis}(Y) = \mathbf{E}(Y - \mathbf{E}Y)^4 / \mathbf{E}^2(Y - \mathbf{E}Y)^2 - 3 \qquad Y = \sum_{i=1}^{10} w_i X_i$$

Green: maximal kurtosis $k = 24.0$; Red: minimal kurtosis $k = -1.43$.

.

How to do it, practically ?

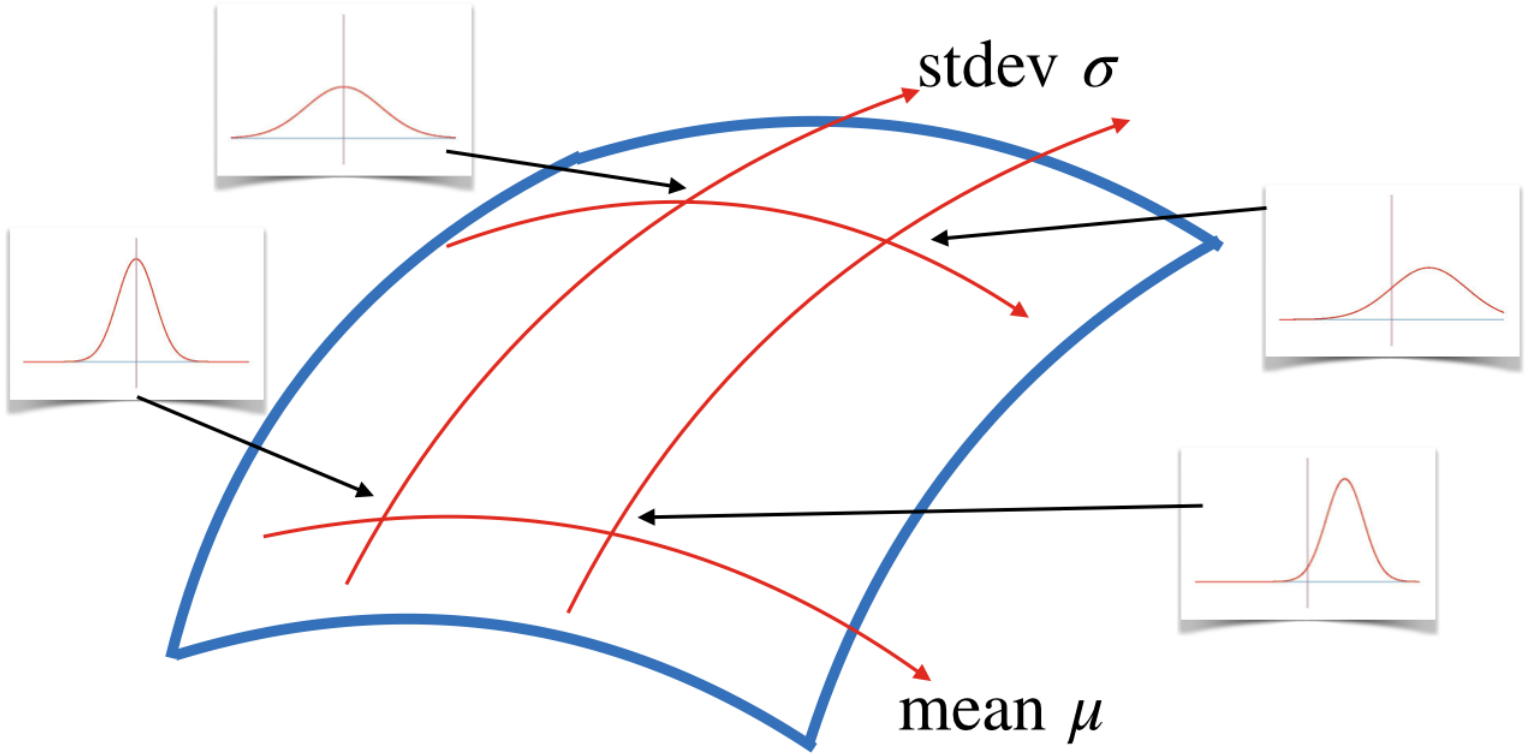
- Ignore time/space structure but use non Gaussianity . . .
  - Minimize dependence between recovered sources . . .
  - . . . as measured by mutual information, *why ? hard !*
  - . . . or approximated using high order cumulants, *clumsy ?*
  - or find the most non Gaussian sources, *why ? how ?*
  - or find sources uncorrelated through non linear functions: *e.g* $\mathbf{Cov}\Big(\phi_i(y_i), \psi_j(y_j)\Big) = 0$ for $i \neq j$. Which functions ϕ_i, ψ_j ?
- . . . or assume Gaussianity but use temporal/spatial structure, such as
  - . . . correlations, or spectral diversity,
  - . . . or non stationarity / inhomogeneities

But how to make sense of all that ? How to do it properly ? *Likelihood!*

Information geometry (a tiny bit, really)

A familiar statistical manifold

Univariate Gaussian with free mean and standard deviation.



.

Kullback-Leibler divergence

- Kullback-Leibler divergence from a distribution P to a distribution Q:

$$K\left[P \,|\, Q\right] = \mathbf{E}_P \log \frac{P(x)}{Q(x)} = \int P(x) \log \frac{P(x)}{Q(x)} dx$$

- It is strictly positive unless P and Q agree P-almost everywhere.

- It is *not* symmetric: in general, $K\left[P \,|\, Q\right] \neq K\left[Q \,|\, P\right]$

- It is invariant: under any invertible transform $f(\cdot)$,

$$K\left[P_X \,|\, P_Y\right] = K\left[P_{f(X)} \,|\, P_{f(Y)}\right]$$

.

Why Kullback ?

A parametric model $p(x;\theta)$ for the distribution of a random variable $X \sim p_*(x)$.

The likelihood loss based on n independent samples:

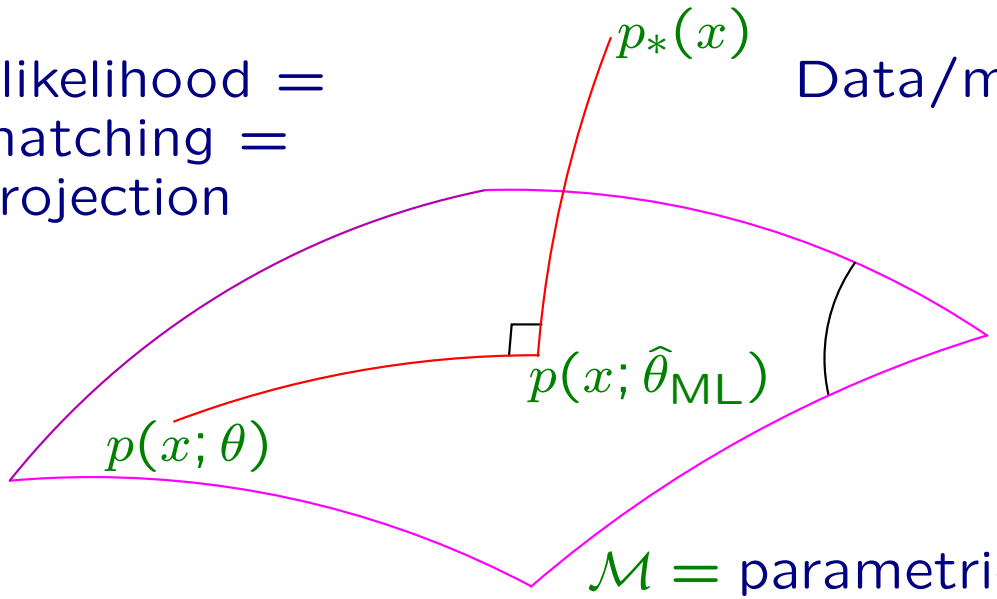
$$\widehat{\mathcal{L}}(\theta) = \frac{1}{n}\sum_i -\log p(x_i;\theta)$$

can be seen as an estimate of the expected loss

$$\mathbf{E}\,\widehat{\mathcal{L}}(\theta) = -\int p_*(x)\log p(x;\theta) = K\left[p_*(x)\,|\,p(x;\theta)\right] + H(p_*)$$

with $H(p_*)$ independent of the model.

Maximum likelihood =
Kullback matching =
Kullback projection



Data/model asymmetry

$\mathcal{M}$ = parametric model

Exponential families of distributions

The exponential segment between 2 distributions with densities $p(x)$ and $q(x)$:

$$\log r(x;\theta) = (1-\theta)\log p(x) + \theta \log q(x) - Z(\theta),\ \ \theta \in [0,\, 1]$$

or

$$r(x;\theta) = p(x)\frac{e^{\theta s(x)}}{Z(\theta)} \qquad s(x) = \log \frac{q(x)}{p(x)}$$

An *exponential family* contains the exponential segment between any two of its members.

Many statistical models are exponential.

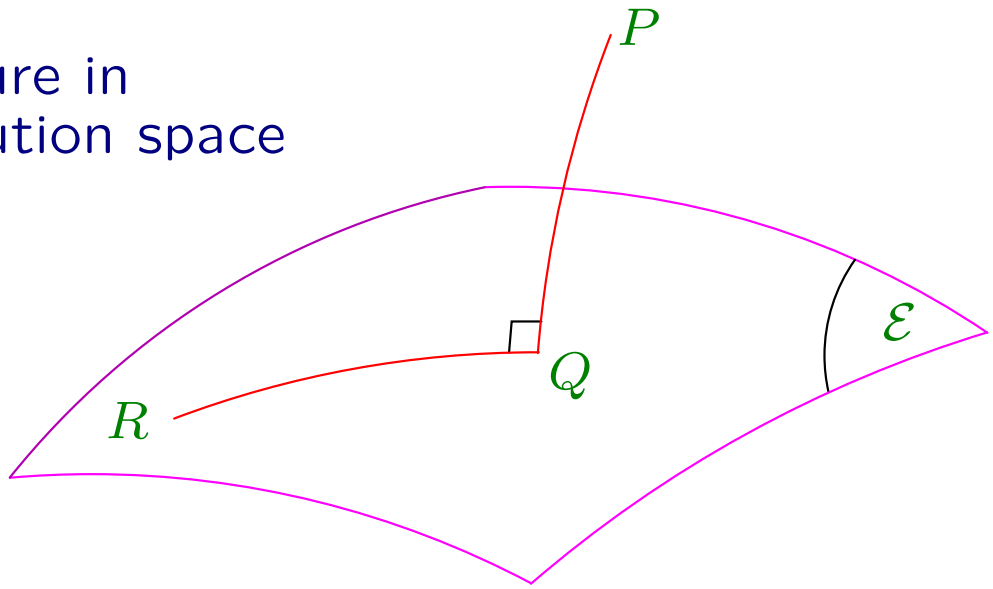
They enjoy very nice properties, such as sufficient statistics, convexity, ...

Stupendous geometrical behavior.

The Pythagorean theorem of information geometry

The Kullback divergence may not be a distance,
it still has its own private Pythagoras theorem.

A picture in
distribution space



- $\mathcal{E}$: an *exponential* family of distributions.
- Q: the *unique* minimizer in $\mathcal{E}$ of $K[P \,|\, \cdot]$. It is a projection of P onto $\mathcal{E}$.
- Then, for any other distribution R of $\mathcal{E}$

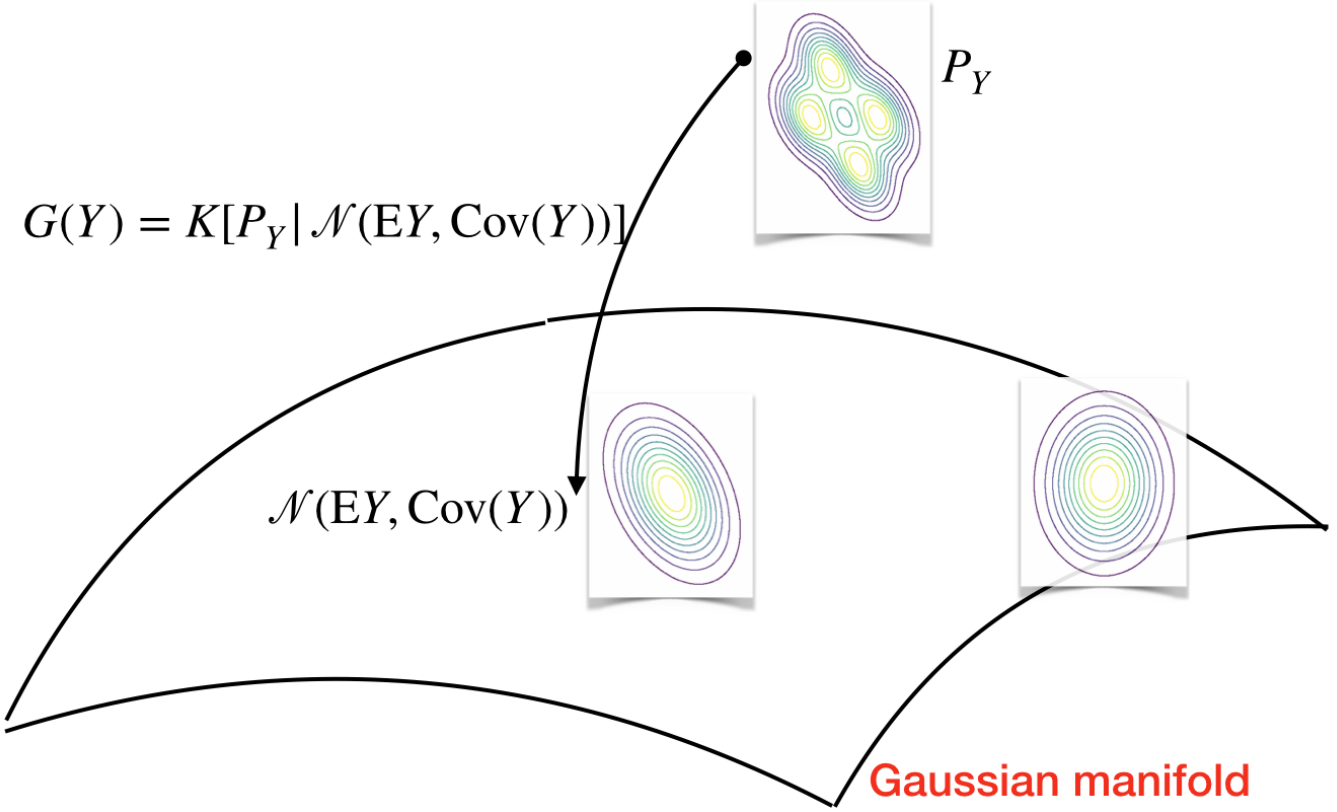
$$K[P \,|\, R] = K[P \,|\, Q] + K[Q \,|\, R]$$

The Kullback divergence behaves here as a squared Euclidean distance.

Exponential families are, somehow, (Kullback)-flat.

A special case: Gaussianity and lack thereof

The Gaussian distributions of an n-vector
with arbitrary mean and covariance matrix form an *exponential family*.



The Kullback projection of P_Y onto it simply is $\mathcal{N}(\mathbf{E}Y, \mathrm{Cov}Y)$.

Non Gaussianity $G(Y)$ of a random vector Y defined as
the Kullback divergence from P_Y to the Gaussian manifold.

.

Information geometry of ICA

The basic ICA model

- An $n \times T$ data set $X = \{x_i(t) \mid 1 \leq i \leq n,\ 1 \leq t \leq T\}$ modelled as $X = AS$ with an $n \times T$ source matrix of *independent* rows.

$$\boxed{X} = \boxed{A} \times \boxed{\begin{matrix} \cdots & S_1 & \cdots \\ & \vdots & \\ \cdots & S_n & \cdots \end{matrix}}$$

For now, we ignore any time/space structure (dependence within a row) and focus on the dependence between rows.

We saw that, assuming independence of the components,
non Gaussianity is enough for an (essentially) unique blind recovery.

In need of statistical guidance, we turn to the likelihood for $X = AS$, $S \sim P_S$:

$$P(X|A) = \frac{1}{|\det A|^T}\ P_S(A^{-1}X) \qquad P_S = \prod_i P_{S_i}$$

ICA likelihood and Kullback matching

The distribution of $X = AS$ is specified by matrix A and $P_S = \prod_i P_{S_i}$:

$$-\mathbf{E} \log P(X|A) \stackrel{c}{=} K\left[P_X \,|\, P_{AS}\right] = K\left[P_{A^{-1}X} \,|\, P_S\right] = K\left[P_Y \,|\, P_S\right]$$

Hence, the likeliest A is the one making P_Y the closest to $P_S = \prod_i P_{S_i}$.

$$P_X \longrightarrow \boxed{A^{-1}} \longrightarrow P_Y \stackrel{K}{\approx} P_S = \prod_i P_{S_i}$$

The ICA parameters:

1. The parameter of interest, the mixing matrix A, lives in the multiplicative group $\mathrm{GL}(n)$.

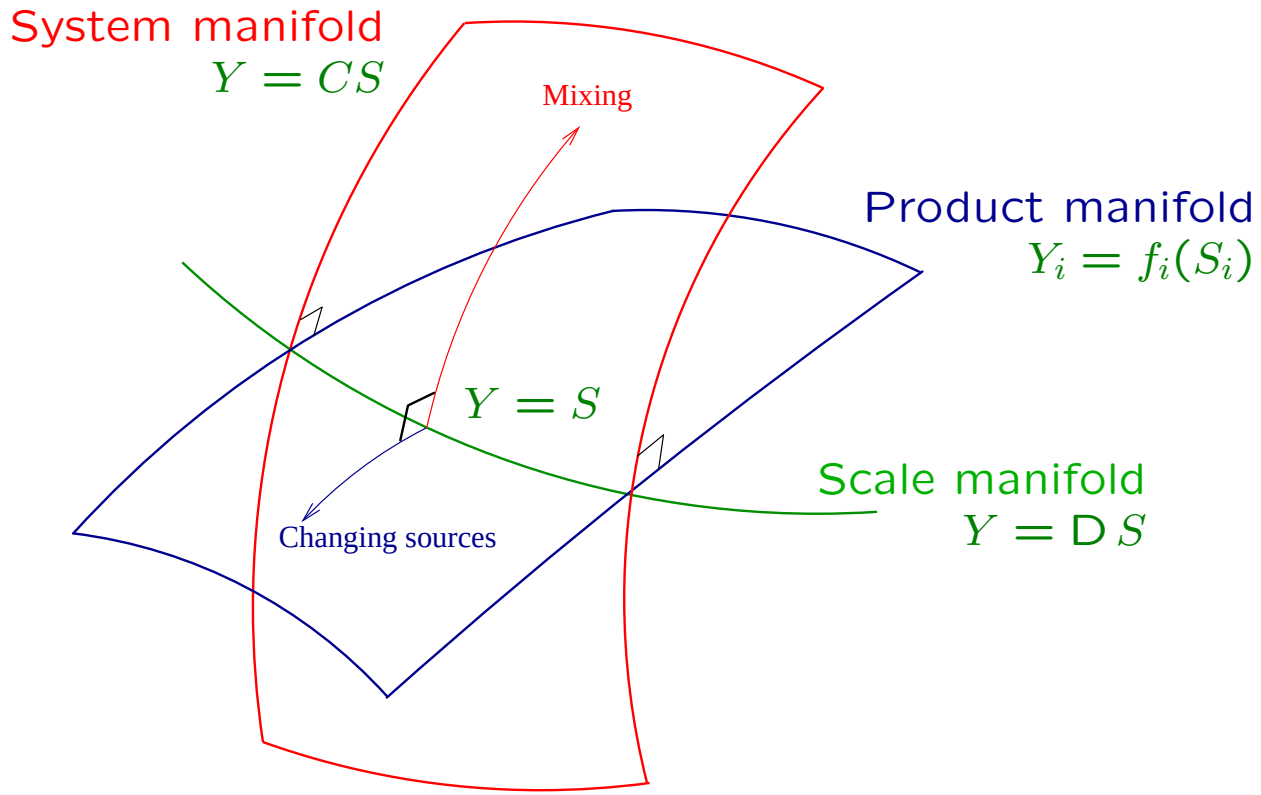
   Theory of **equivariant** estimation in a group (big impact on performance and algorithms, if you pay attention).

   Data X and mixing A enter the likelihood only through $Y = A^{-1}X$.

2. Nuisance parameters: the n source distributions. Hard to estimate, but do we really need it? Not too much of a problem, thanks to orthogonality.

.

Some statistical manifolds of ICA



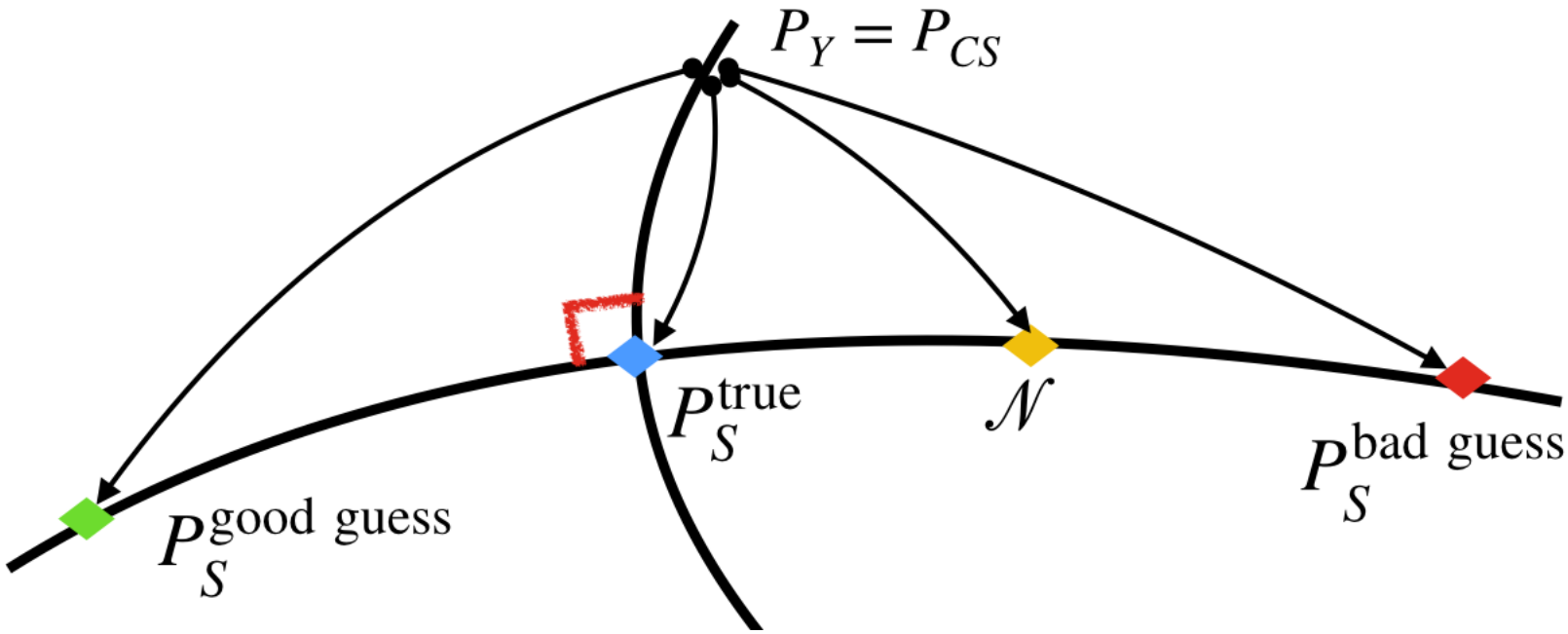
- $Y = S$: the distribution of a vector S with independent entries. Move from there.
- The *system manifold*: distributions of $Y = CS = A^{-1}A_{\text{true}}S$ for all invertible matrices A.
- The *product manifold*: distort the marginals of S, retaining independence.
- The *scale manifold*: change only the scales of the entries of S.

$\rightarrow$ Blind identifiabilty = essential uniqueness of ICA =
system manifold intersects product manifold along scale manifold.

.

Nice orthogonal nuisances

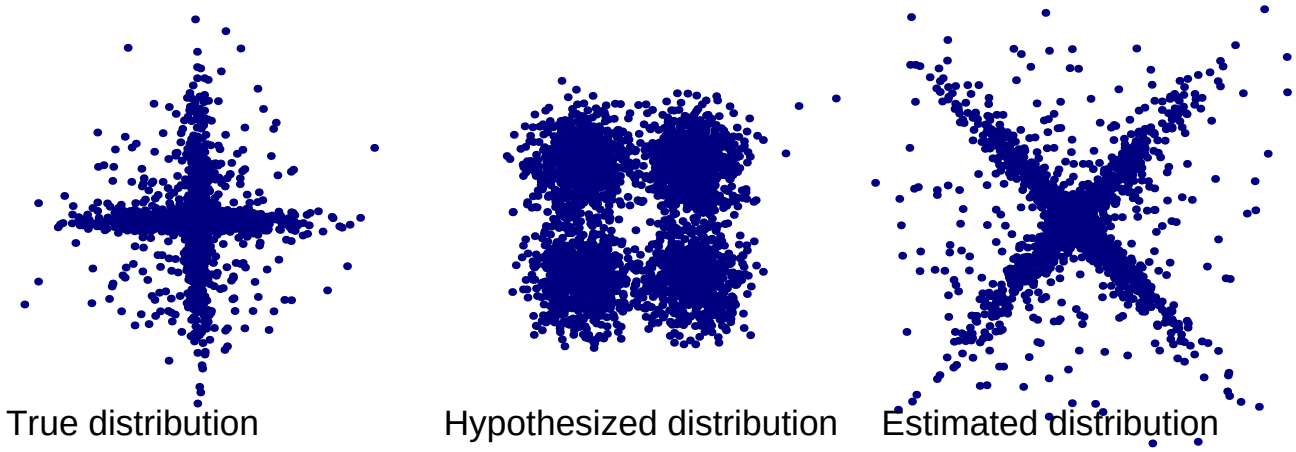
Our likelihood wants to solve $\min_A K[P_Y \,|\, P_S]$ for $Y = A^{-1}X$
but can we do it without knowing the right target? Bias due to $P_S \neq P_S^{\text{true}}$?



For a small transform $Y = e^{\mathcal{E}}S$, $K[P_Y \,|\, P_S] = L(\mathcal{E}) + \frac{1}{2}Q(\mathcal{E}) + o(\|\mathcal{E}\|^2)$ with
a linear term $L(\mathcal{E}) = 0$ and
a quadratic form $Q(\mathcal{E})$, postive if P_S "on the right side of the Gaussian".

Non Gaussianity, sparsity and bad luck

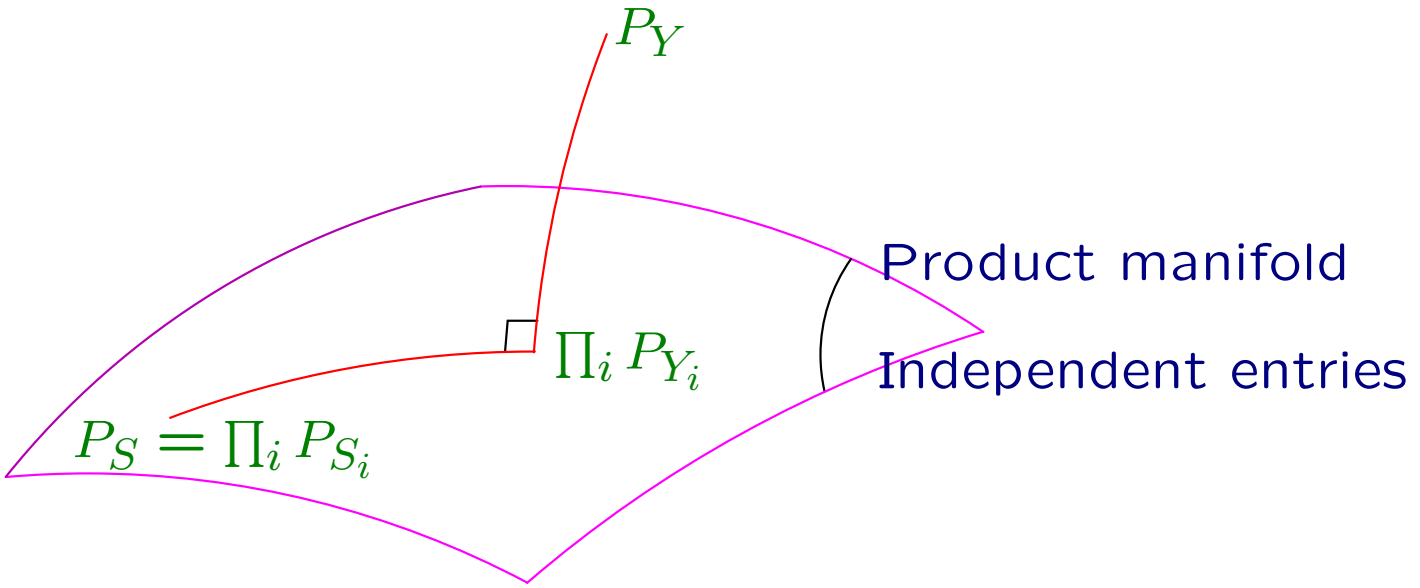
An example of bad guess.



The result is not only wrong; it's maximally wrong!

Likelihood and mutual information

- Maximizing the ICA likelihood amounts to minimizing $K\left[P_Y \,|\, P_S\right]$.
- The target source distribution P_S is usually unknown except for $P_S = \prod_i P_{S_i}$.



$$K\left[P_Y \,|\, P_S\right] = K\left[P_Y \,|\, \prod_i P_{Y_i}\right] + K\left[\prod_i P_{Y_i} \,|\, \prod_i P_{S_i}\right] = I(Y) + \sum_i K\left[P_{Y_i} \,|\, P_{S_i}\right]$$
$$= \text{dependence} + \text{sum of marginal mismatches}$$

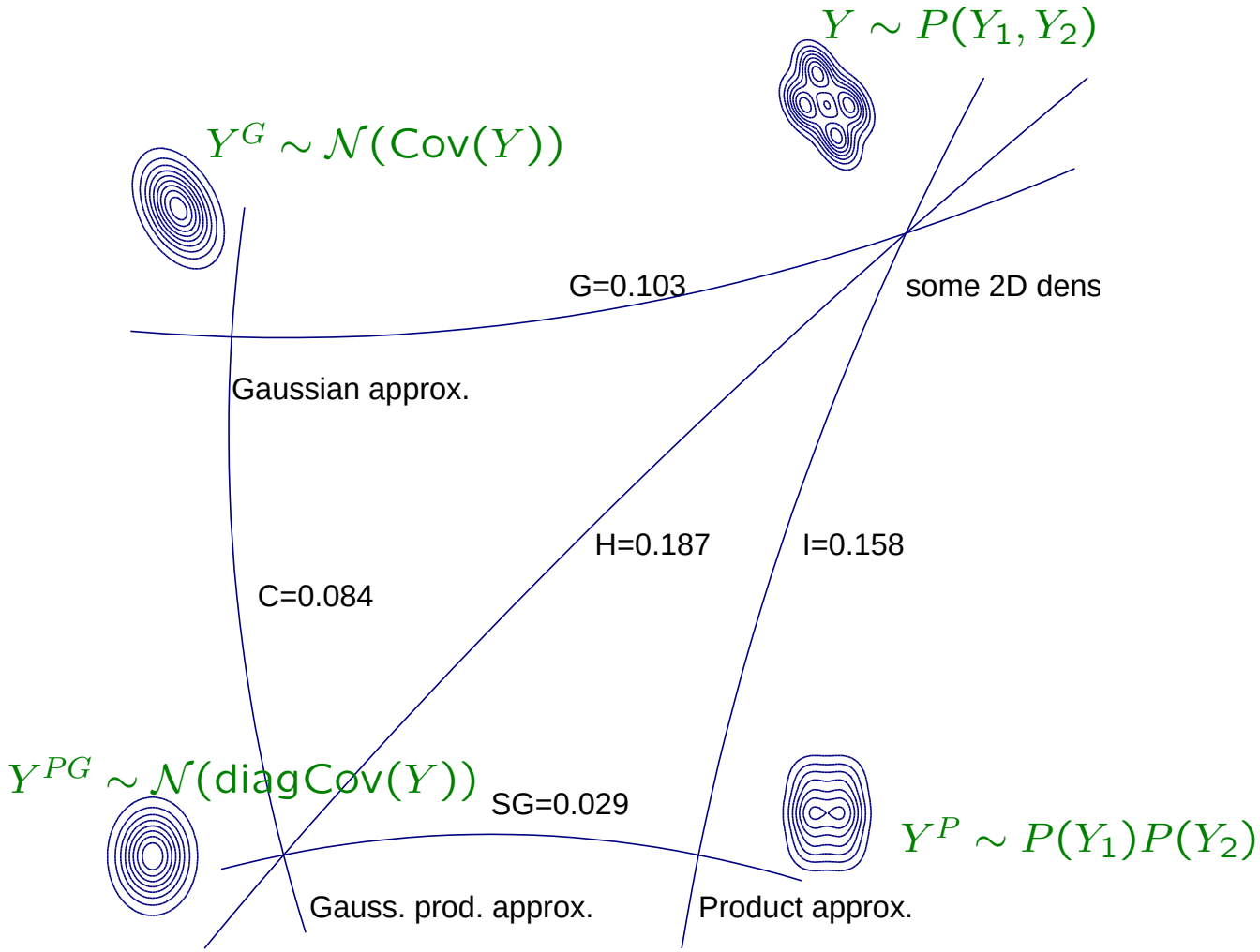
The likelihood says to measure dependence by the mutual information:

$$I(Y) \stackrel{\text{def}}{=} K\left[P_Y \,|\, \prod_i P_{Y_i}\right]$$

Major problem: mutual information is **very** difficult to estimate.

.

Geometry of non Gaussian dependence

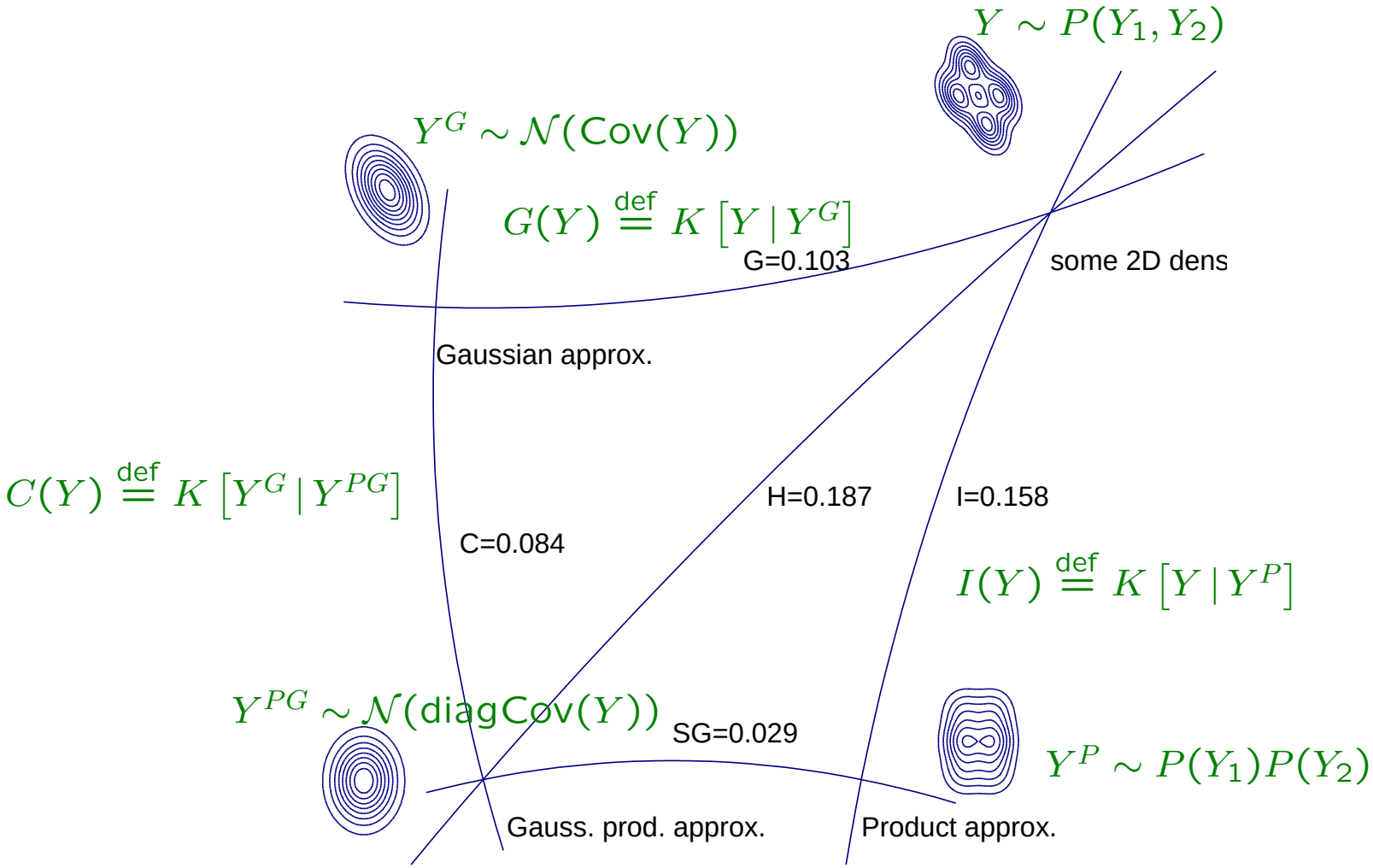


Two familiar statistical approximations can be seen as projections:

Oh, let's assume it's Gaussian...

Ho, let's assume they're independent...

Geometry of non Gaussian dependence



Orthogonal projections $\rightarrow$ two right triangles $\rightarrow$ two Pythagorean theorems

Dependence and non Gaussianity

- *Non Gaussianity*. Define the non Gaussianity $G(Y)$ of Y as

$$G(Y) = K\left[P_Y \,|\, \mathcal{N}(R_Y)\right]$$

i.e. how much the best Gaussian approx. fails to mimic the distrib. of Y.

- The *correlation* $C(Y)$ of Y

$$C(Y) = K\left[\mathcal{N}(R_Y) \,|\, \mathcal{N}(\mathrm{diag} R_Y)\right] = \frac{1}{2} \log \frac{\det \mathrm{diag} R_Y}{\det R_Y}$$

i.e. how much the covariance matrix R_Y of Y fails to be diagonal.

- All these are (geometrically) connected by

$$I(Y) + \sum_i G(Y_i) = C(Y) + G(Y)$$

- Under *linear* transforms, $G(Y)$ is constant. The mutual information then is

$$I(Y) = C(Y) - \sum_i G(Y_i) + \mathrm{cst}$$

.

Dependence and non Gaussianity (cont.)

Repeat: Under linear transforms, making the entries of Y as independent as possible ($\min I(Y)$) is *identical* to making (as much as possible) Y uncorrelated and *each of its entries* non Gaussian.

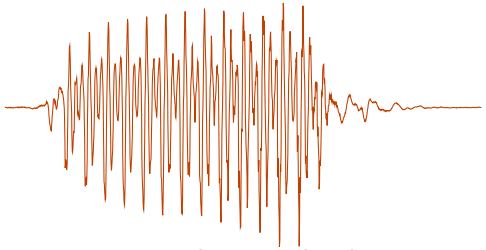
The relation $I(Y) = C(Y) - \sum_i G(Y_i) + G(Y)$ also reads

$$\text{Complicated} = \text{Simple - Simples} + \text{Complicated constant}$$

Note. Some ICA methods *enforce* decorrelation $C(Y) = 0$.
Not unreasonable, but this is not what the likelihood tells us to do.
One is left with a pure degaussianization objective.

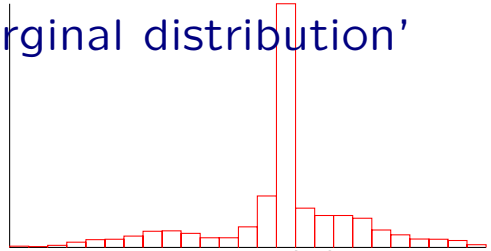
Other geometries: three points of view on a time series.

A random (?!) sequence

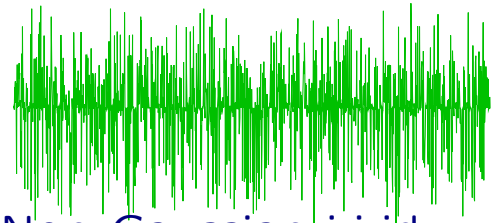


Marginal probability density

Its 'marginal distribution'



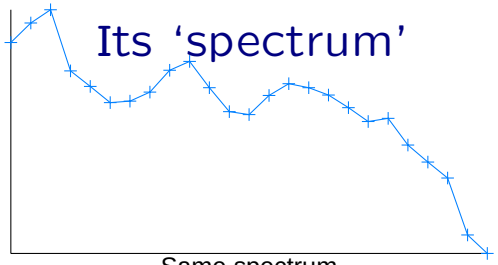
Same marginal



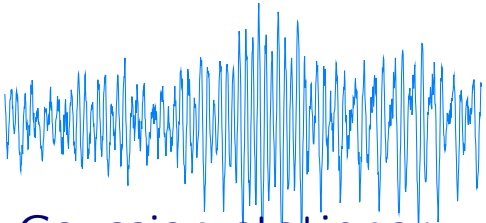
Non Gaussian i.i.id

Spectral energy density

Its 'spectrum'



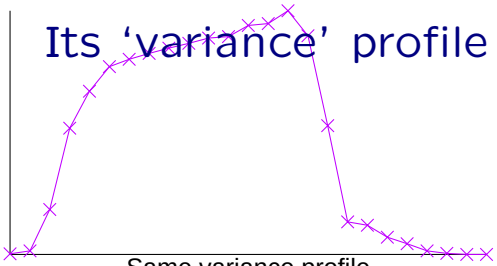
Same spectrum



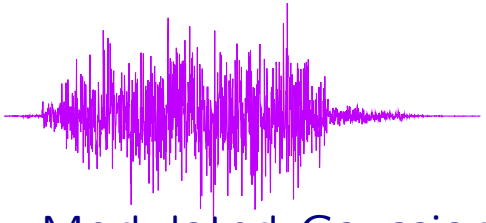
Gaussian stationary

Temporal energy density

Its 'variance' profile



Same variance profile



Modulated Gaussian i.i.d.

Project data onto simple Gaussian models for time series,
such as stationary but colored, or white but non stationary, where...
mutual information is a joint diagonality criterion of covariance matrices.

.

Conclusion

Three applications of the Pythagorean theorem do not cover it all.

Not covered:

1. Equivariance, Lie group, relative gradient,

2. Asymptotic analysis, stability, Fisher efficiency,

3. Extended models:
   noise, convolution, under/over determination, source adaptivity,
   large scale problems, high accurarcy.

4. Algorithms