.

(Outrageously) simple statistical models for blind source separation (BSS)

or

The unreasonable effectiveness of statistical independence

J.-F. Cardoso, CNRS/ENST

`http://tsi.enst.fr/~cardoso`

Int. Workshop on Acoustic Echo and Noise Control. IWAENC 2006. Paris, 12/07/06

Blind separation of audio sources

AUDIO BSS = AUDIO + BSS

- AUDIO is complicated. You know why!

– Long impulse responses
– Time-varying contexts
– Real time constraints
– <Your favorite audio annoyance>

- B.S.S. is complicated.

– it's blind!
– Maybe we don't want to be so blind...
– Statistical independence is easy to understand, hard to express
– <Your favorite statistical annoyance>

→ This talk: focus on the statistical concepts for BSS.
(yes, it means considering non convolutive mixtures).

Outline

- The likelihood approach

- Likelihood and independence

- Outrageously simple models for random processes:

– non Gaussian i.i.d. models

– Gaussian stationary models

– Gaussian non stationary models

- Connections between statistical independence and sparsity.

Who wants to be blind?

Simplest BSS model: $X = AS$ with A unknown and independent rows in S.

We can build a probabilistic model $p(X|A)$ and fit it blindly in the maximum likelihood sense:

$$\widehat{A}_{\mathsf{ML}} = \arg\max_A \log p(X|A)$$

But once we have a likelihood, we can as well incorporate knowledge about the mixing system by defining a prior $p(A)$.

Now, by the standard Bayesian argument the MAP estimate of A is

$$\widehat{A}_{\mathsf{MAP}} = \arg\max_A [\log p(A|X)] = \arg\max_A [\log p(X|A) + \log p(A)]$$

The first term $\log p(X|A)$ is a measure of fit between data and model (see next) while the second term $\log p(A)$ encodes prior knowledge about the system.

There is a continuum between fully blind and fully non blind. The extreme case is hard constraints (or parameterization) of the system: $A = A(\theta)$.

→ In any case, we need a ‘good’ (simple, yet expressive) likelihood $p(X|A)$.
This talk: all about $\log p(X|A)$.

Statistical models for static mixtures

- $X = \{x_i(t) \mid 1 \leq i \leq n,\ 1 \leq t \leq T\}$, an $n \times T$ matrix data set.

- The static ICA model is $X = AS$ with independent rows in S:

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

- We consider fitting this model in the maximum likelihood sense.

- To build a likelihood for A, all that is needed under the independence assumption are the pdf's $P_{S_1}, \ldots, P_{S_n}$ for each row of S.

So we want to look very hard into

$$P(X) = P(X|A, P_S) = P(X|A, P_{S_1}, \ldots, P_{S_n})$$

The easy part is

$$P(X) = \frac{1}{|A|^T} P_S(A^{-1}X) = \frac{1}{|A|^T} \prod_{i=1}^{n} P_{S_i}([A^{-1}X]_i)$$

Likelihood and contrast functions

The shape of the likelihood. In a transformation model $X = AS$ with parameters (A, P_S), the likelihood contrast $-E \log p(X|A, P_S)$ is a distribution mismatch:

$$-E \log p(X|A, P_S) = K\left[P_Y \,|\, P_S\right] + H(X) \qquad\qquad Y = A^{-1}X$$

where $K\left[f \,|\, g\right] = \int f \log f/g$ is the Kullback divergence between probability distributions and H is Shannon differential entropy.

$$\boxed{Y = \widehat{S}} \quad = \quad \boxed{A^{-1}} \quad \times \quad \boxed{X}$$

Conclusion: in our model, the (log)-likelihood is a measure of the fit between the data and the model in the sense that maximizing the likelihood of A is (on average) identical to minimizing the divergence between the observed distribution P_Y of $Y = A^{-1}X$ and the hypothetical distribution P_S of the sources.

It all boils down to studying contrast functions $K\left[P_Y \,|\, P_S\right]$.

Maximum likelihood and minimum dependence

Geometry (repeat). In a transformation model $X = AS$ with parameters (A, P_S), the likelihood contrast $-E \log p(X|A, P_S)$ is a distribution mismatch:

$$-E \log p(X|A, P_S) = K\left[P_Y \,|\, P_S\right] + \text{cst} \qquad Y = A^{-1}X$$

where $K[f \,|\, g] = \int f \log f/g$ is the Kullback divergence.

Statistics. The rows of S are independent: $P_S = \prod_i P_{S_i}$.

$$\begin{aligned} K\left[P_Y \,|\, P_S\right] &= K\left[P_Y \,\middle|\, \textstyle\prod_i P_{Y_i}\right] + \textstyle\sum_i K\left[P_{Y_i} \,\middle|\, P_{S_i}\right] \\ &= \text{dependence} + \text{marginal mismatch} \end{aligned}$$

Optimizing over nuisance parameters P_{S_i} kills each $K\left[P_{Y_i} \,\middle|\, P_{S_i}\right]$ and leaves us with (in)dependence:

$$I(Y) \stackrel{\text{def}}{=} K\left[P_Y \,\middle|\, \textstyle\prod_i P_{Y_i}\right] \qquad \text{a.k.a. } \textit{mutual information}$$

→ Maximum likelihood leads to maximum independence. . .

→ . . . and provides a definition for it.

→ Next: Focus on mutual information or ‘dependence’ $I(Y)$.

The Holy Grail of ICA

The ultimate goal of ICA is just that:
finding components which are ‘as independent as possible’.

In a nutshell, for the static noise-free BSS model $X = AS$, we want to solve

$$\min_A I(Y) \qquad \text{where} \qquad Y \overset{\text{def}}{=} A^{-1}X \qquad \text{and} \qquad I(Y) \overset{\text{def}}{=} K\left[P_Y \mid \prod_i P_{Y_i}\right]$$

Problem
There is no way that mutual information can be estimated in full generality.

Instead, we try to build statistical models (families of probability distributions) in which $I(Y)$ can be estimated from an $n \times T$ data matrix Y.

Correlation

If an $n \times T$ data matrix Y is modelled as the realization of Gaussian, temporally white noise Y^G, its distribution is specified by a single $n \times 1$ mean vector EY and a single $n \times n$ covariance matrix R_Y.

In this model, dependence (or mutual information) boils down to ...

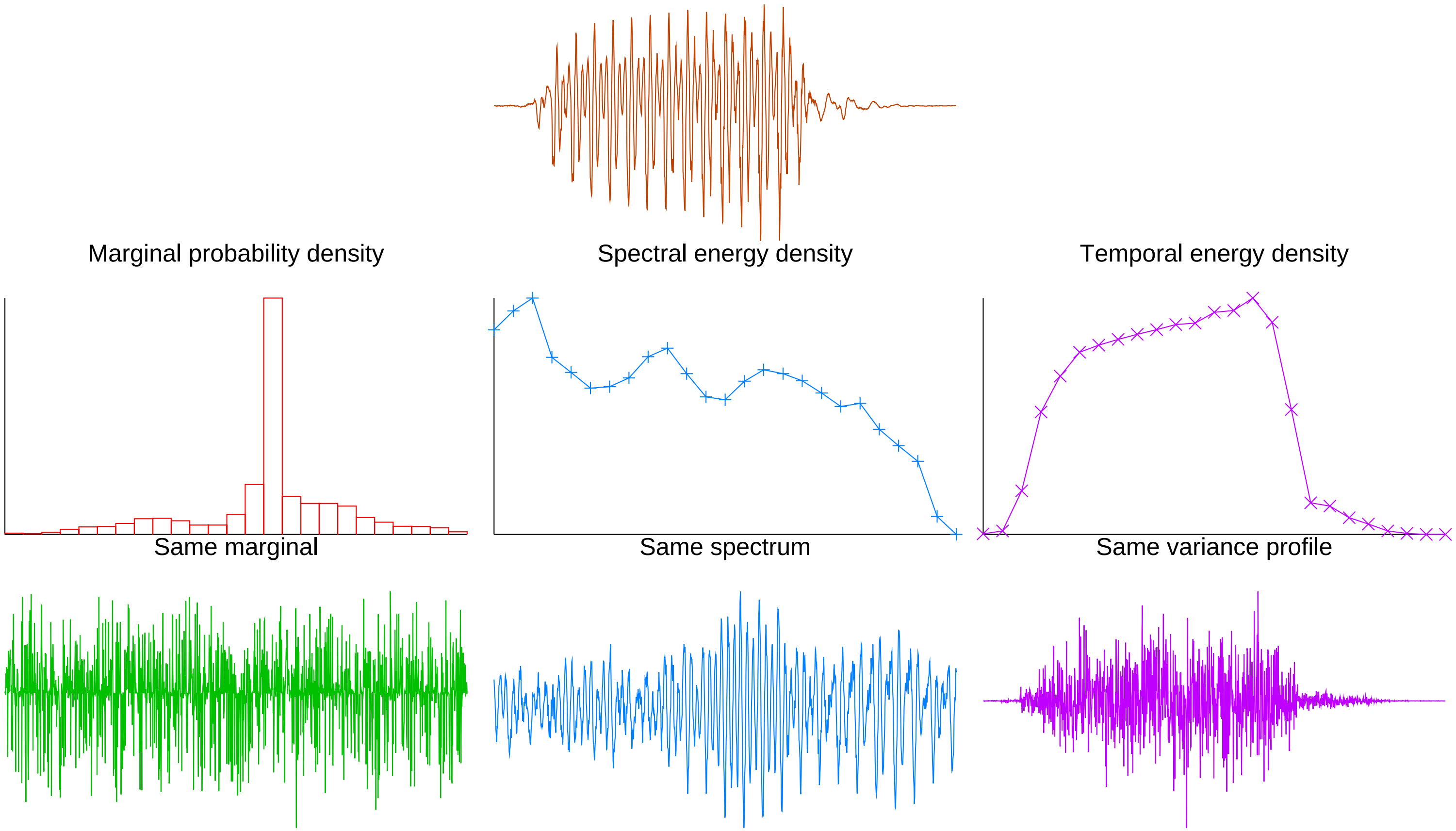
$$\begin{aligned} C(Y) &\overset{\text{def}}{=} I(Y^G) \\ &= K\left[P_Y^G \,|\, \textstyle\prod_i P_{Y_i^G}\right] \\ &= T \times K\left[\mathcal{N}(EY, R_Y) \,|\, \mathcal{N}(EY, \mathrm{diag} R_Y)\right] \\ &= T \times \frac{1}{2} \log \frac{\det \mathrm{diag} R_Y}{\det R_Y} \end{aligned}$$

... correlation which also measures the non-diagonality of R_Y

$$C(Y) = T \times \mathrm{off}(R_Y)$$

But global decorrelation is just too cheap for BSS.

Three points of view on a time series



Using about 20 parameters to capture one out of three possible aspects of a times series.

- *All models are wrong, but some are useful* —George Box
- *. . . especially for BSS.* —JFC.

The non Gaussian i.i.d. case

The (non Gaussian) i.i.d. case: Time structure is ignored.

$$
\begin{aligned}
P(Y) &= P(Y(1), Y(2), \ldots, Y(T)) \\
&= \prod_t P_t(Y(t)) \quad \text{Independently and \ldots} \\
&= \prod_t p(Y(t)) \quad \text{\ldots identically distributed.}
\end{aligned}
$$

where $p()$ is the n-dimensional pdf common (by assumption) to all $Y(t)$, $1 \leq t \leq T$.

Then, divergence between T-long time series distributions:

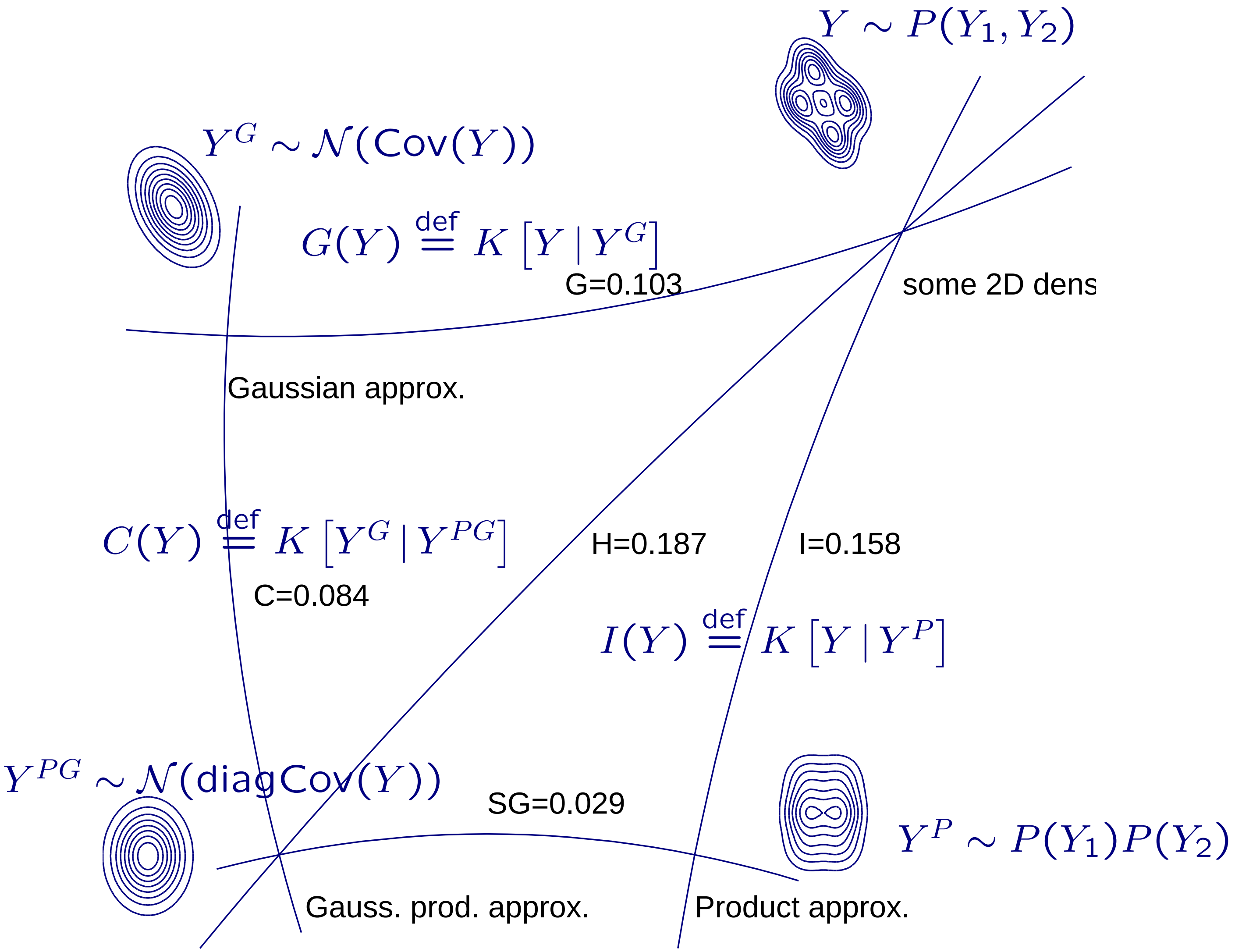
$$
K\left[\prod_t P_t \,|\, \prod_t Q_t\right] = \sum_{t=1}^{T} K\left[P_t \,|\, Q_t\right] = T \times K\left[p \,|\, q\right]
$$

depending only on the n-dimensional distributions p and q.

Yes, my notations are crappy.

Geometry of non Gaussian dependence

The Kullback divergence is not an Euclidean distance but in some cases there is still a Pythagorean theorem. This is happening right in front of your eyes.



$Y \sim P(Y_1, Y_2)$

$Y^G \sim \mathcal{N}(\mathrm{Cov}(Y))$

$G(Y) \overset{\text{def}}{=} K\left[Y \,|\, Y^G\right]$

G=0.103

some 2D dens

Gaussian approx.

$C(Y) \overset{\text{def}}{=} K\left[Y^G \,|\, Y^{PG}\right]$

H=0.187

I=0.158

C=0.084

$I(Y) \overset{\text{def}}{=} K\left[Y \,|\, Y^P\right]$

$Y^{PG} \sim \mathcal{N}(\mathrm{diagCov}(Y))$

SG=0.029

$Y^P \sim P(Y_1)P(Y_2)$

Gauss. prod. approx.

Product approx.

Orthogonal projections → two right triangles → 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.

- Remember the correlation $C(Y)$ of Y

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

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 so

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

→ Hence, 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.

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.

- Note 1: The relation $I(Y) = C(Y) - \sum_i G(Y_i) + G(Y)$ also reads

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

- Note 2: Some people/algos enforce $C(Y) = 0$ (pre-whitening) but even that leaves us with with the not-so-easy-after-all measure of marginal non-Gaussianity $G(Y_i)$. So,how do we do that?

- Note 3: Connection between non Gaussianity and sparsity:

See next slides.

Non Gaussianity and sparsity

Remember that we arrive at the non Gaussianity objective $G(Y_i)$ for each recovered source by assuming that the pdf of the sources are (perfectly) estimated from the data.

If one wants to avoid estimating (explicitly or implicitly) the pdf of the sources, one may try to select a fixed non Gaussian target distribution $P_{S_i} = Q_i$ for each component. The objective would be to minimize

$$K\left[P_Y \mid \textstyle\prod_i Q_i\right].$$

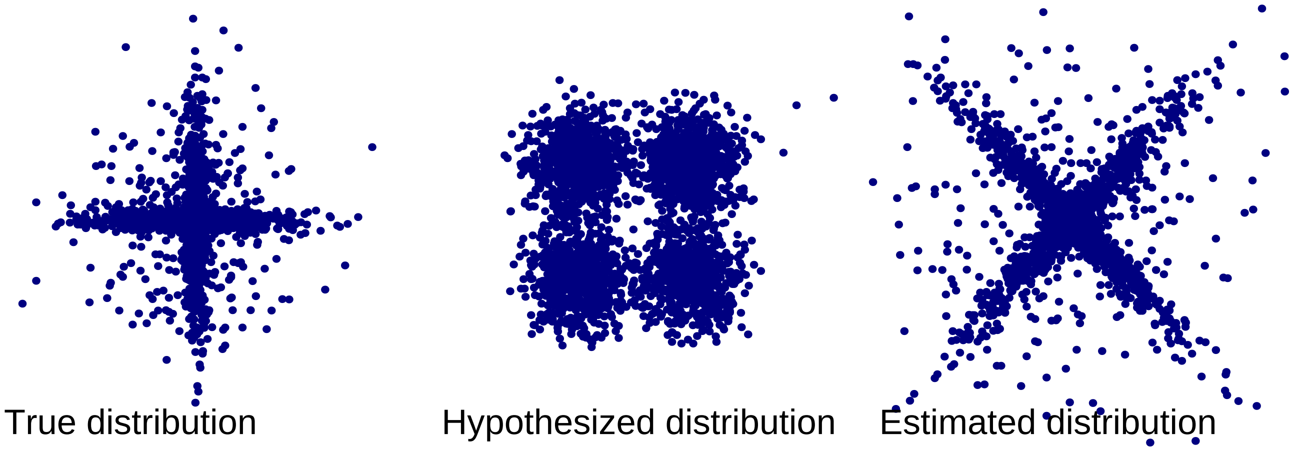
But this is exactly the log-likelihood contrast! So, if the pdf's of the sources are known, one should do just that, indeed.

If the pdf's are not known but are expected to be "sparse", then one could use some sparse guess Q_i in place of the true but unknown distribution. Example: a mixture of two Gaussians $Q_i(u) = (1-\alpha)\mathcal{N}(u;0,\sigma_1^2) + \alpha\mathcal{N}(u;0,\sigma_2^2)$ with zero mean and variance σ_1 "smaller" than σ_2.

This works (provably) provided the mismatch between P_{S_i} and Q_i is not too large. It does not work if a sparse model is used and the true sources are "anti-sparse" and vice-versa. In some technical sense, the truth P_{S_i} and the guess Q_i should be "on the same side of the Gaussian" so that $\min K\left[P_{Y_i} \,|\, Q_i\right]$ works in the same direction as $\max G(Y_i)$.

Bad luck

An example with two sources



Left: Joint histogram for two sparse sources. Each source is a mixture of two zero-mean Gaussians.

Middle: The hypothesized distributions Q_i also are mixtures of Gaussians but anti-sparse $Q_i(u) = \frac{1}{2}\mathcal{N}(u; 1, \sigma^2) + \frac{1}{2}\mathcal{N}(u; -1, \sigma^2)$

Right: The best (maximum likelihood) match $\min K\left[P_Y \mid \prod_i Q_i\right]$ is obtained when Y is a $\frac{\pi}{4}$-rotated version of the orginal sources.
It's not only wrong; it's maximally wrong!

Plan: Everything should be as simple as possible, but not simpler.

- Source separation is impossible if sources are modeled as Gaussian, identically and independently distributed.
- Two simple possibilities to move away from Gaussian i.i.d.:

  (a) Non Gaussian models without time (or space) structure.

  (b) Time structure within a Gaussian framework.

  We have just seen item (a). What about item (b)?

- Simple time structures: in some basis (sample basis, Fourier basis, wavelet basis, . . .), the coefficients of the process are uncorrelated with a smoothly varying variance.
- Questions:
  - Dependence in time-structured Gaussian models.
  - Connections between dependence, correlation, sparseness and other non properties.
  - Algorithms?

For example. . .

- The non Gaussian i.i.d. model.

The probability for an $n \times T$ data batch X in a non Gaussian i.i.d. model reads

$$P(X) = \prod_{t=1}^{T} q(X(t))$$

Hence, it is completely specified by a p.d.f. q on $n \times 1$ vectors.

- A non stationary Gaussian model.

The probability for an $n \times T$ data batch X in a simple non stationary model reads R_t

$$P(X) = \prod_{t=1}^{T} \phi(X(t); R_t) \qquad \phi(x; R) \overset{\text{def}}{=} \frac{1}{\sqrt{2\pi \det R}} \exp\left(-\frac{x^\dagger R^{-1} x}{2}\right)$$

Hence, it is completely specified by a sequence R_t of covariance matrices.

Mutual information

Mutual information $I(Y) = K\left[P_Y \mid \prod_i P_{Y_i}\right]$ takes two different forms (and measures different things) depending on which model we embed the data.

- In the non Gaussian model, we just saw:

$$P(Y) = \prod_{t=1}^{T} q(Y(t)) \quad \longrightarrow \quad \frac{1}{T} I(Y) = K\left[q(\cdot) \mid \textstyle\prod_i q_i(\cdot)\right]$$

- In the non stationary model:

$$P(X) = \prod_{t=1}^{T} \phi(Y(t); R_t) \quad \longrightarrow \quad \frac{1}{T} I(Y) = \frac{1}{T} \sum_{t=1}^{T} C(Y(t))$$

where $C(\cdot)$ is the correlation of a vector.

Hence, mutual information $I(Y)$ appears as a time-averaged correlation in the non stationary Gaussian model.

- Note: if q is a Gaussian distribution or if $\{R_t\}$ is a constant sequence both versions of mutual information boil down to the correlation $C(Y)$.

Algorithm for non stationary signals

Goal: minimize an estimate of $I(Y) = \sum_{t=1}^{T} C(Y(t))$.

1) For a partition $\{I_1, \ldots, I_Q\}$ of the observation interval $[1, T]$ in Q sub-intervals of lengths $\tau_1, \ldots, \tau_Q$, form local sample covariance matrices:

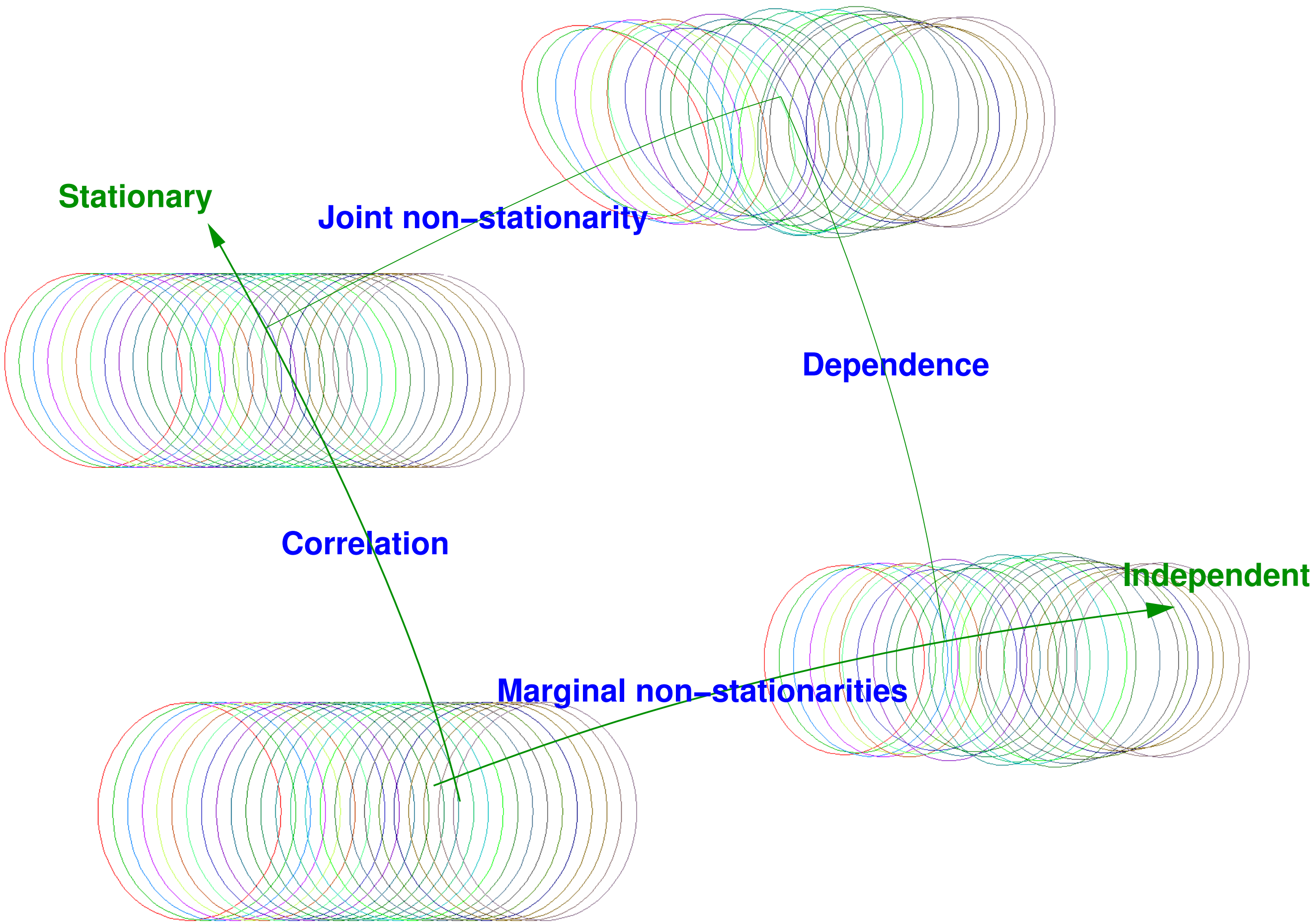
$$\widehat{R}(q) = \frac{1}{\tau_q} \sum_{t \in I_q} X(t) X(t)^\dagger$$

2) Jointly diagonalize the local covariance matrices, using Pham's algorithm

$$\min_B \sum_{q=1}^{Q} \tau_q \, \mathsf{off}(B \widehat{R}(q) B^\dagger)$$

Good things, in no particular order: • It is a maximum likelihood approach, • Smoothing by local averaging, • No need for prewhitening, • Sufficient statistics, • Fast algorithm, • Explicit measure of dependence, • Super-efficiency, • Possible automatic partitioning.

Life in Gaussian non stationary land



Same geometric structure as in the non Gaussian i.i.d. case.

Dependence, correlation and non stationarity

Non stationarity. The unavoidable definition, with $\langle R_t \rangle = \frac{1}{T}\sum_{t=1}^{T} R_t$,

$$S(Y) = K\left[Y \,|\, Y^S\right] = \sum_t K\left[\mathcal{N}(R_t) \,|\, \mathcal{N}(\langle R_t \rangle)\right]$$

For scalar processes, non stationarity boils down to

$$S(Y_i) = \frac{T}{2}\left\{\log\langle \sigma_i^2(t)\rangle - \langle \log \sigma_i^2(t)\rangle\right\}$$

The joint non stationarity $S(Y)$ is constant under linear transforms. Hence, two Pythagoras yield again:

$$I(Y) = C(Y) - \sum_{i=1}^{n} S(Y_i) + \text{cst}$$

Meaning: The most independent linear components are maximally uncorrelated and non stationary.

Heuristic interpretation: the cocktail party.

Sparseness and non properties

- In the non Gaussian model, ICA does not necessary lead to sparse sources because $G(Y_i)$ measures deviation from Gaussianity 'on both sides' of $\mathcal{N}(0, \sigma_i^2)$.

- In a non stationary Gaussian model, the non-stationarity of each component

$$S(Y_i) \propto \log\langle\sigma_i^2(t)\rangle - \langle\log \sigma_i^2(t)\rangle$$

can aslo be read as a sparseness measure.

If the variance profile $\{\sigma_i^2(t)\}$ is constant, then $S(Y_i) = 0$ but is otherwise positive.

Our non-stationarity index $S(Y_i)$ goes to ∞ if there is any silent frame. This also leads to super-efficiency.

Note: Here we are talking about sparsity of the distribution of energy (across time frames). This is not the same thing as the sparsity of the coefficients of the source signals.

Separating stationary signals

For separating stationary signals, do just the same in the Fourier domain: Time intervals become frequency bands.

Rationale: the DFT points of a stationary process are uncorrelated variables with a variance equal to the spectral density.

For Gaussian stationary processes, this procedure is close to optimal for ‘smoothly’ varying source spectra.

Blind separation of component i from j requires spectral diversity:

$$\left(\frac{1}{Q}\sum_{q=1}^{Q}\frac{P_{iq}}{P_{jq}}\right)\left(\frac{1}{Q}\sum_{q=1}^{Q}\frac{P_{jq}}{P_{iq}}\right) > 1.$$

where P_{iq} is the energy of the i-th component in the q-th spectral band.

The same diversity requirement holds for the non stationary model but is very likely to be met there.

Note: SOBI: bad; ML: good.

Still more degrees of freedom

A natural extension is to model the signal in the time-frequency plane or in the wavelet domain.

The implicit model then is that each source has a different pattern distribution of energy in time-frequency domains or in time-scale domains.

The size of the time-frequency ot time-scale domains should be chosen to have a reasonable number of Gabor atoms in each domain so that the energy can be estimated in each of them.

Note: it is also possible to use the likelihood to select the partitioning of the the time-frequency plane which gives the best contrast between the energy distribution of the sources (adaptive partitioning).

Contrasting our models

Model $\mathcal{M}$	non Gaussian $\mathcal{G}$	non stationary $\mathcal{S}$	non flat spectrum $\mathcal{F}$
Source parameter	marg. pdf	variance profile	power spectrum
Diversity required	no	yes	yes
Exhaustive stat.	no (in practice)	yes	yes
Optimization	gradient *et al.*	joint diag.	joint diag.
Estimation of nuisance	possibly	automatic	automatic
# of nuisance parameters	$o(T)$	$O(T)$	$o(T)$
Include noise	not so easy	easy	easy
Super-efficiency	non diff. pdf	silence	spectral holes
Sparsity	possibly	yes	yes
Source extraction	possible	probably	probably not

And, after all, non Gaussianity may just be non stationarity in disguise. . .