THE THREE EASY ROUTES TO INDEPENDENT COMPONENT ANALYSIS; CONTRASTS AND GEOMETRY.

Jean-François Cardoso

CNRS/ENST, 46 rue Barrault, 75634 Paris, France

mailto:cardoso@tsi.enst.fr http://tsi.enst.fr/∼cardoso/stuff.html

ABSTRACT

Blind separation of independent sources can be achieved by exploiting non Gaussianity, non stationarity or time correlation. This paper examines in a unified framework the objective functions associated to these three routes to source separation. They are the 'easy routes' in the sense that the underlying models are the simplest models able to capture the statistical structures which make source separation possible.

A key result is a generic connection between mutual information, correlation and marginal 'non properties': non Gaussianity, non stationarity, non whiteness.

1. INTRODUCTION

It is known that blind separation of instantaneous mixtures of independent sources cannot be achieved using the simplest possible source model, namely that each source signal is a Gaussian i.i.d. (identically and independently distributed) sequence. This paper examines the three simplest departures from the Gaussian i.i.d. model, each corresponding to breaking one of the assumptions of the Gaussian i.i.d. model. Specifically, we will consider the consequences of modeling the sources as 1) non Gaussian i.i.d., 2) Gaussian non stationary, and 3) Gaussian, stationarily correlated in time.

Even though these three models are intended to capture widely different statistical features, they all lead to objective functions with strikingly similar features.

These three models have already been considered in the literature. Indeed, the 'historical approach' is to use a non Gaussian i.i.d. model; non stationarity is considered in contributions [6, 7, 9]; time correlation is used in [10, 11, 2] to name only a few.

This paper uses the language of information geometry [1]: regular families of probability distribution are seen as smooth manifolds embedded on some large distribution space, each point in this space being a probability distribution. In this paper, the 'large space' is the set of distributions of $N \times T$-variates.

2. THREE SOURCE MODELS

In source separation or in ICA, an $N \times T$ data set

$$X = [x(1), \ldots, x(T)]$$

where $x(t)$ is an $N \times 1$ vector is modeled as

$$X = AS \qquad S = [s(1), \ldots, s(T)] \tag{1}$$

where A is an unknown $N \times N$ matrix and the rows of the $N \times T$ source matrix S are modeled as independent:

$$P_S(S) = \prod_{n=1}^{N} P_{S_n}(S_n) \tag{2}$$

The nth row $S_n = [s_n(1), \ldots, s_n(T)]$ is called the nth 'source sequence'. In order to completely specify a model for the distribution of the observed X, one have to choose a model for the distribution of each source sequence.

a) Non Gaussianity The most common source model in the ICA literature is non Gaussian i.i.d.. The probability of a sequence $S_n = [s_n(1), \ldots, s_n(T)]$ then reads:

$$P_{S_n}(S_n) = \prod_{t=1}^{T} r_n(s_n(t))$$

where $r_n(\cdot)$ is a univariate non Gaussian probability density for the nth source. In the following, $\mathcal{G}$ will denote the 'non Gaussian manifold', *i.e.* the set of distributions for X which are i.i.d. in time, that is, $x(t)$ and $x(t')$ are independent if $t \neq t'$ and have the same distribution, possibly non Gaussian. Every point in $\mathcal{G}$ is uniquely identified by specifying the (common) N-variate distribution of $x(t)$ for any t. Let $\mathcal{G}i$ be the submanifold of $\mathcal{G}$ in which the components of each $x(t)$ also are mutually independent for all t. The non Gaussian model is that $P_S \in \mathcal{G}i$.

b) Non stationarity The simplest approach to capture the non stationarity of the nth source sequence probably is to model it as a sequence of T independent Gaussian variables with time-varying variances $\sigma_n^2(1), \ldots, \sigma_n^2(T)$. The probability of a sequence S_n then is

$$P_{S_n}(S_n) = \prod_{t=1}^{T} \phi\left(s_n(t); \sigma_n^2(t)\right)$$

where $\phi(s;\sigma^2) = \exp(-s^2/2\sigma^2)/\sqrt{2\pi\sigma^2}$ is the Gaussian density. This simple non-stationary model is exploited in [9].

In the following, $\mathcal{S}$ will denote the 'non stationary manifold' *i.e.* the set of distributions for X such that $x(t)$ is independent of $x(t')$ ($t' \neq t$) and has a zero-mean Gaussian distribution with covariance matrix $R(t)$, possibly dependent on t. Every point in $\mathcal{G}$ is uniquely identified by specifying a sequence $R(1), \ldots, R(T)$ of covariance matrices. Let $\mathcal{S}i$ be the submanifold of $\mathcal{S}$ in which the components of $x(t)$ also are mutually independent (or, equivalently, $R(t)$ is diagonal) for all t. The non stationary model is that $P_S \in \mathcal{S}i$.

c) Time correlation Several approaches have been proposed to exploit time correlations in stationary source signals. In [10], Pham shows that time-correlated Gaussian sources can be blindly separated provided their spectra are not proportional and also introduces a Whittle approximation to the likelihood. The Whittle approximation uses the fact that the DFT coefficients

$$\tilde{s}_n(l) = \tfrac{1}{\sqrt{T}} \sum_{t=1}^{T} s_n(t) \exp(-2i\pi lt/T)$$

of a stationary sequence are asymptotically (for large enough T) decorrelated with a variance given by the power spectrum

$$D_n(l) = \sum_{\tau} E\{s_n(t)s_n(t+\tau)\} \exp(-2i\pi l\tau/T). \quad (3)$$

In the Whittle approximation, the probability of a source sequence S_n is given by

$$P_{S_n}(S_n) = \textstyle\prod_{l=1}^{T} \phi\left(|\tilde{s}_n(l)|; D_n(l)\right).$$

In the following, $\mathcal{F}$ will denote the 'non flat' manifold *i.e.* the set of distributions for X such that the DFT coefficients of X are independent and have a complex Gaussian distribution with covariance matrix $D(l)$, possibly dependent on l (*i.e.* a possibly non flat spectrum). Every point in $\mathcal{F}$ is uniquely identified by specifying a sequence of spectral covariance matrices. Let $\mathcal{F}i$ be the submanifold of $\mathcal{F}$ in which the components of $x(t)$ also are mutually independent (or, equivalently, the spectral covariance matrices are diagonal). Our model for time correlation is that $P_S \in \mathcal{F}i$.

The blind manifold The three manifolds $\mathcal{G}$, $\mathcal{S}$, and $\mathcal{F}$ intersect along a very specific submanifold of distributions which are Gaussian, stationary and spectrally flat. These are nothing but the Gaussian i.i.d. distributions. They form a manifold denoted by $\mathcal{B}$ and we have in fact

$$\mathcal{G} \cap \mathcal{S} = \mathcal{S} \cap \mathcal{F} = \mathcal{F} \cap \mathcal{G} = \mathcal{B}. \quad (4)$$

The symbol $\mathcal{B}$ for the Gaussian i.i.d. manifold refers to 'blindness' since it is the model which, preventing system identification, makes us 'blind' to (part of) A.

We note that $\mathcal{G}$, $\mathcal{S}$, $\mathcal{F}$, $\mathcal{B}$ are globally invariant under left multiplication *i.e.* if the distribution of X belongs to one of them, so does the distribution of AX for any $N \times N$ matrix A. The same is *not* true of $\mathcal{G}i$, $\mathcal{S}i$ or $\mathcal{F}i$: this is fortunate since it is precisely because left multiplication of an independent sequence S by a transform A breaks the independence that we can expect to recover uniquely the sources by restoring the independence property.

Finally, we must admit that our terminology is a bit abusive since the set of non Gaussian distributions of an $N \times T$ variate is much larger than $\mathcal{G}$ so that $\mathcal{G}$ actually a very restricted non Gaussian manifold (and similarly for $\mathcal{S}$ and $\mathcal{F}$). Again, the idea is to build models which are as simple as possible while still being able to capture some statistical feature, be it non Gaussianity, non stationarity or time correlation. These are the easy routes.

Three likelihood functions We write out the likelihood of X under the three models. It is more convenient rather to work with the negative normalized likelihood function:

$$L(A) \stackrel{\text{def}}{=} -\frac{1}{T} \log p(X|A) \quad (5)$$

where the dependence on the parameters describing the source distributions (nuisance parameters) is implicit; at this stage, we only recall the nuisance parameter for the distribution of the nth source is a univariate probability distribution $r_n(\cdot)$ in the non Gaussian model; it is a variance profile $\{\sigma_n^2(t)\}_{t=1}^T$ in the non stationary case, and a (sampled) power spectrum $\{D_n(l)\}$ in the colored case.

According to the transformation model (1), we have

$$P_X(X) = \frac{1}{|\det A|^T} P_S(A^{-1}X) \quad (6)$$

In the following, we use the notation

$$Y = A^{-1}X$$

without explicitly denoting the dependence of Y on A and $Y_n = [y_n(1), \ldots, y_n(T)]$ will denote the nth row of Y. Thanks to the product form (2), we find

$$L(A) = \textstyle\sum_{n=1}^{N} L_n(Y_n) + \log|\det A|$$

where we set $L_n(\cdot|) \stackrel{\text{def}}{=} -\frac{1}{T} \log P_{S_n}(\cdot)$. The specific form of $L_n(\cdot)$ depends on the source model:

$$L_n^G(Y_n) = \tfrac{1}{T} \textstyle\sum_{t=1}^{T} -\log r_n(y_n(t)) \quad (7)$$

$$L_n^S(Y_n) = \tfrac{1}{2T} \textstyle\sum_{t=1}^{T} \frac{y_n^2(t)}{\sigma_n^2(t)} + \log 2\pi\sigma_n^2(t) \quad (8)$$

$$L_n^F(Y_n) = \tfrac{1}{2T} \textstyle\sum_{l=1}^{T} \frac{|\tilde{y}_n(l)|^2}{D_n(l)} + \log 2\pi D_n(l) \quad (9)$$

for the three models under consideration.

3. CONTRASTS

The previous expressions for the likelihood depend on a model and on the data set X. To go further, we suppose that the data set is the realization of a random process and we denote P_X its 'true' probability distribution (as opposed to the distributions in our models) and P_Y the (true) distribution of $Y = A^{-1}X$. The symbol E denote the mathematical expectations with respect to these 'true' distributions and we look at the expected value $EL(A)$ of the normalized log-likelihood which can be seen as the quantity of which $L(A)$ is an estimator. The resulting deterministic function is sometimes called a 'contrast function'.

Likelihood contrasts Regardless of the specific source model for P_S, the data model is a transformation model of S into $X = AS$, as summarized by eq. (6). This is sufficient to find the form of the likelihood contrast:

$$EL(A) = \frac{1}{T}\left\{\mathrm{K}[P_Y|P_S] + H(X)\right\}. \qquad (10)$$

where $H(X) = -E\log P_X(X)$ is the Shannon entropy of X and where $\mathrm{K}[\cdot|\cdot]$ denotes the Kullback Leibler divergence (KLD). For two distributions with densities p and q, it is defined as

$$\mathrm{K}[p|q] = \int p(x)\log\left(p(x)/q(x)\right)\, dx. \qquad (11)$$

A key point in (10) is that $H(X)$ depends on the distribution of the data but does not depend on the *model* parameters. Therefore, it is a constant term with respect to inference, so that the contrast function associated to the maximum likelihood principle corresponds to the minimization of the Kullback divergence $\mathrm{K}[P_Y|P_S]$ between the reconstructed sources and a source model.

Kullback projection and the Pythagorean theorem Let $\mathcal{M}$ be some manifold of probability distributions. The (Kullback) projection of a distribution P onto $\mathcal{M}$, denoted $P^{\mathcal{M}}$, is the closest distribution to P in $\mathcal{M}$:

$$P^{\mathcal{M}} = \arg\min_{Q\in\mathcal{M}} \mathrm{K}[P|Q] \qquad (12)$$

A nice decomposition property takes place when $\mathcal{M}$ is an exponential manifold. A manifold is said to be *exponential* if for any two of its elements, say $P_0(X)$ and $P_1(X)$, and for any real α such that $z(\alpha) = \int P_0^{1-\alpha}(X)P_1^{\alpha}(X)dX < \infty$, the distribution $P_0^{1-\alpha}(X)P_1^{\alpha}(X)/z(\alpha)$ also belongs to the manifold.

An exponential manifold behaves like a flat space in some respects: when $\mathcal{M}$ is exponential, the Kullback projection exists, is unique and, for any distribution Q of $\mathcal{M}$,

$$\mathrm{K}[P|Q] = \mathrm{K}[P|P^{\mathcal{M}}] + \mathrm{K}[P^{\mathcal{M}}|Q] \qquad (13)$$

which is to be understood as a Pythagorean theorem in distribution space with the KLD playing the role of a squared Euclidean distance.

The manifolds $\mathcal{G}, \mathcal{G}i, \mathcal{S}, \mathcal{S}i, \mathcal{F}, \mathcal{F}i$ and $\mathcal{B}$ all are exponential manifolds, as is easily checked thanks to the characteristic form of the densities in each of them. Many significant decompositions are obtained thanks the 'orthogonal' decomposition (13) applied to the manifolds of interest.

Projecting onto the models Whenever the source model P_S belongs to $\mathcal{M}i = \mathcal{G}i, \mathcal{S}i$ or $\mathcal{F}i$, decomposition (13) yields

$$\mathrm{K}[P_Y|P_S] = \mathrm{K}[P_Y|P_Y^{\mathcal{M}}] + \mathrm{K}[P_Y^{\mathcal{M}}|P_S] \qquad (14)$$

because for $\mathcal{M} = \mathcal{G}, \mathcal{S}, \mathcal{F}$, manifold $\mathcal{M}$ is exponential and $P_S \in \mathcal{M}i \subset \mathcal{M}$. Distribution $P_Y^{\mathcal{M}}$ being the best approximation to P_Y in $\mathcal{M}$, the divergence $\mathrm{K}[P_Y|P_Y^{\mathcal{M}}]$ measures how good the approximation is. However, it is not relevant to the estimation of A because $\mathrm{K}[P_Y|P_Y^{\mathcal{M}}]$ does *not* depend on A: if $Y = AX$, we have

$$\mathrm{K}[P_Y|P_Y^{\mathcal{M}}] = \mathrm{K}[P_X|P_X^{\mathcal{M}}], \qquad (15)$$

The reason why $\mathrm{K}[P_Y|P_Y^{\mathcal{M}}]$ is invariant under left (invertible) multiplication is that $\mathcal{M}$ itself is globally invariant under left multiplication (as mentioned above). Hence, if Y is transformed by left multiplication, its best approximation $P_Y^{\mathcal{M}}$ undergoes the same transformation and (15) results from the invariance of the KLD under invertible transforms.

Given the invariance (15), decomposition (14) shows that, as soon as a particular source model $\mathcal{M}i$ (for $\mathcal{M} = \mathcal{G}, \mathcal{S}, \mathcal{F}$) is selected, the only aspects of the data distribution relevant to maximum likelihood estimation are those captured by the approximation $P_Y^{\mathcal{M}}$. In the following, we do focus on $\mathrm{K}[P_Y^{\mathcal{M}}|P_S]$.

We now the Kullback projection $P_Y^{\mathcal{M}}$ of P_Y for $\mathcal{M} = \mathcal{G}, \mathcal{S}, \mathcal{F}$ which is easy, give the simple structure of these manifolds. For the non Gaussian model, one finds that $P_Y^{\mathcal{G}}$ is the i.i.d. distribution $P_Y^{\mathcal{G}}(Y) = \prod_{t=1}^{T} P_y(y(t))$ where $P_y(\cdot)$ is the distribution of an $N \times 1$ obtained by marginalizing P_Y over time, *i.e.*

$$P_y(\cdot) \stackrel{\mathrm{def}}{=} \frac{1}{T}\sum_{t=1}^{T} P_{y(t)}(\cdot) \qquad (16)$$

For the non stationary model, distribution $P_Y^{\mathcal{S}}$ is the distribution of $\mathcal{S}$ which have, for each t, the same covariance matrix $R_Y(t) = Ey(t)y(t)^\dagger$ as $y(t)$. Similarly, for the time correlated model, distribution $P_Y^{\mathcal{F}}$ is the distribution of $\mathcal{F}$ which have the same spectral covariance matrix $D^Y(l)$ as Y at each frequency lag l.

The resulting expressions of $\mathrm{K}[P_Y^{\mathcal{M}}|P_S]$ are

$$\mathrm{K}[P_Y^{\mathcal{G}}|P_S] = T\,\mathrm{K}[P_y|P_s] \qquad (17)$$

$$\mathrm{K}[P_Y^{\mathcal{S}}|P_S] = \sum_{t=1}^{T}\mathrm{K}\{R^Y(t)|R^S(t)\} \qquad (18)$$

$$\mathrm{K}[P_Y^{\mathcal{F}}|P_S] = \sum_{l=1}^{T}\mathrm{K}\{D^Y(l)|D^S(l)\} \qquad (19)$$

where we denote $\mathrm{K}\{R_1|R_2\} = \mathrm{K}[\mathcal{N}(R_1)|\mathcal{N}(R_1)]$, the KLD between two N-variate zero-mean Gaussian distributions with covariance matrices R_1 and R_2.

The non Gaussian form (17) is well known and simply expresses the Kullback mismatch between the marginal (in time) distribution of $Y = A^{-1}X$ and the model source distribution, P_s denoting the (common) distribution of $s(t)$ for $P_S \in \mathcal{G}$. The other two forms (18) and (19) could be obtained by taking expectations on (8) and (9). Since they stem from Gaussian models, the Gaussian form $\mathrm{K}\{\cdot|\cdot\}$ of the divergence appears there quite naturally.

Similar to the non Gaussian case, the non stationary case expresses a 'Kullback mismatch' between the distribution of Y and the model source distribution, but the divergence now is an average (18) over time of mismatches between second-order moments. The obvious notation is $R^S(t) = E(s(t)s(t)^\dagger) = \mathrm{diag}(\sigma_1^2(t), \ldots, \sigma_n^2(t))$.

The time-correlated case is, as before, the 'Fourier dual' of the non stationary case with $D^S(l) = \mathrm{diag}(D_1(l), \ldots, D_n(l))$. The Kullback mismatch becomes an average through the frequency spectrum of the mismatches between the spectral covariance matrices of Y and those of the model S.

4. MUTUAL INFORMATIONS

We proceed to further decompose the likelihood mismatch $\mathrm{K}[P_Y^{\mathcal{M}}|P_S]$ for the three choices of $\mathcal{M}$. We shall isolate the nuisance parameters (the model source distributions) and three definitions of mutual information will result.

Isolating the nuisance parameters. In our next step, we project $P_Y^{\mathcal{M}}$ onto $\mathcal{M}i$ at point $P_Y^{\mathcal{M}i}$ for $\mathcal{M} = \mathcal{G}, \mathcal{S}, \mathcal{F}$. If $P_S \in \mathcal{M}i$, the Pythagorean theorem applies since $\mathcal{M}i$ is exponential, resulting in

$$\mathrm{K}[P_Y^{\mathcal{M}}|P_S] = \mathrm{K}[P_Y^{\mathcal{M}}|P_Y^{\mathcal{M}i}] + \mathrm{K}[P_Y^{\mathcal{M}i}|P_S]. \qquad (20)$$

The first term of this decomposition depends only on the (joint) distribution of $P_Y^{\mathcal{M}}$ and not on the nuisance parameters. We denote it by

$$I_{\mathcal{M}}(Y) \stackrel{\text{def}}{=} \mathrm{K}[P_Y^{\mathcal{M}}|P_Y^{\mathcal{M}i}] \qquad (21)$$

The second term is a KLD between two distributions with independent components. Therefore it decomposes as a sum of KLDs between marginal distributions:

$$\mathrm{K}[P_Y^{\mathcal{M}i}|P_S] = \textstyle\sum_{n=1}^N \mathrm{K}[P_{Y_n}^{\mathcal{M}i}|P_{S_n}]. \qquad (22)$$

Mutual informations The quantity $I_{\mathcal{M}}(Y)$, being the KLD from $P_Y^{\mathcal{M}}$ to the corresponding independent manifold $\mathcal{M}i$, is a measure of the independence between the rows of Y *in the particular model* $\mathcal{M}$. Therefore, we obtain three forms for the *mutual information*, depending on the choice of $\mathcal{M}$:

$$I_{\mathcal{G}}(Y) = T\,\mathrm{K}[P_y|\textstyle\prod_{n=1}^N P_n] \qquad (23)$$

$$I_{\mathcal{S}}(Y) = \textstyle\sum_{t=1}^T \mathrm{K}\{R^Y(t)|\mathrm{diag}(R^Y(t))\} \qquad (24)$$

$$I_{\mathcal{F}}(Y) = \textstyle\sum_{l=1}^T \mathrm{K}\{D^Y(l)|\mathrm{diag}(D^R(l))\} \qquad (25)$$

These three forms are derived from (17,18,19) by minimizing over the nuisance parameters.

The first expression is the familiar non Gaussian i.i.d. form for the mutual information. In the other two expressions (24) and (25), the mutual information appears as a measure of the mean diagonality of (spectral) covariance matrices. They lead to simple separation techniques since an efficient algorithm exists for the joint diagonalization of a set of positive matrices. See [9] for the non stationary case and [8] for the time correlated case.

Marginal mismatches The second term of decomposition (20) is a sum (22) of marginal mismatches. Their respective forms in the three models are

$$\mathrm{K}[P_{Y_n}^{\mathcal{G}i}|P_{S_n}] = T\,\mathrm{K}[P_{y_n}|P_{s_n}] \qquad (26)$$

$$\mathrm{K}[P_{Y_n}^{\mathcal{S}i}|P_{S_n}] = \mathrm{K}(\{Ey_n^2(t)\}|\{\sigma_n^2(t)\}) \qquad (27)$$

$$\mathrm{K}[P_{Y_n}^{\mathcal{F}i}|P_{S_n}] = \mathrm{K}(\{E|\tilde{y}_n(l)|^2\}|\{D_n(l)\}) \qquad (28)$$

Expression (26) only involves the time marginals of P_y defined at (16) and those of P_s, defined similarly. The marginal mismatches for $\mathcal{S}$ (resp. $\mathcal{F}$) involve only the variance profiles (resp. spectral profiles) between Y and S. We have used in (27) and in (28) a divergence between two sequences $\{a(t)\}_{t=1}^T$ and $\{b(t)\}_{t=1}^T$ of positive numbers, defined as

$$\mathrm{K}(\{a(t)\}|\{b(t)\}) \stackrel{\text{def}}{=} \textstyle\sum_{t=1}^T \frac{a_t}{b_t} - \log\frac{a_t}{b_t} - 1 \qquad (29)$$

which is nothing but $\mathrm{K}\{\mathrm{diag}(a_1, \ldots, a_T)|\mathrm{diag}(b_1, \ldots, b_T)\}$ so this is hardly a new notation.

5. DEPENDENCE AND CORRELATION

This section unveils the relationships between the three forms of dependence, the correlation, and three 'non-properties' of the source distributions: non Gaussianity, non stationarity and spectral coloration (non flatness).

Non properties The blind manifold $\mathcal{B}$ contains only Gaussian i.i.d. distributions which are simplistic models: projecting $P_Y^{\mathcal{G}}$ onto $\mathcal{B}$ 'erases' the non Gaussian structure of $P_Y^{\mathcal{G}}$ by retaining only the second-order structure; projecting $P_Y^{\mathcal{S}}$ onto $\mathcal{B}$ similarly 'erases' the time structure by averaging it out; projecting $P_Y^{\mathcal{F}}$ onto $\mathcal{B}$ similarly 'erases' the spectral structure by 'stationarizing it'. We take the divergence from

$P_Y^{\mathcal{G}}$ (resp. $P_Y^{\mathcal{S}}$, resp. $P_Y^{\mathcal{F}}$) to $P_Y^{\mathcal{B}}$ as measures of non Gaussianity (resp. non stationarity, resp. non flatness). To each $\mathcal{M} = \mathcal{G}, \mathcal{S}, \mathcal{F}$, corresponds a 'non property' $\mathcal{M}(Y)$ of Y:

$$\begin{aligned} \mathcal{M}(Y) &\stackrel{\text{def}}{=} \mathrm{K}[P_Y^{\mathcal{M}}|P_Y^{\mathcal{B}}] : \text{ non property in } \mathcal{M} \quad (30) \\ \mathcal{G}(Y) &= \mathrm{K}[P_Y^{\mathcal{G}}|P_Y^{\mathcal{B}}] : \text{ non Gaussianity of } Y \quad (31) \\ \mathcal{S}(Y) &= \mathrm{K}[P_Y^{\mathcal{S}}|P_Y^{\mathcal{B}}] : \text{ non stationarity of } Y \quad (32) \\ \mathcal{F}(Y) &= \mathrm{K}[P_Y^{\mathcal{F}}|P_Y^{\mathcal{B}}] : \text{ non (spectral) flatness} \quad (33) \end{aligned}$$

These non properties are invariant under invertible left multiplication of Y because the KLD as well as both manifolds $\mathcal{B}$ and $\mathcal{M} = \mathcal{G}, \mathcal{S}, \mathcal{F}$ are globally invariant such transforms. Thus $\mathcal{M}(Y) = \mathcal{M}(X)$ if $Y = BX$ for any invertible B.

Marginal non properties The above definitions of 'non properties' also apply to each component Y_n (each row) of Y, that is, to $1 \times T$ variables. In this case, the just mentioned invariance property reduces to a simple scale invariance.

In the non Gaussian case, $\mathcal{G}(Y_n)$ is T times the KLD from P_{y_n} to its best Gaussian approximation. This quantity is sometimes called *neguentropy* [4]. In the non stationary (resp. non flat) case, $\mathcal{S}(Y_n)$ (resp. $\mathcal{F}(Y_n)$) measures the deviation of $\{Ey_n^2(t)\}$ (resp. $\{E\tilde{y}_n^2(l)\}$) from a constant variance profile (resp. from a flat spectrum). Using the Kullback-like measure (29) of divergence between two positive sequences, we define the 'Kullback dispersion' $\mathcal{D}(\{a(t)\})$ of a sequence $\{a(t)\}_{t=1}^T$ of T positive numbers as the divergence to the closest constant sequence:

$$\mathcal{D}(a) \stackrel{\text{def}}{=} \min_c \mathrm{K}((a(1), \ldots, a(T))|(c, \ldots, c)) \geq 0 \quad (34)$$

which is equal to 0 if and only if a is a constant sequence. The minimizer is easily found to be $c = \frac{1}{T}\sum_t a(t)$. Inserting this value back in (29) and rearranging yields

$$\mathcal{D}(a) = T \ \left(\log(\tfrac{1}{T}\textstyle\sum_t a(t)) - \tfrac{1}{T}\textstyle\sum_t \log a(t)\right) \quad (35)$$

(which could be an alternate definition of the dispersion). With this, the Gaussian 'non properties' appear as dispersions, as in this summary:

$$\begin{aligned} \mathcal{G}(Y_n) &= T\ \mathrm{K}[P_{y_n}|\phi(\cdot; Ey_n^2)] \quad (36) \\ \mathcal{S}(Y_n) &= \mathcal{D}(\{Ey_n^2(t)\}) \quad (37) \\ \mathcal{F}(Y_n) &= \mathcal{D}(\{E\tilde{y}_n^2(l)\}). \quad (38) \end{aligned}$$

Correlation The Gaussian i.i.d. manifold $\mathcal{B}$ is a poor manifold but we can still define a mutual information $I_{\mathcal{B}}(Y)$ with respect to it by (21). We use a special notation $C(Y)$ for it since one finds

$$C(Y) \stackrel{\text{def}}{=} I_{\mathcal{B}}(Y) = \mathrm{K}[P_Y^{\mathcal{B}}|P_Y^{\mathcal{B}i}] = T\ \mathrm{K}\{R_Y|\mathrm{diag}(R_Y)\}$$

where R_Y is the covariance matrix

$$R_Y \stackrel{\text{def}}{=} \tfrac{1}{T}\textstyle\sum_{t=1}^T R_Y(t). \quad (39)$$

so that the Gaussian i.i.d. mutual information $C(Y) = I_{\mathcal{B}}(Y)$ is nothing but a *correlation* measure since it measures the diagonality of R_Y. We will plainly call it 'the correlation' in the following.

Connecting everything All our distributional measures are connected by applying twice the Pythagorean theorem. First, we can apply it to triangle $(P_Y^{\mathcal{M}}, P_Y^{\mathcal{B}}, P_Y^{\mathcal{B}i})$ for each $\mathcal{M} = \mathcal{G}, \mathcal{S}, \mathcal{F}$ because $P_Y^{\mathcal{M}}$ projects at point $P_Y^{\mathcal{B}}$ onto $\mathcal{B}$ which is exponential and contains $P_Y^{\mathcal{B}i}$. We get

$$\mathrm{K}[P_Y^{\mathcal{M}}|P_Y^{\mathcal{B}i}] = \mathrm{K}[P_Y^{\mathcal{M}}|P_Y^{\mathcal{B}}] + \mathrm{K}[P_Y^{\mathcal{B}}|P_Y^{\mathcal{B}i}]. \quad (40)$$

Second, we can apply it to triangle $(P_Y^{\mathcal{M}}, P_Y^{\mathcal{M}i}, P_Y^{\mathcal{B}i})$ as in eq. (20) because $P_Y^{\mathcal{B}i}$ belongs to $\mathcal{M}i$. We get

$$\mathrm{K}[P_Y^{\mathcal{M}}|P_Y^{\mathcal{B}i}] = \mathrm{K}[P_Y^{\mathcal{M}}|P_Y^{\mathcal{M}i}] + \mathrm{K}[P_Y^{\mathcal{M}i}|P_Y^{\mathcal{B}i}]. \quad (41)$$

On the right-hand side of (40), the first term is the 'non property' $\mathcal{M}(Y)$ associated to $\mathcal{M}$ and the second term is the correlation $C(Y)$. On the right hand side of (41), the first term is the dependence $I_{\mathcal{M}}(Y)$ defined at (21); the second term, being a KLD between distributions with independent rows, is the sum of the marginal KLDs. More to the point, we have $\mathrm{K}[P_Y^{\mathcal{M}i}|P_Y^{\mathcal{B}i}] = \sum_{n=1}^N \mathcal{M}(Y_n)$ since each of the marginal KLDs is recognized as the 'non property' associated to manifold $\mathcal{M}$. We have thus obtained two independent expressions for $\mathrm{K}[P_Y^{\mathcal{M}}|P_Y^{\mathcal{B}i}]$ using twice the Pythagorean theorem over the same hypotenuse while the other sides of both triangles have been identified with meaningful properties. Combining (40) and (41) yields

$$I_{\mathcal{M}}(Y) + \sum_{n=1}^N \mathcal{M}(Y_n) = \mathcal{C}(Y) + \mathcal{M}(Y), \quad (42)$$

which connects independence, correlation and joint and marginal 'non properties' in a given model. These connections are depicted in the figure below for $\mathcal{M} = \mathcal{G}$.

Of particular relevance to ICA, is the fact that $\mathcal{M}(Y)$ is constant under left multiplication (see above). Therefore, if $Y = BX$, then for $\mathcal{M} = \mathcal{G}, \mathcal{S}, \mathcal{F}$:

$$\boxed{I_{\mathcal{M}}(Y) = \mathcal{C}(Y) - \textstyle\sum_{n=1}^N \mathcal{M}(Y_n) + \mathcal{M}(X)} \quad (43)$$

In other words,

> under linear transforms, the dependence $I_{\mathcal{M}}(Y)$ measured in some model $\mathcal{M} = \mathcal{G}, \mathcal{S}, \mathcal{F}$ is, up to a constant term $\mathcal{M}(X)$, the correlation $\mathcal{C}(Y)$, minus the marginal 'non-properties' associated to the model.

6. DISCUSSION

Summary Three simple models $\mathcal{M} = \mathcal{G}, \mathcal{S}, \mathcal{F}$ can be used in ICA. In a model $\mathcal{M}$, The likelihood objective (5) is an estimator of the likelihood contrast (10) which amounts, by (14,15) to matching the distribution $P_Y^{\mathcal{M}}$ of $Y = A^{-1}X$ as seen by model $\mathcal{M}$ to the source distribution P_S. This mismatch itself decomposes as mutual information $I_{\mathcal{M}}(Y)$ plus marginal source mismatches (20,21,22). Mutual information itself, whose form depends on the model (23,24),25), can be decomposed as a pure correlation measure $C(Y)$ plus 'marginal non properties (43) All these quantities have different expressions depending on model $\mathcal{M}$ but retain the same structure across all models.

The bottom line is that looking for components which are 'as independent as possible' is equivalent, in a given model, to look for components which are maximally decorrelated and non Gaussian, or non stationary, or spectrally colored. This is quantitatively expressed by eq. (43) and illustrated by fig. 1.

Relation with previous works In the non Gaussian i.i.d. case, Comon had already noticed in his seminal paper [4], that *if one enforces decorrelation*, (that is, $C(Y) = 0$), then the mutual information boils down to the sum of the marginal 'neguentropies'. The present paper generalizes Comon's insight in two directions. First, equation (43) shows that mutual information nicely balances correlation and non Gaussianity. This is to compared to two other extreme approaches: the pre-whitening or sphering approach to ICA —which enforces decorrelation— amounts to give an infinite weight to the correlation term; on the opposite side, the one-unit approach (see *e.g.* [5]) starts with the idea of finding the most non Gaussian marginal distribution, amounting in effect to give zero weight to the correlation term. Second, equation (43) shows that, depending on the chosen source model, the marginal non Gaussianity (36) is replaced by a measure of non flatness of the variance profiles either in time (37), or in frequency (38).

Due to space limitations, most proofs have been omitted paper. However most of the present material revisits other publications. See [3] for the geometry of the non Gaussian i.i.d. case, see [9] for mutual information in non stationary models and [8] for mutual information in temporally correlated models.

Diversity This paper emphasizes the structural similarities between three models. The comparison could be pushed further in terms of estimating equations, adaptivity (estimation of the nuisance parameters) and stability conditions. However, as a closing remark, a significant difference can be pointed out: blind identifiability in model $\mathcal{S}$ (resp. $\mathcal{F}$) requires some diversity: if two sources have proportional variance profiles (resp. power spectra) they cannot be blindly separated (in the given model). No such diversity is required in the non Gaussian model.

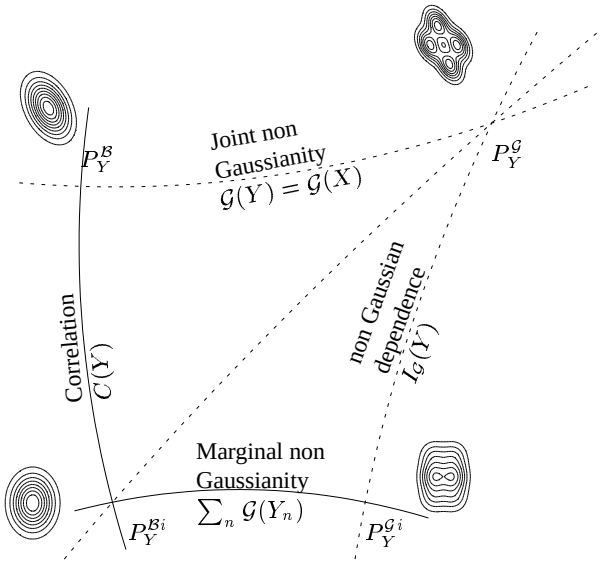

Figure 1: Connections between mutual information, correlation and non Gaussianity in the case of N=2 components. Solid lines: $\mathcal{G}i$ (horizontal) and $\mathcal{B}$ (vertical). The whole picture is embedded in $\mathcal{G}$.

7. REFERENCES

[1] S.-I. Amari. *Differential-Geometrical Methods in Statistics*. Number 28 in Lect. Notes in Stat. Springer-Verlag, 1985.

[2] A. Belouchrani *et al.* A blind source separation technique based on second order statistics. *IEEE Trans. on Sig. Proc.*, 45(2):434–44, Feb. 1997.

[3] J.-F. Cardoso. Entropic contrasts for source separation: geometry and stability, in *Unsupervised adaptive filters*, volume 1, pages 139–190. John Wiley & sons, 2000.

[4] P. Comon. Independent component analysis, a new concept ? *Signal Processing, Elsevier*, 36(3):287–314, Apr. 1994.

[5] A. Hyvrinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[6] K. Matsuoka, M. Ohya, and M. Kawamoto. A neural net for blind separation of nonstationary signals. *Neural networks*, 8(3):411–419, 1995.

[7] L. Parra and C. Spence. Convolutive blind source separation of non-stationary sources. *IEEE Trans. on Speech and Audio Processing*, pages 320–327, may 2000.

[8] D. Pham. Blind separation of instantaneous mixture of sources via the Gaussian mutual information criterion. *Signal Processing*, (4):855–870, 2001.

[9] D.-T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. on Sig. Proc.*, 2001. to appear.

[10] D.-T. Pham and P. Garat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Tr. SP*, 45(7):1712–1725, July 1997.

[11] L. Tong, V. Soon, Y. Huang, and R. Liu. AMUSE: a new blind identification algorithm. In *Proc. ISCAS*, 1990.