# On Blind Methods in Signal Processing

Christian Jutten, Massoud Babaie-Zadeh, Shahram Hosseini

*Abstract*— **Blind methods are powerful tools when very weak information is necessary. Although many algorithms can be called blind, in this paper, we focus on blind source separation (BSS) and independent component analysis (ICA). After a discussion concerning the blind nature of these techniques, we review three main points: the separability, the criteria, the algorithms.**

## I. Introduction

Basically, the term *blind* in signal processing is associated to methods in which weak information about the signal is required. The methods are then efficient when no information is available. For instance, blind equalization techniques avoid to regularly send known sequences for identifying the channel; the blind techniques are then more efficient in term of communication rate. Another point of view is that blind techniques use criteria (or cost functions) which can be directly computed from the data. The algorithms are sometimes said unsupervised. For instance, in Kohonen's self-organizing maps (SOM) [35], the criterion is a mean square error of quantization, which only depends on the data and on the quantifiers.

In this paper, we focus on recent [1] techniques called blind source separation (BSS) or independent component analysis (ICA)[10], [30]. The main difference is that BSS concerns signals (with time properties) while ICA concerns more generally data. Anyway, BSS as well as ICA are driven by statistical independence which can be computed from the observations (*i.e.* blindly). In the context of signal processing, BSS is then a much more natural concept, and leads to many applications.

BSS in instantaneous nonlinear mixtures consists in estimating $n$ unknown sources from $p$ observations $\mathbf{x}(t) = [x_1(t), x_2(t), \cdots, x_p(t)]^T$, which are an unknown mapping of $n$ unknown sources $\mathbf{s}(t) = [s_1(t), s_2(t), \cdots, s_n(t)]^T$ viewed by a set of $p$ sensors:

$$\mathbf{x}(t) = \mathcal{F}(\mathbf{s}(t)),\ t = 1, \ldots, T \qquad (1)$$

where $\mathcal{F}$ is a one-to-one mapping. In the following, for sake of simplicity, we consider $n = p$.

The authors are with the Signals and Images Laboratory (CNRS UMR 5083), INPG-LIS, 46 avenue Félix Viallet, 38031 Grenoble Cedex, E-mail: Christian.Jutten@inpg.fr. C. Jutten is professor with University Joseph Fourier, Sciences and Techniques Institute of Grenoble (ISTG). This work has been partly funded by the European project BLISS (IST-1999-14190) and by the SASI project (ELESA-IMAG).

[1] The first works have been published in 1985 in a conference [28] and then in a journal in 1991 [32], [19]

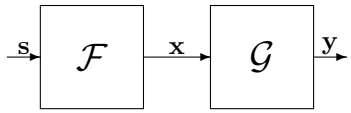


Fig. 1. General mixture and separation models.

The main assumption is the statistical independence of the sources. Although it is physically plausible, it is often considered as a strong assumption. In fact, it rather contests Gaussian assumption and relevance of decorrelation which have been so usual and convenient for many years, and are much more simpler to handle than non Gaussian and independence. The term blind emphasizes on the fact that (i) the sources are not observed, (ii) source distribution is unknown and *a priori* non Gaussian. Often, people claims that it is also blind since there is no information concerning the mapping $\mathcal{F}$. As we will see in section II, strong assumptions on the mapping structure are used, and are actually necessary for insuring separability. Finally, it is often interesting to exploit very general signal properties, like temporal correlation (colored signals)[50], [9], [29] or non stationarity [38], [43].

In this review paper, we mainly discuss on three points: separability, independence criteria, algorithms.

## II. Separability

The basic idea of BSS consists in estimating some inverse mapping $\mathcal{G}$ (Figure 1) of the transformation $\mathcal{F}$ such that each component of the output vector $\mathbf{y}(t) = \mathcal{G}(\mathbf{x}(t))$, where $\mathbf{y}(t) = [y_1(t), y_2(t), \cdots, y_n(t)]^T$, is a function of one source. In the general case, the goal of separation is to obtain

$$y_i(t) = h_i(s_{\sigma(i)}(t)),\ i = 1, \ldots, n \qquad (2)$$

where $h_i$ is any invertible mapping which is associated to a residual distortion.

Since the main assumption is the source independence, one suggests to estimate $\mathcal{G}$, such that the estimated outputs $\mathbf{y}(t) = \mathcal{G}\mathbf{x}(t)$ become statistically independent. The key question is the following: does output independence always insure separation, *i.e.* $\mathbf{y}(t) = \mathbf{s}(t)$ ?

*A. Indeterminacies*

First, recall briefly the definition of independent random vector.

*Definition II-A.1:* A random vector $\mathbf{x}$ is statistically independent if its joint probability density function (pdf) $p_{\mathbf{x}}(\mathbf{u})$ satisfies $p_{\mathbf{x}}(\mathbf{u}) = \prod_i p_{x_i}(u_i)$, where $p_{x_i}(u_i)$ are the marginal pdf of the random variables $x_i$.

Then, it is clear that the product of a permutation matrix $\mathbf{P}$ by any diagonal mapping both preserves independence and insures separability.

*Definition II-A.2:* A one-to-one mapping $\mathcal{H}$ is called *trivial*, if it transforms *any* random vector $\mathbf{s}$ with independent components into a random vector with independent components.

The set of trivial transformations will be denoted by $\mathfrak{T}$. Trivial mappings are then mapping preserving the independence property of *any* random vector. One can easily show that a one-to-one mapping $\mathcal{H}$ is trivial if and only if it writes as:

$$\mathcal{H}_i(u_1, u_2, \ldots, u_n) = h_i(u_{\sigma(i)}),\ i = 1, 2, \ldots, n \quad (3)$$

where $h_i$ are arbitrary functions and $\sigma$ is any permutation over $\{1, 2, \ldots, n\}$.

This result establishes a link between the independence assumption and the objective of source separation. In fact, it becomes clear in the following that the source separation objective is, using the independence assumption, to impose that the global transformation $\mathcal{H} = \mathcal{G} \circ \mathcal{F}$ is trivial.

However, from (3) it is clear that sources can only be separated up to a permutation and a nonlinear function. In fact, for any invertible mappings $\mathbf{F}(\mathbf{x}) = [f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})]^T$ whose each component is a scalar nonlinear mappings $f_i(\mathbf{x}) = f_i(x_i)$, $i = 1, \ldots, n$, it is evident that if $p_{\mathbf{x}}(\mathbf{u}) = \prod_i p_{x_i}(u_i)$, then $p_{\mathbf{F}(\mathbf{x})}(\mathbf{v}) = \prod_i p_{f_i(x_i)}(v_i)$. Moreover, this is not possible without imposing additional structural constraints on $\mathcal{H}$, as we shall see in the next section.

*B. Results from factor analysis*

In the general case, *i.e.* the mapping $\mathcal{H}$ has no particular form, a well known statistical result shows that the independence conservation constraint is not strong enough to insure the separability in the sense of equation (2). This result has been established, early in the 50's, by Darmois [21] where he used a simple constructive method for decomposing any random vector as a non trivial mapping of independent variables.

This result is negative, in the sense that it shows the existence of non trivial transformations $\mathcal{H}$ which still *"mix"* the variables while preserving their statistical independence. Hence, for general nonlinear transformations and without constraints on the transformation model, source separation is simply *impossible* by only using the statistical independence.

In the conclusion of [21], Darmois clearly states: *"These properties [...] clarify the general problem of factor analysis by showing the large indeterminacies it presents as soon as one leaves the field, already very wide, of linear diagrams."*.

B.1 A simple example

A simple example derived from [47] is the following: suppose $s_1 \in \mathcal{R}^+$ is a Rayleigh distributed variable with pdf $p_{s_1}(s_1) = s_1 exp(-s_1^2/2)$ and $s_2 \in [0, 2\pi)$ is uniform and independent of $s_1$. Consider the nonlinear mapping

$$\begin{aligned}[y_1, y_2] &= \mathcal{H}(s_1, s_2) \\ &= [s_1 \cos(s_2), s_1 \sin(s_2)] \end{aligned} \quad (4)$$

which has a non diagonal Jacobian matrix

$$\mathbf{J} = \begin{pmatrix} cos(s_2) & -s_1 sin(s_2) \\ sin(s_2) & s_1 cos(s_2) \end{pmatrix}. \quad (5)$$

The joint pdf of $y_1$ and $y_2$ is:

$$\begin{aligned} p_{y_1,y_2}(y_1, y_2) &= \frac{p_{s_1,s_2}(s_1, s_2)}{|\mathbf{J}|} \\ &= \frac{1}{2\pi} exp(\frac{-y_1^2 - y_2^2}{2}) \\ &= (\frac{1}{\sqrt{2\pi}} exp\frac{-y_1^2}{2})(\frac{1}{\sqrt{2\pi}} exp\frac{-y_2^2}{2}) \end{aligned}$$

The previous relation shows that the two random variables $y_1$ and $y_2$ are independent (although they are completely different of the sources) and Gaussian.

Other examples can be found in the literature (see for example Lukacs [37]) or can be easily constructed.

*C. Specific model*

The previous negative result is due to the fact that we assume no constraints on the transformation $\mathcal{H}$. Constraining the transformation $\mathcal{H}$ in a certain set of transformations $\mathfrak{Q}$ can reduce these large indeterminacies.

To characterize the indeterminacies for a specific model $\mathfrak{Q}$, one must solve the independence preservation equation which writes as:

$$\begin{aligned} &\forall E \in \mathfrak{M}_n \\ &\int_E dF_{s_1} dF_{s_2} \cdots dF_{s_n} = \int_{\mathcal{H}(E)} dF_{y_1} dF_{y_2} \cdots dF_{y_n} \end{aligned} \quad (6)$$

where $\mathfrak{M}_n$ is a $\sigma$-algebra on $\mathbb{R}^{\kappa}$.

Let $\mathfrak{P}$ denote the set :

$$\begin{aligned} \mathfrak{P} = \{&(F_{s_1}, F_{s_2}, \ldots, F_{s_n}), / \exists \mathcal{H} \in \mathfrak{Q} \setminus (\mathfrak{T} \cap \mathfrak{Q}) : \\ &\mathcal{H}(\boldsymbol{s}) \text{ has independent components}\} \end{aligned} \quad (7)$$

This set $\mathfrak{P}$ contains all source distributions for which there exists a non trivial mapping $\mathcal{H}$ belonging to the model $\mathfrak{Q}$ and preserving the independence of the $\mathbf{s}$ components.

An ideal model will be such that $\mathfrak{P}$ is empty and $\mathfrak{T} \cap \mathfrak{Q}$ contains the identity as a unique element. However, in general this is not fulfilled, we then say that source separation is possible when the sources distribution is in $\bar{\mathfrak{P}}$, the complement of $\mathfrak{P}$, sources are then restored up to a trivial transformation belonging to $\mathfrak{T} \cap \mathfrak{Q}$.

C.1 Example: Linear models

In the case of linear models, the transformation $\mathcal{F}$ is linear and can be represented by an $n \times n$ matrix $\mathbf{A}$, the observed signals write then as $\mathbf{e} = \mathbf{A}\mathbf{s}$. Source separation consists then in estimating a matrix $\boldsymbol{B}$ such that $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{e} = \boldsymbol{H}\boldsymbol{s}$ has independent components.

The set of *linear* trivial transformations $\mathfrak{T} \cap \mathfrak{Q}$ is the set of matrices equal to the product of a permutation and a diagonal matrix.

Considering two linear functions of $n$ independent random variables $s_i$, $i = 1, \ldots, n$:

$$x_1 = a_1 s_1 + \ldots + a_n s_n$$
$$x_2 = b_1 s_1 + \ldots + b_n s_n$$

the Darmois-Skitovich theorem [21] states that, if $a_i b_i \neq 0$, independence of $x_1$ and $x_2$ implies that $s_i$ is Gaussian. From this theorem, it is clear that the set $\mathfrak{P}$ contains the distributions having at least two Gaussian components.

We then conclude that source separation is possible whenever we have at most one Gaussian source, sources are then restored up to a permutation and a diagonal matrix [17].

*D. Separation of PNL mixtures*

A postnonlinear model (PNL) consists in nonlinear observations of the following form:

$$x_i(t) = f_i(\sum_{j=1}^{n} a_{ij} s_j(t)),\ i = 1, \ldots, n, \qquad (8)$$

Figure 2 shows what this model looks like. One can see that this model is a cascade of a linear mixture and a component-wise nonlinearity, *i.e.* acting on each output independently from the others. The nonlinear functions (distortions) $f_i$ are supposed invertible.

Besides its theoretical interest, this model, belonging to the L-ZMNL[2] family, suits perfectly for a lot of real world applications. For instance, such models can be found in sensors arrays [41], satellite and microwave communications [46], and in a lot of biological systems [36].

[2]L stands for Linear and ZMNL stands for Zero-Memory NonLinearity.

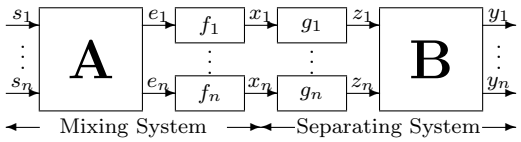


Fig. 2. The mixing-separating system for PNL mixtures.

As discussed before, the most important thing when dealing with nonlinear mixtures is the separability problem. First, we must think about the separation structure $\mathcal{G}$ which has as constraints:
1. **Can invert the mixing system** in the sense of equation (2): this constraint is quite obvious because that is what we want!
2. **Be as simple as possible**: In fact we want to reduce, in case we are successful, the residual distortions $h_i$ which are the blind spot of the independence assumption.
By defining these two constraints, we have no other choice that selecting for the separating system $\mathcal{G}$ the mirror structure of the mixing system $\mathcal{F}$ (Fig. 2).

*E. Other separable non linear mixtures*

Due to the interesting Darmois's result for linear mixtures, it is clear that nonlinear mixtures which could be reduced to linear mixtures with a simple mapping would be separable.

E.1 A simple example

As an example, one can consider multiplicative mixtures:

$$x_j(t) = \prod_{i=1}^{n} s_i^{\alpha_i}(t),\ j = 1, \ldots, n \qquad (9)$$

where $s_i(t)$ are positive independent sources. Taking the logarithm leads to:

$$\ln x_j(t) = \sum_{i=1}^{n} \alpha_i \ln s_i(t),\ j = 1, \ldots, n \qquad (10)$$

which is a linear model of the new independent random variables (since ln is monotonous) $\ln s_i(t)$. For instance, this type of mixtures can be used for modeling the gray-level images as the product of incident light and reflected light [25], or the cross-dependency between temperature and magnetic field in Hall silicon sensor. Considering in more details the latter example, the Hall voltage [45] is equal to:

$$V_H = kBT^{\alpha} \qquad (11)$$

where $\alpha$ depends on the semiconductor type, since the temperature effect is related to the mobility of the majority carriers. In fact, in this model, the temperature $T$ is positive, but the sign of the magnetic field $B$ can vary. Then, using two types (N

and P) of sensors, we have:

$$V_{H_N}(t) = k_N B(t) T^{\alpha_N}(t) \quad (12)$$
$$V_{H_P}(t) = k_P B(t) T^{\alpha_P}(t) \quad (13)$$

For simplifying the equations, in the following, we drop the variable $t$. Then, taking the logarithm:

$$\ln \mid V_{H_N} \mid = \ln k_N + \ln \mid B \mid + \alpha_N \ln T \quad (14)$$
$$\ln \mid V_{H_P} \mid = \ln k_P + \ln \mid B \mid + \alpha_P \ln T \quad (15)$$

The above equations can be easily solved, even with simple decorrelation approaches. However, in that case, it is still simpler to directly compute the ratio of the two above equations :

$$R = \frac{V_{H_N}}{V_{H_P}} = \frac{k_N}{k_P} T^{\alpha_N - \alpha_P} \quad (16)$$

which only depends on the temperature. Then, since $R$ is a temperature reference (up to a scale factor), for separating the magnetic field, it is sufficient to estimate the parameter $k$ such that $V_{H_N} R^k$ becomes uncorrelated with $R$. We then deduce $B(t)$ up to a mutiplicative constant. Finally, absolute estimations of $B$ and $T$ require calibration steps.

### E.2 Generalization to a class of mappings

This extention of the Darmois-Skitovic theorem to more general nonlinear functions has been addressed in beginning of 70's by Kagan *et al.* [33]. Their results have recently been revisited in the framework of source separation in nonlinear mixtures by Eriksson and Koivunen [25]. The main idea is to consider particular mappings $\mathcal{F}$ satisfying an *addition theorem* in the sense of the theory of functional equations. As a simple example of such a mapping, consider:

$$x_1 = \frac{s_1 + s_2}{1 + s_1 s_2}$$
$$x_2 = \frac{s_1 - s_2}{1 - s_1 s_2}$$

where $s_1$ and $s_2$ are two independent random variables. Now, with the variable transformations $u_1 = \tan^{-1} s_1$ and $u_2 = \tan^{-1} s_2$, the above nonlinear model becomes:

$$x_1 = \tan(u_1 + u_2)$$
$$x_2 = \tan(u_1 - u_2)$$

Then, applying again the variable transformation on $x_1$ and $x_2$ leads to:

$$v_1 = \tan^{-1}(x_1) = u_1 + u_2$$
$$v_2 = \tan^{-1}(x_2) = u_1 - u_2$$

which is now a linear mixture of two independent variables. As explained Kagan *et al.*, the nice result is due to the the fact that $\tan(a+b)$ is a function of $\tan a$ and $\tan b$:

$$\tan(a+b) = \frac{\tan a + \tan b}{1 + \tan a \tan b} \quad (17)$$

More generally, this property will hold provided than we consider mappings $\mathcal{F}$ satisfying an addition theorem like:

$$f(s_1 + s_2) = \mathcal{F}[f(s_1), f(s_2)] \quad (18)$$

Let the range of $u \in \mathfrak{S}$ be in the range $[a, b]$, the basic properties required for the mapping (in the case of two variables, but extension is straightforward) are the following:

- $\mathcal{F}$ is continuous at least separately for the 2 variables,
- $\mathcal{F}$ is commutative, *i.e.* $\forall (u, v) \in \mathfrak{S}^2$, $\mathcal{F}(u, v) = \mathcal{F}(v, u)$,
- $\mathcal{F}$ is associative, *i.e.* $\forall (u, v, w) \in \mathfrak{S}^3$, $\mathcal{F}(\mathcal{F}(u, v), w) = \mathcal{F}(u, \mathcal{F}(v, w))$
- It exists an identity element $e \in \mathfrak{S}$ such that $\forall u \in \mathfrak{S}$, $\mathcal{F}(u, e) = \mathcal{F}(e, u) = u$
- $\forall u \in \mathfrak{S}$, it exists an inverse element $u^{-1} \in \mathfrak{S}$ such that $\mathcal{F}(u, u^{-1}) = \mathcal{F}(u^{-1}, u) = e$

In other words, denoting $u \circ v = \mathcal{F}(u, v)$, it means that the set $(\mathfrak{S}, \circ)$ is an Abelian group. Under this condition, Aczel [1] proved that it exists a monotonic and continuous function $f : \mathbb{R} \to [a, b]$ such that:

$$f(x + y) = \mathcal{F}(f(x), f(y)) = f(x) \circ f(y) \quad (19)$$

Clearly, applying $f^{-1}$ (which exists since $f$ is monotonic) to the above equation leads to:

$$x + y = f^{-1}(\mathcal{F}(f(x), f(y))) = f^{-1}(f(x) \circ f(y)) \quad (20)$$

Using the above property (19), one can define a product $\star$ with integer and extend it to real:

$$f(cx) = c \star f(x) \quad (21)$$

or, taking $f^{-1}$ and denoting $f(x) = u$,

$$c f^{-1}(u) = f^{-1}(c \star u) \quad (22)$$

Then, for any constants $c_1, \ldots, c_n$ and random variables $u_1, \ldots, u_n$, the following relation holds:

$$c_1 f^{-1}(u_1) + \ldots c_n f^{-1}(u_n) = f^{-1}(c_1 \star u_1 \circ \ldots \circ c_n \star u_n) \quad (23)$$

Finally, Kagan *et al.* stated the following theorem:

*Theorem II-E.1:* Let $u_1, \ldots, u_n$ be independent random variables such that

$$x_1 = a_1 \star u_1 \circ \ldots \circ a_n \star u_n$$
$$x_2 = b_1 \star u_1 \circ \ldots \circ b_n \star u_n$$

are independent, and where the operators $\star$ and $\circ$ satisfy the above conditions. Then, denoting $f$ the

function defined by the operator $\circ$, $f^{-1}(u_i)$ is Gaussian if $a_i b_i \neq 0$.

Practically, with such mixtures, the separation can be done in 3 steps:
- Apply $f^{-1}$ to the nonlinear observations for providing linear mixtures in $s_i = f^{-1}(u_i)$
- Solve the linear mixtures in $s_i$ by any BSS method
- Restore the actual independent sources by applying $u_i = f(s_i)$

If the function $f$ is known, it is very easy. If $f$ is not known, but if you suspect that the nonlinear mixtures has this particular form, you can constraint the separation structure to consist of identical nonlinear component-wise blocs (able to approximate $f_{-1}$) followed by a linear matrix $\mathbf{B}$ able to separate sources in linear mixtures, and followed itself by identical non-linear component-wise blocs (which approximate $f$) for restoring the actual sources. We remark that the 2 first blocs of the structure are identical (in fact, a little bit simpler, since all the nonlinear blocs are the same) to the separation structure of PNL mixtures. We can then estimate the independent distorted sources $s_i$ with a PNL algorithm. Then, after computing $f$ from the nonlinear bloc estimations, one can restore the actual sources.

Finally, one can remark that the PNL mixture is close to these mappings. It is in fact more general since the nonlinear functions $f_i$ can be different. Other examples of such mappings are given in [33], [25], but realistic mixtures belonging to this class seems unusual.

## III. Independence criterion

The previous section on separability points out that output independence, under a few structural constraints on the mixtures, is strong enough for driving the estimation of the separating structure $\mathcal{G}$. In this section, we propose to study in more details the key concept of source separation: the statistical independence.

### *A. Kullback-Leibler divergence*

Let $\mathbf{y}$ denote the output random vector, independence of $\mathbf{y}$ is defined by:

$$p_{\mathbf{y}}(\mathbf{u}) = \prod_i p_{y_i}(u_i) \tag{24}$$

but this relation between pdf's is not easily tractable. A more convenient (scalar) measure of independence is derived from the Kullback-Leibler divergence [3] which measures the similarity between two distributions $p$ and $q$ of the same variable $\mathbf{y}$:

$$KL(p \parallel q) = \int p(\mathbf{u}) \log \frac{p(\mathbf{u})}{q(\mathbf{u})} d\mathbf{u} \tag{25}$$

[3] it is not a distance, since it is not commutative

It can be shown that $KL(p \parallel q)$ is greater of equal to 0, and the equality holds if and only if $p = q$. Then, independence can be computed like the KL divergence between $p_{\mathbf{y}}$ and $\prod_i p_{y_i}$:

$$KL(p_{\mathbf{y}} \parallel \prod_i p_{y_i}) = \int p_{\mathbf{y}}(\mathbf{u}) \log \frac{p_{\mathbf{y}}(\mathbf{u})}{\prod_i p_{y_i}(u_i))} d\mathbf{u} \tag{26}$$

which vanishes if and only if $p_{\mathbf{y}} = \prod_i p_{y_i}$, *i.e.* $\mathbf{y}$ is independent. This quantity is often called mutual information and denoted $I(\mathbf{y})$ [20]. Defining $H(\mathbf{y})$ and $H(y_i)$ the joint and marginal entropies, respectively:

$$\begin{aligned} H(\mathbf{y}) &= -\int p_{\mathbf{y}}(\mathbf{u}) \log p_{\mathbf{y}}(\mathbf{u}) d\mathbf{u} \\ H(y_i) &= -\int p_{y_i}(u_i) \log p_{y_i}(u_i) du_i \end{aligned}$$

the mutual information can be written:

$$I(\mathbf{y}) = KL(p_{\mathbf{y}} \parallel \prod_i p_{y_i}) = \sum_i H(y_i) - H(\mathbf{y}) \tag{27}$$

### *B. Estimating equations*

Due to MI properties, estimation of the separation structure can then be obtained by minimizing MI. Although the general relation (27) can be directly derived for providing the estimation equations, usually, this equation is simplified taking into account the separation structure.

#### B.1 MI optimization for linear mixtures

For linear mixtures, $\mathcal{F}$ is a $n \times n$ invertible matrix $\mathcal{F} = \mathbf{A}$, and the separation structure is constrained to be a $n \times n$ invertible matrix $\mathbf{B}$. Since $\mathbf{y} = \mathbf{Bx}$ and $\mathbf{x} = \mathbf{As}$, the relation between $\mathbf{x}$ and $\mathbf{y}$ pdf's writes:

$$p_{\mathbf{y}}(\mathbf{u}) = p_{\mathbf{x}}(\mathbf{B}^{-1}\mathbf{u}) / \mid \det \mathbf{B} \mid \tag{28}$$

and the MI becomes:

$$I(\mathbf{y}) = \sum_i H(y_i) - H(\mathbf{x}) - \log \mid \det \mathbf{B} \mid \tag{29}$$

Then MI optimization with respect to the separation matrix $\mathbf{B}$ requires the gradient:

$$\frac{\partial I(\mathbf{y})}{\partial \mathbf{B}} = \frac{\partial}{\partial \mathbf{B}} \sum_i H(y_i) - \mathbf{B}^{-T} \tag{30}$$

Finally, since $y_i = \sum_i b_{ik} x_k$, the derivative of MI with respect to $\mathbf{B}$ is:

$$\frac{\partial I(\mathbf{y})}{\partial \mathbf{B}} = E\Psi_{\mathbf{y}} \mathbf{x}^T - \mathbf{B}^{-T} \tag{31}$$

where $E$ denotes mathematical expectation and $\Psi_{\mathbf{y}} = [\Psi_{y_1} \dots \Psi_{y_n}]^T$, whose component $\Psi_{y_i}$ is the score function of $y_i$ defined as:

$$\Psi_{y_i}(y_i) = -\frac{\partial \log p_{y_i}(y_i)}{\partial y_i} = -\frac{p'_{y_i}}{p_{y_i}}(y_i) \tag{32}$$

After right multiplication by $\mathbf{B}^T$, one gets the estimation equation:

$$E\Psi_{\mathbf{y}}\mathbf{y}^T - \mathbf{I} = 0 \tag{33}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix. The expression (33) points out the relevance of score functions. In fact, the complete knowledge on each estimated sources $y_i$ is summarized in the score function $\Psi_{y_i}$. Moreover, the score functions are related to the optimal statistics. In fact, if $y_i$ is zero-mean Gaussian with unit variance, its score function is $\Psi_{y_i}(y_i) = y_i$, and the estimating equation entries reduce to $Ey_iy_j = \delta_{ij}$, *i.e.* simple decorrelation equations (for $i \neq j$). Conversely, if $y_i$ is non Gaussian, its score function is a nonlinear function and estimating equation entries $E\psi_{y_i}(y_i)y_j = \delta_{ij}$ involve cancellation of higher (than 2) order statistics.

B.2 MI optimization for nonlinear mixtures

For nonlinear mixtures, we assume that $\mathcal{F}$ and $\mathcal{G}$ are both $n$-variate invertible mappings. Since $\mathbf{y} = \mathcal{G}(\mathbf{x})$ and $\mathbf{x} = \mathcal{F}(\mathbf{s})$, the relation between $\mathbf{x}$ and $\mathbf{y}$ pdf's writes:

$$p_{\mathbf{y}}(\mathbf{u}) = p_{\mathbf{x}}(\mathcal{G}^{-\infty}(\mathbf{u}))/ \mid \det \mathbf{J}_{\mathcal{G}} \mid \tag{34}$$

and the MI becomes:

$$I(\mathbf{y}) = \sum_i H(y_i) - H(\mathbf{x}) - \log \mid \det \mathbf{J}_{\mathcal{G}} \mid \tag{35}$$

In the special case of PNL mixtures, MI simplifies to:

$$I(\mathbf{y}) = \sum_i H(y_i) - H(\mathbf{x}) - \log \mid \Pi_i g_i'(x_i) \mid - \log \mid \det \mathbf{B} \mid \tag{36}$$

Consequently, MI minimization leads to two equations for estimating the nonlinear part and the nonlinear part of the separating structure. Equations related to linear part is similar to (33). Conversely, denoting $z_i = g_i(x_i)$, equations related to the nonlinear part is:

$$E[\sum_i b_{ij}\Psi_{y_i}(y_i) \mid z_j] = \Psi_{z_j}(z_j),\ j = 1, \ldots, n \tag{37}$$

Again, one remarks that optimal minimization of MI requires knowledge of the score functions.

B.3 Direct MI optimization

One also can optimize directly the MI $I(\mathbf{y}) = \sum_i H(y_i) - H(\mathbf{y})$ with respect to $\mathbf{B}$:

$$\frac{\partial I(\mathbf{y})}{\partial \mathbf{B}} = E\Psi_{\mathbf{y}}\mathbf{x}^T - E\Phi_{\mathbf{y}}\mathbf{x}^T = E\beta_{\mathbf{y}}\mathbf{x}^T \tag{38}$$

where

- $E$ denotes mathematical expectation,
- $\Psi_{\mathbf{y}} = [\Psi_{y_1} \ldots \Psi_{y_n}]^T$ whose component $\Psi_{y_i}$ is the marginal score function of $y_i$ defined as in (32)
- $\Phi_{\mathbf{y}}(\mathbf{y}) = [\Phi_1(\mathbf{y}) \ldots \Phi_n(\mathbf{y})]^T$ whose component $\Phi_i(\mathbf{y})$ is the joint score function of $\mathbf{y}$ defined as:

$$\Phi_i(\mathbf{y}) = -\frac{\partial \log p_{\mathbf{y}}(\mathbf{y})}{\partial y_i} \tag{39}$$

- $\beta_{\mathbf{y}} \triangleq \mathbf{\Psi}_{\mathbf{y}} - \mathbf{\Phi}_{\mathbf{y}}$ is the difference between marginal and joint score functions of $\mathbf{y}$, and we call it SFD (Score Function Difference).

After right multiplication by $\mathbf{B}^T$, one gets the estimation equation:

$$E\Psi_{\mathbf{y}}\mathbf{y}^T - \mathbf{I} = 0 \quad \text{or} \quad E\beta_{\mathbf{y}}\mathbf{y}^T = \mathbf{0} \tag{40}$$

B.4 MI optimization for linear convolutive mixtures

For linear convolutive mixtures, $\mathcal{F}$ is a $n \times n$ invertible matrix of filters $\mathcal{F} = \mathbf{A}(z)$. For instance, if $\mathbf{A}(z)$ is a finite impulse response filter of order $L$, the observation is:

$$\mathbf{x}(k) = [\mathbf{A}(z)]\mathbf{s}(k) = \sum_{l=0}^{L} \mathbf{A}_l\mathbf{s}(k-l) \tag{41}$$

The separation structure is then constrained to be a $n \times n$ invertible matrix of filters $\mathcal{G} = \mathbf{B}(z)$:

$$\mathbf{y}(k) = [\mathbf{B}(z)]\mathbf{x}(k) = \sum_{l=0}^{p} \mathbf{B}_l\mathbf{x}(k-l) \tag{42}$$

Unfortunately, due to the filter, there is no simple relationship between $\mathbf{x}$ and $\mathbf{y}$ pdf's. So, we have to derive the basic MI equation

$$I(\mathbf{y}) = \sum_i H(y_i) - H(\mathbf{y}) \tag{43}$$

In fact, it can be shown [5] that up to first order terms, we have:

$$I(\mathbf{y} + \mathbf{\Delta}) - I(\mathbf{y}) = E\mathbf{\Delta}^T\beta_{\mathbf{y}} \tag{44}$$

where $\mathbf{\Delta}$ stands for a 'small' random vector. This equation proposes that SFD is the stochastic gradient of the mutual information.

Moreover, it must be noted that signal independence in convolutive mixtures means stochastic process independence. Then, we have to take into account independence between delayed signals. For sake of simplicity, consider only the above term and derive it with respect to $\mathbf{B_k}$. To do that, let $\hat{\mathbf{B}}_k = \mathbf{B}_k + \varepsilon$, where $\varepsilon$ represents a 'small' matrix. Then we have:

$$\hat{\mathbf{y}}(n) \triangleq [\mathbf{B}(z)]\mathbf{x}(n) = \mathbf{y}(n) + \varepsilon\mathbf{x}(n-k) \tag{45}$$

Now by combining the above equation with (44), and doing a little algebra, we find that:

$$\frac{\partial I(\mathbf{y}(n))}{\partial \mathbf{B}_k} = E\beta_{\mathbf{y}}(\mathbf{y}(n))\,\mathbf{x}^T(n-k) \tag{46}$$

However, we need the gradient of the MI of the delayed versions of outputs, that is, the gradient of $I(y_1(n), y_2(n-m))$ (in the case of 2 mixtures of 2 sources). This gradient can be found with a similar approach, which is:

$$\frac{\partial I(y_1(n), y_2(n-m))}{\partial \mathbf{B}_k} = E\beta_m(n)\,\mathbf{x}^T(n-k) \quad (47)$$

where $\beta(n)$ is obtained by the following procedure:

$$\begin{pmatrix} y_1(n) \\ y_2(n) \end{pmatrix} \xrightarrow{\text{Shift}} \begin{pmatrix} y_1(n) \\ y_2(n-m) \end{pmatrix} \xrightarrow{\text{SFD}} \begin{pmatrix} \beta_1^\star(n) \\ \beta_2^\star(n) \end{pmatrix} \xrightarrow{\text{Shift back}} \begin{pmatrix} \beta_1^\star(n) \\ \beta_2^\star(n+m) \end{pmatrix} \triangleq \beta_m(n) \quad (48)$$

Hence, the complete set of estimating equations is:

$$E\beta_m(n)\,\mathbf{x}^T(n-k) = \mathbf{0} \qquad ,\forall k, m \quad (49)$$

It is equivalent to using the separation criterion:

$$J = \sum_m I(y_1(n), y_2(n-m)) \quad (50)$$

However, minimizing $J$ is very cost consuming. Hence, in practice, we use $I(y_1(n), y_2(n-m))$, but in each iteration a (randomly chosen) different $m$ is used [3].

B.5 MI optimization for convolutive PNL (CPNL) mixtures

The approaches of the sections III-B.2 and III-B.4 can be combined for separating convolutive PNL (CPNL) mixtures [4]. In these mixtures, a linear convolutive mixture is followed by component-wise invertible non-linearities (corresponding to sensors). The separation criterion is as (50). This results in the same estimating equation (49) for estimating the convolutive part, and the estimating equation:

$$E[\alpha_i \mid z_i] = 0, \qquad i = 1, 2 \quad (51)$$

for non-linear part. In this equation $\alpha = (\alpha_\mathbf{1}, \alpha_\mathbf{2})^\mathbf{T}$ is defined by:

$$\alpha(n) \triangleq \sum_{k=0}^{p} \mathbf{B}_k^T \beta_m(n+k) = \left[\mathbf{B}^T\left(\frac{1}{z}\right)\right]\beta_m(n) \quad (52)$$

where $\beta_\mathbf{m}(\mathbf{n})$ is given by (48).

*C. Criteria for constrained separation structures*

As we explained, constraining the separation structures can reduce the indeterminacies and even insuring separability. It also can have a direct influence on the independence criterion. In this section, we show how the MI is modified by two usual constraints.

C.1 Separating structure with pre-whitening

Many methods [15], [31] are based on the following decomposition of the separation matrix $\mathbf{B}$:

$$\mathbf{B} = \mathbf{UW} \quad (53)$$

where $\mathbf{W}$ is a whitening matrix and $\mathbf{U}$ is an orthogonal matrix. Let us denote $\mathbf{z} = \mathbf{Bx}$, $\mathbf{z}$ is whitened means that it satisfies $E\mathbf{zz^T} = \mathbf{I}$. The decomposition also forces the output variance to be equal to 1 (since $\mathbf{U}$ is orthogonal, $\mathbf{y} = \mathbf{Uz}$ satisfies $E\mathbf{yy^T} = \mathbf{I}$) which relaxes the scale indeterminacies. Then, the MI of the outputs can be written:

$$I(\mathbf{y}) = \sum_i H(y_i) - H(\mathbf{z}) - \log \mid \det \mathbf{U} \mid \quad (54)$$

Moreover, after computing (at the second order) the whitening matrix $\mathbf{W}$, the joint entropy $H(\mathbf{z})$ is a constant and, since the determinant of an orthogonal matrix is equal to 1, MI reduces to:

$$I(\mathbf{y}) = \sum_i H(y_i) + \text{cst} \quad (55)$$

Minimizing the MI is then equivalent to minimizing the sum of the marginal output entropies. It is well known that, for random variables with a given variance, the entropy is maximal for Gaussian variables. Consequently, since output variance is constant $E\mathbf{yy^T} = \mathbf{I}$, (55) can be seen as a measure of gaussianity. In other words, minimizing MI is equivalent to minimizing the gaussianity of the estimated sources.

C.2 Infomax

Another approach, initially proposed by Bell and Sejnowski [7], consists in computing transformed output signals $\mathbf{z}$, obtained by component-wise invertible nonlinear mapping $F_{y_i}$:

$$z_i = F_{y_i}(y_i),\ i = 1, \ldots, n \quad (56)$$

where $F_{y_i}$ is the cumulative probability density function of the random variable $y_i$. Since MI is not modified by component-wise invertible mappings, one can write:

$$I(\mathbf{z}) = I(\mathbf{y}) = \sum_i H(z_i) - H(\mathbf{z}) \quad (57)$$

Each transformed variable $z_i$ being uniformly distributed in $[0, 1]$, MI becomes:

$$I(\mathbf{z}) = I(\mathbf{y}) = \text{cst} - H(\mathbf{z}) \quad (58)$$

Consequently, minimizing the MI is equivalent to maximazing the joint entropy of the transformed outputs $\mathbf{z}$: the algorithm associated to this constraint is called Infomax.

*D. Estimating the estimation equations*

As we explained after deriving the gradient of MI for various mixtures in III-B, the key information concerning the estimated sources is their score functions, *i.e* the opposite of the derivatives of their log-densities. If the score functions are *a priori* chosen, the statistics involved in the estimating equations are not optimal.

However, for estimating the separation matrix (in linear mixtures), one can remark that a very accurate estimation of score functions is not necessary. In fact, the estimating equation (33) leads to the set of equations:

$$E\psi_{y_i}(y_i)y_j = 0,\ i \neq j \tag{59}$$

since diagonal terms of (33) are only normalization equations. If the $y_i$'s are statistically independent zero-mean variables (*i.e.* $Ey_j = 0,\ i = 1, \ldots, n$), it is clear that:

$$E\psi_{y_i}(y_i)y_j = E\psi_{y_i}(y_i)Ey_j = 0,\ i \neq j \tag{60}$$

even with a poor estimation equation of $\psi_{y_i}$. However, a coarse estimation of the score function leads to slower convergence or even to algorithm divergence.

Conversely, solving equation (37), associated to nonlinear part estimation, requires an accurate estimation of the score functions which appear both in left and right side terms on the equation. A comparison between 4-th order estimations obtained with a Gram-Charlier expansion and a MSE criterion is presented in [47]. It emphasizes on the weak accuracy of 4-th Gram-Charlier expansion, and explained the difficulties for separating hard nonlinear mixtures when deriving MI with this expansion [53] .

In any case, the score function estimation is very important for determining the optimal statistics and for designing efficient separation algorithms.

D.1 Estimating the score function from the density

Since the score function is related to the probability density function, a natural approach is to first estimate the pdf (or the logarithm of the pdf), and then to deduce the score function with:

$$\hat{\psi}_{y_i}(y_i) = \frac{\hat{p}'_{y_i}}{\hat{p}_{y_i}}(y_i) \tag{61}$$

Various methods for estimating pdf's are very usual in statistics. A first approach is based on Gram-Charlier or Edgeworth expansions of the density $p_{y_i}$ around Gaussian [34]. Without considering details, we can note that these expansions explicitely involve high-order cumulants.

Another approach is based on kernel estimator of pdf's [26]:

$$\hat{p}_X(u) = \frac{1}{T}\sum_{i=1}^{T} K_h(u - X_i) \tag{62}$$

where the $X_i$'s are samples of the random variable $X$, and $h$ is the bandwidth which determines the width of each kernel. Larger is $h$, smoother is the estimation. Of course, $h$ depends on the sample number $T$. Experimentally, it seems that an optimal choice of $h$, based on cost-computing cross-validation, is not required, and a simple rule [47] is efficient.

D.2 Direct estimation of score functions

Score functions can also be directly estimated for minimizing a mean square error. In fact, denoting $\hat{\psi}(\mathbf{w}, u)$ a parametric estimation of $\psi(u)$, and using the definition of the score function, the MSE criterion:

$$J(\mathbf{w}) = \frac{1}{2}E(\hat{\psi}(\mathbf{w}, u) - \psi(u))^2 \tag{63}$$

leads to the gradient:

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = E[\hat{\psi}(\mathbf{w}, u)\frac{\partial \hat{\psi}}{\partial \mathbf{w}}(\mathbf{w}, u) + \frac{\partial^2 \hat{\psi}}{\partial u \partial \mathbf{w}}(\mathbf{w}, u)] \tag{64}$$

The above gradient can be used for estimating any type of model, *e.g.* based on a basis of nonlinear functions [44], [12], on neural networks [47], or on other nonlinear models like splines, polynomials, etc.

*E. Contrast functions*

Constrast functions, initially defined by Donoho [24] for blind deconvolution, have been introduced by Comon [16], [17] in the framework of blind source separation.

E.1 Definition

Basically, a contrast function is a real function of a distribution $\mathbf{y}$, which should be minimal if $\mathbf{y}$ is independent. For linear mixtures, the definition is the following:

*Definition III-E.1:* A function $\phi[\mathbf{s}]$ of the distribution $\mathbf{s}$ is a contrast function if, for any matrix $\mathbf{C}$ and any random variable $\mathbf{s}$, $\phi[\mathbf{Cs}] \geq \phi[\mathbf{s}]$, with equality if and only if $\mathbf{C} = \mathbf{PD}$, where $\mathbf{P}$ and $\mathbf{D}$ are a permutation matrix and a diagonal matrix, respectively.

The above definition has been extended to convolutive mixtures [18] or nonlinear mixtures, $\mathbf{PD}$ being replaced by trivial filters $\mathbf{PD}(z)$ or by trivial nonlinear mappings $h_i(s_{\sigma(i)})$. Contrast functions are then good candidates for driving source separation algorithms, since they should be minimum when source separation is achieved.

Then, a main issue is to find contrast functions as simple as possible.

E.2 A few examples

First, it is easy to verify that MI is a contrast function. However, this function is complex enough, since it requires the estimation of score functions. Conversely, it is easy to see that $E\mathbf{y}\mathbf{y}^{\mathbf{T}}$ is not a contrast, which is not surprising since we know that second order statistics does not generally insure independence. In fact, adding high order statistics up to order 3 or 4, can lead to the simplest contrast functions. However, these contrast functions are often associated to restrictive conditions on the sources. For instance, the simple function:

$$\phi[\mathbf{y}] = \epsilon \sum_{i=1}^{n} \mid Cumy_i, y_i, y_i, y_i \mid \qquad (65)$$

is a contrast for sources with the same kurtosis sign $\epsilon$. If not, the algorithm diverges! Moreover, simpler contrasts can be derived under structural constraints. For instance, the decomposition of the separation matrix $\mathbf{B} = \mathbf{UW}$ leads to the class of orthogonal contrasts. An detailed and very interesting discussion on contrast functions can be found in [10].

*F. Performance*

In the case of linear mixtures, the choice of the contrast functions influence convergence speed, but asymptotic performance is weakly dependent of the contrast function, provided than the algorithm converges. On the contrary, Cardoso proved [10] that asymptotic performance is penalized by using a prewhitening.

## IV. Using time structure

When the sources are time signals, they may contain more structure than simple random variables. This time structure (correlation, non-stationarity, etc.) may be used for improving the estimation. It may even make the separation possible in cases where the basic ICA methods fail, for example, if the sources are Gaussian but correlated or non-stationary over time.

*A. Correlated sources*

The time dependence between successive samples of the signals may be explored in different manners. While simple second order approaches consider only the time-lagged covariance matrices, more complicated methods may consider the higher order time-lagged statistics or even the joint pdf of successive time samples, for example using Markov models.

A.1 Second order approaches

It is known that the instantaneous (zero-lagged) covariance matrix does not contain enough parameters for solving the ICA problem up to classical indeterminacies. If $\mathbf{x}(t)$ is the vector of observations (mixtures), there exists an infinity of different matrices $\mathbf{V}$ so that the components of the vector $\mathbf{y}(t) = \mathbf{V}\mathbf{x}(t)$ are decorrelated. This is why in basic ICA, we must use higher order statistics. However, if the sources are correlated in time, using the time-lagged covariance matrices,

$$\mathbf{C}_{\tau}^{\mathbf{x}} = E\left\{\mathbf{x}(t)\mathbf{x}(t-\tau)^T\right\}, \qquad (66)$$

we can obtain the complementary information for separating the sources without using higher order statistics. In the simplest case, we can find a separating matrix which diagonalizes both the instantaneous and the first lagged covariance matrices. It has been shown that such a matrix reconstructs the sources up to classical indeterminacies, if the first lagged covariances are different for all the sources. The algorithms presented in [49] and [39] are examples of first lagged covariance diagonalization algorithms. The method may be extended by using the joint diagonalization of several time lagged covariances [9], [55], [22]. A good review about second order approaches can be found in chapter 18 of [30]. The main drawback of these methods is that they are not able to separate the sources with same power spectra.

A.2 Markov models

Another possibility is to exploit the complete time dependence structure of the sources, which is modeled as $q$-th order Markov sequences for simplicity. Such models can be completely characterized by the joint probability density of $q+1$ successive samples of each source. Then, a quasi maximum likelihood (ML) method may be used for estimating the separating matrix $\mathbf{B}$.

Basically, the ML method consists in maximizing the joint pdf of all the $T$ samples of all the components of the vector $\mathbf{x}$ (all the observations), with respect to $\mathbf{B}$. We denote this pdf:

$$f(x_1(1), \cdots, x_n(1), \cdots\cdots, x_1(T), \cdots, x_n(T)) \qquad (67)$$

Under the assumption of source independence, this function is equal to:

$$\left(\frac{1}{|\det(\mathbf{B}^{-1})|}\right)^T \prod_{i=1}^{n} f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(1), \cdots, \mathbf{e}_i^T\mathbf{B}\mathbf{x}(T)) \qquad (68)$$

where $f_{s_i}(.)$ represents the joint density of $T$ samples of the source $s_i$ and $\mathbf{e}_i$ is the $i$-th column of the identity matrix. We suppose now that the sources are $q$-th order Markov sequences, *i.e.*:

$$f_{s_i}(s_i(t)|s_i(t-1), \cdots, s_i(1)) = f_{s_i}(s_i(t)|s_i(t-1), \cdots, s_i(t-q)) \qquad (69)$$

Using (69), equation (68) reduces to:

$$(\frac{1}{|\det(\mathbf{B}^{-1})|})^T \prod_{i=1}^{n}[f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(1),\cdots,\mathbf{e}_i^T\mathbf{B}\mathbf{x}(q)) \prod_{t=q+1}^{T} f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t)|\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-1),\cdots, \mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-q))] \quad (70)$$

Taking the logarithm of (70), one obtains the log-likelihood function which must be maximized to estimate the separating matrix $\mathbf{B}$:

$$L_1 = T\log(|\det(\mathbf{B})|) + \sum_{i=1}^{n}[\log(f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(1),\cdots, \mathbf{e}_i^T\mathbf{B}\mathbf{x}(q))) + \sum_{t=q+1}^{T}\log(f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t)|\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-1) ,\cdots,\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-q)))] \quad (71)$$

In practice, the density functions of the true sources, $f_{s_i}$ in (71), are unknown and must be replaced with the estimated density functions of reconstructed sources, $f_{y_i}$. Then, the criterion (71) becomes asymptotically:

$$L_2 = -\log|\det(\mathbf{B})| - \sum_{i=1}^{n} E[\log f_{y_i}(y_i(t)|y_i(t-1), \cdots, y_i(t-q))] \quad (72)$$

To estimate the matrix $\mathbf{B}$, we need to compute the gradient of the criterion (72) with respect to $\mathbf{B}$:

$$\frac{\partial L_2}{\partial \mathbf{B}} = -\mathbf{B}^{-T} - E[\frac{\partial}{\partial \mathbf{B}}\sum_{i=1}^{n}\log f_{y_i}(y_i(t)|y_i(t-1) ,\cdots,y_i(t-q))] \quad (73)$$

Using some computations, the following estimating equations can be derived (for $i \neq j$):

$$E[\sum_{l=0}^{q}\psi_{y_i}^{(l)}(y_i(t)|y_i(t-1),\cdots,y_i(t-q))y_j(t-l)] = 0 \quad (74)$$

which determine $\mathbf{B}$ up to a scaling and a permutation. In these equations, $\psi_{y_i}^{(l)}(y_i(t)|y_i(t-1),\cdots,y_i(t-q)) = -\frac{\partial}{\partial y_i(t-l)}\log f_{y_i}(y_i(t)|y_i(t-1),\cdots,y_i(t-q))$ is the $l$-th component of the conditional score function, which is a vector of size $q+1$. The conditional score functions are estimated using a kernel method.

The algorithm is theoretically able to separate every linear mixture unless there are at least two Gaussian sources with same spectral densities. It provides an asymptotically efficient (unbiased, minimum variance) estimation of the separating matrix. It is however rather slow because estimating the conditional score functions is time-consuming.

### *B. Non stationary sources*

The non stationarity of the sources can be also exploited to achieve separation. In particular, the non stationarity of the source variances has been used for separating Gaussian or non Gaussian sources with no temporal correlation. The simplest approach consists in considering the second order statistics, *i.e.* correlation between different sources. It has been shown [38] that if $\sigma_i^2/\sigma_j^2$ are not constant for $\forall i \neq j$, a matrix $\mathbf{B}$ so that the components of $\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t)$ are uncorrelated *at every time instant* $t$, is a separating matrix. A practical algorithm consists in estimating the instantaneous covariance matrix on short time intervals over which the signals are supposed stationary. Because of the non stationarity of the sources, this matrix depends on the interval label. Thus, an adaptive algorithm may be used for diagonalizing this matrix on all the intervals. Other algorithms based on the non stationarity of the source variances can be found in the literature [8], [43]. From a theoretical point of view, in [43], Pham and Cardoso derived a Gaussian form of the MI in that case and proved very interesting properties of the criterion and of the associated algorithms. Practically, the algorithms are still based on joint diagonalization of covariance matrices, computed on successive time windows.

## V. Algorithms

Many algorithms have been developed in the literature since 1985 (see in [30] for a review and most of the references). However, a few principles are very important and will be pointed out in this section.

### *A. Equivariant algorithms*

The convergence speed of the first algorithms [28], [14] was depending on the mixture $\mathbf{A}$. For overcoming this problem, Cichocki *et al.* proposed the first robust (equivariant [4]) algorithm [13]. Independently, instead additive algorithms of the form $\mathbf{B} \leftarrow \mathbf{B} - \mu\frac{\partial J(\mathbf{y})}{\partial \mathbf{B}}$, Cardoso introduced algorithms (for separating linear mixtures) based on multiplicative updates: $\mathbf{B} \leftarrow (\mathbf{I} - \mu\nabla_{\mathbf{B}}(J))\mathbf{B}$. Left-multiplying this relation by $\mathbf{A}$ leads to:

$$\mathbf{C} \leftarrow (\mathbf{I} - \mu\nabla_{\mathbf{B}}(\mathbf{J}(\mathbf{C}\mathbf{s}))\mathbf{C} \quad (75)$$

which clearly does no longer depend on the mixing matrix $\mathbf{A}$. The algorithm is then based on the differential of $J((\mathbf{I}+\varepsilon)\mathbf{y})$, which has been derived independently by Cardoso and Laheld [11], and Amari

[4] The word "equivariant" has been introduced later by Cardoso and Laheld

*et al.* [2], under different names, relative and natural gradients, respectively.

### *B. Joint diagonalization*

In the previous section, we explained how, relaxing the i.i.d. assumption, with colored signals (first "i" no longer holds) or with non stationary sources ("i.d." non longer holds), MI (or ML) criterion leads to joint diagonalization of variance-covariance matrices. Then, a special interest has been focused on the designing of efficient joint diagonalization algorithms [52], [42], [54]. We especially recommend the algorithm proposed by Pham [42] for its simplicity, its high speed and its performance.

### *C. Minimizing Mutual Information*

Finally, deriving quasi-optimal [5] algorithms minimizing MI is possible. The algorithm is then based on (i) the estimations $\hat{\psi}_{y_i}$ of the score functions of the estimated sources $y_i$, (ii) the estimation of the separation structure with estimating equations where the true score functions $\psi_{y_i}$ are replaced by the estimations $\hat{\psi}_{y_i}$ [44], [12], [47]. An interactive demonstration of such algorithms, developed in Java (and the Matlab codes), is available on the Web page:

`http://www.lis.inpg.fr/demos/sep_sourc/ICAdemo/index.html.`

It shows the efficacy of these algorithms for separating linear as well as post-nonlinear mixtures and for achieving blind deconvolution and blind inversion of Wiener systems.

## VI. Conclusion

In this paper, we reviewed a few key points on the blind source separation and independent component analysis.
Concerning the separability, one has to remember that statistical independence cannot generally insures separation. In fact, a large class of non diagonal mapping preserves independence, and adding structural constraints is a good approach for restricting the solutions to trivial mappings. The methods are also strongly related to blind deconvolution or blind inversion of Wiener (nonlinear) systems: it is easy to see that the problems are very similar to source separation in linear or post-nonlinear mixtures, respectively [48].
These blind methods, based on statistical independence, are driven by the minimization of the mutual information (MI). Simple approximations of the MI leads to algorithms generally having a few restrictions. Conversely, approximations based on score function estimation can provide quasi-optimal algorithms.

However, this new and very active field of research emphasizes still many challenging questions. As examples, designing efficient methods for separating sources (i) from realistic MIMO convolutive mixtures, (ii) from noisy mixtures, (iii) from mixtures with more sources than sensors, (iv) from nonlinear mixtures, are among the most relevant. The development of independent component analysis (ICA) for sparsely representing complex data like speech or images [23], [51] (with a basis whose elements are as independent as possible) and for understanding how the brain could sparsely code such complex signals [6], [27], [40], is also an promising topics of research.
Finally, although promising applications has been developed in many fields (EEG and MEG signal processing, communications, smart sensor array design, "cocktail party" processing, etc.), a lot of works remains to do for considering these methods as an usual tools in signal processing or statistics toolboxes.

[5] 'quasi' since they used estimated score functions instead the true (unknown) ones

## References


[1] J. Aczel. *Lectures on functional equations and their applications*. Academic Press, New-York, 1966.

[2] S. Amari, A. Cichocki, and H. Yang. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, pages 757–763, Denver (Colorado), December 1996.

[3] M. Babaie-Zadeh, C. Jutten, and K. Nayebi. Separating convolutive mixtures by mutual information minimization. In *Proceedings of IWANN 2001*, Granada, Spain, 2001.

[4] M. Babaie-Zadeh, C. Jutten, and K. Nayebi. Separating convolutive post non-linear mixtures. In *Proceedings of ICA 2001*, San Diego (California, USA), 2001.

[5] M. Babaie-Zadeh, C. Jutten, and K. Nayebi. Differential of mutual information function. *IEEE Signal Processing Letters*, 2002. Submitted in May.

[6] H. B. Barlow. Unsupervised learning. *Neural Computation*, 1:295–311, 1989.

[7] A. Bell and T. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6), 1995.

[8] A. Belouchrani and M. Amin. Blind source separation based on time-frequency signal representation. *IEEE Trans. on Signal Processing*, 46(11):2888–2897, 1998.

[9] A. Belouchrani, K. Abed Meraim, J. F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on Signal Processing*, 45(2):434–444, 1997.

[10] J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings IEEE*, 9(10):2009–2025, 1998.

[11] J.-F. Cardoso and B. Laheld. Equivariant adaptive source separation. *IEEE Trans. on SP*, 44(12):3017–3030, 1996.

[12] N. Charkani and Y. Deville. Optimization of the asymptotic performance of time-doamin convolutive source separation algorithms. In *Proc. ESANN*, pages 273–278, Bruges, Belgium, April 1997.

[13] A. Cichoki, Unbehauen R., and E. Rummert. Robust learning algorithm for blind separation of signals. *Electronics Letters*, 30(17):1386–1387, 1994.

[14] P. Comon. Séparation de mélanges de signaux. In *Actes du XII ème colloque GRETSI*, pages 137–140, Juan-Les-Pins (France), Juin 1989.

[15] P. Comon. Separation of sources using higher-order cumulants. In *SPIE Vol. 1152 Advanced Algorithms and Architectures for Signal Processing IV*, San Diego (CA), USA, August 8-10 1989.

[16] P. Comon. Independent component analysis. In J.-L. Lacoume, M. A. Lagunas, and C. L. Nikias, editors, *In-*

*ternational Workshop on High Order Statistics*, pages 111–120, Chamrousse, France, July 1991.

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

[18] P Comon. Contrasts for multichannel blind deconvolution. *IEEE Signal Processing Letters*, 3(7):209–211, 1996.

[19] P. Comon, C. Jutten, and J. Hérault. Blind separation of sources, Part II: Statment problem. *Signal Processing*, 24(1):11–20, 1991.

[20] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications, 1991.

[21] G. Darmois. Analyse des liaisons de probabilité. In *Proceedings Int. Stat. Conferences 1947*, volume III A, page 231, Washington (D.C.), 1951.

[22] S. Degerine and R. Malki. Second order blind separation of sources based on canonical partial innovations. *IEEE Trans. on Signal Processing*, 48(3):629–641, 2000.

[23] D. L. Donoho. Nature vs. math.:interpreting independent component analysis in light of computational harmonic analysis. In *Proceedings of ICA 2000*, Helsinki (Finland), 2000.

[24] D.L. Donoho. On minimum entropy deconvolution. In *Applied Time Series Analysis II*, pages 565–608, Tulsa, 1980.

[25] J. Eriksson and V. Koivunen. Blind identifiability of class of nonlinear instantaneous ICA models. In *EUSIPCO 2002*, Toulouse (France), September 2002.

[26] W. Härdle. *Smoothing Techniques, with implementation in S*. Springer-Verlag, 1990.

[27] G. H. Harpur and R. W. Prager. Development of low entropy coding in a recurrent network. *Networks*, 7:277–284, 1996.

[28] J. Hérault, C. Jutten, and B. Ans. Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. In *Actes du X ème colloque GRETSI*, pages 1017–1022, Nice (France), 20-24 Mai 1985.

[29] S. Hosseini, C. Jutten, and D. T. Pham. Blind separation of temporally correlated sources using a quasi-maximum likelihood approach. In *Proceedings of ICA 2001*, San Diego (California, USA), 2001.

[30] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley, New York, 2001.

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

[32] C. Jutten and J. Hérault. Blind separation of sources, Part I: an adaptive algorithm based on a neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

[33] A.M. Kagan, Y.V. Linnik, and C.R. Rao. Extension of darmois-skitovic theorem to functions of random variables satisfying an addition theorem. *Communications in Statistics*, 1(5):471–474, 1973.

[34] M. Kendall and A. Stuart. *The Advanced Theory of Statistics, Distribution Theory*, volume 1. Griffin, 1977.

[35] T. Kohonen. *Self-Organizing Maps*. Springer, 1995.

[36] M.J. Korenberg and I.W. Hunter. The identification of nonlinear biological systems: LNL cascade models. *Biological Cybernetics*, 43(12):125–134, December 1995.

[37] E. Lukacs. A characterization of the Gamma distribution. *Ann. Math. Statist.*, (26):319–324, 1955.

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

[39] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlation. *Physical Review Letters*, 72:3634–3636, 1994.

[40] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural image. *Nature*, 381:607–609, 1996.

[41] A. Parashiv-Ionescu, C. Jutten, A.M. Ionescu, A. Chovet, and A. Rusu. High performance magnetic field smart sensor arrays with source separation. In *MSM 98*, pages 666–671, Santa Clara (California, USA), April 1998.

[42] D. T. Pham. Joint approximate diagonalization of positive definite hermitian matrices. *SIAM Journal on Matrix analysis and Application*, 2001. To appear.

[43] D. T. Pham and J. F. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. In *Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA'2000*, pages 187–193, Helsinki (Finland), 2000.

[44] D. T. Pham, Ph. Garat, and C. Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In *Proc. European Signal Processing Conf. EUSIPCO 92*, pages 771–774, Brussels (Belgium), August 1992.

[45] R. S. Popovic. *Hall-effect devices*. Adam Hilger, Bristol, 1991.

[46] S. Prakriya and D. Hatzinakos. Blind identification of lti-zmnl-lti nonlinear channel models. *IEEE trans. S.P.*, 43(12):3007–3013, December 1995.

[47] A. Taleb and C. Jutten. Source separation in post nonlinear mixtures. *IEEE Tr. on SP*, 47(10):2807–2820, 1999.

[48] A. Taleb, J. Sole i Casals, and C. Jutten. Quasi-nonparametric blind inversion of Wiener systems. *IEEE Tr. on Signal Processing*, 49(5):917–924, 2001.

[49] L. Tong and V. Soon. Indeterminacy and identifiability of blind identification. *IEEE Tr. on CS*, 38:499–509, May 1991.

[50] L. Tong, V. Soon, Y. Huang, and R. Liu. AMUSE: a new blind identification algorithm. In *Proc. ISCAS*, New Orleans (USA), 1990.

[51] J. H. Van Hateren and D. L. Ruderman. Independent component analysis of natural image sequences yields spatiotemporal filters similar to simple cells in primary visual cortex. *Proc. R. Soc. London B*, 1998.

[52] M. Wax and J. Sheinvald. A least-square approach to joint diagonalization. *IEEE Signal Processing Letters*, 4(2), 1997.

[53] H. H. Yang, S. Amari, and A. Cichocki. Information-theoritic approach to blind separation of sources in non-linear mixture. *Signal Processing*, 64(3):291–300, 1998.

[54] A. Yeredor. Approximate joint diagonalization using non-orthogonal matrices. In *Proc. Second Int. Workshop on Independent Component Analysis and Blind Signal Separation*, pages 33–38, Helsinki, Finland, 1990.

[55] A. Ziehe and K. R. Müller. TDSEP: an efficient algorithm for blind separation using time structure. In *Int. Conf. on Artificial Neural Networks (ICANN'98)*, pages 675–680, Skvde (Sweden), 1998.