THIS IS *VERY CLOSE* THE OFFICIAL VERSION, PUBLISHED AS: PROCEEDINGS OF THE IEEE, VOL. 9, NO 10, PP. 2009-2025, OCT. 1998 1

Blind signal separation: statistical principles

Jean-François Cardoso, C.N.R.S. and E.N.S.T.
cardoso@tsi.enst.fr and http://tsi.enst.fr/~cardoso.html

Abstract— **Blind signal separation (BSS) and independent component analysis (ICA) are emerging techniques of array processing and data analysis, aiming at recovering unobserved signals or 'sources' from observed mixtures (typically, the output of an array of sensors), exploiting only the assumption of mutual independence between the signals. The weakness of the assumptions makes it a powerful approach but requires to venture beyond familiar second order statistics. The objective of this paper is to review some of the approaches that have been recently developed to address this exciting problem, to show how they stem from basic principles and how they relate to each other.**

Keywords— **Signal separation, blind source separation, independent component analysis.**

I. Introduction

Blind signal separation (BSS) consists in recovering unobserved signals or 'sources' from several observed mixtures. Typically, the observations are obtained at the output of a set of sensors, each sensor receiving a different combination of the 'source signals'. The adjective 'blind' stresses the fact that **i)** the source signals are not observed and **ii)** no information is available about the mixture. This is a sound approach when modeling the transfer from the sources to the sensors is too difficult; it is unavoidable when no *a priori* information is available about the transfer. The lack of *a priori* knowledge about the mixture is compensated by a statistically strong but often physically plausible assumption of *independence* between the source signals. The so-called 'blindness' should not be understood negatively: the weakness of the prior information is precisely the strength of the BSS model, making it a versatile tool for exploiting the 'spatial diversity' provided by an array of sensors. Promising applications can already be found in the processing of communications signals *e.g.* [24], [64], [68], [6], biomedical signals[1] like ECG [31] and EEG [51], [47] monitoring [38], [36], or as an alternative to principal component analysis, see *e.g.* [47], [10], [53], [7].

The simplest BSS model assumes the existence of n independent signals $s_1(t), \ldots, s_n(t)$ and the observation of as many mixtures $x_1(t), \ldots, x_n(t)$, these mixtures being linear and instantaneous, *i.e.* $x_i(t) = \sum_{j=1}^{n} a_{ij} s_j(t)$ for each $i = 1, n$. This is

[1]See the ICA page of the CNL group at http://www.cnl.salk.edu/~tewon/ica_cnl.html for [1] several biomedical applications.

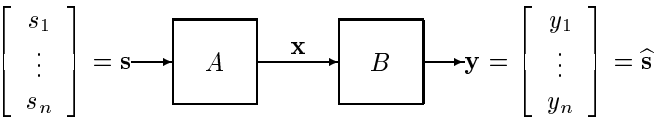


Fig. 1. Mixing and separating. Unobserved signals: $\mathbf{s}$; observations: $\mathbf{x}$, estimated source signals: $\mathbf{y}$.

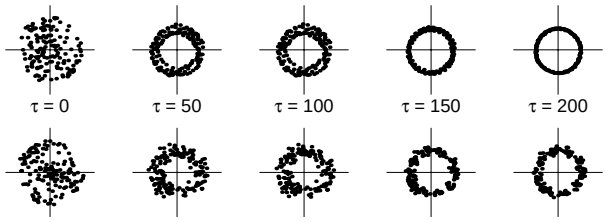


Fig. 2. Outputs $y_1(t)$ (top row) and $y_2(t)$ (bottom row) when using the separating matrix obtained after adaptation based on $0, 50, 100, 150, 200$ samples of a 2×2 mixture of constant modulus signals. Each subplot is in the *complex* plane: the clustering around circles shows the restoration of the constant modulus property.

compactly represented by the mixing equation

$$\mathbf{x}(t) = A\mathbf{s}(t) \tag{1}$$

where $\mathbf{s}(t) = [s_1(t), \ldots, s_n(t)]^\dagger$ is an $n \times 1$ column vector collecting the source signals, vector $\mathbf{x}(t)$ similarly collects the n observed signals and the square $n \times n$ 'mixing matrix' A contains the mixture coefficients. Here as in the following, † denotes transposition. The BSS problem consists in recovering the source vector $\mathbf{s}(t)$ using only the observed data $\mathbf{x}(t)$, the assumption of independence between the entries of the input vector $\mathbf{s}(t)$ and possibly some *a priori* information about the probability distribution of the inputs. It can be formulated as the computation of an $n \times n$ 'separating matrix' B whose output $\mathbf{y}(t)$

$$\mathbf{y}(t) = B\mathbf{x}(t) \tag{2}$$

is an estimate of the vector $\mathbf{s}(t)$ of the source signals.

Figure 2 shows an example of adaptive separation of (real) digital communications signals: a two-sensor array collects complex-valued noisy mixtures of two 'sources signals' which both have a constant modulus envelope. Successful separation upon adaptation is evidenced by the restoration of the constant modulus at each output. In figure 2, the underlying BSS algorithm optimizes a cost function composed of two penalty terms: one for correlation between outputs and one for deviation of

the modulus from a constant value. This example introduces several points to be developed below:

- A penalty term involving only pairwise decorrelation (second order statistics) would not lead to separation: source separation must go beyond second-order statistics (see section II);
- Source separation can be obtained by optimizing a 'contrast function' *i.e.* a scalar measure of some 'distributional property' of the output $\mathbf{y}$. The constant modulus property is very specific; more general contrast functions are based on other measures: entropy, mutual independence, high-order decorrelations, divergence between the *joint* distribution of $\mathbf{y}$ and some model,.... Contrast functions are discussed in sec. III where we show how they relate to each other and can be derived from the maximum likelihood principle.
- Fast adaptation is possible, even with simple algorithms (see secs. IV and V) and blind identification can be accurate even with a small number of samples (see sec. VI on performance analysis).

The basic BSS model can be extended in several directions. Considering, for instance, more sensors than sources, noisy observations, complex signals and mixtures, one obtains the standard narrow band array processing/beam-forming model. Another extension is to consider convolutive mixtures: this results in a multichannel blind deconvolution problem. These extensions are of practical importance, but this paper is restricted to the simplest model: real signals, as many sensors as sources, non-convolutive mixtures, noise free observations because it captures the essence of the BSS problem and because our objective is to present the basic statistical ideas, focusing on principles. Some pointers are nonetheless provided in the last section to papers addressing more general models.

The paper is organized as follows: section II discusses blind identifiability; section III and IV present contrast functions and estimating functions, starting from information-theoretic ideas and moving to suboptimal high order approximations; adaptive algorithms are described in section V; section VI addresses some performance issues.

II. Can it be done? Modeling and identifiability.

When is source separation possible? To which extent can the source signals be recovered? What are the properties of the source signals allowing for partial or complete blind recovery? These issues are addressed in this section.

A. The BSS model

Source separation exploits primarily 'spatial diversity', that is the fact that different sensors receive different mixtures of the sources. Spectral diversity, if it exists, could also be exploited but the approach of source separation is essentially 'spatial': looking for structure across the sensors, not across time. The consequence of ignoring any time structure is that the information contained in the data is exhaustively represented by the sample distribution of the observed vector $\mathbf{x}$ (as graphically depicted in fig. 3 for instance). Then, BSS becomes *the problem of identifying the probability distribution of a vector* $\mathbf{x} = A\mathbf{s}$ *given a sample distribution.* In this perspective, the statistical model has two components: the mixing matrix A and the probability distribution of the source vector $\mathbf{s}$.

- *Mixing matrix.* The mixing matrix A is the parameter of interest. Its columns are assumed to be linearly independent (see [14] for the discussion of a more general case) so that it is invertible.

There is something special about having an invertible matrix as the unknown parameter, because matrices represent linear transformations. Indeed, model (1) is a particular instance of a *transformation model.* Furthermore, the set of all $n \times n$ invertible matrices forms a multiplicative group. This simple fact has a profound impact on source separation because it allows to design algorithms with uniform performance *i.e.* whose behavior is completely independent of the particular mixture (sec. V-A and sec. VI-C).

- *Source distribution.* The probability distribution of each source is a 'nuisance parameter': it means that we are not primarily interested in it, even though knowing or estimating these distributions is necessary to estimate *efficiently* the parameter of interest. Even if we say nothing about the distribution of each source, we say a lot about their *joint* distribution by the key assumption of mutual *source independence.* If each source $i = 1, n$ is assumed to have a probability density function (pdf) denoted $q_i(\cdot)$, the independence assumption has a simple mathematical expression: the (joint) pdf $q(\mathbf{s})$ of the source vector $\mathbf{s}$ is:

$$q(\mathbf{s}) = q_1(s_1) \times \cdots \times q_n(s_n) = \prod_{i=1,n} q_i(s_i). \quad (3)$$

i.e. it is the product of the densities for all sources (the 'marginal' densities). Source separation techniques differ widely by the (explicit or implicit) assumptions made on the individual distributions of the sources. There is a whole range of options:

1. The source distributions are known in advance.
2. Some features are known (moments, heavy tails, bounded support,...)
3. They belong to a parametric family.
4. No distribution model is available.

A priori, the stronger the assumption, the narrower the applicability. However, well designed approaches are in fact surprisingly robust even to gross errors in modeling the source distributions, as shown below. For ease of exposition, zero mean

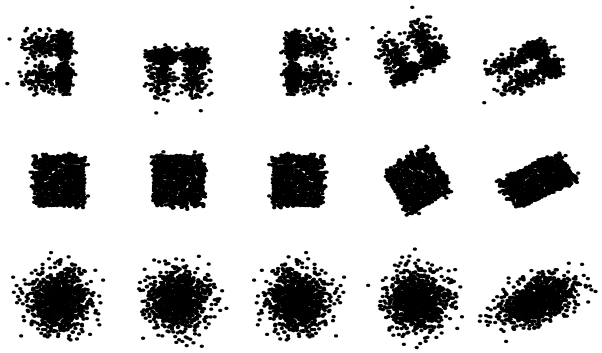


Fig. 3. Sample distributions of (x_1, x_2) when $\mathbf{x} = A\mathbf{s}$ for 5 different transformation matrices A, and 3 pairs of distributions for (s_1, s_2). From left to right: the identity transform, permutation of the sources, sign change, a $\pi/3$ rotation, a 'generic' linear transform.

sources are assumed throughout:

$$\mathrm{E}\mathbf{s} = 0 \quad i.e. \quad \mathrm{E}s_i = 0 \quad 1 \le i \le n. \tag{4}$$

B. Blind identifiability

The issue of blind identifiability is to understand to which extent matrix A is determined from the sole distribution of the observed vector $\mathbf{x} = A\mathbf{s}$. The answer depends on the distribution of $\mathbf{s}$ *and* on what is known about it.

A square matrix is said to be *non-mixing* if it has one and only one non-zero entry in each row and each column. If C is non-mixing then $\mathbf{y} = C\mathbf{s}$ is a *copy* of $\mathbf{s}$ *i.e.* its entries are identical to those of $\mathbf{s}$ up to permutations and changes of scales and signs. Source separation is achieved if such a copy is obtained. When the distribution of $\mathbf{s}$ is unknown, one cannot expect to do any better than signal copy but the situation is a bit different if some prior information about the distribution of $\mathbf{s}$ is available: if the sources have distinct distributions, a possible permutation can be detected; if the scale of a given source is known, the amplitude of the corresponding column of A can be estimated, etc...

Some intuition about identifiability can be gained by considering simple examples of 2×2 mixing. Each row of figure 3 shows (sample) distributions of a pair (s_1, s_2) of independent variables after various linear transforms. The columns successively show (s_1, s_2), (s_2, s_1), $(-s_1, -s_2)$ and the effect of a $\pi/3$ rotation and of a nondescript linear transform. Visual inspection of the transformed distribution compared to the original one gives a feeling of how well the transform matrix A can be identified based on the observation of a mixture. The first row of fig. 3 shows a case where the second column of A can be identified only up to sign because s_2 is symmetrically distributed about the origin (and therefore has the same distribution as $-s_2$). The second row shows a more severe indetermination: there, s_1 and s_2 have the same symmetric distribution, the transform can be determined only up to arbitrary changes of sign and a permutation. The last row shows the most severe case: there s_1 and s_2 are normally distributed with equal variance so that their joint distribution is invariant under rotation.

These simple examples suggest that A can be blindly identified indeed —possibly up to some indeterminations induced by the symmetries in the distribution of the source vector— in the case of known source distributions. However, this knowledge is not necessary: the eye certainly can capture the distortion in the last columns of figure 3 even without reference to the undistorted shapes in first column. This is because the graphical 'signature of independence' (the pdf shape in the first column) clearly appears as distorted in the last colum. This intuition is supported by the following statement (adapted from Comon [26] after a theorem of Darmois. See also [14]). *For a vector* $\mathbf{s}$ *of independent entries with at most one Gaussian entry and for any invertible matrix* C*, if the entries of* $\mathbf{y} = C\mathbf{s}$ *are independent, then* $\mathbf{y}$ *is a copy of* $\mathbf{s}$ *(C is non-mixing).* Thus, unless a linear transform is non-mixing, it turns a vector of independent entries (at most one being Gaussian) into a vector whose entries are *not* independent. This is a key result because it entails that blind signal separation can be achieved by restoring statistical independence. This is not only a theoretical result about blind identifiability: it also suggests that BSS algorithms could be devised by maximizing the independence between the outputs of a separating matrix. Section III shows that the maximum likelihood principle does support this idea and leads to a specific measure of independence.

Independence and decorrelation. Blind separation can be based on independence but independence can *not* be reduced to the simple decorrelation conditions that $\mathrm{E}y_i y_j = 0$ for all pairs $1 \le i \ne j \le n$. This is readily seen from the fact that there are, by symmetry, only $n(n-1)/2$ such conditions (one for each pair of sources) while there are n^2 unknown parameters.

Second order information (decorrelation), however, can be used to reduce the BSS problem to a simpler form. Assume for simplicity that the source signals have unit variance so that their covariance matrix is the identity matrix: $\mathrm{E}\mathbf{s}\mathbf{s}^\dagger = I$; vector $\mathbf{s}$ is said to be *spatially white*. Let W be a 'whitening matrix' for $\mathbf{x}$, that is $\mathbf{z} \triangleq W\mathbf{x}$ is spatially white. The composite transform WA necessarily is a *rotation matrix* because it relates two spatially white vectors $\mathbf{s}$ and $\mathbf{z} = WA\mathbf{s}$. Therefore, 'whitening' or 'sphering' the data reduces the mixture to a rotation matrix. It means that a separating matrix can be found as a product $B = UW$ where W is a

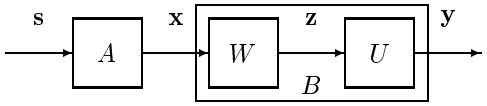


Fig. 4. Decorrelation leaves an unknown rotation.

whitening matrix and U is a rotation matrix. Note that any further rotation of $\mathbf{z}$ into $\mathbf{y} = U\mathbf{z}$ preserves spatial whiteness, so that two equivalent approaches to exploiting source decorrelation are **i)** find B as $B = UW$ with W a spatial whitener and U a rotation or **ii)** find B under the *whiteness constraint*: $\mathrm{E}\mathbf{y}\mathbf{y}^\dagger = I$. For further reference, we write the whiteness constraint as

$$\mathrm{E}H_w(\mathbf{y}) = 0 \qquad \text{where} \qquad H_w(\mathbf{y}) \triangleq \mathbf{y}\mathbf{y}^\dagger - I. \quad (5)$$

Spatial whiteness imposes $n(n+1)/2$ constraints, leaving $n(n-1)/2$ unknown (rotation) parameters to be determined by other than second order information: second order information is able to do 'about half the BSS job'.

The prewhitening approach is sensible from an algorithmic point of view but it is not necessarily statistically efficient (see sec. VI-B). Actually, enforcing the whiteness constraint amounts to believe that second order statistics are 'infinitely more reliable' than any other kind of statistics. This is, of course, untrue.

C. Likelihood

This section examines in a simple graphical way the likelihood of source separation models. The likelihood, in a given model, is the probability of a data set as a function of the parameters of the model. The simple model $\mathbf{x} = A\mathbf{s}$ for vector $\mathbf{x}$ discussed in sec. II-A is parameterized by the pair (A, q) made of the mixing matrix A and of the density q for the source vector $\mathbf{s}$. The density of $\mathbf{x} = A\mathbf{s}$ for a given pair (A, q) is classically given by

$$p(\mathbf{x}; A, q) = |\det A|^{-1} q(A^{-1}\mathbf{x}) \quad (6)$$

If T samples $X_{1:T} \triangleq [\mathbf{x}(1), \ldots, \mathbf{x}(T)]$ of $\mathbf{x}$ are modeled as independent, then $p(X_{1:T}) = p(\mathbf{x}(1)) \times \cdots \times p(\mathbf{x}(T))$. Thus the normalized (*i.e.* divided by T) log-likelihood of $X_{1:T}$ for the parameter pair (A, q) is

$$\frac{1}{T}\log p(X_{1:T}; A, q) = \frac{1}{T}\sum_{t=1}^{T} \log q(A^{-1}\mathbf{x}(t)) - \log|\det A|. \quad (7)$$

Figures 5 to 7 show the 'likelihood landscape' when A is varied while q is kept fixed. For each figure, $T = 1000$ independent realizations of $\mathbf{s} = [s_1, s_2]^\dagger$ are drawn according to some pdf $r(\mathbf{s}) = r_1(s_1)r_2(s_2)$ and are mixed with a 2×2 matrix $\underline{A}$ to produce T samples of $\mathbf{x}$. Therefore, this data set

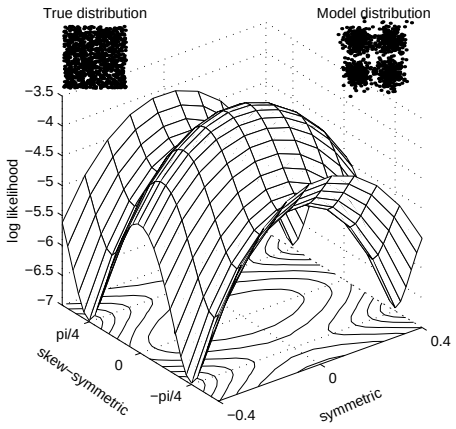


Fig. 5. Log-likelihood with a slightly misspecified model for source distribution: maximum is reached close to the true value.

$X_{1:T}$ follows exactly model (1) with a 'true mixing matrix' $\underline{A}$ and a 'true source distribution' $r(\mathbf{s})$. The figures show the log-likelihood when A is varied around its true value $\underline{A}$ while model density $q(\mathbf{s})$ is kept fixed. These figures illustrate the impact of the choice of a particular model density.

In each of these figures, the matrix parameter A is varied in two directions in matrix space according to $A = \underline{A}M(u, v)$ where $M(u, v)$ is the matrix

$$M(u,v) \triangleq \begin{bmatrix} \cosh u & \sinh u \\ \sinh u & \cosh u \end{bmatrix} \cdot \begin{bmatrix} \cos v & -\sin v \\ \sin v & \cos v \end{bmatrix}. \quad (8)$$

This is just a convenient way to generate a neighborhood of the identity matrix. For small u and v:

$$M(u,v) \approx I + u\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + v\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}. \quad (9)$$

Therefore u and v are called symmetric and skew-symmetric parameters respectively. Each one controls a particular deviation of $M(u, v)$ away from the identity.

In fig. 5, the true source distributions r_1 and r_2 are uniform on $[-1, +1]$ but the model takes q_1 and q_2 to be each a mixture of two normal distributions with same variance but different means (as in second column of fig. 11). True and hypothesized sample distributions of $\mathbf{s} = (s_1, s_2)$ are displayed in upper left and right corners of the plot. Even though an incorrect model is used for the source distribution: $q \neq r$, the figure shows that the likelihood is maximal around $(u, v) = (0, 0)$ *i.e.* the most likely mixing matrix given the data and the model is close to $\underline{A}$.

In fig. 6, the true sources are 'almost binary' (see upper left corner) but a Gaussian model is used: $q_1(s) = q_2(s) \propto \exp -s^2/2$. The figure shows that the likelihood of $A = \underline{A}M(u, v)$ does not depend on

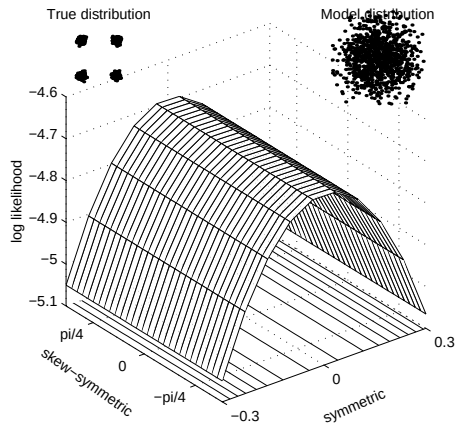


Fig. 6. Log-likelihood with a Gaussian model for source distribution: no 'contrast' in the skew-symmetric direction.

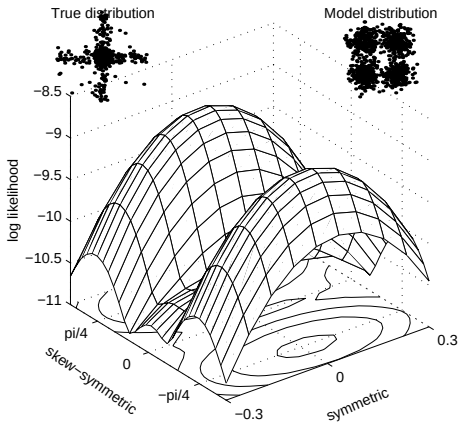


Fig. 7. Log-likelihood with a widely misspecified model for source distribution: maximum is reached for a mixing system.

the skew-symmetric parameter v, again evidencing the insufficiency of Gaussian modelling.

In fig. 7, the source are modeled as in fig. 5 but the true (and identical) source distributions r_1 and r_2 now are mixtures of normal distributions with the same mean but different variances (as in second column of fig. 11). A disaster happens: the likelihood is no longer maximum for A in the vicinity of $\underline{A}$. Actually, if the value $\widehat{A}$ of A maximizing the likelihood is used to estimate the source signals as $\mathbf{y} = \widehat{s} = \widehat{A}\mathbf{x}$, one obtains maximally mixed sources! This is explained in section III-A and fig. 8.

The bottom line of this informal study is the necessity of non-Gaussian modeling (fig. 6); the possibility of using only an approximate model of the sources (fig. 5); the existence of a limit to the misspecification of the source model (fig. 7). How wrong can the source distribution model be? This is quantified in section VI-A.

III. Contrast functions

This section introduces 'contrast functions' which are objective functions for source separation. The maximum likelihood principle is used as a starting point, suggesting several information-theoretic objective functions (sec. III-A) which are then shown to be related to another class of objective functions based on high-order correlations (sec. III-B).

Minimum contrast estimation is a general technique of statistical inference [58] which encompasses several techniques like maximum likelihood or least squares. It is relevant for blind deconvolution (see the inspiring paper [37] and also [12]) and has been introduced in the related BSS problem by Comon [26]. In both instances, a *contrast function* is a real function of a *probability distribution*. To deal with such functions, a special notation will be useful: for $\mathbf{x}$ a given random variable, $f(\mathbf{x})$ generically denotes a function of $\mathbf{x}$ while $f[\mathbf{x}]$ denotes a function *of the distribution* of $\mathbf{x}$. For instance, the mean value of $\mathbf{x}$ is denoted $m[\mathbf{x}] \triangleq \mathrm{E}\mathbf{x}$.

Contrast functions for source separation (or 'contrasts', for short) are generically denoted $\phi[\mathbf{y}]$. They are real valued functions of the distribution of the output $\mathbf{y} = B\mathbf{x}$ and they serve as objectives: they must be designed in such a way that source separation is achieved when they reach their minimum value. In other words, a valid contrast function should, for any matrix C, satisfy $\phi[C\mathbf{s}] \geq \phi[\mathbf{s}]$ with equality only when $\mathbf{y} = C\mathbf{s}$ is a copy of the source signals. Since the mixture can be reduced to a rotation matrix by enforcing the whiteness constraint $\mathrm{E}\mathbf{y}\mathbf{y}^\dagger = I$ (sect. II-B), one can also consider 'orthogonal contrast functions': these are denoted $\phi^\circ[\mathbf{y}]$ and must be minimized under the whiteness constraint $\mathrm{E}\mathbf{y}\mathbf{y}^\dagger = I$.

A. Information theoretic contrasts

The maximum likelihood (ML) principle leads to several contrasts which are expressed via the *Kullback divergence*. The Kullback divergence between two probability density functions $f(\mathbf{s})$ and $g(\mathbf{s})$ on $\mathbb{R}^n$ is defined as

$$\mathbf{K}(f|g) \triangleq \int_{\mathbf{s}} f(\mathbf{s}) \log\left(\frac{f(\mathbf{s})}{g(\mathbf{s})}\right) \, d\mathbf{s} \qquad (10)$$

whenever the integral exists [28]. The divergence between the distributions of two random vectors $\mathbf{w}$ and $\mathbf{z}$ is concisely denoted $\mathbf{K}[\mathbf{w}|\mathbf{z}]$. An important property of $\mathbf{K}$ is that $\mathbf{K}[\mathbf{w}|\mathbf{z}] \geq 0$ with equality if and only if $\mathbf{w}$ and $\mathbf{z}$ have the same distribution. Even though $\mathbf{K}$ is not a distance (it is not symmetric), it should be understood as a 'statistical way' of quantifying the closeness of two distributions.

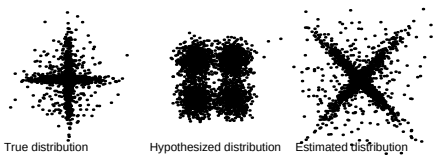


Fig. 8. How the maximum likelihood estimator is misled.

A.1 Matching distributions: likelihood and infomax

The likelihood landscapes displayed in figures 5-7 assumes a particular pdf $q(\cdot)$ for the source vector. Denoting $\underline{\mathbf{s}}$ a random vector with distribution q, simple calculus shows that

$$\frac{1}{T}\log p(X_{1:T};A,q) \stackrel{T\to\infty}{\longrightarrow} -\mathbf{K}[A^{-1}\mathbf{x}|\underline{\mathbf{s}}] + \text{cst.} \quad (11)$$

Therefore, figures 5-7 approximately display (up to a constant term) minus the Kullback divergence between the distribution of $\mathbf{y} = A^{-1}\mathbf{x}$ and the hypothesized distribution of the sources. This shows that the maximum likelihood principle is associated with a contrast function

$$\phi_{ML}[\mathbf{y}] = \mathbf{K}[\mathbf{y}|\underline{\mathbf{s}}] \quad (12)$$

and the normalized log-likelihood can be seen, via (11) as an estimate of $-\mathbf{K}[\mathbf{y}|\underline{\mathbf{s}}]$ (up to a constant). The ML principle thus says something very simple when applied to the BSS problem: *find matrix A such that the distribution of $A^{-1}\mathbf{x}$ is as close as possible (in the Kullback divergence) to the hypothesized distribution of the sources.*

The instability problem illustrated by fig. 7 may now be understood as follows: in this figure, the likelihood is maximum when $M(u,v)$ is a $\pm\pi/4$ rotation because the true source distribution is closer to the hypothesized source distribution *after* it is rotated by $\pm\pi/4$. As figure 8 shows, after such a rotation the areas of highest density of $\mathbf{y}$ correspond to the points of highest probability *of the hypothesized source model.*

A different approach to derive the contrast function (12) is very popular among the neural network community. Denote $g_i(\cdot)$ the distribution function

$$g_i(s) \triangleq \int_{-\infty}^{s} q_i(t)dt \ \in [0,1] \qquad 1 \le i \le n \quad (13)$$

so that $g_i' = q_i$ and denote $g(\mathbf{s}) = [g_1(s_1),\ldots,g_n(s_n)]^\dagger$. An interpretation of the infomax principle (see[9], [55], and references therein) suggests the contrast function

$$\phi_{IM}[\mathbf{y}] \triangleq -\mathbf{H}[g(\mathbf{y})] \quad (14)$$

where $\mathbf{H}[\cdot]$ denotes the Shannon entropy (for a random vector $\mathbf{u}$ with density $p(\mathbf{u})$, this is $\mathbf{H}[\mathbf{u}] = -\int p(\mathbf{u})\log p(\mathbf{u})d\mathbf{u}$ with the convention $0\log 0 = 0$). This idea can be understood as follows: on one hand, $g(\mathbf{s})$ is uniformly distributed on $[0,1]^n$ if $\mathbf{s}$ has pdf q; on the other hand, the uniform distribution has the highest entropy among all distributions on $[0,1]^n$ [28]. Therefore $g(C\underline{\mathbf{s}})$ has the highest entropy when $C = I$. The infomax idea, however, yields the same contrast as the likelihood *i.e.* in fact $\phi_{IM}[\mathbf{y}] = \phi_{ML}[\mathbf{y}]$. The connection between maximum likelihood and infomax was noted by several authors (see [57], [19], [50]).

A.2 Matching the structure: mutual information

The simple likelihood approach described above is based on a *fixed* hypothesis about the distribution of the sources. This becomes a problem if the hypothesized source distributions differ too much from the true ones, as illustrated by fig. 7 and 8. This remark suggests that the observed data should be modeled by adjusting both the unknown system *and* the distributions of the sources. In other words, one should minimize the divergence $\mathbf{K}[\mathbf{y}|\underline{\mathbf{s}}]$ with respect to A (via the distribution of $\mathbf{y} = A^{-1}\mathbf{x}$) *and* with respect to the model distribution of $\underline{\mathbf{s}}$. The last minimization problem has a simple and intuitive theoretical solution. Denote $\tilde{\mathbf{y}}$ a random vector with **i)** independent entries and **ii)** each entry distributed as the corresponding entry of $\mathbf{y}$. A classic property (see *e.g.* [28]) of $\tilde{\mathbf{y}}$ is that

$$\mathbf{K}[\mathbf{y}|\underline{\mathbf{s}}] = \mathbf{K}[\mathbf{y}|\tilde{\mathbf{y}}] + \mathbf{K}[\tilde{\mathbf{y}}|\underline{\mathbf{s}}] \quad (15)$$

for any vector $\underline{\mathbf{s}}$ with independent entries. Since $\mathbf{K}[\mathbf{y}|\tilde{\mathbf{y}}]$ does not depend on $\underline{\mathbf{s}}$, eq. (15) shows that $\mathbf{K}[\mathbf{y}|\underline{\mathbf{s}}]$ is minimized in $\underline{\mathbf{s}}$ by minimizing its second term *i.e.* $\mathbf{K}[\tilde{\mathbf{y}}|\underline{\mathbf{s}}]$; this is simply achieved by taking $\underline{\mathbf{s}} = \tilde{\mathbf{y}}$ for which $\mathbf{K}[\tilde{\mathbf{y}}|\underline{\mathbf{s}}] = 0$ so that $\min_{\underline{\mathbf{s}}} \mathbf{K}[\mathbf{y}|\underline{\mathbf{s}}] = \mathbf{K}[\mathbf{y}|\tilde{\mathbf{y}}]$. Having minimized the likelihood contrast $\mathbf{K}[\mathbf{y}|\underline{\mathbf{s}}]$ with respect to the source distribution, leading to $\mathbf{K}[\mathbf{y}|\tilde{\mathbf{y}}]$, our program is completed if we minimize the latter with respect to $\mathbf{y}$, *i.e.* if we minimize the contrast function

$$\phi_{MI}[\mathbf{y}] \triangleq \mathbf{K}[\mathbf{y}|\tilde{\mathbf{y}}]. \quad (16)$$

The Kullback divergence $\mathbf{K}[\mathbf{y}|\tilde{\mathbf{y}}]$ between a distribution and the closest distribution with independent entries is traditionally called the *mutual information* (between the entries of $\mathbf{y}$). It satisfies $\phi_{MI}[\mathbf{y}] \ge 0$ with equality if and only if $\mathbf{y}$ is distributed as $\tilde{\mathbf{y}}$. By definition of $\tilde{\mathbf{y}}$, this happens when the entries of $\mathbf{y}$ are independent. In other words, $\phi_{MI}[\mathbf{y}]$ measures the independence between the entries of $\mathbf{y}$. Thus, mutual information apears as the quantitative measure of independence associated to the maximum likelihood principle.

Note further that $\mathbf{K}[\tilde{\mathbf{y}}|\underline{\mathbf{s}}] = \sum_{i=1}^n \mathbf{K}[y_i|\underline{s}_i]$ (because both $\tilde{\mathbf{y}}$ and $\underline{\mathbf{s}}$ have independent entries).

Therefore,

$$\phi_{ML}[\mathbf{y}] = \phi_{MI}[\mathbf{y}] + \sum_{i=1}^{n} \mathbf{K}[y_i|\underline{s}_i] \qquad (17)$$

so that the decomposition (15) or (17) of the 'global' distribution matching criterion $\phi_{ML}[\mathbf{y}] = \mathbf{K}[\mathbf{y}|\underline{\mathbf{s}}]$ should be understood as

$$\begin{pmatrix} \text{Total} \\ \text{mismatch} \end{pmatrix} = \begin{pmatrix} \text{Deviation from} \\ \text{independence} \end{pmatrix} + \begin{pmatrix} \text{Marginal} \\ \text{mismatch} \end{pmatrix}$$

Therefore, maximizing the likelihood with fixed assumptions about the distributions of the sources amounts to minimize a sum of two terms: the first term is the true objective (mutual information as a measure of independence) while the second term measures how far the (marginal) distributions of the outputs $y_1, \ldots, y_n$ are from the assumed distributions.

A.3 Orthogonal contrasts

If the mixing matrix has been reduced to a rotation matrix by whitening, as explained in sect. II-B, contrast functions like ϕ_{ML} or ϕ_{MI} can still be used. The latter takes an interesting alternative form under the whiteness constraint $\mathrm{E}\mathbf{y}\mathbf{y}^\dagger = I$: one can show then that $\phi_{MI}[\mathbf{y}]$ is, up to a constant term, equal to the sum of the Shannon entropies of each output. Thus, under the whiteness constraint, minimizing the mutual information between the entries of $\mathbf{y}$ is equivalent to minimizing the sum of the entropies of the entries of $\mathbf{y}$ and we define

$$\phi^{\circ}_{MI}[\mathbf{y}] \triangleq \sum_i \mathbf{H}[y_i] \qquad (18)$$

There is a simple interpretation: mixing the entries of $\mathbf{s}$ 'tends' to increase their entropies; it seems natural to find separated source signals as those with *minimum* marginal entropies. It is also interesting to notice that $-\mathbf{H}[y_i]$ is (up to a constant) the Kullback divergence between the distribution of y_i and the zero-mean unit-variance normal distribution. Therefore, minimizing the sum of the marginal entropies is also equivalent to driving the marginal distributions of $\mathbf{y}$ as far away as possible from normality. Again, the interpretation is that mixing 'tends' to gaussianize the marginal distributions so that a separating technique should go in the opposite direction. Figure 9 is a visual illustration of the tendency to normality by mixing. The first column shows histograms for two independent variables s_1 and s_2 with a bimodal distribution and, superimposed to it as a solid line, the best Gaussian approximation. The following columns shows the histograms after rotations by steps of $\pi/16$, going from 0 to $\pi/4$ where mixing is maximal. The tendency to normality is very apparent.

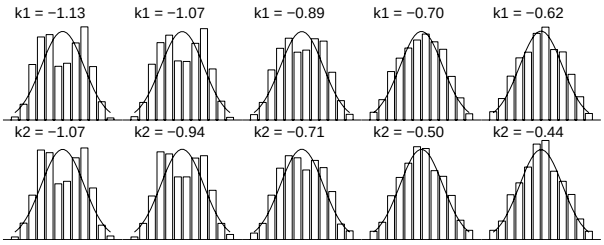


Fig. 9. Gaussianization by mixing. Histograms of y_1 (top row) and y_2 (bottom row) when $\mathbf{y}$ rotated by $\alpha\pi/4$ for $\alpha = 0, 1/4, 1/2, 3/4, 1$. Each subplot also shows the estimated kurtosis k_1 and k_2 (defined at eq. (21)) decreasing (in absolute value) upon mixing.

The entropic form (18) of the mutual information was used as starting point by Comon [26]; it remains a valid contrast under the weaker constraint that B is a volume-preserving transformation [56].

A.4 Discussion

The 'canonical' contrast for source separation is the mutual information ϕ_{MI} because it expresses the key property of source independence and nothing else: it does not include any explicit or implicit assumption about the distributions of the sources. On the other hand, if the source distributions are known, ϕ_{ML} is more appropriate because it expresses directly the fit between data and model. Also, ϕ_{ML} is easier to minimize because its gradient is easily estimated (see eq. (31)) while estimating the gradient of ϕ_{MI} is computationally demanding [60]. Even when the source distributions are unknown, one may use ϕ_{ML} with hypothesized source distributions which only need to be 'close enough' to the true distributions: recall sec. II-C for a qualitative explanation and see sec. VI-A for a quantitative statement and sec. V-B about adapting the model distributions). Another approach is to approximate the Kullback-based contrasts using high-order statistics, as examined next.

B. High order approximations

High order statistics can be used to define contrast functions which are simple approximations to those derived from the ML approach. High order information is most simply expressed by using *cumulants*. The discussion being limited to cumulants of order 2 and 4, only the following definitions are needed. For zero-mean random variables a, b, c, d, 2nd order cumulants are identical to 2nd order moments $\mathrm{Cum}[a, b] \triangleq \mathrm{E}ab$ and 4th order cumulants are

$$\mathrm{Cum}[a, b, c, d] \triangleq \mathrm{E}abcd - \mathrm{E}ab\mathrm{E}cd - \mathrm{E}ac\mathrm{E}bd - \mathrm{E}ad\mathrm{E}bc. \qquad (19)$$

Whenever the random variables a, b, c, d can be split in two groups which are mutually independent, their cumulant is zero. Therefore, indepen-

dence beyond second-order decorrelation can be easily tested using high order cumulants.

For simplicity, the following notation for the cumulants of the elements of a given vector $\mathbf{y}$ is used throughout:

$$\mathcal{C}_{ij}[\mathbf{y}] \triangleq \mathrm{Cum}[y_i, y_j], \quad \mathcal{C}_{ijkl}[\mathbf{y}] \triangleq \mathrm{Cum}[y_i, y_j, y_k, y_l].$$

Since the source vector $\mathbf{s}$ has independent entries, all its cross-cumulants vanish:

$$\mathcal{C}_{ij}[\mathbf{s}] = \sigma_i^2 \delta_{ij} \quad \mathcal{C}_{ijkl}[\mathbf{s}] = k_i \delta_{ijkl} \tag{20}$$

where δ is the Kronecker symbol and we have defined the variance σ_i^2 and the *kurtosis* k_i of the i-th source as the second and fourth order 'auto-cumulants' of s_i:

$$\sigma_i^2 \triangleq \mathcal{C}_{ii}[\mathbf{s}] = \mathrm{E}s_i^2 \qquad k_i \triangleq \mathcal{C}_{iiii}[\mathbf{s}] = \mathrm{E}s_i^4 - 3\mathrm{E}^2 s_i^2. \tag{21}$$

The likelihood contrast $\phi_{ML}[\mathbf{y}] = \mathbf{K}[\mathbf{y}|\underline{\mathbf{s}}]$ is 'the' measure of mismatch between output distribution and a model source distribution. A cruder measure can be defined from the quadratic mismatch between the cumulants:

$$\phi_2[\mathbf{y}] \triangleq \sum_{ij} \left(\mathcal{C}_{ij}[\mathbf{y}] - \mathcal{C}_{ij}[\mathbf{s}]\right)^2 = \sum_{ij} \left(\mathcal{C}_{ij}[\mathbf{y}] - \sigma_i^2 \delta_{ij}\right)^2$$

$$\phi_4[\mathbf{y}] \triangleq \sum_{ijkl} \left(\mathcal{C}_{ijkl}[\mathbf{y}] - \mathcal{C}_{ijkl}[\mathbf{s}]\right)^2 = \sum_{ijkl} \left(\mathcal{C}_{ijkl}[\mathbf{y}] - k_i \delta_{ijkl}\right)^2$$

Are ϕ_2 and ϕ_4 contrast functions as introduced in the beginning of this section? Clearly ϕ_2 is not a contrast because $\phi_2[\mathbf{y}] = 0$ expresses only the decorrelation between the entries of $\mathbf{y}$. On the contrary, one can show that $\phi_4[\mathbf{y}]$ is a contrast if all the sources have known non-zero kurtosis. Even though fourth order information is sufficient by itself to solve the BSS problem, it is interesting to use ϕ_2 and ϕ_4 in conjunction because they jointly provide an approximation to the likelihood contrast: if $\mathbf{s}$ and $\mathbf{y}$ are symmetrically distributed with distributions 'close enough' to normal, then

$$\mathbf{K}[\mathbf{y}|\mathbf{s}] \approx \phi_{24}[\mathbf{y}] \triangleq \frac{1}{48}\left(12\phi_2[\mathbf{y}] + \phi_4[\mathbf{y}]\right). \tag{22}$$

Room is lacking to discuss the validity of this approximation (which stems from an Edgeworth expansion, see sec. V-B). The point however is not to determine how closely $\phi_{24}[\mathbf{y}]$ approximates $\mathbf{K}[\mathbf{y}|\mathbf{s}]$ but rather to follow the suggestion that second and fourth order information could be used jointly.

Orthogonal contrasts. We consider cumulant-based orthogonal contrasts. The orthogonal approach, which enforces whiteness *i.e.* $\phi_2[\mathbf{y}] = 0$, thus corresponds to replacing the factor 12 in eq. (22) by an 'infinite weight' (optimal weighting is considered in [20]; see also sec. V-B) or equivalently to minimizing $\phi_4[\mathbf{y}]$ under the whiteness constraint $\phi_2[\mathbf{y}] = 0$. Simple algebra shows that, if $\phi_2[\mathbf{y}] = 0$, then $\phi_4[\mathbf{y}]$ is equal (up to a constant additive term) to

$$\phi_4^\circ[\mathbf{y}] \triangleq -2\sum_{i=1}^{n} k_i \mathcal{C}_{iiii}[\mathbf{y}] = \mathrm{E}f_4(\mathbf{y}) \tag{23}$$

where we have defined $f_4(\mathbf{y}) \triangleq -2\sum_{i=1}^n k_i(y_i^4 - 3)$. This is a pleasant finding: this contrast function being the expectation of a function of $\mathbf{y}$, it is particularly simple to estimate by a sample average.

Recall that the contrast function ϕ_{ML} defined at eq. (12) depends on a source model *i.e.* it is defined using an hypothetical density $q(\cdot)$ for the source distribution. Similarly, the fourth-order approximation ϕ_4° requires an hypothesis about the sources but it is only a 'fourth-order hypothesis' in the sense that only the kurtosis k_i for each source must be specified in definition (23). In the same manner as minimizing ϕ_{ML} over the source distribution yields the mutual information contrast ϕ_{ML}, minimizing $\phi_4^\circ[\mathbf{y}]$ (which approximates ϕ_{ML}) over the kurtosis k_i of each source yields an approximation to the mutual information. One finds $\phi_{MI}[\mathbf{y}]$ by

$$\phi_{\mathrm{ICA}}^\circ[\mathbf{y}] = \sum_{ijkl \neq iiii} \mathcal{C}_{ijkl}^2[\mathbf{y}] = -\sum_i \mathcal{C}_{iiii}^2[\mathbf{y}] + \mathrm{cst} \tag{24}$$

as such an orthogonal fourth-order approximation. This was obtained first by Comon [26] (along a slightly different route) and by Lacoume *et al.* [41] by approximating the likelihood by a Gram-Charlier expansion. This contrast is similar to ϕ_{MI} also in that its first form involves only terms measuring the (4th order) independence between the entries of $\mathbf{y}$. Its second form stems from the fact that $\sum_{ijkl} \mathcal{C}_{ijkl}^2[\mathbf{y}]$ is constant if $\phi_2[\mathbf{y}] = 0$ holds (see *e.g.* [26]). It is also similar to (23) when $k_i \approx \mathcal{C}_{iiii}[\mathbf{y}]$ which is indeed the case close to separation.

One benefit of considering 4th-order orthogonal contrasts like $\phi_{\mathrm{ICA}}^\circ$ is that they can be optimized by the Jacobi technique: the 'unknown rotation' (sec. II-B) can be found as sequence of 2×2 rotations applied in sequence to all pairs (y_i, y_j) for $i \neq j$ with the optimal angle at each step being often available in close form. Comon [26] has such a formula for ϕ_{ICA} in the case of real signals.

Independence can also be tested on a smaller subset of cross-cumulants with:

$$\phi_{\mathrm{JADE}}^\circ[\mathbf{y}] \triangleq \sum_{ijkl \neq ijkk} \mathcal{C}_{ijkl}^2[\mathbf{y}]. \tag{25}$$

The motivation for using this specific subset is that ϕ_{JADE} also is a 'joint diagonalization' criterion, entailing that it can be optimized by Jacobi technique for which the rotation angles can be found in close form even in the complex case [23]. A similar technique is described in [32].

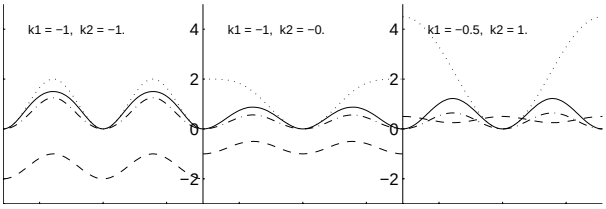


Fig. 10. Variation of orthogonal contrast functions (Solid: ϕ°_{ICA}, dash-dots: ϕ°_{JADE}, dashes: ϕ°_m, dots: ϕ°_4) when two sources with kurtosis k_1 and k_2 are rotated between $-\pi/2$ and $\pi/2$. Left: $(k_1, k_2) = (-1, -1)$, center: $(k_1, k_2) = (-1, 0)$, right: $(k_1, k_2) = (-0.5, 1)$.

Simpler contrasts can be used if the kurtosis of the sources are known. For instance, eq. (23) suggests, for negative kurtosis ($k_i < 0\ \forall i$), a very simple contrast:

$$\phi^\circ_m[\mathbf{y}] = \sum_{i=1}^{n} \mathrm{E} y_i^4 \tag{26}$$

(see [22], [54], [48], [45]) Actually, the condition that $k_i + k_j < 0$ for all pairs of sources is sufficient for the stationary points of this orthogonal contrast function to be locally stable (see sec VI-A).

Some properties of the fourth-order contrasts discussed above are illustrated by fig. 10 displaying the variation of some orthogonal contrast functions in the two-source case: a 2×1 source vector $\mathbf{s}$ with kurtosis (k_1, k_2) is rotated into $\mathbf{y}$ by an angle $\theta \in (-\pi/2, \pi/2)$. On the left panel, the sources have identical kurtosis $k_1 = k_2 = -1$: all the four contrasts are minimized at integer multiples of $\pi/2$. On the center panel, one source is Gaussian ($k_2 = 0$): the contrasts show smaller variations except for ϕ°_4. Note that ϕ_4 'knows' that one source has zero kurtosis, thus distinguishing between even and odd multiplies of $\pi/2$. On the right panel, $k_1 = -0.5$ and $k_2 = 1$ so that $k_1 + k_2 > 0$ which violates the condition for ϕ°_m to be a contrast: its minima become maxima and vice versa. This is the same phenomenon as illustrated by figure 7.

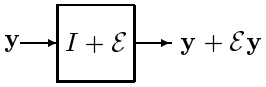
IV. Estimating functions

By design, all valid contrast functions reach their minima at a separating point when the model holds; in this sense, no one is better than another. In practice, however, contrasts are only estimated from a finite data set: sample-based contrasts depend not on the distribution of $\mathbf{y}$ but on its *sample* distribution. Estimation from a finite data set introduces stochastic errors depending on the available samples and also on the contrast function. Thus a statistical characterization of the minima of sample-based contrast functions is needed and will provide a basis for comparing contrast functions. For this purpose, the notion of *estimating function* is introduced; it is also closely related to gradient algorithms for BSS (sec. V-A).

A. Relative gradient

The variation of a contrast function $\phi[\mathbf{y}]$ under a linear transform of $\mathbf{y}$ is may be expressed by defining a 'relative gradient'. This specific notion builds on the fact that the parameter of interest is a square matrix.

Definition. An infinitesimal transform of $\mathbf{y}$ is $\mathbf{y} \to (I + \mathcal{E})\mathbf{y} = \mathbf{y} + \mathcal{E}\mathbf{y}$ where $\mathcal{E}$ is a 'small' matrix.



If ϕ is smooth enough, $\phi[\mathbf{y} + \mathcal{E}\mathbf{y}]$ can be expanded as

$$\phi[\mathbf{y} + \mathcal{E}\mathbf{y}] = \phi[\mathbf{y}] + \sum_{i,j=1}^{n} G_{ij}\mathcal{E}_{ij} + o(\|\mathcal{E}\|) \tag{27}$$

with G_{ij} the partial derivative of $\phi[\mathbf{y} + \mathcal{E}\mathbf{y}]$ with respect to $\mathcal{E}_{ij}$ at $\mathcal{E} = 0$. These coefficients form a $n \times n$ matrix, denoted $\nabla\phi[\mathbf{y}]$, called the *relative gradient* [22] of $\phi[\mathbf{y}]$ at $[\mathbf{y}]$. In matrix form, expansion (27) reads

$$\phi[\mathbf{y} + \mathcal{E}\mathbf{y}] = \phi[\mathbf{y}] + \langle \nabla\phi[\mathbf{y}] \mid \mathcal{E} \rangle + o(\|\mathcal{E}\|) \tag{28}$$

where $\langle \cdot | \cdot \rangle$ is the Euclidean scalar product between matrices: $\langle M | N \rangle = \mathrm{trace}\left(MN^\dagger\right) = \sum_{i,j=1}^{n} M_{ij}N_{ij}$.

Note that the relative gradient is defined *without* explicit reference to the possible dependence of $\mathbf{y}$ on B as $\mathbf{y} = B\mathbf{x}$; thus it actually characterizes the first order variation of the contrast function itself. It is of course possible to relate $\nabla\phi[\mathbf{y}]$ to a 'regular' gradient with respect to B if $\mathbf{y} = B\mathbf{x}$. Elementary calculus yields

$$\nabla\phi[B\mathbf{x}] = B^\dagger \frac{\partial\phi[B\mathbf{x}]}{\partial B}. \tag{29}$$

The notion of *natural gradient* was independently introduced by Amari [2]. It is distinct in general from the relative gradient: the latter is defined in any continuous group of transformation while the former is defined in any smooth statistical model. However, for the BSS model which, as a statistical transformation model combines both features, the two ideas yield the same class of algorithms (sec. V-A).

Score functions. The source densities $q_1, \ldots, q_n$ used in (3) and (7) to define the likelihood of a BSS model enter in the estimating function via their log-derivatives: the so-called 'score functions' $\varphi_1, \ldots, \varphi_n$, defined as

$$\varphi_i \triangleq -(\log q_i)' \quad \text{or} \quad \varphi_i(\cdot) = -\frac{q_i(\cdot)'}{q_i(\cdot)}. \tag{30}$$

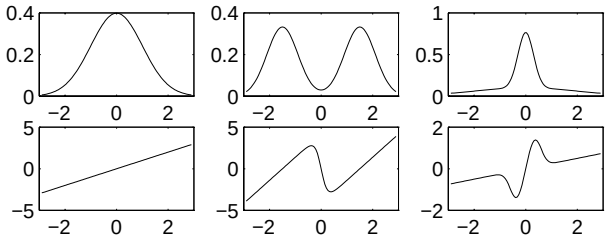


Fig. 11. Some densities and their associated scores.

Figure 11 displays some densities and their associated score functions. Note that the score for 'the most basic distribution' is 'the most basic function': if s is a zero-mean unit-variance Gaussian variable: $q(s) = (2\pi)^{-1/2} \exp -\frac{s^2}{2}$, then the associated score is $\varphi(s) = s$. Actually, Gaussian densities precisely are these densities associated with *linear* score functions. Thus, the necessity of non Gaussian modeling (recall section II) translates in the necessity of considering *non-linear* score functions.

Relative gradient of the likelihood contrast. At the core of the BSS contrast functions is $\phi_{ML}[\mathbf{y}]$ associated with the likelihood given the source densities $q_1, \ldots, q_n$. Its relative gradient is found to be [62]

$$\nabla \phi_{ML}[\mathbf{y}] = \mathrm{E} H_\varphi(\mathbf{y}) \tag{31}$$

where $H_\varphi : \mathbb{R}^n \mapsto \mathbb{R}^{n \times n}$ is

$$H_\varphi(\mathbf{y}) \triangleq \varphi(\mathbf{y})\mathbf{y}^\dagger - I \tag{32}$$

with $\varphi : \mathbb{R}^n \mapsto \mathbb{R}^n$ the entry-wise non-linear function

$$\varphi(\mathbf{y}) \triangleq [\varphi_1(y_1), \ldots, \varphi_n(y_n)]^\dagger \tag{33}$$

collecting the score functions related to each source. This is a remarkably simple result: this relative gradient merely is the expected value of a fixed function H_φ of $\mathbf{y}$.

Interpretation. The ML contrast function ϕ_{ML} is minimum at points where its (relative) gradient cancels, *i.e.* by (31), at these points which are solutions of the matrix equation $\mathrm{E} H_\varphi(\mathbf{y}) = 0$. This is interpreted by examining the (i,j)-th entry of this matrix equation. For $i = j$, we find $\mathrm{E}\varphi_i(y_i)y_i = 1$ which depends only on y_i and determines the scale of the i-th source estimate. For $i \neq j$, the (i,j)-th entry of $\mathrm{E} H_\varphi(\mathbf{y}) = 0$ reads $\mathrm{E}\varphi_i(y_i)y_j = 0$ meaning that the jth output y_j should be uncorrelated to a non-linear version $\varphi_i(y_i)$ of the ith output. Because φ_i and φ_j are *non-linear* functions, the conditions for the pairs (i,j) and the pair (j,i) are (in general) not equivalent. Note that if the source signals are modeled as zero-mean unit-variance normal variables, then $\varphi_i(y_i) = y_i$ for all i and $H_\varphi(\mathbf{y}) = \mathbf{y}\mathbf{y}^\dagger - I = H_w(\mathbf{y})$ (recalling def. (5)). Then $\phi_{ML}[\mathbf{y}]$ is minimum at points where $\mathrm{E}\mathbf{y}\mathbf{y}^\dagger = I$: we only obtain the whiteness condition. Again, this is not sufficient to determine a separating solution; score functions must be non-linear (the source model must be non Gaussian).

The idea of using non-linear functions to obtain a sufficient set of independence conditions can be traced back to the seminal paper of Hérault and Jutten [44] (see [46] for a reference in English) but the choice of the non-linear functions was somewhat ad hoc; Féty [39] gave an interpretation of the non-linear functions as 'amplifiers' for the signals of interest; Bar-Ness also produced early work using non-linear functions [8]. However, the ML principle makes it clear that the non-linear functions are related via (30) to a non-Gaussian model of the source distributions.

B. Estimating functions

An *estimating function* for the BSS problem is a function $H : \mathbb{R}^n \mapsto \mathbb{R}^{n \times n}$. It is associated to an *estimating equation*

$$\frac{1}{T} \sum_{t=1}^{T} H(\mathbf{y}(t)) = 0 \tag{34}$$

thus called because, H being matrix-valued, equation (34) specifies *a priori* as many constraints as unknown parameters in the BSS problem. Many BSS estimates can be characterized via an estimating function [18], [3].

A simple instance of estimating function is $H_w(\mathbf{y})$, used in eq. (5) to express that decorrelation between the entries of $\mathbf{y}$. Equation (34) with $H(\mathbf{y}) = H_w(\mathbf{y})$ is equivalent to $\frac{1}{T}\sum_t \mathbf{y}(t)\mathbf{y}(t)^\dagger = I$ *i.e.* it expresses the *empirical* whiteness of a batch of T samples of $\mathbf{y}$ as opposed to the 'actual' whiteness *i.e.* $\mathrm{E} H_w(\mathbf{y}) = 0$. The estimating function $H_w(\mathbf{y})$, however, is *not* appropriate for BSS, since whitening (or decorrelation) is not sufficient to determine a separating matrix.

The simplest example of estimating function for BSS is obtained in the ML approach. The gradient of the likelihood (7) may be shown [62] to cancel at points A_{ML} which are exactly characterized by eq. (34) with $\mathbf{y} = A_{ML}^{-1}\mathbf{x}$ and $H = H_\varphi$ as defined in (32). In other words, maximum likelihood estimates correspond exactly to the solution of an estimating equation. This equation is nothing but the sample counterpart of $\mathrm{E} H_\varphi(\mathbf{y}) = 0$ which characterizes the stationary points of $\phi_{ML}[\mathbf{y}]$. Recall that the latter is obtained (at eqs. (11) and (12)) as a limit of the log-likelihood.

Because the value of an estimating function is a square matrix, it can be decomposed into a symmetric part (equal to its transpose) and a skew symmetric part (opposite to its transpose). This

decomposition simply is

$$H(\mathbf{y}) = \frac{H(\mathbf{y}) + H(\mathbf{y})^\dagger}{2} + \frac{H(\mathbf{y}) - H(\mathbf{y})^\dagger}{2}. \quad (35)$$

If the optimization of some regular contrast function corresponds to an estimating function $H(\mathbf{y})$, it is found that the optimization of the same contrast under the whiteness constraint corresponds to an estimating function $H^\circ\mathbf{y})$ given by

$$H^\circ(\mathbf{y}) = H_w(\mathbf{y}) + \frac{1}{2}\left(H(\mathbf{y}) - H(\mathbf{y})^\dagger\right). \quad (36)$$

Thus, the symmetric part of $H(\mathbf{y})$ replaced by $H_w(\mathbf{y}) = \mathbf{y}\mathbf{y}^\dagger - I$, already introduced at eq. (5), whose effect is to enforce the whiteness constraint. In particular, maximum likelihood estimates under the whiteness constraint are (again) solutions of eq. (34) with the estimating function $H = H^\circ_\varphi$:

$$H^\circ_\varphi(\mathbf{y}) \triangleq \mathbf{y}\mathbf{y}^\dagger - I + \varphi(\mathbf{y})\mathbf{y}^\dagger - \mathbf{y}\varphi(\mathbf{y})^\dagger \quad (37)$$

Other orthogonal contrast functions are associated to similar estimating functions. For instance, the simple 4th-order contrasts $\phi^\circ_4[y]$ and $\phi^\circ_m[y]$ (eqs. (23) and (26) repsectively) yield estimating equations in the form (37) with non-linear functions respectively given by

$$\varphi_i(y_i) = -k_i y_i^3 \quad \text{and} \quad \varphi_i(y_i) = y_i^3 \quad (38)$$

Recall that using the contrast function (26) supposes sources with *negative* kurtosis k_i. Thus the two functions in (38) 'agree' on the sign to be given to a cubic distortion (as was to be expected).

Some contrast functions, like ϕ°_{ICA} and ϕ°_{JADE}, when estimated from T samples are minimized at points which cannot be represented *exactly* as the solution of (34) for a *fixed* estimating function. However, one can often find, as in [37], an 'asymptotic' estimating function in the sense that the solution of the associated estimating equation is very close to the minimizer of the estimated contrast. For instance, the contrast ϕ°_{ICA} and ϕ°_{JADE} are asymptotically associated to the same estimating function as ϕ°_4. This implies that minimizing ϕ°_{ICA}, ϕ°_{JADE} or ϕ°_4 with cumulants estimated from T samples yields estimates which are equivalent (they differ by a term which is smaller than the estimation error) for large enough T.

Which functions are appropriate as estimating functions? One could think of using *any* H such that $\mathrm{E}H(\mathbf{s}) = 0$ as an estimating function because the estimating equation (34) would just be the sample counterpart of $\mathrm{E}H(\mathbf{y}) = 0$ and would *a priori* provide as many scalar equations as unknown parameters. However, the ML principle suggests the specific forms (32) and (37) with the non-linear functions in $\varphi(\mathbf{y})$ being (approximations of) the score functions for the probability densities of the signals to be separated.

V. Adaptive algorithms

A simple generic technique for optimizing an objective function is gradient descent. In most optimization problems, its simplicity is at the expense of performance: more sophisticated techniques —such as 'Newton-like' algorithms using second derivatives in addition to the gradient— can often significantly speed up convergence. For the BSS problem, however, it turns out that a simple gradient descent offers 'Newton-like' performance (see below). This surprising and fortunate result is obtained by descending along the *relative* gradient defined in sec. IV-A.

A. Relative gradient techniques

Relative gradient descent. We first describe a 'generic' *relative gradient descent.* Generally, the steepest descent technique of minimization consists in moving by a small step in a direction opposite to the gradient of the objective function. The relative gradient of a contrast $\phi[\mathbf{y}]$ is defined (sec. IV-A) with respect to a 'relative variation' of $\mathbf{y}$ by which $\mathbf{y}$ is changed into $(I + \mathcal{E})\mathbf{y}$. The resulting variation of $\phi[\mathbf{y}]$ is (at first order) the scalar product $\langle \nabla\phi[\mathbf{y}] \mid \mathcal{E}\rangle$ between the relative variation $\mathcal{E}$ and the relative gradient $\nabla\phi[\mathbf{y}]$ as in eq. (27) or (28). Aligning the direction of change in the direction opposite to the gradient is to take $\mathcal{E} = -\mu\nabla\phi[\mathbf{y}]$ for a 'small' positive step μ. Thus, one step of a relative gradient descent can be formally described as

$$\mathbf{y} \leftarrow (I - \mu\nabla\phi[\mathbf{y}])\mathbf{y} = \mathbf{y} - \mu\nabla\phi[\mathbf{y}]\ \mathbf{y}. \quad (39)$$

According to (28), the resulting variation of $\phi[\mathbf{y}]$ is $\delta\phi \approx \langle\nabla\phi[\mathbf{y}]|\mathcal{E}\rangle = \langle\nabla\phi[\mathbf{y}]| - \mu\nabla\phi[\mathbf{y}]\rangle = -\mu\|\nabla\phi[\mathbf{y}]\|^2$ which is negative for positive μ.

The formal description (39) can be turned into off-line and on-line algorithms as described next.

Off-line relative gradient descent. Consider the separation of a batch $\mathbf{x}(1), \ldots, \mathbf{x}(T)$ of T samples based on the minimization of a contrast function $\phi[\mathbf{y}]$ with relative gradient $\nabla\phi[\mathbf{y}] = \mathrm{E}H(\mathbf{y})$. One looks for a linear transform of the data satisfying the corresponding estimating equation $\frac{1}{T}\sum_{t=1}^{T} H(\mathbf{y}(t)) = 0$. The relative gradient descent to solve it goes as follows: Set $\mathbf{y}^{(0)}(t) = \mathbf{x}(t)$ for $1 \le t \le T$ and iterate through the following two steps

$$\widehat{\mathcal{H}} \leftarrow \frac{1}{T}\sum_{t=1}^{T} H(\mathbf{y}(t)), \quad (40)$$

$$\mathbf{y}(t) \leftarrow \mathbf{y}(t) - \mu\widehat{\mathcal{H}}\mathbf{y}(t) \quad (1 \le t \le T). \quad (41)$$

The first step computes an estimate $\widehat{\mathcal{H}}$ of the relative gradient for the current values of the data; the second step updates the data in the (relative) direction opposite to the relative gradient as in (39).

The algorithm stops when $\frac{1}{T}\sum_{t=1}^{T} H(\mathbf{y}(t)) = 0$ *i.e.* when the estimating equation is solved. It is amusing to note that this implementation does not need to maintain a separating matrix: it directly operates on the data set itself with the source signals emerging during the iterations.

On-line relative gradient descent. On-line algorithms update a separating matrix B_t upon reception of a new sample $\mathbf{x}(t)$. The (relative) linear transform $\mathbf{y} \leftarrow (I + \mathcal{E})\mathbf{y}$ corresponds to changing B into $(I + \mathcal{E})B = B + \mathcal{E}B$. In the on-line mode, one uses the *stochastic* gradient technique where the gradient $\nabla\phi[\mathbf{y}] = \mathrm{E}H[\mathbf{y}]$ is replaced by its instantaneous value $H(\mathbf{y}(t))$. Hence the stochastic relative gradient rule

$$B_{t+1} = B_t - \mu_t H(\mathbf{y}(t))B_t \tag{42}$$

where μ_t is a sequence of positive learning steps.

Uniform performance of relative gradient descent. A striking feature of BSS model is that the 'hardness' (in a statistical sense discussed in section VI-C) of separating mixed sources does not depend on the particular value of the mixing matrix A: the problem is 'uniformly hard in the mixture'. Very significantly, the device of relative updating produces algorithms which also behave uniformly well in the mixture. Right-multiplying the updating rule (42) by matrix A and using $\mathbf{y} = B\mathbf{x} = BA\mathbf{s}$, one readily finds that the trajectory of the global system $C_t \triangleq B_t A$ which combines mixing and unmixing matrices is governed by

$$C_{t+1} = C_t - \mu_t H(C\mathbf{s}(t))C_t. \tag{43}$$

This trajectory is expressed here as a sole function of the global system C_t: the only effect of the mixing matrix A itself is to determine (together with B_0) the initial value $C_0 = B_0 A$ of the global system. This is a very desirable property: it means that the algorithms can be studied and optimized without reference to the actual mixture to be inverted. This is true for any estimating function $H(y)$; however uniformly *good* performance can only be expected if the $H(y)$ is correctly adjusted to the distribution of the source signals, for instance by deriving it from a contrast function. Algorithms based on an estimating function in the form (32) are described in [5] for the on-line version and in [62] for an off-line version; those based on form (37) are studied in detail in [22]. The uniform performance property was also obtained in [25].

Regular gradient algorithms. It is interesting to compare the relative gradient algorithm to the algorithm obtained by a 'regular' gradient, that is by applying a gradient rule to the entries of B for the minimization of $f(B) \triangleq \phi[B\mathbf{x}]$. This is

$$B_{t+1} = B_t - \mu_t H(\mathbf{y}(t))B_t^{-\dagger}. \tag{44}$$

Not only is this form more costly because it requires (in general) the inversion of B_t at each step, but it lacks the uniform performance property: the trajectory of the global system depends on the particular mixture A to be inverted.

B. Adapting to the sources

The iterative and adaptive algorithms described above require the specification of an estimating function H, for which two forms H_φ and H°_φ (eqs. (32) and (37)) are suggested by the theory. These forms, in turn, depend on non-linear functions $\varphi_1, \ldots, \varphi_n$ which, ideally, should be the score functions associated to the distributions of the sources (sec. IV-B). When the source distributions are unknown, one may try to estimate them from the data (for instance using some parametric model as in [57]) or to directly estimate 'good' non-linear functions.

A first idea is to use Edgeworth expansions (see e.g. [52]) which provide approximations to probability densities in the vicinity of a Gaussian density. The simplest non trivial Edgeworth approximation of a symmetric pdf q in the vicinity of the standard normal distribution is

$$q(s) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s^2}{2}\right)\left(1 + \frac{k}{24}(s^4 - 6s^2 + 3) + \cdots\right)$$

where k is the kurtosis of q. The corresponding approximate score function then is

$$\varphi(s) = s - \frac{k}{6}(s^3 - 3s) + \cdots. \tag{45}$$

Thus the Edgeworth expansion suggests that in a linear-cubic approximation to the score function the coefficient of the cubic part should be $-k_i/6$ for the ith source. Asymptotic analysis shows that such a choice at least guarantees the local stability (sec. VI-A). There are other possibilities for deriving score functions by a density expansion: see for instance [69] for a different proposal involving odd and even terms in φ.

A more direct approach than pdf expansion is proposed by Pham [62] who considers approximating ψ by a linear combination

$$\varphi_\alpha(s) \triangleq \sum_{l=1}^{L} \alpha_l f_l(s) \tag{46}$$

of a fixed set $\{f_1, \ldots, f_L\}$ of arbitrary basis functions. Rather surprisingly, the set $\{\alpha_1, \ldots, \alpha_L\}$ of coefficients minimizing the mean square error $\mathrm{E}(\varphi_\alpha(s) - \psi(s))^2$ between the true score ψ and its

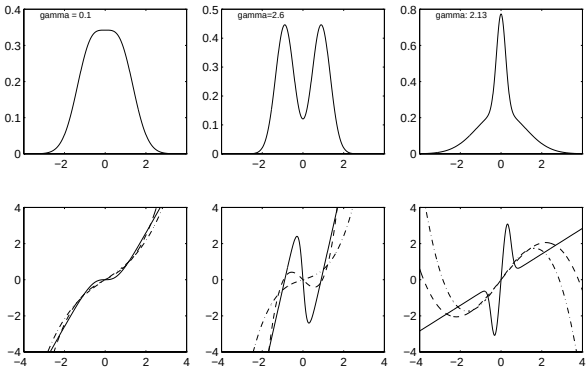


Fig. 12. Top row: three distributions and the values of $\gamma_\star$ as a measure of non-Gaussianity (see sec. VI-B). Bottom row: the score function (solid) and its linear-cubic approximations: based on Edgeworth expansion (dashes-dots) and optimal (dashes).

approximation can be found without knowing ψ: the best mean-square approximation involves only the expectation operator. It is:

$$\varphi_*(s) = \left(\mathrm{E}F'(s)\right)^\dagger \left(\mathrm{E}F(s)F(s)^\dagger\right)^{-1} F(s) \qquad (47)$$

where $F(s) \triangleq [f_1(s), \ldots, f_L(s)]^\dagger$ is the $L \times 1$ column vector of basis functions and $F'(s)$ is the column vector of their derivatives. This is a nice result because the expression of φ_* can be simply estimated by replacing in (47) expectations by sample averages and the values of s by the estimated source signals.

The two approaches of Edgeworth expansion and mean-square fit, respectively leading to the approximations (45) and (47), are compared in figure 12. Three pdf's are displayed in the top row; the bottom row shows the corresponding score function (solid line), the linear-cubic approximation by (45) (dash-dotted line) and the Pham approximation (dashed line) obtained from (47) with $F(s) = [s, s^3]$. Both approximations are similar in the first example when the pdf is close to Gaussian; in the second case, the optimal approximation fits much better the true score in the area of highest probability. None of the approximations seem really good in the third example for the simple reason that the true score there cannot be well approximated by a linear-cubic function. However, the two approximations fit the score well enough to guarantee the stability of the gradient algorithms (see sect. VI-A).

VI. Performance issues

This section is concerned with the performance of BSS algorithms: it presents some asymptotic analysis results. It has been repeatedly stressed that it was not necessary to know the source distributions (or equivalently: the associated score functions) to a great accuracy to obtain *consistent* BSS algorithms. There is however a limit to the misspecification of the source distributions as illustrated by fig. 7; this is elucidated at sec. VI-A which gives explicit stability limits. Even if an hypothesized distribution is good enough to preserve stability, one may expect a loss of estimation accuracy due to misspecification when a finite number of samples are available; this is quantified at sec. VI-B which also describes the ultimate achievable separation performance. The concluding section VI-C discusses the general property of 'equivariance' which governs the performance of BSS algorithms.

A. Local stability

A stationary point (or equilibrium point) B of the learning rule (42) is characterized by $\mathrm{E}H(\mathbf{y}) = \mathrm{E}H(B\mathbf{x}) = 0$ *i.e.* the mean value of the update is zero. We have seen that separating matrices (with the proper scale) are equilibrium points; we are now interested in finding when they are locally stable *i.e.* when a small deviation from the equilibrium is pulled back to the separating point. In other words, we want the separating matrix to a (local) attractor for the learning rule (42). In the limit of small learning steps, it exists a simple criterion for testing local stability which depends on the derivative of $\mathrm{E}H(B\mathbf{x})$ with respect to B. For both the symmetric form H°_φ and for the asymmetric form H_φ the stability condition can be worked out exactly. They are found to depend only the following non-linear moments

$$\kappa_i \triangleq \mathrm{E}\varphi_i'(s_i)\ \mathrm{E}s_i^2 - \mathrm{E}\varphi_i(s_i)s_i \qquad (48)$$

where each s_i is rescaled according to $\mathrm{E}H(\mathbf{s}) = 0$, that is $\mathrm{E}\varphi_i(s_i)s_i = 1$ for $H = H_\varphi$ or $\mathrm{E}s_i^2 = 1$ for $H = H^\circ_\varphi$.

Leaving aside the issue of stability with respect to scale, the stability conditions for the symmetric form (37) are [22]

$$(1+\kappa_i)(1+\kappa_j) > 1 \quad \text{for} \quad 1 \le i < j \le n \qquad (49)$$

and for the asymmetric form (32), the conditions are [4] that $1 + \kappa_i > 0$ for $1 \le i \le n$ and that

$$\kappa_i + \kappa_j > 0 \quad \text{for} \quad 1 \le i < j \le n. \qquad (50)$$

Therefore stability appears to depend on pairwise conditions. The stability domains for a given pair of sources are displayed on fig. 13 in the (κ_i, κ_j) plane. Note that the stability domain is larger for the symmetric form (37): this is a consequence of letting the second order information (the whiteness constraint) do 'half the job' (see sec. II-B).

Some comments are in order. First, it appears that in both cases, a *sufficient* stability condition is $\kappa_i > 0$ for all the sources. Thus, regarding stability, tuning the non-linear functions φ_i's to the

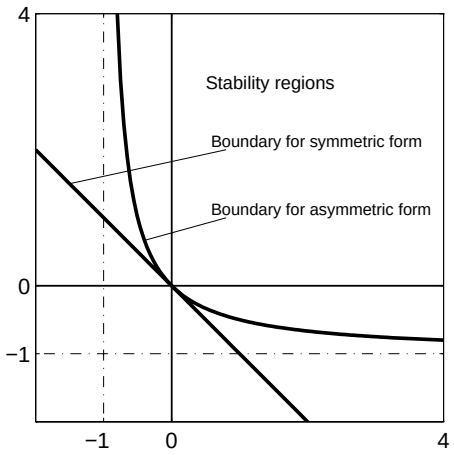


Fig. 13. Stability domains in the (κ_i, κ_j) plane.

source distributions should be understood as making the κ_i's positive. Second, one can show that if s_i is Gaussian, then $\kappa_i = 0$ for any function φ_i. Therefore the stability conditions can never be met if there is more than one Gaussian source, in agreement with the identifiability statements of sec. II. Third, it can also be shown that if φ_i is taken to be the score function for the true density of s_i, then $\kappa_i \geq 0$ with equality only if s_i is Gaussian.

Section II-C illustrated the fact that the hypothesized source distributions should be 'close enough' to the true distributions for the likelihood to still show a maximum around a separating point. The definition of κ_i provides a quantitative measure of how wrong the hypothesis can be: they should not allow κ_i to become negative.

We also note that it is not necessary that *all* the κ_i's are positive: if $\kappa_i < 0$ for at most one source, this can be compensated if the moments κ_j are large enough for all $j \neq i$. As seen from the stability domains (fig. 13), one source at most can have an arbitrarily negative κ_i if the symmetric form is used while the stability of the asymmetric form requests that $\kappa_i > -1$.

We have considered linear-cubic score functions in secs. IV and V. If $\varphi_i(s_i) = \alpha_i s_i + \beta_i s_i^3$ for two constants α_i and β_i, then $\kappa_i = \beta_i(3\mathrm{E}^2 s_i^2 - \mathrm{E}s_i^4) = -\beta_i k_i$ where, as above, k_i denotes the kurtosis. Note that the linear part of φ_i does not affect the stability and that stability is guaranteed if the coefficient β_i of the cubic part has a sign *opposite* to the sign of the kurtosis. Quite naturally, the functions in eq. (38) and (45) come up naturally with the right sign. Therefore, if one wishes to use cubic non-linearities, it is sufficient to know the sign of the kurtosis of each source to make separating matrices stable. For other than cubic scores, stability depends on the sign of κ_i, not on the sign of the kurtosis.

B. Accuracy of estimating equations

This section characterizes the accuracy of signal separation obtained by solving an estimating equation (34) with T independent realizations of $\mathbf{x}$.

If a matrix B is used for separation, the pth entry of $\mathbf{y} = B\mathbf{x} = BA\mathbf{s}$ contains the signal of interest s_p at power $(BA)_{pp}^2 \sigma_p^2$ and the qth interfering signal s_q at power $(BA)_{pq}^2 \sigma_q^2$. Therefore, for a given matrix B the quantity

$$\rho_{pq}(B) \triangleq \frac{(BA)_{pq}^2 \mathrm{E}s_q^2}{(BA)_{pp}^2 \mathrm{E}s_p^2} \qquad p \neq q \tag{51}$$

measures the interference-to-signal ratio (ISR) provided by B in rejecting the qth source in the estimate of the pth source. Let $\widehat{B}_T$ be the separating matrix obtained via a particular algorithm using T samples. In general, the estimation error in regular statistical models decreases as $1/\sqrt{T}$ so that the limit

$$\mathrm{ISR}_{pq} \triangleq \lim_{T\to\infty} T\ \mathrm{E}\ \rho_{pq}(\widehat{B}_T) \tag{52}$$

usually exists provides an asymptotic measure of performance of separation of a given off-line BSS technique. When $H = H_\varphi$ or $H = H_\varphi^\circ$ are used in the estimating equation, the asymptotic ISR depends on the moments κ_i in (48) and also on:

$$\gamma_i \triangleq \mathrm{E}\varphi_i^2(s_i)\mathrm{E}s_i^2 - \mathrm{E}^2\left(\varphi_i(s_i)s_i\right) \geq 0. \tag{53}$$

For simplicity, we consider identically distributed signals and identical non-linear functions: $\varphi_i(\cdot) = \varphi(\cdot)$, so that $\kappa_i = \kappa$ and $\gamma_i = \gamma$ for $1 \leq i \leq n$. With a symmetric estimating function H_φ°, one finds

$$\mathrm{ISR}_{pq}^\circ = \mathrm{ISR}^\circ = \frac{1}{2}\left(\frac{\gamma}{\kappa^2} + \frac{1}{2}\right) \qquad p \neq q. \tag{54}$$

Note that ISR° is lower bounded by $1/4$ regardless of the value of γ: this is a general property of orthogonal BSS techniques [16] and is the 'price to pay' for blindly trusting second order statistics *i.e.* for whitening. Thus rejection rates obtained under the whiteness constraint cannot be (asymptotically) better that $\frac{1}{4T}$.

For an asymmetric estimating function H_φ the ISR does not take such a simple form unless the common score φ is obtained by Pham's method (sec. V-B). One then finds $\mathrm{ISR}_{pq} = \mathrm{ISR}$ and $\mathrm{ISR}_{pq}^\circ = \mathrm{ISR}^\circ$ as

$$\mathrm{ISR} = \frac{1}{2}\left(\frac{1}{\gamma} + \frac{1}{\gamma+2}\right), \qquad \mathrm{ISR}^\circ = \frac{1}{2}\left(\frac{1}{\gamma} + \frac{1}{2}\right) \tag{55}$$

where the last equation stems from (54) because Pham's method guarantees $\gamma = \kappa$. These expressions show that both ISR and ISR° are minimized by maximizing γ; not surprisingly, γ can be shown to reach its maximum value $\gamma_\star$ precisely

when $\varphi = \psi$ where ψ is the score function corresponding to the true density of the sources:

$$\gamma_* \triangleq \mathrm{E}\psi^2(s)\mathrm{E}s^2 - \mathrm{E}^2[\psi(s)s]. \tag{56}$$

Note that the solution of (34) with $H = H_\psi$ then is the ML estimator based on the true model. It follows that the expression of ISR in (55) also is the asymptotic Cramér-Rao bound for source separation *i.e.* the best achievable ISR rate with T independent samples (see [59], [70], [62]).

Since the achievable performance depends on the magnitude of γ_*, this moment characterizes the hardness of the BSS problem with respect to source distribution. Not surprisingly, we can relate it to the non-Gaussianity of the sources as follows. As above, denote ψ the score function for the (true) distribution of s and denote ψ_n the score function for the *Gaussian* distribution with the same variance as s (this is just $\psi_n(s) = s/\mathrm{E}s^2$). A 'large' non-Gaussianity translates into a large difference between ψ and ψ_n. As we just saw, the measure of non-Gaussianity from the asymptotic point of view is measured by $\gamma_\star$. Indeed one finds:

$$\gamma_\star = \frac{\mathrm{E}(\psi(s) - \psi_n(s))^2}{\mathrm{E}(\psi_n(s))^2}. \tag{57}$$

See fig. 12 for the values of $\gamma_\star$ in three examples. For close-to-Gaussian sources, $\gamma_\star$ is (arbitrarily) small: in this case, according to (55) the best achievable rejection rates are about $\frac{1}{2\gamma_\star T}$ for both the symmetric and the asymmetric forms. This gives an idea of the minimum number of samples required to achieve a given separation. The other extreme is for sources which are far away from normality: the moment $\gamma_\star$ is not bounded above. In particular, it tends to ∞ when the source distributions tend to have a discrete or a bounded support. In the case of discrete sources, deterministic (error-free) blind identification is possible with a finite number of samples. In the case of sources with bounded support, the MSE of blind identification decreases at a much faster rate than the $1/T$ rate obtained for finite values of γ (see in particular [42]).

C. Equivariance and uniform performance

At first thought, the hardness of the BSS problem seems to depend on the distributions of the source signals and on the mixing matrix, with harder problems when sources are nearly Gaussian and when the mixing matrix is poorly conditioned. This is not correct however: the BSS problem is 'uniformly hard in the mixing matrix'. Let us summarize the instances where this property appeared: the ultimate separation performance depends only on $\gamma_\star$ (eq. (55)); the asymptotic performance index in eqs. (54) and (55) depend only on some statistical moments; the stability of the adaptive algorithms (42) also depends only on the values of κ_i's; even better, the trajectory (43) of the global system $C_t = B_t A$ does not depend on A whose sole effect is to determine the initial point.

Therefore, not only does the problem appears to be 'uniformly hard in the mixing matrix', but it exists estimation techniques with a statistical behavior (regarding signal separation) which is independent of the particular value of the system to be inverted. This is a very desirable property: such algorithms can be studied and tuned independently of the particular mixture to be inverted; their performance can also be predicted independently of the mixture [17]. This is an instance of 'equivariance', a property holding more generally in transformation models.

There is a simple prescription to design algorithms with uniform performance: adjust freely (*i.e.* without constraint) the separating matrix according to a rule expressed *only* in terms of the output $\mathbf{y}$. To understand why the 'output only' prescription ensures uniform performance, consider for instance using a particular estimating function $H(\cdot)$ to separate a mixture of T samples $[\mathbf{s}(1), \ldots, \mathbf{s}(T)]$. If the source signals are mixed by a given matrix A, then a solution of (34) is a matrix B such that $BA = \widehat{C}$ where matrix $\widehat{C}$ is a solution of $T^{-1}\sum_{t=1}^{T} H(\widehat{C}\mathbf{s}(t)) = 0$. Matrix $\widehat{C}$ does not depend on A so that the global system $BA = \widehat{C}$ is itself independent of A and the estimated signals are $\mathbf{y}(t) = \widehat{C}\mathbf{s}(t)$ regardless of A. In particular, the recovered signals are exactly identical to those that would be obtained with $A = I$ *i.e.* when there is *no mixing at all*. This argument, based on estimating equations, extends to the minimizers of contrast functions since the latter are defined as functions of the distribution of the output (the argument also apply to orthogonal contrast functions because the whiteness constraint is expressed only in terms of $\mathbf{y}$). The argument also justifies the specific definition of the 'relative gradient': a device was needed to express the first-order variations of a contrast function $\phi[\mathbf{y}]$ in terms of a variation of $\mathbf{y}$ itself *i.e.* *without* reference to B. Finally, it must be stressed that the argument does not involve asymptotics: equivariance is exactly observed for any finite value of T.

Not all BSS algorithms are equivariant. For instance, the original algorithm of Jutten and Hérault imposes constraints on the separating matrix resulting in a greatly complicated analysis (and behavior) (see [35], [40], [49]). Other instances of non equivariant techniques is to be found in most of the algebraic approaches (see sec. VII) based on the structure of the cumulants of the *observed* vector $\mathbf{x}$. Precisely because the identification is based on

$\mathbf{x}$ and not on $\mathbf{y}$, such approaches are not equivariant in general unless they can be shown to be equivalent to the optimization of a contrast function of $\mathbf{y}$.

A word of caution is necessary before concluding: equivariance holds exactly in the noise-free model which we have considered so far. In practice, there is always some kind of noise which must be taken into account. Assume that a better model is $\mathbf{x} = A\mathbf{s} + \mathbf{n}$ where $\mathbf{n}$ represents an additive noise. This can be rewritten as $\mathbf{x} = A(\mathbf{s} + A^{-1}\mathbf{n})$. As long as $A^{-1}\mathbf{n}$ can be neglected with respect to $\mathbf{s}$, this is a noise-free situation. This shows the limit of equivariance: a poorly conditioned matrix A has a large inverse which amplifies the effect of the noise. More precisely, we can expect equivariance in the high SNR domain *i.e.* when the covariance matrix of $\mathbf{s}$ remains 'larger' than the covariance matrix of $A^{-1}\mathbf{n}$.

VII. Conclusions

Due to limited space, focus was given to principles and many interesting issues have been left out: discussion of the connections between BSS and blind deconvolution; convergence rates of adaptive algorithms; design of consistent estimators based on noisy observations, detection of the number of sources, etc... Before concluding, we briefly mention some other points.

Algebraic approaches. The 4th order cumulants of $\mathbf{x}$ have a very regular structure in the BSS model:

$$\mathcal{C}_{ijkl}[\mathbf{x}] = \sum_{p=1}^{n} k_p A_{ip} A_{jp} A_{kp} A_{lp} \qquad 1 \leq i,j,k,l \leq n. \tag{58}$$

Given sample estimates of the cumulants, the equation set (58) (or some subset of them) can be solved in A in a least square sense. This is a cumulant matching approach [71][43] which does not yield equivariant estimates. Optimal matching, though, can be shown to correspond to a contrast function [20]. However, the specific form of (58) also calls for algebraic approaches. Simple algorithms can be based on the eigen-structure of 'cumulant matrices' built from cumulants [63], [65]. An exciting direction of research is to investigate high-order decompositions that would generalize matrix factorizations like SVD or EVD to 4th order cumulants [33], [30], [15], [27], [21].

Using temporal correlation. The approaches to BSS described above exploit only properties of the distribution of $\mathbf{x}(t)$. If the source signals are *temporally correlated*, time structures can also be exploited. It is possible to achieve separation if all the source signals have distinct spectra even if each source signal is a Gaussian process [67]. Simple algebraic techniques can be devised (see [66], [11]); the Whittle approximation to the likelihood is investigated in [61]. Cyclostationary properties, when they exist, can also be exploited [13].

Deterministic identification. As indicated in sec. VI-B, sources with discrete support allow for deterministic identification (infinite Fisher information). Specific contrast functions can be devised [42] to take advantage of discreteness. There is a rich domain of application with digital communication signals coding information with discrete symbols by which deterministic identification is possible. See the review by Van der Veen [68] and the papers on CMA in this issue.

Open problems and perspectives

1. *Learning source distributions.* In the BSS problem, source distributions are a nuisance parameter. For large enough sample size, it is possible to *estimate* the distributions and still obtain the same asymptotic performance as if the distributions were known in advance[3]; the design of practical algorithms achieving 'source adaptivity' still is an open question.
2. *Dealing with noise.* BSS techniques remaining consistent in presence of additive noise have not been described here. For additive Gaussian noise, such techniques may resort to high-order cumulants or to noise modeling. It is not clear however that it is worth combating the noise. As a matter of fact, one may argue that taking noise effects into account is unnecessary at high SNR and futile at low SNR (because the BSS problem becomes too difficult anyway). Therefore, we believe it is still an open question to determine which application domains would really benefit from noise modeling.
3. *Global convergence.* Some cumulant based contrast functions can be proved to be free of spurious local minima in the two-source case (see *e.g.* [29]) or in a 'deflation approach' (successive extractions of the source signals) [34][45]. There is however a lack of general understanding of the global shape of contrast functions in the general case.
4. *Multidimensional independent components.* An interesting original variation of the basic ICA model would be to decompose a random vector in a sum of independent components with the requirement that the components are linearly independent but not necessarily one-dimensional. In the BSS model, this would be equivalent to grouping the source signals in subsets with independence between the subsets but not *within* the subsets. This more general decomposition could be called 'multidimensional independent component analysis' (MICA).
5. *Convolutive mixtures.* The most challenging open problem in BSS probably is the extension to *convolutive mixtures.* This is a very active area of research, mainly motivated by applications in the audio-frequency domain where the BSS is of-

ten termed the 'cocktail-party problem'. The convolutive problem is significantly harder than the instantaneous problem: even input-output (*i.e.* non blind) identification is a very challenging because of the large number of parameters usually necessary to describe audio channels.

6. *When the model does not hold.* The introduction mentioned successful applications of BSS to biomedical signals. When examining these data, it is very striking to realize that the extracted source signals seem to be very far to obeying the simple BSS model. The fact that BSS still yields apparently meaningful (to the experts) results is worth of consideration. A partial explanation stems from basing separation on contrast functions: even if the model does not hold (there are no independent source signals and no system A to be inverted), the algorithms still try to produce output which are 'as independent as possible'. This does not tell the whole story though because for many data sets a stochastic description does not seem appropriate. We believe it will a very interesting challenge to understand the behavior of BSS algorithms when applied 'outside the model'.

We are indebted to the anonymous reviewers whose constructive comments helped us improving on a first version of this paper.

References

[1] The ICA-CNL group at the Salk Institute. `http://www.cnl.salk.edu/~tewon/ica_cnl.html`.

[2] S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

[3] S.-I. Amari and J.-F. Cardoso. Blind source separation — semiparametric statistical approach. *IEEE Trans. on Sig. Proc.*, 45(11):2692–2700, Nov. 1997. Special issue on neural networks.

[4] S.-I. Amari, T.-P. Chen, and A. Cichocki. Stability analysis of adaptive blind source separation. *Neural Networks*, 10(8):1345–1351, 1997.

[5] S.-I. Amari, A. Cichocki, and H. Yang. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems, 8*, pages 757–763, Denver, 1996. MIT Press.

[6] K. Anand, G. Mathew, and V. Reddy. Blind separation of multiple co-channel BPSK signals arriving at an antenna array. *IEEE Signal Proc. Letters*, 2(9):176–178, Sept. 1995.

[7] A. Back and A. Weigend. A first application of independent component analysis to extracting structure from stock returns. *Int. Journal of Neural Systems*, vol. 8, no 4, pp. 473-484, Aug. 1997.

[8] Y. Bar-Ness. Bootstraping adaptive interference cancelers: Some practical limitations. In *Proc. Globecom*, pages 1251–1255, Nov. 1982.

[9] A. J. Bell and T. J. Sejnowski. An information-maximisation approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1004–1034, 1995.

[10] A. J. Bell and T. J. Sejnowski. Edges are the 'independent components' of natural scenes. In *Advances in Neural Information Processing Systems, 9*. MIT Press, 1996.

[11] A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and Éric Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on Sig. Proc.*, 45(2):434–44, Feb. 1997.

[12] A. Benveniste, M. Goursat, and G. Ruget. Robust identification of a non-minimum phase system. Blind adjustment of a linear equalizer in data communication. *IEEE Tr. on AC*, 25(3):385–399, 1980.

[13] B.G. Agee and S.V. Schell and W.A. Gardner. Spectral self-coherence restoral: A new approach to blind adaptive signal extraction using antenna arrays. *Proceedings of the IEEE*, pages 753–766, Apr. 1990.

[14] X.-R. Cao and R.-W. Liu. General approach to blind source separation. *IEEE Trans. Signal Processing*, 44(3):562–571, Mar. 1996.

[15] J.-F. Cardoso. Super-symmetric decomposition of the fourth-order cumulant tensor. Blind identification of more sources than sensors. In *Proc. ICASSP*, pages 3109–3112, 1991.

[16] J.-F. Cardoso. On the performance of orthogonal source separation algorithms. In *Proc. EUSIPCO*, pages 776–779, Edinburgh, Sept. 1994.

[17] J.-F. Cardoso. The equivariant approach to source separation. In *Proc. NOLTA*, pages 55–60, 1995.

[18] J.-F. Cardoso. Estimating equations for source separation. In *Proc. ICASSP'97*, pages 3449–52, 1997.

[19] J.-F. Cardoso. Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4(4):112–114, Apr. 1997.

[20] J.-F. Cardoso, S. Bose, and B. Friedlander. On optimal source separation based on second and fourth order cumulants. In *Proc. IEEE Workshop on SSAP, Corfou, Greece*, 1996.

[21] J.-F. Cardoso and P. Comon. Tensor based independent component analysis. In *Proc. EUSIPCO*, pp. 673-676, 1990.

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

[23] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6):362–370, Dec. 1993.

[24] E. Chaumette, P. Comon, and D. Muller. ICA-based technique for radiating sources estimation: application to airport surveillance. *IEE Proceedings-F*, 140(6):395–401, Dec. 1993.

[25] A. Cichocki, R. Unbehauen, and E. Rummert. Robust learning algorithm for blind separation of signals. *Electronic letters*, 30(17):1386–87, 1994.

[26] P. Comon. Independent component analysis, a new concept ? *Signal Processing, Elsevier*, 36(3):287–314, Apr. 1994. Special issue on Higher-Order Statistics.

[27] P. Comon and B. Mourrain. Decomposition of quantics in sums of powers of linear forms. *Signal Processing*, 53(2):93–107, Sept. 1996.

[28] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley series in telecommunications. John Wiley, 1991.

[29] A. Dapena, L. Castedo, and C. Escudero. An unconstrained single stage criterion for blind source separation. In *Proc. ICASSP*, volume 5, pages 2706–9, 1996.

[30] L. De Lathauwer, B. De Moor, and J. Vandewalle. Blind source separation by higher order singular value decomposition. In *Proc. EUSIPCO*, volume 1, pages 175–178, 1994.

[31] L. De Lathauwer, B. De Moor, and J. Vandewalle. Fetal electrocardiogram extraction by source subspace separation. In *Proc. HOS'95*, pages 134–8, Aiguablava, Spain, June 1995.

[32] L. De Lathauwer, B. De Moor, and J. Vandewalle. Blind source separation by simultaneous third-order tensor diagonalization. In *Proc. EUSIPCO*, Trieste, pp. 2089-2092, 1996.

[33] L. De Lathauwer, B. De Moor, and J. Vandewalle. Independent component analysis based on higher-order statistics only. In *Proc. IEEE SSAP workshop, Corfu*, pages 356–359, 1996.

[34] N. Delfosse and P. Loubaton. Adaptative blind separation of independent sources: a deflation approach. *Signal Processing*, 45(1):59–83, 1995.

[35] Y. Deville. A unified stability analysis of the Hérault–

Jutten source separation neural network. *Signal Processing*, 51(3):229–233, June 1996.

[36] Y. Deville and L. Andry. Application of blind source separation techniques to multi-tag contactless identification system. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E79-A(10):1694–99, 1996.

[37] D. Donoho. On minimum entropy deconvolution. In *Applied time-series analysis II*, pages 565–609. Academic Press, 1981.

[38] G. d'Urso, P. Prieur, and C. Vincent. Blind identification methods applied to EDF civil works and power plants monitoring. In *Proc. HOS'97*, pages –, Banff, Canada, June 1997.

[39] L. Féty. Méthodes de traitement d'antenne adaptées aux radio-communications. *Thèse de doctorat. Télécom Paris*, June 1988.

[40] J.-C. Fort. Stability of the source separation algorithm of Jutten and Hérault. In T. Kohonen, Makasira, Simula, and Kangas, editors, *Artificial Neural Networks*, pages 937–941. Elsevier, 1991.

[41] M. Gaeta and J.-L. Lacoume. Source separation without a priori knowledge: the maximum likelihood solution. In *Proc. EUSIPCO*, pages 621–624, 1990.

[42] F. Gamboa and Élizabeth Gassiat. Source separation when the input sources are discrete or have constant modulus. *IEEE Trans. Signal Processing*, 45(12):3062–72, 1997.

[43] G. Giannakis and S. Shamsunder. Modelling of non Gaussian array data using cumulants: DOA estimation of more sources with less sensors. *Signal Processing*, 30(3):279–297, July 1993.

[44] 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 *Proc. GRETSI*, pages 1017–1020, Nice, France, 1985.

[45] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–92, 1997.

[46] C. Jutten and J. Herault. Blind separation of sources I. An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, July 1991.

[47] J. Karhunen, A. Hyvarinen, R. Vigario, J. Hurri, and E. Oja. Applications of neural blind separation to signal and image processing. In *Proc. ICASSP*, volume 1, pages 131–4, 1997.

[48] J. Karhunen, E. Oja, L. Wang, R. Vigário, and J. Joutsensalo. A class of neural networks for independent component analysis. *IEEE Transactions on Neural Networks*, 8(3):486–504, May 1997.

[49] O. Macchi and Éric Moreau. Self-adaptive source separation, Part I: Convergence analysis of a direct linear network controled by the Hérault-Jutten algorithm. *IEEE Trans. Signal Processing*, 45(4):918–926, Apr. 1997.

[50] D. J. C. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. in preparation, 1996.

[51] S. Makeig, A. Bell, T.-P. Jung, and T. J. Sejnowski. Independent component analysis of electroencephalographic data. In *Advances in Neural Information Processing Systems, 8*. MIT Press, 1995.

[52] P. McCullagh. *Tensor Methods in Statistics*. Monographs on Statistics and Applied Probability. Chapman and Hall, 1987.

[53] J. Moody and L. Wu. What is the 'true price'?— State space models for high frequency financial data. In *Progress in Neural Information Processing. Proceedings of the International Conference on Neural Information Processing*, volume 2, pages 697–704. Springer-Verlag, 1996.

[54] E. Moreau and O. Macchi. High order contrasts for self-adaptive source separation. *International Journal of Adaptive Control and Signal Processing*, 10(1):19–46, jan 1996.

[55] J.-P. Nadal and N. Parga. Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer. *NETWORK*, 5:565–581, 1994.

[56] D. Obradovic and G. Deco. An information theory based learning paradigm for linear feature extraction. *Neurocomputing*, 12:203–221, 1996.

[57] B. A. Pearlmutter and L. C. Parra. A context-sensitive generalization of ICA. In *International Conference on Neural Information Processing*, Hong Kong, 1996.

[58] J. Pfanzagl. Asymptotic expansions related to minimum contrast estimators. *The Annals of Statistics*, 1(6):993–1026, 1973.

[59] D.-T. Pham. Blind separation of instantaneous mixture of sources via an independent component analysis. Research report RT 119, LMC IMAG, Grenoble, France, 1994.

[60] D.-T. Pham. Blind separation of instantaneous mixture of sources via an independent component analysis. *IEEE Trans. on Sig. Proc.*, 44(11):2768–2779, Nov. 1996.

[61] D.-T. Pham and P. Garat. Séparation aveugle de sources temporellement corrélées. In *Proc. GRETSI*, pages 317–320, 1993.

[62] 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.

[63] V. C. Soon, L. Tong, Y. F. Huang, and R. Liu. An extended fourth order blind identification algorithm in spatially correlated noise. In *Proc. ICASSP*, pages 1365–1368, 1990.

[64] A. Swindlehurst, M. Goris, and B. Ottersten. Some experiments with array data collected in actual urban and suburban environments. In *IEEE workshop on Signal Processing Advances in Wireless Communications*, pages 301–304, Apr. 1997.

[65] L. Tong, Y. Inouye, and R. Liu. Waveform preserving blind estimation of multiple independent sources. *IEEE Tr. on SP*, 41(7):2461–2470, July 1993.

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

[67] L. Tong, V. Soon, Y. Huang, and R. Liu. A necessary and sufficient condition for the blind identification of memoryless systems. In *Proc. ISCAS*, volume 1, pages 1–4, Singapore, 1991.

[68] A.-J. Van der Veen. Blind beamforming. *This issue*, 1998.

[69] H. H. Yang and S. Amari. Adaptive on-line learning algorithms for blind separation — maximum entropy and minimum mutual information. *Neural Computation*, 9(7):1457–1482, oct 1997.

[70] D. Yellin and B. Friedlander. Multi-channel system identification and deconvolution: performance bounds. In *Proc. IEEE SSAP workshop, Corfu*, pages 582–585, 1996.

[71] N. Yuen and B. Friedlander. Asymptotic performance analysis of blind signal copy using fourth-order cumulants. *International Journal of Adaptive Control and Signal Processing*, pages 239–65, mar 1996.