Equivariant adaptive source separation

Jean-François Cardoso and Beate Laheld.

Abstract— **Source separation consists in recovering a set of independent signals when only mixtures with unknown coefficients are observed. This paper introduces a class of adaptive algorithms for source separation which implements an adaptive version of equivariant estimation and is henceforth called EASI (Equivariant Adaptive Separation via Independence). The EASI algorithms are based on the idea of** *serial updating*: **this specific form of matrix updates systematically yields algorithms with a simple, parallelizable structure, for both real and complex mixtures. Most importantly, the performance of an EASI algorithm does not depend on the mixing matrix. In particular, convergence rates, stability conditions and interference rejection levels depend only on the (normalized) distributions of the source signals. Close form expressions of these quantities are given via an asymptotic performance analysis. This is completed by some numerical experiments illustrating the effectiveness of the proposed approach.**

Keywords— **Source separation, blind array processing, multichannel equalization, signal copy, adaptive signal processing, high order statistics, equivariant estimation.**

Introduction

The problem of *blind separation of sources* has received some attention in the recent signal processing literature, sometimes under different names: blind array processing, signal copy, independent component analysis, waveform preserving estimation... In all these instances, the underlying model is that of n statistically independent signals whose m (possibly noisy) linear combinations are observed; the problem consists in recovering the original signals from their mixture.

The 'blind' qualification refers to the coefficients of the mixture: no *a priori* information is assumed to be available about them. This feature makes the blind approach extremely versatile because it does not rely on modeling the underlying physical phenomena. In particular, it should be contrasted with standard narrow band array processing where a similar data model is considered but the mixture coefficients are assumed to depend in a known fashion on the location of the sources. When the propagation conditions between sources and sensors, the sensor locations, or the receivers characteristics are subject to unpredictable variations or are too difficult to model with accuracy (think of multipaths in urban environment), it may be safer to resort to a blind procedure for recovering the source signals.

This paper addresses the issue of *adaptive* source separation and consider the case where any additive noise can be neglected. The signal model then is that of a m-dimensional time series $\mathbf{x}_t$ in the form :

$$\mathbf{x}_t = A\mathbf{s}_t \quad t = 1, 2, \cdots \tag{1}$$

where $\mathbf{x}_t$ and $\mathbf{s}_t$ are column vectors of sizes m and n respectively and A is a $m \times n$ matrix. The idea here is that vector $\mathbf{x}_t$ results from measurements by n sensors receiving contributions from n sources. Hence, the components of $\mathbf{s}_t$ are often termed 'source signals'. Matrix A is called the 'mixing matrix'.

Adaptive source separation consists in updating an $n \times m$ matrix B_t such that its output $\mathbf{y}_t$:

$$\mathbf{y}_t = B_t\mathbf{x}_t \tag{2}$$

is as close as possible to the vector $\mathbf{s}_t$ of the source signals (see fig. 1). Consider the global system denoted C_t, obtained by chaining the mixing matrix A and the separating matrix B_t, that is :

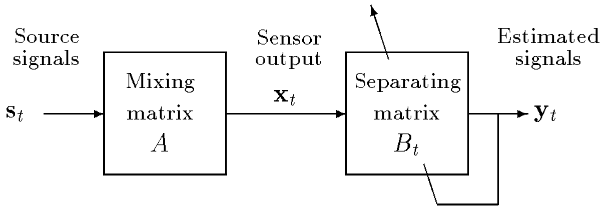

Fig. 1. Adapting a separating matrix

$$C_t \overset{\text{def}}{=} B_tA. \tag{3}$$

Ideally, an adaptive source separator should converge to a matrix $B_\star$ such that $B_\star A = I$, or, equivalently, the global system C_t should converge to the $n \times n$ identity matrix I.

Outline of the paper. The main point of this paper is to introduce and study 'serial updating' algorithms. Defining a serial updating algorithm consists in specifying an $n \times n$ matrix-valued function $\mathbf{y} \rightarrow H(\mathbf{y})$ which is used for updating B_t according to

$$B_{t+1} = B_t - \lambda_t H(\mathbf{y}_t)B_t \tag{4}$$

where, as above, $\mathbf{y}_t$ is the output of B_t and λ_t is sequence of positive adaptation steps.

After some background on the source separation problem in section I, the serial updating scheme is investigated in section II: it is shown to yield adaptive algorithms whose performance is *independent of the mixing matrix A*. When the algorithm is intended to optimize an objective function $c(B)$, we show that the required function $H(\cdot)$ may be obtained as the 'relative gradient' of the objective function. In section III, a particular function $H(\cdot)$ is obtained from a cumulant based approach to blind identification This is then generalized in section IV, into a family of adaptive source separation algorithms (35), called EASI for Equivariant Adaptive Separation via Independence, whose stability and asymptotic convergence are studied in section V. Section VI extends all the results to the complex case. This is completed in section VII by some numerical experiments illustrating the effectiveness of the approach.

I. Source separation

A. Assumptions and notations.

Some notational conventions are: scalars in lower case, matrices in upper case, and vectors in boldface lowercase. The i-th component of a vector, say $\mathbf{x}$, is denoted x_i. The expectation operator is E and transposition is indicated by supscript T. The identity matrix is denoted I; throughout, it is the $n \times n$ identity.

The following assumptions hold throughout.

Assumption 1. *Matrix A is full rank with $n \leq m$.*

Assumption 2. *Each component of $\mathbf{s}_t$ is a stationary zero-mean process.*

Assumption 3. *At each t, the components of $\mathbf{s}_t$ are mutually statistically independent.*

Assumption 4. *The components of $\mathbf{s}_t$ have unit variance.*

Some comments are in order. Assumption 3 is the key ingredient for source separation. It is a strong statistical hypothesis but a physically very plausible one since it is expected to be verified whenever the source signals arise from physically separated systems. Regarding assumption 4, we note that it is only a *normalization convention* since the amplitude of each source signal can be incorporated into A. We note that assumptions 2, 3 and 4 combine into

$$R_s \stackrel{\text{def}}{=} \mathrm{E}\left[\, \mathbf{s}_t \mathbf{s}_t^{\mathrm{T}} \right] = I. \tag{5}$$

Assumption 1 is expected to hold 'almost surely' in any physical situation. More important is the existence of A itself *i.e.* the possibility of observing instantaneous mixtures.

Instantaneous mixtures occur whenever the difference of time of arrival between two sensors can be neglected or approximated by a phase shift so that the propagation from sources to sensors can be represented by a scalar factor : the relation between the emitted signals and the signals received on the sensors then amounts to a simple matrix multiplication as in (1). This kind of instantaneous mixtures is the standard model in narrow-band array processing. In this context, one must then consider *complex* analytic signals and a *complex* mixing matrix A. For ease of exposition, most of the results are derived in the real case; extension to the complex case is is straightforward and described in section VI.

Finally, for source separation to be possible, there are conditions on the probability distribution of the source signals. Since this condition is algorithm-dependent, its formulation is deferred to section V-A. Anticipating a bit, we mention that at most one source signal may be normally distributed.

Before starting, it is important to mention a technical difficulty, due to the following fact: without additional information (such as spectral content, modulation scheme, etc...), the outputs of a separating matrix cannot be ordered since the ordering of the source signals is itself immaterial (conventional): the individual source signals can be estimated up to an indetermination. Also a scalar factor can be exchanged between each source signal and the corresponding column of matrix A without modifying the observations. Hence, even with the normalization convention implied by assumption 4, the sign (real case) or the phase (complex case) of each signal remains unobservable. This may be formalized using the following definitions: any matrix which is the product of a permutation matrix with a diagonal matrix with unit-norm diagonal elements is called a *quasi-identity* matrix; any matrix $B_\star$ is said to be a *separating matrix* if the product $B_\star A$ is a quasi-identity matrix.

The adaptive source separation problem then consists in updating an $n \times m$ matrix B_t such that it converges to a separating matrix or, equivalently, such that the global system $C_t = B_t A$ converges to a quasi-identity matrix. The issue of indetermination is addressed at length in [24].

B. Approaches to source separation

The seminal paper on source separation is [17]. Therein, the separating matrix B is parameterized as $B = (I+W)^{-1}$ and the off-diagonal entries of W are updated with a rule like $w_{ij} \leftarrow w_{ij} - \lambda f(y_i) g(y_j)$ where f and g are odd functions. If separation is achieved, each y_i is proportional to some s_j so that by the independence assumption: $\mathrm{E}[f(y_i)g(y_j)] = \mathrm{E}f(y_i)\mathrm{E}g(y_j)$ which cancels for symmetrically distributed sources. Hence, any separating matrix is an equilibrium point of the algorithm. This kind of equilibrium condition also appears in [12]. The Jutten-Hérault algorithm is inspired by a neuromimetic approach; this line is further followed by Karhunen [18] and Chicocki [7].

Nonlinear distortions of the output $\mathbf{y}$ also appear when the equilibrium condition stems from minimization of some measure of independence between the components of $\mathbf{y}$. When independence is measured by the cancelation of some 4th-order cumulants of the output, cubic nonlinearities show up, as in [11], [19].

When the sources have a known differentiable density of probability (ddp), the maximum likelihood (ML) estimator is easily obtained in the i.i.d. case; the (asymptotically optimal) nonlinearities are the log derivatives of the ddp's [20]. See also [2] for an ML approach for with discrete sources in unknown Gaussian noise.

Our starting point for finding a $H(\cdot)$ function required for serial updating is the idea of 'orthogonal contrast functions'. In the context of source separation, these were introduced by Comon [9] as functions of the distribution of $\mathbf{y}$ which are to be optimized under a whiteness constraint: $R_y = \mathrm{E}\mathbf{y}\mathbf{y}^{\mathrm{T}} = I$. Comon suggest minimizing the squared cross-cumulants of the components of $\mathbf{y}$. This orthogonal contrast is also arrived at by Gaeta and Lacoume [14] as a Gram-Charlier approximation of the likelihood. A similar (and asymptotically equivalent) contrast which can be efficiently optimized by a Jacobi-like algorithm, especially in the complex case, is described in [6].

When the sources have kurtosis of identical signs, simpler orthogonal contrasts may be exhibited. For instance, if all the sources have a negative kurtosis, the minimization of

$$\phi_4(B) \stackrel{\text{def}}{=} \mathrm{E}[\sum_{i=1,n} |y_i|^4] \tag{6}$$

subject to $R_y = I$ is achieved only when B is a separating matrix. This is a strongly reminiscent of 4th-order objectives used in blind equalization [23]. This contrast lends itself more easily to adaptive minimization since it is the expectation of a function of the output vector $\mathbf{y}$. It is used in [11] where it is optimized by a deflation technique. The resulting adaptive algorithm can be proved to be asymptotically free of spurious attractors, but the implementation is not simple.

Before closing this section, other batch estimation techniques may be mentionned: higher-order cumulants are used together with a prewhitening strategy in Tong and al. [24], [25]; fourth-order-only is investigated in [5], [4]; purely second-order is possible if the sources have different spectra as investigated in [13], [21], [1], [24] and also in [15] in an adaptive implementation.

C. *Equivariant source separation.*

Our approach to adaptive source separation may be motivated by first considering *batch* estimation. Consider the problem of estimating matrix A form T samples $X_T = [\mathbf{x}(1), \ldots, \mathbf{x}(T)]$ where we assume for simplicity that $n = m$ (as many sources as 'sensors'). A blind estimator of A is, by definition, a function of X_T only. This may be denoted by:

$$\widehat{A} = \mathcal{A}(X_T). \tag{7}$$

A particular estimator is said to be *equivariant* if it satisfies

$$\mathcal{A}(MX_T) = M\mathcal{A}(X_T) \tag{8}$$

for any invertible $n \times n$ matrix M. Equivariant estimation is in fact a broader notion which is relevant whenever the parameters to be estimated form a group. This is indeed the case here with the multiplicative group of invertible matrices.

The equivariance property is quite natural in the context of source separation. For instance, M-estimators [16] which compute $\widehat{A}$ as the solution of an estimation equation in the form

$$T^{-1} \sum_{t=1,T} H(A^{-1}\mathbf{x}(t)) = 0 \tag{9}$$

are easily seen to be equivariant. The ML estimator in the i.i.d. case is an instance of M-estimator. In equation (9), the vector-to-matrix function H is as in (4): the serial algorithm (4) is a stochastic solver of equation $\mathrm{E}H(\mathbf{y}) = 0$, while the M-estimator defined by eq. (9) solves the sample version of $\mathrm{E}H(\mathbf{y}) = 0$.

The point to be made here is that, in the context of source separation, equivariant estimators exhibit *uniform performance.* This is to be understood in the following sense. Assume that the source signals are estimated as $\widehat{\mathbf{s}}(t) = (\widehat{A})^{-1}\mathbf{x}(t)$ where $\widehat{A}$ is obtained from an equivariant estimator. Then

$$\widehat{\mathbf{s}}(t) = [\mathcal{A}(X_T)]^{-1}\mathbf{x}(t) = [\mathcal{A}(AS_T)]^{-1}A\mathbf{s}(t) = [\mathcal{A}(S_T)]^{-1}\mathbf{s}(t) \tag{10}$$

The last equality is obtained thanks to the equivariance property (8) and reveals that the source signals estimated by an equivariant equivariant estimator $\mathcal{A}$ for a particular realization $S_T = [\mathbf{s}(1), \ldots, \mathbf{s}(T)]$ depend only on S_T but *do not depend on the mixing matrix A*. It follows that, in terms of signal separation, the performance of an equivariant algorithm does not depend at all on the mixing matrix.

That the performance of a batch algorithm may not depend on the 'hardness' of the mixture is a very desirable property. However, *adaptive* source separation is addressed here: next section actually shows how 'uniform performance properties' can be inherited by an adaptive algorithm from a batch estimation procedure.

II. Serial matrix updating

A. *Serial updates*

The adaptation rule (4) is termed a 'serial update', because it reads equivalently $B_{t+1} = (I - \lambda_t H_t)B_t$. This later form evidences that B_t is updated by 'plugging' matrix $I - \lambda_t H_t$ at the *output* of the current system B_t to get the updated system B_{t+1} (see fig. 2). This could be opposed to 'parallel updating' which would consist in adding a small matrix to B_t rather than multiplying it with a matrix close to the identity. Of course, any serial update also *is* a parallel update where B_t is updated by (formally) plugging $-\lambda_t H_t B_t$ between its input and output. However, not every parallel update can be seen as a serial update because we specifically require that the variation of B_t is in the form $-\lambda_t H_t B_t$ where H_t depends *only* on the output vector $\mathbf{y}_t$.

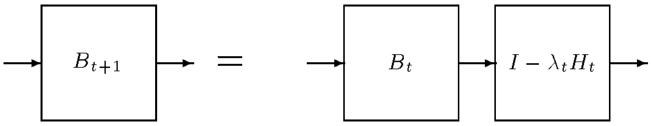

Fig. 2. Serial update

Note the following two facts. On one hand, uniform performance of equivariant batch algorithms is a direct consequence of (8) which is a *multiplicative* equation. On the other hand, the system B_t is serially updated by left *multiplication* by matrix $I - \lambda_t H_t$. Thus, the group structure underlying equivariance is turned into an updating rule. We show below that this simple fact actually leads to uniform performance adaptive algorithms. This is then further specialized to the case of gradient descent algorithm. Again, we take advantage of the existence of the matrix product to define a 'relative gradient' which is consistent with the notion of serial updating. The idea is that when *matrices* are to be updated, specific rules may be considered which have no equivalent for a generic adaptive system with an unstructured vector of parameters.

B. *Serial updates and uniform performance*

The benefits of serial updating are revealed by considering the global mixing-unmixing system $C_t = B_t A$. Its evolution under the updating rule (4) is readily obtained by right multiplication of (4) by matrix A, yielding

$$C_{t+1} = C_t - \lambda_t H(C_t \mathbf{s}_t) C_t \tag{11}$$

where we used $\mathbf{y} = B\mathbf{x} = BA\mathbf{s} = C\mathbf{s}$. Hence, the global system C_t also undergoes serial updating, an obvious fact anyway in the light of figure 2. This is a trivial but remarkable result because it means that, under serial updating, the evolution law of the global system is independent of the mixing matrix A in the sense described below. The reader will notice that the argument parallels the one used in previous section regarding batch algorithms.

Assume the algorithm is initialized with some matrix B_o so that the global system has initial value $C_o = B_oA$. By equation (11), the subsequent trajectory $\{C_t|t > 1\}$ of the global system will be *identical* to the trajectory that would be observed for another mixing matrix A', provided the initial point is $B'_o = B_oAA'^{-1}$. This is pretty obvious since in both cases, the *global* system starts from the same initial condition and evolves according to (11) which involves only the *source* signals and C_t. Hence, with respect to the global system C_t, changing the mixing matrix A is tantamount to changing the initial condition B_0.

The key point here is that, since the issue is the separation of the source signals, the performance of a separating algorithm is completely characterized by the global system C_t and not by the individual values of B_t and A; this is because the amplitude of the j-th source signal in the estimate of the i-th source signal at time t is determined only by the (i,j)-th entry of C_t.

It follows that it is only needed to study the convergence of C_t to a quasi-identity matrix under the stochastic rule (11) to completely characterize a serial source separation algorithm.

In summary, serial updating is the only device needed to transfer the uniform performance of equivariant batch algorithms to an adaptive algorithm.

C. The relative gradient

A serial algorithm is determined by the choice of a specific function H. To obtain such a function, the notion of 'relative gradient' is instrumental. In this section, we denote $< \cdot|\cdot >$ the Euclidian scalar product of matrices:

$$< M|N >= \mathrm{Trace}[M^{\mathrm{T}}N] \qquad < M|M >= ||M||^2_{\mathrm{Fro}}. \quad (12)$$

Let $\phi(B)$ be an objective function of the $n \times m$ matrix B, differentiable with respect to the entries of B. The gradient of ϕ at point B is denoted $\frac{\partial\phi}{\partial B}$; it is the $n \times m$ matrix, depending on B, whose (i,j)th entry is $\frac{\partial\phi}{\partial b_{ij}}$. The first order expansion of ϕ at B then reads

$$\phi(B+\mathcal{E}) = \phi(B) + < \frac{\partial\phi}{\partial B}|\mathcal{E} > + o(\mathcal{E}). \quad (13)$$

In order to be consistent with the perturbation of B induced by the serial serial updating rule (4), we define the *relative gradient* of ϕ at B as the $n \times n$ matrix, denoted $\nabla\phi$, such that:

$$\phi(B+\mathcal{E}B) = \phi(B) + < \nabla\phi|\mathcal{E} > + o(\mathcal{E}). \quad (14)$$

There is no profound difference with the 'absolute gradient' though: one easily finds that $\nabla\phi = \frac{\partial\phi}{\partial B}B^{\mathrm{T}}$, but that the relative gradient is the appropriate quantity is confirmed in th following.

To illustrate the relevance of considering the relative gradient, we now compute it in the case where $\phi(B)$ is in the form $\phi(B) = \mathrm{E}f(\mathbf{y}) = \mathrm{E}f(B\mathbf{x})$. If function f is differentiable everywhere, one has

$$f(\mathbf{y}+\delta\mathbf{y}) = f(\mathbf{y}) + \mathbf{f}'(\mathbf{y})^{\mathrm{T}}\delta\mathbf{y} + o(\delta\mathbf{y}) \quad (15)$$

where $\mathbf{f}'(\mathbf{y})$ is the gradient of f at $\mathbf{y}$, *i.e.* it is the column vector whose i-th component is the partial derivative of $f(\mathbf{y})$ with respect to y_i. Computing the first order expansion in matrix $\mathcal{E}$ of $\phi(B+\mathcal{E}B)$ and comparing with (14) yields, after elementary manipulations, the relative gradient:

$$\nabla\mathrm{E}f(\mathbf{y}) = \nabla\mathrm{E}f(B\mathbf{x}) = \mathrm{E}\ [\ \mathbf{f}'(\mathbf{y})\mathbf{y}^{\mathrm{T}}]. \quad (16)$$

Note that this relative gradient depends only on the distribution of $\mathbf{y}$. This was to be expected since modifying B in to $B+\mathcal{E}B$ amounts to modifying $\mathbf{y}$ into $\mathbf{y}+\mathcal{E}\mathbf{y}$, regardless of the particular values of $\mathbf{x}$ or B. In view of (13), the gradient rule for minimizing $\phi(B)$ is to modify B into $B+\mathcal{E}B$ with $\mathcal{E} = -\lambda\nabla\phi$ because then the variation of ϕ is $< \nabla\phi|\mathcal{E} > + o(\mathcal{E}) = -\lambda||\nabla\phi||^2_{\mathrm{Fro}} + o(\lambda)$ which is negative if λ is a small enough positive scalar as long as $\nabla\phi \neq 0$. A stochastic relative gradient is obtained by deleting the expectation operator in (16), leading to the adaptation rule:

$$B_{t+1} = B_t - \lambda\ \mathbf{f}'(\mathbf{y}_t)\mathbf{y}_t^{\mathrm{T}}\ B_t. \quad (17)$$

for the stochastic minimization of $\mathrm{E}f(\mathbf{y})$.

The key point here is that the adaptation rule (17) actually is serial updating algorithm in the form (4) with $H(\mathbf{y}) = \mathbf{f}'(\mathbf{y})\mathbf{y}^{\mathrm{T}}$. According to the discussion of the previous section, it will enjoy uniform performance. The conclusion is that stochastic relative gradient algorithm yields adaptive algorithm in the serial form. Had we used the absolute gradient rather than the relative one, we would have found an updating rule *not* meeting the conditions for uniform performance, namely that H should depend on $\mathbf{y}$ only.

The process of obtaining function H via a relative gradient computation is not limited to the optimization of objectives in the form $\phi(B) = \mathrm{E}f(\mathbf{y})$. Recall in particular that equation (6) defines an 'orthogonal' contrast function for source separation, *i.e.* it is to be optimized under the constraint that the output of B is (spatially) white. Next section shows how the previous approach is easily adapted to yield the required $H(\cdot)$ function for orthogonally constrained optimization.

III. Serial updates for orthogonal contrasts

The contrast function ϕ_4 defined in (6) is in the form $\phi_4 = \mathrm{E}f(\mathbf{y})$ but must be optimized under the decorrelation constraint $R_y = \mathrm{E}\mathbf{y}\mathbf{y}^{\mathrm{T}} = I$. Batch procedures for optimizing contrast functions under this constraint have been described in [6], [9], [8]; they are based on factoring the separating matrix as $B = UW$ where W an $n \times m$ whitening matrix and U is an $n \times n$ orthogonal matrix: there is an

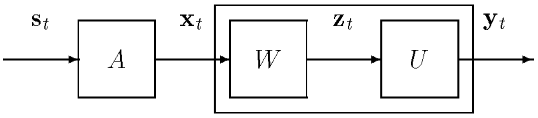

Fig. 3. A two-stage separation in batch processing

intermediate vector $\mathbf{z}_t = W\mathbf{x}_t$ and the estimated source signal vector is $\mathbf{y}_t = U\mathbf{z}_t$ (see figure 3). By definition, W is a whitening matrix if its output is spatially white *i.e.*:

$$I = R_z \overset{\text{def}}{=} \mathrm{E}\left[\, \mathbf{z}_t\mathbf{z}_t^{\mathrm{T}}\right] = WR_xW^{\mathrm{T}}. \tag{18}$$

The constraint $R_y = I$ is then satisfied if, and only if, U is orthogonal. Thus, after whitening of $\mathbf{x}$ into $\mathbf{z}$, the problem of minimizing a contrast function $\mathrm{E}f(\mathbf{y}) = \mathrm{E}f(B\mathbf{x})$ over B under the constraint $R_y = I$ becomes that of minimizing $\mathrm{E}f(\mathbf{y}) = Ef(U\mathbf{z})$ over U under the constraint that U is orthogonal.

We now show how this program is completed in the adaptive context with serial updates: serial updates of a whitening matrix W and of an orthogonal matrix U are first obtained and then combined into a unique serial updating rule for B.

A. Serial update of a whitening matrix

It is desired to adapt a matrix W such that it converges to a point where $R_z = I$. This is obtained by minimizing a 'distance' between R_z and I. The Kullback–Leibler divergence [10] between two zero-mean normal distributions with covariance matrices equal to R_z and I respectively is

$$K(R_z) \overset{\text{def}}{=} \mathrm{Trace}(R_z) - \log\det(R_z) - n. \tag{19}$$

Hence a whitening matrix is a minimizer of

$$\phi_2(W) \overset{\text{def}}{=} K(WR_xW^{\mathrm{T}}). \tag{20}$$

Computing the relative gradient is easily done in two steps. First, if W is modified into $W + \delta W = W + \mathcal{E}W$, the corresponding variation of $R_z = WR_xW^{\mathrm{T}}$ is

$$\delta R_z = \delta WR_xW^{\mathrm{T}} + WR_x\delta W^{\mathrm{T}} = \mathcal{E}R_z + R_z\mathcal{E}^{\mathrm{T}} \tag{21}$$

Second, the differential of function K is known to be

$$K(R_z + \delta R_z) = K(R_z) + \mathrm{Trace}\{(I - R_z^{-1})\delta R_z\} + o(\delta R_z). \tag{22}$$

Combining (21) and (22) yields, after simplification:

$$\nabla\phi_2 = 2(R_z - I) = 2\mathrm{E}[\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}} - I] \tag{23}$$

The serial adaptive whitener is obtained by deleting the expectation operator:

$$W_{t+1} = W_t - \lambda_t\, [\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}} - I]\, W_t. \tag{24}$$

Interestingly enough, this rule can be shown to correspond the first order (in λ) approximation of the Potter formula [22] for the recursive computation of the inverse square root of a covariance matrix estimated with an exponential window. In this instance, the serial approach is seen to correspond to an optimal solution.

B. Serial update of an orthogonal matrix

It is desired to adapt an orthogonal matrix U such that $\phi_4(U) = \mathrm{E}f(\mathbf{y}) = \mathrm{E}f(U\mathbf{z})$ is minimized. Unconstrained minimization of such an objective leads to the updating rule (17) which does *not* preserve the orthogonality of U. Orthogonality could be preserved by some parameterization of the orthogonal matrices (as product of Givens rotations for instance), but this solution is to be discarded because it would result in losing the uniform performance property of serial adaptation and also because we ultimately want to get rid of the factorization of B into two distinct matrices W and U. Hence, we rather stick to the idea that U should be updated in the form $U + \mathcal{E}U$ but note that if U is orthogonal, *i.e.* $UU^{\mathrm{T}} = I$, then

$$(U + \mathcal{E}U)(U + \mathcal{E}U)^{\mathrm{T}} = I + \mathcal{E} + \mathcal{E}^{\mathrm{T}} + \mathcal{E}\mathcal{E}^{\mathrm{T}} \tag{25}$$

so that the orthogonality of $U + \mathcal{E}U$ is preserved at first-order, *i.e.* $(U + \mathcal{E}U)(U + \mathcal{E}U)^{\mathrm{T}} = I + o(\mathcal{E})$, if $\mathcal{E}$ is skew-symmetric, *i.e.* verifies $\mathcal{E}^{\mathrm{T}} = -\mathcal{E}$.

Thus the (relative) gradient rule, which consists in aligning $-\mathcal{E}$ along the (relative) gradient $\mathrm{E}[\mathbf{f}'(\mathbf{y})\mathbf{y}^{\mathrm{T}}]$ cannot be followed since this gradient is not skew-symmetric. In order to satisfy the orthogonality constraint, matrix $-\mathcal{E}$ must be aligned along the orthogonal projection of the relative gradient onto the space of skew-symmetric matrices. This choice guarantees that matrix $-\mathcal{E}$ makes an acute angle with the relative gradient matrix, still resulting in a decrease of the objective function if λ is small enough. The orthogonal projection of $\nabla\phi_4$ onto the skew-symmetric matrix set is just $(\nabla\phi_4 - \nabla\phi_4^{\mathrm{T}})/2$ leading to the serial update:

$$U_{t+1} = U_t - \lambda_t\, [\mathbf{f}'(\mathbf{y}_t)\mathbf{y}_t^{\mathrm{T}} - \mathbf{y}_t\mathbf{f}'^{\mathrm{T}}(\mathbf{y}_t)]\, U_t. \tag{26}$$

Of course, such an updating rule does not preserve *exactly* unitarity, but only at first order in λ. Next section shows that this problem disappears when the whitening stage and the orthogonal stage are considered altogether.

C. The one-stage solution

A global updating rule for matrix $B = UW$ is obtained by computing $B_{t+1} = U_{t+1}W_{t+1}$ where W_{t+1} is given by (24) and U_{t+1} by (26). From (26), we readily obtain

$$U_{t+1}W_t = B_t - \lambda_t[\mathbf{f}'(\mathbf{y}_t)\mathbf{y}_t^{\mathrm{T}} - \mathbf{y}_t\mathbf{f}'^{\mathrm{T}}(\mathbf{y}_t)]B_t. \tag{27}$$

From (24) and using $U_t^{\mathrm{T}}U_t = I$ and $\mathbf{y}_t = U_t\mathbf{z}_t$, we get

$$\begin{aligned} U_tW_{t+1} &= B_t - \lambda_tU_t[\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}} - I]W_t \\ &= B_t - \lambda_t[\mathbf{y}_t\mathbf{y}_t^{\mathrm{T}} - I]B_t. \end{aligned} \tag{28}$$

There is no reason to use the same step size in (27) and (28), but since a ratio different from 1 could be integrated in f, we do assume here an identical value, and the resulting adaptation for B_t, dropping the term in λ_t^2, then just is

$$B_{t+1} = B_t - \lambda_t\; H(\mathbf{y}_t)\; B_t \tag{29}$$

where function $H(\mathbf{y})$ appears to be:

$$H(\mathbf{y}) = \mathbf{y}\mathbf{y}^{\mathrm{T}} - I + \mathbf{f}'(\mathbf{y})\mathbf{y}^{\mathrm{T}} - \mathbf{y}\mathbf{f}'(\mathbf{y})^{\mathrm{T}}. \tag{30}$$

Hence, we do arrive at an algorithm for updating a separating matrix B in the serial form. This completes the program of this section.

IV. The EASI algorithms

In the previous section, the notion of serial update applied to a 4th-order contrast function provided us with a specific form (30) for the function $H(\mathbf{y})$ required in the serial approach. The source separation algorithms to be considered in this paper improves on (30) by modifying it in two respects. First, we consider using functions other than $\mathbf{f}'(\mathbf{y})$ for increased flexibility. Second, stabilizing factors are introduced which are needed since finite adaptation steps are used in practice. This is discussed in the next two subsections and yields a family of adaptive source separation algorithms as summarized by eqs. (35) and (36).

A. Stationarity and non-linearities

An stationary point for a serial updating algorithm is any matrix B such that $\mathrm{E}H(\mathbf{y}) = 0$. For the serial algorithm derived in the previous section, *i.e.* for $H(\cdot)$ given by (30), this equation can be decomposed into symmetric and skew-symmetric parts, yielding:

$$\begin{aligned} \mathrm{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}] &= I & (31)\\ \mathrm{E}[\mathbf{f}'(\mathbf{y})\mathbf{y}^{\mathrm{T}} - \mathbf{y}\mathbf{f}'(\mathbf{y})^{\mathrm{T}}] &= 0. & (32) \end{aligned}$$

The condition (31) is that the output $\mathbf{y}$ is spatially white and matches the normalization convention (5). This condition ensures the *second-order* independence (*i.e.* decorrelation) of the separated signals. It is however clearly not sufficient for determining a separating matrix since, if the output $\mathbf{y}$ is further rotated by some orthogonal matrix, the condition $R_y = I$ is preserved but source separation is no longer achieved. Hence, other than second order conditions are required and these are provided by (32). If the components of $\mathbf{y}$ are mutually independent, then, for $i \neq j$, one has $\mathrm{E}[y_i f_j'(y_j)] = \mathrm{E}y_i \ \mathrm{E}f_j'(y_j)$ which cancels by the zero mean assumption, Thus condition (32) is satisfied if B is a separating matrix. This conclusion reached using only the fact that $\mathbf{f}'$ acts componentwise. Thus, defining a componentwise nonlinear function $\mathbf{g}$:

$$\mathbf{g}(\mathbf{y}) = [g_1(y_1), \ldots, g_n(y_n)]^{\mathrm{T}}, \qquad (33)$$

the form (30) may be generalized into

$$H(\mathbf{y}) = \mathbf{y}\mathbf{y}^{\mathrm{T}} - I + \mathbf{g}(\mathbf{y})\mathbf{y}^{\mathrm{T}} - \mathbf{y}\mathbf{g}(\mathbf{y})^{\mathrm{T}} \qquad (34)$$

with the separating matrices remaining stationary points of the rule (4). To any componentwise nonlinear function $\mathbf{g}$, we thus associate a corresponding EASI algorithm:

EASI algorithms for adaptive source separation

$$B_{t+1} = B_t - \lambda_t \left[\mathbf{y}_t\mathbf{y}_t^{\mathrm{T}} - I + \mathbf{g}(\mathbf{y}_t)\mathbf{y}_t^{\mathrm{T}} - \mathbf{y}_t\mathbf{g}(\mathbf{y}_t)^{\mathrm{T}}\right] B_t \qquad (35)$$

We note that the functions g_i must be nonlinear: if any two functions g_i and g_j are linear, then the corresponding entries in the skew-symmetric part of $H(\mathbf{y})$ provide only second-order equilibrium conditions which are redundant with those provided by the symmetric part of $H(\mathbf{y})$.

B. Normalization

In some applications like digital communications, fast convergence is required, implying the use of 'large' adaptation steps (say $\lambda > 10^{-2}$) which may cause explosive behavior if no special provisions are taken. We note that a stabilization procedure should not be based on clipping the entries of the separating matrix or renormalizing its rows. In fact, stabilization should *not* involve any action on the separating matrix itself, because this would spoil the uniform performance property. Hence, stabilization should rather be implemented by preventing $H(\cdot)$ to take too large values, suggesting the following normalized form:

Normalized EASI algorithms for adaptive source separation

$$B_{t+1} = B_t - \lambda_t \left[\frac{\mathbf{y}_t\mathbf{y}_t^{\mathrm{T}} - I}{1 + \lambda_t \ \mathbf{y}_t^{\mathrm{T}}\mathbf{y}_t} + \frac{\mathbf{g}(\mathbf{y}_t)\mathbf{y}_t^{\mathrm{T}} - \mathbf{y}_t\mathbf{g}(\mathbf{y}_t)^{\mathrm{T}}}{1 + \lambda_t \ |\mathbf{y}_t^{\mathrm{T}}\mathbf{g}(\mathbf{y}_t)|}\right] B_t \qquad (36)$$

which is very similar to the modification of the LMS algorithm into the 'normalized LMS'. It offers the following advantages. It entails very little extra computation with respect to (30) and it does not introduce additional parameter. Also, when the system is close to a stationary point, the covariance of $\mathbf{y}$ is close to the identity matrix so that, for reasonably small λ, the normalized version is expected to behave like the raw version (as confirmed in section VII) for which a detailed performance analysis is possible. Finally, the choice of the denominators is such that a natural protection against the outliers is granted. Finally, the normalized form has proved very satisfactory in the numerical experiments.

C. Discussion

Stability and permutations. The choice of the nonlinear function $\mathbf{g}$ is of course crucial to the performance of the algorithm. For any choice of $\mathbf{g}$, a separating matrix a stationary point but the real issue is the *stability* of the separating matrices. The stability condition is (48), established below by an asymptotic analysis which also give some clues as how to choose and scale the nonlinear functions $g_1, \ldots, g_n$. We note here that this analysis is led for C_t being close to the identity matrix, but the case where C_t converges to another permutation matrix reduces to the previous case by permuting accordingly the nonlinear functions acting at the output of B_t.

Uniform performance and the noise. The uniform convergence property rigorously holds if model (1) is verified exactly, as discussed above. In particular, one can deal with arbitrarily ill conditioned mixtures, a fact which may appear paradoxical : the intrinsic hardness of array processing is known to depend on the conditioning of matrix A. This is not true, though, in the specific case of model (1) which ignores any additive noise. In practice, some noise is always present and the claim of uniform performance may be more cautiously restated as: matrix A determines an upper limit to the noise level, under which the performance of EASI does not depend on A.

On the scale indetermination. Because of the scaling indeterminations inherent to the source separation problem, some parameters have to be arbitrarily fixed. Quite often, this is achieved by constraining the separating matrix. For instance, its diagonal elements or those of its inverse are fixed to unity [17], [19] or the rows of B_t are normalized [20]. In contrast, EASI does not constrain the separating matrix; indeterminations are dealt with by requiring that the output signals have unit variance. This solution is necessary to get uniform performance but also offers another important benefit: knowing in advance the range of the output signals allows to properly scale the non linearities. Assume for instance, that the hyperbolic tangent is used at the first output, *i.e.* $g_1(y_1) = \tanh(\alpha y_1)$. Here α is a real parameter which should not be chosen too small because this would make the tangent to work in its linear domain. However, the choice of α depends on the scale of y_1 which is known in advance when indeterminations are fixed by requiring unit variance output signals. In contrast, if indeterminations are fixed by constraining B, the range of y_1 may be arbitrarily large or small, depending on the mixing matrix A.

V. Asymptotic analysis

In this section, we evaluate the quantities governing the stability and the performance of serial adaptive algorithms. Since theoretical results are mainly available in the limit of arbitrarily small step size, we use the form (34) of function $H(\cdot)$ rather than the normalized version of (36). This approximation has negligible impact as checked in the experimental section VII.

We informally recall some definitions and results (see [3]) about stochastic algorithms in the form

$$\theta_{t+1} = \theta_t - \lambda_t \psi(\theta_t, \mathbf{x}_t) \tag{37}$$

where $\mathbf{x}_t$ is a stationary sequence of random variables and λ_t a sequence of positive numbers. A stationary point $\theta_\star$ verifies $\mathrm{E}\psi(\theta_\star, \mathbf{x}) = 0$ and is said to be *asymptotically stable* if all the eigenvalues of matrix Γ defined as

$$\Gamma \overset{\text{def}}{=} \frac{\partial \mathrm{E}\psi(\theta, \mathbf{x})}{\partial \theta}|_{\theta=\theta_\star} \tag{38}$$

have positive real parts.

When $\theta_\star$ is the unique global attractor, then for large t, small enough fixed step size $\lambda_t = \lambda$, and under rather restrictive conditions, the covariance matrix of θ_t is approximately given, in the i.i.d. case, by the solution of the Lyapounov equation:

$$\Gamma \mathrm{Cov}(\theta_t) + \mathrm{Cov}(\theta_t)\Gamma^{\mathrm{T}} = \lambda P \tag{39}$$

where P denotes the covariance matrix of ψ for $\theta = \theta_\star$:

$$P \overset{\text{def}}{=} \mathrm{Cov}(\psi(\theta_\star, \mathbf{x})) = \mathrm{E}[\psi(\theta_\star, \mathbf{x})\psi^{\mathrm{T}}(\theta_\star, \mathbf{x})]. \tag{40}$$

Clearly, this result does not apply in full rigor to the source separation problem where there are several basins of attraction. However, in practical applications, the step size is chosen to ensure that the probability of jumps from one separating matrix to another is sufficiently small. The close form solution of equation (39) is given below and is indeed found to predict with great accuracy the residual error of source separation observed in numerical simulations (see section VII).

We recall that it is only needed to study the dynamics of the global system C_t as given by (11). The above results apply to our algorithm by the identifications $\theta_t \rightarrow C_t$ and, according to (11), $\psi(\theta, \mathbf{x}) \rightarrow H(C\mathbf{s})C$. It is also needed to vectorize these matrices. The following convention turns out convenient: an $n \times n$ matrix is turned into a $n^2 \times 1$ vector by first stacking the (i, j)-th and (j, i)-th entries for each $1 \leq i < j \leq n$. and then appending the diagonal terms of the matrix, For instance, matrix C corresponds to vector θ:

$$\theta = [\underbrace{\cdots, c_{ij}, c_{ji}, \cdots}_{1\leq i<j\leq n}, \underbrace{\cdots, c_{ii}, \cdots}_{1\leq i\leq n}]^{\mathrm{T}} \tag{41}$$

and similarly for matrix $H(C\mathbf{s})C$.

A. Asymptotic stability

The 'mean field' of an adaptive algorithm at point θ is the vector $\mathrm{E}\psi(\theta, \mathbf{x})$. In our setting, the mean field is denoted $\mathcal{H}(C)$ and is

$$\mathcal{H}(C) \overset{\text{def}}{=} \mathrm{E}[H(C\mathbf{s}_t)C]. \tag{42}$$

Simple calculations (see appendix) reveal that its linear approximation in the neighborhood of $C_\star = I$ is

$$\mathcal{H}_{ii}(I + \mathcal{E}) \;=\; 2\mathcal{E}_{ii} + o(\mathcal{E}) \tag{43}$$

$$\begin{bmatrix} \mathcal{H}_{ij}(I + \mathcal{E}) \\ \mathcal{H}_{ji}(I + \mathcal{E}) \end{bmatrix} \;=\; DJ^{ij}D^{-1} \begin{bmatrix} \mathcal{E}_{ij} \\ \mathcal{E}_{ji} \end{bmatrix} + o(\mathcal{E}) \tag{44}$$

where the 2×2 matrices D and J^{ij} are

$$D \overset{\text{def}}{=} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \qquad J^{ij} \overset{\text{def}}{=} \begin{bmatrix} 2 & 0 \\ \xi_i - \xi_j & \kappa_i + \kappa_j \end{bmatrix} \tag{45}$$

with the non-linear moments of the source signals:

$$\kappa_i \;\overset{\text{def}}{=}\; \mathrm{E}\left[g_i'(s_i) - s_i g_i(s_i)\right] \tag{46}$$

$$\xi_i \;\overset{\text{def}}{=}\; \mathrm{E}\left[g_i'(s_i) + s_i g_i(s_i)\right]. \tag{47}$$

The significant fact in eq. (43) (holding for $1 \leq i \leq n$) and in eq. 44) (holding for $1 \leq i < j \leq n$) is the pairwise decoupling. It means that, with the vectorization (41), matrix Γ is block diagonal: there are $n(n-1)/2$ blocks of size 2×2 equal to $DJ^{ij}D^{-1}$ for $1 \leq i < j \leq n$ and n 'blocks' of size 1×1 with entries equal to 2. Since the eigenvalues of J^{ij} are 2 and $\kappa_i + \kappa_j$, we get the following

$$\text{Stability condition:} \quad \kappa_i + \kappa_j > 0 \quad \text{for} \quad 1 \leq i < j \leq n \tag{48}$$

for a separating matrix B such that $BA = I$.

The stability conditions for these separating matrices B such that BA is a permutation are very similar. Indeed, if the source signal s_i is present at the $\sigma(i)$-th output of B, then it undergoes the non-linearity $g_{\sigma(i)}$. Hence, the stability of this separating B is subject to condition (48) provided the moments κ_i are understood as $\mathrm{E}\,[\,g'_{\sigma(i)}(s_i) - s_i g_{\sigma(i)}(s_i)]$. Obviously when identical functions g_i are used or when sources with identical distributions have to be separated, the stability condition is verified for $C_\star$ being any permutation if it is verified for $C_\star = I$. The case where $C_\star$ is a permutation matrix with some 1's changed to -1, *i.e.* when $C_\star$ is any quasi-identity, leads again to the same condition when the g_i's are odd functions because the moments κ_i are then invariant under a change of sign.

The non-linear moments κ_i deserve some comments. First note that if g_i is a cubic distortion : $g_i(s_i) = s_i^3$, then $\kappa_i = 3 - \mathrm{E}|s_i|^4$ since $\mathrm{E}|s_i|^2 = 1$. This just the opposite of the fourth-order cumulant (or kurtosis) of s_i. The stability condition for cubic non-linearities then is that the sum of the kurtosis of any two sources must be negative. This condition (48) is weaker than the requirement that all source signals have a negative kurtosis. In particular, the stability condition (48) is verified if one source is Gaussian (in which case, its kurtosis is zero) and the other sources have negative kurtosis. Also note that, integrating by parts the definition of κ_i, it is easily seen that $\kappa_i = 0$ if s_i is a Gaussian variable, *independently* of the non-linear function g_i. This shows that the stability condition (48) can never be met if there is more than one Gaussian source signal. Finally, if g_i is a linear function, then $\kappa_i = 0$: it is seen that all the functions g_i but possibly one must be non linear to make a separating matrix stable.

B. Asymptotic covariance and rejection rates

In this section, we give close form expressions for the rejection rates obtained after convergence with a 'small' fixed step size λ. When the global system is $C = I + \mathcal{E}$, the i-th estimated source signal (the i-th output of C) is

$$\widehat{s}_i = y_i = [(I + \mathcal{E})\mathrm{s}]_i = (1 + \mathcal{E}_{ii})s_i + \sum_{j \neq i} \mathcal{E}_{ij} s_j . \tag{49}$$

Since the signals are independent with unit variance and since $\mathcal{E}$ is of order $\sqrt{\lambda}$, equation (49) shows that the ratio of the variance of the (undesired) j-th signal to the variance of the i-th signal (of interest) is approximately equal to $|\mathcal{E}_{ij}|^2$. Hence, we are interested in computing pairwise rejection rates, which correspond to intersymbol interference in the terminology of equalization, and are defined by:

$$\mathrm{ISI}_{ij} = \mathrm{E}|(C_t - I)_{ij}|^2 . \tag{50}$$

If C_t is 'vectorized' in a n^2-dimensional parameter vector, these quantities are the diagonal elements of matrix $\mathrm{Cov}(\theta)$. The computations are deferred to appendix B as well as the results for sources with different distributions.

For signals with identical distributions and a single nonlinearity $g(\cdot) = g_i(\cdot)$, there is only one extra moment involved:

$$\gamma \stackrel{\mathrm{def}}{=} \mathrm{E}g^2(s)\ \mathrm{E}s^2 - \mathrm{E}^2[g(s)s] \tag{51}$$

where s is any of the s_i's. The rejection rates are (necessarily) identical and given by

$$\mathrm{ISI} = \mathrm{ISI}_{ij} = \lambda \left(\frac{1}{4} + \frac{\gamma}{2\kappa} \right) . \tag{52}$$

Note that γ is positive by the Cauchy-Schwartz inequality and κ is positive by the stability condition. Hence, we have

$$\mathrm{ISI} \geq \frac{\lambda}{4} \tag{53}$$

and this bound is reached when $s = \pm 1$ with equal probability and g is an odd function because then $\gamma = 0$.

C. Tuning the nonlinearities

The analytical results obtained above provide us with guidelines for choosing the nonlinearities in $\mathbf{g}(\cdot)$. We do not intend to address this issue in full generality and will discuss here only the simplest case, often encountered in practice, where the sources have identical distributions. Since there is no reason in this case to use different nonlinearities, we take $g_1(\cdot) = \cdots = g_n(\cdot) = g(\cdot)$ and all the nonlinear moments are then also equal: we denote $\kappa = \kappa_i$ and $\gamma = \gamma_i$. Three points are discussed below.

Local convergence. The mean field $\mathcal{H}(C)$ then has a very simple local structure when C is close to any quasi-identity attractor $C_\star$: equations (44) and (43) combine into

$$\mathcal{H}(C_\star + \mathcal{E}) = (\mathcal{E} + \mathcal{E}^{\mathrm{T}}) + \kappa(\mathcal{E} - \mathcal{E}^{\mathrm{T}}) + o(\mathcal{E}) \tag{54}$$

showing that symmetric and skew-symmetric deviations of C_t from $C_\star$ are pulled back with a mean strength proportional to 2 and to κ respectively. When the moment κ is known in advance or can be (even roughly) estimated, expression (54) suggests to normalize the non-linearity $g(\cdot)$ into $\tilde{g}(\cdot) = g(\cdot)/\kappa$ because then the nonlinear moment $\tilde{\kappa}$ associated to $\tilde{g}$ is $\tilde{\kappa} = 1$. With such a choice, the mean field in the neighborhood of an attractor becomes

$$\mathcal{H}(C_\star + \mathcal{E}) = 2\mathcal{E} + o(\mathcal{E}), \tag{55}$$

meaning that all the deviations to a separator are locally *isotropically* pulled back, a benefit usually reserved to Newton-like algorithms.

Rejection rates. The nonlinear function g can be chosen to minimize the rejection rates under the constraint that its amplitude is fixed by the requirement of isotropic local convergence. In view of (52), the optimal nonlinearity should minimize γ under the constraint that $\kappa = 1$. This optimization problem is easily solved by the Lagrange multiplier method when the source signals are identically distributed with a differentiable probability density function $p(s)$. The optimal nonlinearity is found to be

$$g_{\mathrm{opt}}(s) = \frac{\psi(s)}{\mathrm{E}\psi^2(s) - 1} \quad \text{where} \quad \psi(s) \stackrel{\mathrm{def}}{=} -\frac{p'(s)}{p(s)} . \tag{56}$$

The resulting minimal rejection rate may be computed to be

$$\mathrm{ISI}_{\min} = \lambda \left(\frac{1}{4} + \frac{1}{2(\mathrm{E}\psi^2(s) - 1)} \right) . \tag{57}$$

As a final comment, we note that the various nonlinear moments appearing during performance analysis are not homogeneous and, unlike cumulants, cannot generally be normalized. This is an unavoidable effect when arbitrary nonlinearities are used. They are defined for unit variance random variables and, in any application, the source signals should be normalized to unit variance before the corresponding formulas are theoretically or empirically evaluated. It should be clear that our results giving the stability conditions and the rejection rates are valid regardless of the 'true' scale of the source signals.

VI. The complex case

At this stage, the processing of complex valued signals and mixtures is obtained straightforwardly from the real case by understanding the transposition operator $\cdot^{\mathrm{T}}$ as the transpose-conjugation operator and understanding 'unitary' in place of 'orthogonal'. The discussion in section IV-A on stationarity of the separating matrices carries over to the complex case with only one restriction: the diagonal terms of the skew-symmetric part of $\mathrm{E}H(\mathbf{s})$ are not necessarily zero unless the scalar-to-scalar functions g_i are restricted to be phase-preserving, *i.e.* of the form

$$g_i(y_i) = y_i l_i(|y_i|^2) \quad 1 \le i \le n \tag{58}$$

where the l_i's are real-valued functions. In order to easily extend the performance analysis to the complex case, it must be assumed that the source signals are 'circularly distributed', *i.e.* we assume:

Assumption 5. (Circularity): $\mathrm{E}\,[\,s_i(t)^2] = 0,\ 1 \le i \le n$. The modifications with respect to the real case are then mainly cosmetic and the results are given below without proof.

Regarding the stability of the separating matrices, the computations are very similar to the real case: it is found that

$$\begin{bmatrix} \mathcal{H}_{ij}(I+\mathcal{E}) \\ \mathcal{H}^*_{ji}(I+\mathcal{E}) \end{bmatrix} = DJ^{ij}D^{-1} \begin{bmatrix} \mathcal{E}_{ij} \\ \mathcal{E}^*_{ji} \end{bmatrix} + o(\mathcal{E}) \tag{59}$$

where matrices D and J^{ij} are as in (45), but the nonlinear moments are slightly different:

$$\kappa_i \overset{\text{def}}{=} \mathrm{E}[|s_i|^2 l_i'(|s_i|^2) + l_i(|s_i|^2) - |s_i|^2 l_i(|s_i|^2)], \tag{60}$$

$$\xi_i \overset{\text{def}}{=} \mathrm{E}[|s_i|^2 l_i'(|s_i|^2) + l_i(|s_i|^2) + |s_i|^2 l_i(|s_i|^2)]. \tag{61}$$

Hence, the stability condition (48) is unchanged provided κ_i is defined according to (60). For cubic non-linearities, *i.e.* for $l_i(s) = s$, one has $\kappa_i = 2 - \mathrm{E}\,|s_i|^4$ and $-\kappa_i$ again is the fourth-order cumulant of s_i in the circular case.

Regarding the asymptotic covariance, it is governed by the nonlinear moments

$$\gamma_i \overset{\text{def}}{=} \mathrm{E}[|s_i|^2 l_i^2(|s_i|^2)] - [\mathrm{E}|s_i|^2 l_i(|s_i|^2)]^2 \quad \mu_i \overset{\text{def}}{=} \mathrm{E}[|s_i|^2 l_i(|s_i|^2)] \tag{62}$$

which are direct complex counterparts of those defined in (72). With these definitions, the rejections rates take the very same form, either in the i.i.d. case, as given by the simple formula (52) or in the general case as given by the general expression (81).

VII. Numerical experiments

This section illustrates some properties of EASI and investigates the accuracy of the theoretical results, since these are only asymptotics (small λ). All the experiments are done in the complex case (but in figure 7). Figures 4 to 6 display trajectories of the modulus of the coefficients of the global system C_t. Hence, an experiment with n sources is illustrated by a plot with n^2 trajectories, with n of them getting close to 1 and the other getting close to zero.

Fast convergence is first illustrated by figure 4 for two i.i.d. QAM16 sources using the basic cubic nonlinearity $g_i(\mathbf{y}) = |y_i|^2 y_i$ for $1 \le i \le n$. The dashed lines represent ± two standard deviations computed from (52) and (75).

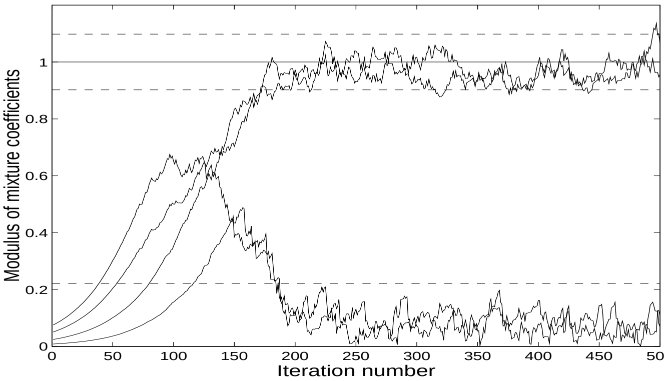

Fig. 4. A sample run. Convergence to 0 or 1 of the moduli of the coefficients of the global system B_tA. Fixed step size : $\lambda = 0.03$. Two QAM16 sources, cubic non-linearities : $g_i(\mathbf{y}) = |y_i|^2 y_i$.

Figure 5 is similar but three QAM16 sources are involved and the step size is decreased according to the cooling scheme: $\lambda_t = 2/t$.

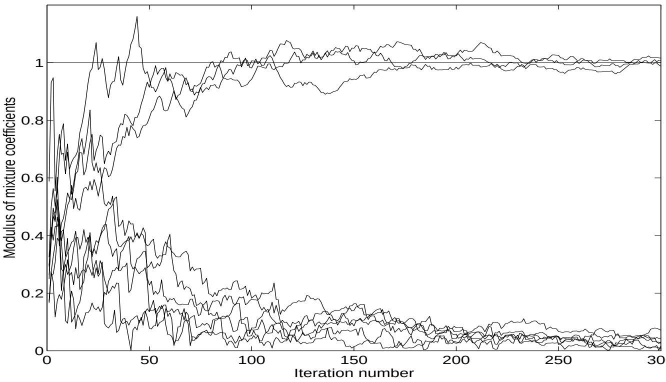

Fig. 5. A sample run. Convergence to 0 or 1 of the moduli of the coefficients of the global system B_tA. Three QAM16 sources. Decreasing step size : $\lambda_t = 2/t$.

Figure 6 is concerned with the effect of normalization. With the same QAM16 input, two serial algorithms are run with $\lambda = 0.01$ one with the normalized algorithm (36), the other with the raw algorithm (35); Both trajectories are displayed and show little discrepancy (see also table I).

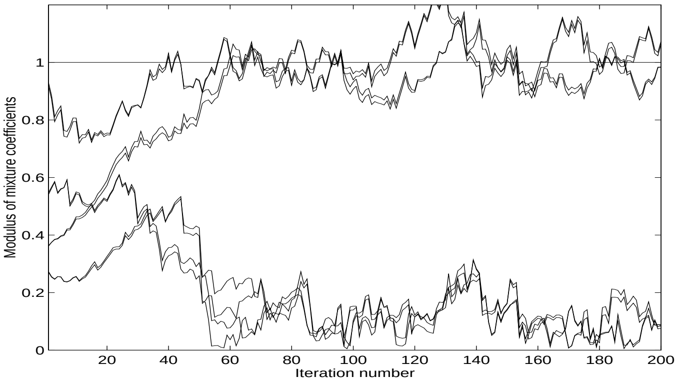

Fig. 6. Effect of normalization.

Figure 7 illustrates the isotropic convergence with a balanced nonlinearity. It displays the evolution of a logarithmic distance of C_t to the identity, namely $20 \log_{10} ||C_t - I||_{\mathrm{Fro}}$, with a constant step size. Each curve corresponds to a different initial condition. These initial conditions are randomly chosen but are at a fixed Frobenius distance from the identity matrix. Both panels are for cubic nonlinearities and uniformly distributed sources which have a normalized kurtosis equal to $-6/5$ (this is the only experiment with real signals). Isotropic convergence is achieved by taking $g(s) = 5/6\ s^3$, so that $\kappa = 1$ as suggested in the discussion of section V-C. The resulting trajectories are displayed in the lower panel, where the dashed line corresponds to a distance varying as $\exp(-2\lambda t)$. The upper panel displays trajectories for $g(s) = 0.2\ s^3$: they are sandwiched between two dashed lines corresponding to $\exp(-2\lambda t)$ and $\exp(-2\frac{0.2}{5/6}\lambda t)$ which are the mean decaying rates for the symmetric and skew-symmetric parts respectively. Hence, according to the respective proportion of symmetric and skew-symmetric errors in C_0, various decaying rates are observed, while the lower panel shows logarithmic slopes which are essentially independent of the initial condition.

The rejection rates predicted by (52) have been experimentally measured in the case of $n = 2$ sources. Results are reported in table I. The following fixed step sizes are used: $\lambda = 0.1, 0.3, 0.01, 0.003$. For each step size, $N_{MC} = 500$ trajectories are simulated. The initial point is $C_o = I$ and the sample estimate of ISI_{12} is computed over a trajectory in the range $5/\lambda < t < 35/\lambda$ (the scaling with $1/\lambda$ is adopted to get a constant relative precision). The resulting N_{MC} values are further averaged and also used to determine an experimental standard deviation. The table presents the mean plus and minus two standard deviations of $\lambda^{-1}\mathrm{ISI}_{12}$. There are no results presented for $\lambda = 0.1$ and QAM16 signals for the non-normalized algorithms because a significant fraction of divergent trajectories have been observed. In all the other cases, representing 15×500 trajectories, no divergence have been observed. It appears that asymptotic analysis correctly predicts the rejection rates for step sizes as large as $\lambda = 0.01$. We also note that normalization does not affect much the empirical performance.

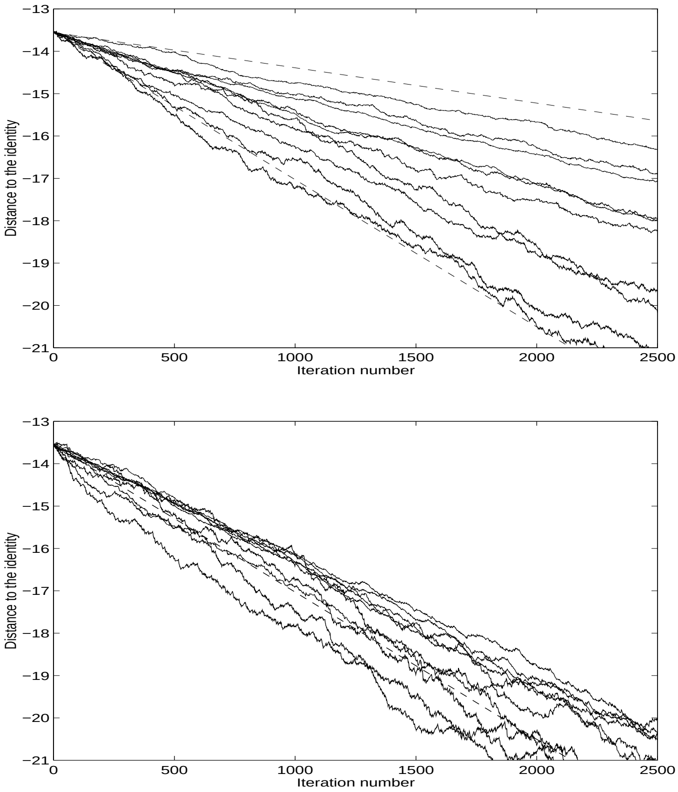

Fig. 7. Vertical axis: $20 \log_{10} ||C_t - I||_{\mathrm{Fro}}$. Lower panel: convergence rate depends on the starting point; unbalanced nonlinearity. Upper panel: isotropic convergence with a balanced nonlinearity.

	Rejection rate $\lambda^{-1}\mathrm{ISI}_{12}$ with QAM4 sources		
λ	Theoretical	Non normalized	Normalized
0.100	0.250	0.229 ± 0.003	0.213 ± 0.003
0.030	0.250	0.240 ± 0.003	0.233 ± 0.003
0.010	0.250	0.249 ± 0.003	0.246 ± 0.003
0.003	0.250	0.248 ± 0.003	0.247 ± 0.003

	Rejection rate $\lambda^{-1}\mathrm{ISI}_{12}$ with QAM16 sources		
λ	Theoretical	Non normalized	Normalized
0.100	0.410	Non convergent	0.417 ± 0.008
0.030	0.410	0.435 ± 0.006	0.410 ± 0.006
0.010	0.410	0.417 ± 0.005	0.411 ± 0.005
0.003	0.410	0.412 ± 0.005	0.410 ± 0.005

TABLE I

Empirical and theoretical rejection rates.

VIII. Conclusion

A class of adaptive algorithms for the blind separation of sources has been introduced. It is based on the idea of serial updating by which the uniform performance property of equivariant estimators is directly inherited by the corresponding adaptive serial algorithms. For adaptive algorithms, the uniform performance property means that changing the mixing matrix is equivalent to changing the initial condition. As a result, the characteristics of a serial algorithm, such as the stability conditions, the convergence rates or the residual errors, do not depend on the mixing matrix.

A serial algorithm is defined by specifying a vector-to-matrix mapping H verifying $\mathrm{E}H(\mathbf{s}) = 0$ if the random vector $\mathbf{s}$ has independent components. While many such mappings may be devised, we have considered a specific class, where the symmetric part of H corresponds to a second order condition of independence (decorrelation) while the skew-symmetric part involves nonlinear functions. This structure allows a simple, very regular implementation in the real case as well as in the complex case. By its very structure, the algorithm can be used 'as is' when more sensors than sources are available.

The asymptotic analysis for arbitrary nonlinearities reveals a pairwise decoupling, pairwise stability conditions and yields the rejection rates in close form. These results allow the symmetric and skew-symmetric parts to be balanced in order obtain isotropic local convergence and the non linearity to be shaped in order to maximize interference rejection.

Appendix

I. Derivative of the mean field

We compute the first-order expansion of the mean field in the neighborhood of the identity matrix. This amounts to finding the linear term in $\mathcal{E}$ in $\mathcal{H}(I+\mathcal{E})$. First note that the definition (42) rewrites

$$\mathcal{H}(I+\mathcal{E}) = \mathrm{E}[\ H(\mathbf{s}+\mathcal{E}\mathbf{s})(I+\mathcal{E})]. \tag{63}$$

Since the identity is a stationary point, we have $\mathrm{E}H(\mathbf{s}) = 0$ so that the mean field also is

$$\mathcal{H}(I+\mathcal{E}) = \mathrm{E}H(\mathbf{s}+\mathcal{E}\mathbf{s}) + o(\mathcal{E}). \tag{64}$$

The hermitian part of $\mathrm{E}H(\mathbf{s}+\mathcal{E}\mathbf{s})$ is readily obtained as :

$$\mathrm{E}\ [\ (\mathbf{s}+\mathcal{E}\mathbf{s})(\mathbf{s}+\mathcal{E}\mathbf{s})^{\mathrm{T}} - I] = \mathcal{E} + \mathcal{E}^{\mathrm{T}} + o(\mathcal{E}) \tag{65}$$

since our normalization convention is $\mathrm{E}\ [\ \mathbf{s}\mathbf{s}^{\mathrm{T}}] = I$. In order to compute the antisymmetric part of $\mathcal{H}(I+\mathcal{E})$ that is $\mathrm{E}\ [\ \mathbf{g}(\mathbf{y})\mathbf{y}^{\mathrm{T}} - \mathbf{y}\mathbf{g}(\mathbf{y}^{\mathrm{T}}]$ with $\mathbf{y} = \mathbf{s}+\mathcal{E}\mathbf{s}$, we have to go down to the component level. We start by evaluating the (i,j)-th entry of $\mathrm{E}\ [\ \mathbf{y}\mathbf{g}(\mathbf{y})^{\mathrm{T}}]$. Using $y_i = s_i + \sum_a \mathcal{E}_{ia}s_a$, we get

$$\begin{aligned} y_i g_j(y_j) &= s_i g_j(s_j) + \sum_a \mathcal{E}_{ia} s_a g_j(s_j) \\ &+ \sum_b \mathcal{E}_{jb} s_i s_b g_j'(s_j) + o(\mathcal{E}) \end{aligned} \tag{66}$$

There is no need evaluating the terms for $i = j$ because these disappear in the anti-symmetrization. Focusing on the terms with $i \neq j$, we next find that

$$\mathrm{E}s_a g_j(s_j) = \delta(j,a)\ \mathrm{E}s_j g_j(s_j) \tag{67}$$

$$\mathrm{E}s_i s_b g_j'(s_j) = \delta(i,b)\ \mathrm{E}s_i^2 \mathrm{E}g_j'(s_j) \quad \text{for } i \neq j \tag{68}$$

because the source signals are independent with zero mean. It follows that, for $i \neq j$,

$$\mathrm{E}y_i g_j(y_j) = \mathcal{E}_{ij}\mathrm{E}s_j g_j(s_j) + \mathcal{E}_{ji}\mathrm{E}s_i^2\mathrm{E}g_j'(s_j) + o(\mathcal{E}). \tag{69}$$

Expectations (63), (65) and (69) then combine into :

$$\begin{aligned} \mathcal{H}_{ij}(I+\mathcal{E}) &= \mathcal{E}_{ij}(1 + \mathrm{E}s_j^2\mathrm{E}g_i'(s_i) - \mathrm{E}s_j g_j(s_j)) \\ &+ \mathcal{E}_{ji}(1 - \mathrm{E}s_i^2\mathrm{E}g_j'(s_j) + \mathrm{E}s_i g_i(s_i)) + o(\mathcal{E}) \end{aligned}$$

which, after symmetrization yields (44).

II. Asymptotic covariance

To solve (39), we must first evaluate matrix P. Using source independence, it is easily checked most of the entries of $H(\mathbf{s})$ are uncorrelated. The non vanishing terms can be computed to be

$$\mathrm{Cov}(H_{ii}(\mathbf{s})) = \mathrm{E}|s_i|^4 - 1 \tag{70}$$

$$\mathrm{Cov}\left(\begin{bmatrix} H_{ij}(\mathbf{s}) \\ H_{ji}(\mathbf{s}) \end{bmatrix}\right) = DQ^{ij}D^{\mathrm{T}} \tag{71}$$

with the following definitions

$$Q^{ij} \stackrel{\mathrm{def}}{=} \begin{bmatrix} 1 & \mu_i - \mu_j \\ \mu_i - \mu_j & \gamma_i + \gamma_j + (\mu_i - \mu_j)^2 \end{bmatrix} \tag{72}$$

$$\gamma_i \stackrel{\mathrm{def}}{=} \mathrm{E}[g_i^2(s_i)] - [\mathrm{E}s_i g_i(s_i)]^2 \tag{73}$$

$$\mu_i \stackrel{\mathrm{def}}{=} \mathrm{E}[g_i(s_i)s_i]. \tag{74}$$

This is a pleasant finding since it means that P has the same block diagonal structure as Γ, allowing the Lyapounov equation (39) to be solved blockwise. Further, the blocks having sizes 1 and 2, close form solutions can be worked out.

Solving for the 1×1 blocks is immediate: each scalar equation yields

$$\mathrm{Cov}(C_{ii}) = \lambda\frac{\mathrm{E}s_i^4 - 1}{4}. \tag{75}$$

The 2×2 Lyapounov equation extracted from (39) for a pair $i \neq j$ is

$$(DJ^{ij}D^{-1})R^{ij} + R^{ij}(DJ^{ij}D^{-1})^{\mathrm{T}} = \lambda DQ^{ij}D^{\mathrm{T}} \tag{76}$$

where we set

$$R^{ij} \stackrel{\mathrm{def}}{=} \mathrm{Cov}\left(\begin{bmatrix} C_{ij}(\mathbf{s}) \\ C_{ji}(\mathbf{s}) \end{bmatrix}\right). \tag{77}$$

Left and right multiplication by D^{-1} and $D^{-\mathrm{T}}$ respectively yields

$$J^{ij}(D^{-1}R^{ij}D^{-\mathrm{T}}) + (D^{-1}R^{ij}D^{-\mathrm{T}})J^{ij\mathrm{T}} = \lambda Q^{ij}. \tag{78}$$

Matrix J^{ij} being lower triangular, this 2×2 Lyapounov equation is easily worked out. We find the following intermediate result: If (x, y, t) is the solution of

$$\begin{bmatrix} a & 0 \\ c & b \end{bmatrix} \begin{bmatrix} x & t \\ t & y \end{bmatrix} + \begin{bmatrix} x & t \\ t & y \end{bmatrix} \begin{bmatrix} a & c \\ 0 & b \end{bmatrix} = \begin{bmatrix} \alpha & \gamma \\ \gamma & \beta \end{bmatrix} \tag{79}$$

then, the northwest entry of $D \begin{bmatrix} x & t \\ t & y \end{bmatrix} D^{\mathrm{T}}$ is

$$x + y + 2t = \frac{\alpha}{2a} + \frac{\beta}{2b} + \frac{(2b-c)(2a\gamma - c\alpha)}{2ab(a+b)}. \tag{80}$$

From this, an explicit expression for $\mathrm{Cov}(C_{ij})$ is readily obtained. We skip some additional uninspiring algebraic reorganization which yields the form most appropriate for our concerns:

$$\mathrm{E}|C_{ij}|^2 = \mathrm{Cov}(C_{ij}) = \lambda(\frac{1}{4} + \frac{1}{2}\frac{\gamma_i + \gamma_j}{\kappa_i + \kappa_j} + \beta_{ij}^+ + \beta_{ij}^-) \tag{81}$$

where β_{ij}^+ and β_{ij}^- cancel for identical sources and nonlinearities. They are respectively symmetric and skew-symmetric in the exchange $i \leftrightarrow j$:

$$\beta_{ij}^+ \stackrel{\text{def}}{=} \frac{2(\kappa_i + \kappa_j)(\mu_i - \mu_j)^2 + (\kappa_i - \kappa_j)^2}{4(\kappa_i + \kappa_j)(2 + \kappa_i + \kappa_j)} \tag{82}$$

$$\beta_{ij}^- \stackrel{\text{def}}{=} \frac{(2\mu_i - \kappa_i) - (2\mu_j - \kappa_j)}{2(2 + \kappa_i + \kappa_j)}. \tag{83}$$

References

[1] Karim Abed Meraim, Adel Belouchrani, Jean-François Cardoso, and Eric Moulines. Asymptotic performance of second order blind source separation. In *Proc. ICASSP*, volume 4, pages 277–280, April 1994.

[2] Adel Belouchrani and Jean-François Cardoso. Maximum likelihood source separation for discrete sources. In *Proc. EUSIPCO*, pages 768–771, Edinburgh, September 1994.

[3] A. Benveniste, M. Métivier, and P Priouret. *Adaptive algorithms and stochastic approximations.* Springer Verlag, 1990.

[4] Jean-François Cardoso. Fourth-order cumulant structure forcing. Application to blind array processing. In *Proc. 6th SSAP workshop on statistical signal and array processing*, pages 136–139, October 1992.

[5] Jean-François Cardoso. Iterative techniques for blind source separation using only fourth order cumulants. In *Proc. EUSIPCO*, pages 739–742, 1992.

[6] Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6):362–370, December 1993.

[7] A. Chichocki and L Moszczynski. New learning alforithms for blind separation of sources. *Electronic Letters*, 28:1986–1987, 1992.

[8] P. COMON. Independent Component Analysis, a new concept ? *Signal Processing, Elsevier*, 36(3):287–314, April 1994. Special issue on Higher-Order Statistics.

[9] Pierre Comon. Independent component analysis. In *Proc. Int. Workshop on Higher-Order Stat., Chamrousse, France*, pages 111–120, 1991.

[10] Thomas M. Cover and Joy A. Thomas. *Robust statistics.* Wiley series in telecommunications. John Wiley, 1991.

[11] Nathalie Delfosse and Philippe Loubaton. Adaptive separation of independent sources: a deflation approach. In *Proc. ICASSP*, volume 4, pages 41–44, 1994.

[12] A. Dinç and Y. Bar-Ness. Bootstrap: A fast blind adaptive signal separator. In *Proc. ICASSP*, volume 2, pages 325–328, 1992.

[13] Luc Féty and J. P. Van Uffelen. New methods for signal separation. In *Proc. of 4th Int. Conf. on HF radio systems and techniques*, pages 226–230, London, April 1988. IEE.

[14] Michel Gaeta and Jean-Louis Lacoume. Source separation without a priori knowledge: the maximum likelihood solution. In *Proc. EUSIPCO*, pages 621–624, 1990.

[15] Stefan Van Gerven and Dirk Van Compernolle. On the use of decorrelation in scalar signal separation. In *Proc. ICASSP*, Adelaide, Australia., 1994.

[16] Peter J. Huber. *Robust statistics.* Wiley series in probability and mathematical statistics. John Wiley, 1981.

[17] Christian Jutten and J. Hérault. Independent component analysis versus PCA. In *Proc. EUSIPCO*, pages 643–646, 1988.

[18] Juha Karhunen and Jyrki Joutsensalo. Representation and separation of signals using nonlinear PCA type learning. *Neural Networks*, 7(1):113–127, 1993.

[19] Eric Moreau and Odile Macchi. New self-adaptive algorithms for source separation based on contrast functions. In *Proc. IEEE SP Workshop on Higher-Order Stat., Lake Tahoe, USA*, pages 215–219, 1993.

[20] Dinh Tuan Pham, Philippe Garrat, and Christian Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In *Proc. EUSIPCO*, pages 771–774, 1992.

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

[22] J. E. Potter. New statistical formulas. Technical report, Instrumentationn Laboratory, Mass. Inst. of Technology, 1963.

[23] O. Shalvi and E. Weinstein. New criteria for blind deconvolution of nonminimum phase systems (channels). *IEEE Tr. on IT*, 36(2):312–321, 1990.

[24] L. Tong, R. Liu, V.C. Soon, and Y. Huang. Indeterminacy and identifiability of blind identification. *IEEE Tr. on CS*, 38(5):499–509, May 1991.

[25] Lang Tong, Yujiro Inouye, and Ruey-wen Liu. Waveform preserving blind estimation of multiple independent sources. *IEEE Tr. on SP*, 41(7):2461–2470, July 1993.