Equivariant Adaptive Source Separation

Jean-François Cardoso, *Member, IEEE*, and Beate Hvam Laheld

***Abstract*— Source separation consists of 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 that implements an adaptive version of equivariant estimation and is henceforth called equivariant adaptive separation via independence (EASI). The EASI algorithms are based on the idea of *serial updating*: This specific form of matrix updates systematically yields algorithms with a simple 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. Closed-form expressions of these quantities are given via an asymptotic performance analysis.**

The theme of equivariance is stressed throughout the paper. The source separation problem has an underlying multiplicative structure: The parameter space forms a (matrix) multiplicative group. We explore the (favorable) consequences of this fact on implementation, performance, and optimization of EASI algorithms.

I. Introduction

BLIND *separation of sources* is receiving 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 of 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 narrowband 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 an urban environment), it may be safer to resort to a blind procedure for recovering the source signals.

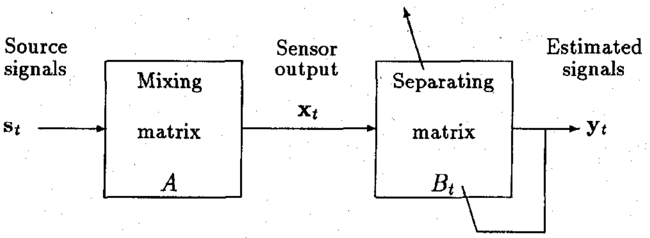


Fig. 1. Adapting a separating matrix.

This paper addresses the issue of *adaptive* source separation and considers the case where any additive noise can be neglected. The signal model then reduces to that of observations $\mathbf{x}_t$ in the form

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

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 m 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, which is obtained by chaining the mixing matrix A and the separating matrix B_t, that is

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

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 update 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 a sequence of positive adaptation steps.

After some background on the source separation problem in Section II, the serial updating scheme is investigated in Section

Manuscript received November 3, 1994; revised June 17, 1996. This work was supported by a grant from the Norwegian government. The associate editor coordinating the review of this paper and approving it for publication was Dr. Tyseer Aboulnasr.

J.-F. Cardoso is with the Centre National de la Recherche Scientifique (C.N.R.S.) and École Nationale Supérieure des Télécommunications (E.N.S.T.), 46 rue Barrault, 75634 Paris Cedex 13, France (e-mail: cardoso@sig.enst.fr).

B. H. Laheld was with the École Nationale Supérieure des Télécommunications (E.N.S.T.), 46 rue Barrault, 75634 Paris Cedex 13, France. She is now with France Télécom Mobiles.

Publisher Item Identifier S 1053-587X(96)09053-8.

III; 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 IV, a particular function $H(\cdot)$ is obtained from a cumulant-based approach to blind identification. This is then generalized in Section V into a family of equivariant adaptive source separation algorithms, whose stability and asymptotic convergence are studied in Section VI. Section VII extends all the results to the complex case. This is completed in Section VIII by some numerical experiments illustrating the effectiveness of the approach and the accuracy of asymptotic analysis.

II. Source Separation

A. Assumptions and Notations

Some notational conventions are as follows: Scalars are in lower case, matrices in upper case, and vectors in boldface lowercase. The ith component of a vector, say $\mathbf{x}$, is denoted x_i. The expectation operator is $\mathbf{E}$ and transposition is indicated by superscript T. The $n \times n$ identity matrix is denoted I.

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

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

Assumption 1 is expected to hold 'almost surely' in any physical situation. More important is the existence of A itself i.e., the plausibility 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 mixture is the standard model in narrowband 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 straightforward and described in Section VII.

Finally, for source separation to be possible, there are conditions on the probability distributions of the source signals. Since this condition is algorithm-dependent, its formulation is deferred to Section VI-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); source signals can be at best recovered up to a permutation. In addition, 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 that is the product of a permutation matrix with a diagonal matrix with unit-norm diagonal elements is called a *quasiidentity* matrix; any matrix $B_\star$ is said to be a *separating matrix* if the product $B_\star A$ is a quasiidentity matrix.

The adaptive source separation problem then consists of 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 quasiidentity matrix. The issue of indetermination is addressed at length in [1].

B. Approaches to Source Separation

The seminal work on source separation is by Jutten and Hérault [2], [3]. 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, $\mathbf{E}[f(y_i)g(y_j)] = \mathbf{E}f(y_i)\mathbf{E}g(y_j)$, 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 [4]. The Jutten-Hérault algorithm is inspired by a neuromimetic approach; this line is further followed by Karhunen *et al.* [5] and Cichocki *et al.* [6], [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 fourth-order cumulants of the output, cubic nonlinearities show up, as in [8] and [9]. Other nonlinearities are considered in [10] based on information-theoretic ideas.

When the sources have known differentiable probability distribution functions (p.d.f.'s), the maximum likelihood (ML) estimator is easily obtained in the i.i.d. case; the (asymptotically optimal) nonlinearities are the log-derivatives of the p.d.f.'s [11]. See also [12] for an ML approach for discrete sources in unknown Gaussian noise.

A generic approach to source separation is based on 'orthogonal contrast functions.' In the context of source separation, these were introduced by Comon [13] as functions of the distribution of $\mathbf{y}$, which are to be optimized under the whiteness constraint: $R_y = \mathbf{E}\mathbf{y}\mathbf{y}^{\mathrm{T}} = I$. Comon suggests minimizing the squared cross-cumulants 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 that can be efficiently optimized by a Jacobi-like algorithm, especially in the complex case, is described in [15].

When the sources have kurtosis of identical signs, simpler orthogonal contrasts may be exhibited. For instance, if all the sources have a negative kurtosis, it is easily proved that minimizing

$$\phi_4(B) \overset{\text{def}}{=} \mathbf{E}f(\mathbf{y}) \quad \text{with } f(\mathbf{y}) = \sum_{i=1,n} |y_i|^4 \tag{6}$$

subject to $R_y = I$ is achieved only when B is a separating matrix. This contrast is strongly reminiscent of fourth-order objectives used in blind equalization [16] and lends itself easily to adaptive minimization. It is considered in [8] where it is optimized by a deflation technique. The resulting adaptive algorithm can be proved to be asymptotically free of spurious attractors.

Before closing this section, other batch estimation techniques may be mentioned: Higher order cumulants are used together with a prewhitening strategy in Tong *et al.* [1], [17]; fourth-order-only approaches are investigated in [18] and [19]; second-order-only approaches are possible if the sources are nonstationary [20] or have different spectra as investigated in [21]–[23], [1], and [24] in an adaptive implementation.

C. Equivariant Source Separation

The equivariant approach to adaptive source separation introduced in this paper is best motivated by first considering *batch* estimation. Assume for simplicity that $n = m$ (as many sources as 'sensors'), and consider the problem of estimating matrix A from a batch of T samples $X_T = [\mathbf{x}(1), \ldots, \mathbf{x}(T)]$. A blind estimator of A is, by definition, a function of X_T only. This may be denoted by

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

According to (1), the $m \times T$ data matrix X_T can be factored as $X_T = AS_T$ with $S_T = [\mathbf{s}(1), \ldots, \mathbf{s}(T)]$. A trivial observation is that multiplying the data by some matrix M has the same effect as multiplying A itself by M since one has $M(X_T) = M(AS_T) = (MA)S_T$.

When a transformation on the data is equivalent to a transformation of the parameter, the notion of *equivariance* is of relevance (see, for instance, [25]). Note that we are dealing here with a simple case where the transformations of both the parameter A and on the data set are implemented by the same algebraic operation: left multiplication by a matrix. Equivariance theory becomes interesting when a whole *group* of transformations can be considered. In the following, we consider the group of left multiplication by invertible matrices. An estimator behaves 'equivariantly' if it produces estimates that, under data transformation, are transformed accordingly. In the case of interest, a particular estimator $\mathcal{A}$ for A 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. The equivariance property arises quite naturally in the context of source separation. For instance, the ML estimator and estimators based on optimizing contrast functions are equivariant [26].

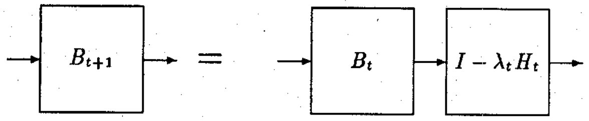


Fig. 2. Serial update of a matrix.

The key property shared by equivariant batch estimators for source separation is that they offer *uniform performance*. This is to be understood in the following sense. Assume that the source signals are estimated as $\hat{\mathbf{s}}(t) = \hat{A}^{-1}\mathbf{x}(t)$, where $\hat{A}$ is obtained from an equivariant estimator. Then,

$$\begin{aligned}\hat{\mathbf{s}}(t) &= (\mathcal{A}(X_T))^{-1}\mathbf{x}(t) = (\mathcal{A}(AS_T))^{-1}A\mathbf{s}(t)\\ &= (A\mathcal{A}(S_T))^{-1}A\mathbf{s}(t) = \mathcal{A}(S_T)^{-1}\mathbf{s}(t)\end{aligned} \tag{9}$$

where we have only used the equivariance property (8). The last equality reveals that source signals estimated by an equivariant estimator $\mathcal{A}$ for a particular realization $S_T = [\mathbf{s}(1), \ldots, \mathbf{s}(T)]$ are given by $\hat{\mathbf{s}}(t) = \mathcal{A}(S_T)^{-1}\mathbf{s}(t)$, i.e., they 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; the next section actually shows how 'uniform performance properties' can be inherited by an adaptive algorithm from a batch estimation procedure.

III. Serial Matrix Updating

A. Serial Updates

The adaptation rule (4) is termed a 'serial update' because it reads equivalently $B_{t+1} = (I - \lambda_t H(\mathbf{y}_t))B_t$. This latter multiplicative form evidences that B_t is updated by 'plugging' matrix $I - \lambda_t H(\mathbf{y}_t)$ at the *output* of the current system B_t to get the updated system B_{t+1} (see Fig. 2). On one hand, uniform performance of equivariant batch algorithms is a direct consequence of (8), which is a *multiplicative* equation. On the other hand, by the learning rule (4), the system B_t is serially updated by left *multiplication* by matrix $I - \lambda_t H(\mathbf{y}_t)$. In this sense, serial updating is consistent with the multiplicative structure and the key result of Section III-B is that serially updated systems inherit the uniform performance property from their batch counterparts. Gradient algorithms are ubiquitous in system adaptation. A fully consistent theory of equivariant adaptive separation should include an appropriate definition of the gradient with respect to a matrix; this is the 'relative gradient' introduced in Section III-C.[1]

[1] While this paper was being revised, we became aware of the work [7] using a similar updating rule. A 'natural gradient' identical to our 'relative gradient' has also been introduced independently by Amari (private communication).

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, immediately yielding

$$C_{t+1} = C_t - \lambda_t H(C_t \mathbf{s}_t) C_t \tag{10}$$

where we used $\mathbf{y}_t = B_t \mathbf{x}_t = B_t A \mathbf{s}_t = C_t \mathbf{s}_t$. Hence, the global system C_t also undergoes serial updating (compare with (4)), which is an obvious fact in light of Fig. 2. This is a trivial but remarkable result because it means that under serial updating, the evolution 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 $C_o = B_o A$. By (10), 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' if it is initialized with B_o' such that $B_o A = B_0' A'$. This is pretty obvious since in both cases, the *global* system starts from the same initial condition and evolves according to (10), 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 value B_0 of the separator.

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 jth source signal in the estimate of the ith source signal at time t is determined only by the (i,j)th entry of C_t.

It follows that it is only necessary to study the convergence of C_t to a quasiidentity matrix under the stochastic rule (10) 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 $H(\cdot)$. To obtain such a function, the notion of 'relative gradient' is instrumental. In this section, we denote $\langle \cdot \mid \cdot \rangle$ to be the Euclidean scalar product of matrices

$$\langle M \mid N \rangle = \mathrm{Trace}[M^{\mathrm{T}} N] \quad \langle M \mid M \rangle = \|M\|^2_{\mathrm{Fro}}. \tag{11}$$

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 $\frac{\partial \phi}{\partial B}(B)$; 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

$$\phi(B + \mathcal{E}) = \phi(B) + \left\langle \frac{\partial \phi}{\partial B}(B) \mid \mathcal{E} \right\rangle + o(\mathcal{E}). \tag{12}$$

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

$$\phi(B + \mathcal{E}B) = \phi(B) + \langle \nabla\phi(B) \mid \mathcal{E} \rangle + o(\mathcal{E}). \tag{13}$$

There is no profound difference between usual ('absolute') and relative gradient since comparing (12) and (13) shows that $\nabla\phi(B) = \frac{\partial \phi}{\partial B}(B)\, B^{\mathrm{T}}$. However, the relative gradient is consistent with the notion of serial update, and its appropriateness is confirmed in the following.

The relevance of considering the relative gradient is first illustrated in the case where $\phi(B)$ is in the form $\phi(B) = \mathbf{E}f(\mathbf{y}) = \mathbf{E}f(B\mathbf{x})$ as in (6). If function f is differentiable everywhere, one has

$$f(\mathbf{y} + \delta\mathbf{y}) = f(\mathbf{y}) + \mathbf{f}'(\mathbf{y})^{\mathrm{T}} \delta\mathbf{y} + o(\delta\mathbf{y}) \tag{14}$$

where $\mathbf{f}'(\mathbf{y})$ is the gradient of f at $\mathbf{y}$, i.e., it is the column vector whose ith component is the partial derivative of $f(\mathbf{y})$ with respect to y_i. We then have

$$\begin{aligned} \phi(B + \mathcal{E}B) &= \mathbf{E}f((B + \mathcal{E}B)\mathbf{x}) = \mathbf{E}f(\mathbf{y} + \mathcal{E}\mathbf{y}) \\ &= \mathbf{E}f(\mathbf{y}) + \mathbf{E}\mathbf{f}'(\mathbf{y})^{\mathrm{T}} \mathcal{E}\mathbf{y} + o(\mathcal{E}) \\ &= \phi(B) + \mathbf{E}\mathrm{Trace}\{\mathbf{y}\mathbf{f}'(\mathbf{y})^{\mathrm{T}} \mathcal{E}\} + o(\mathcal{E}) \\ &= \phi(B) + \langle \mathbf{E}\mathbf{f}'(\mathbf{y})\mathbf{y}^{\mathrm{T}} \mid \mathcal{E} \rangle + o(\mathcal{E}). \end{aligned} \tag{15}$$

Identifying the last expression with (13) yields the result

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

The point is that the relative gradient at point B depends only on the distribution of $\mathbf{y} = B\mathbf{x}$ and not on B itself. This was to be expected since modifying B into $B + \mathcal{E}B$ amounts to modifying $\mathbf{y}$ into $\mathbf{y} + \mathcal{E}\mathbf{y}$, regardless of the particular values of $\mathbf{x}$ or B.

As any gradient rule, the 'relative gradient rule' is to align the 'relative variation' $\mathcal{E}$ in a direction opposite to the relative gradient. In other words, it consists of modifying B into $B + \mathcal{E}B$ with $\mathcal{E} = -\lambda\nabla\phi(B)$ and λ a positive scalar. Indeed, by (11) and (13),

$$\begin{aligned} &\phi(B - \lambda\nabla\phi(B)B) \\ &\quad = \phi(B) + \langle \nabla\phi(B) \mid -\lambda\nabla\phi(B) \rangle + o(\lambda\nabla\phi(B)) \quad (17) \\ &\quad = \phi(B) - \lambda\|\nabla\phi(B)\|^2_{\mathrm{Fro}} + o(\lambda) \quad (18) \end{aligned}$$

so that, for small enough positive λ, as long as $\nabla\phi(B) \neq 0$, the objective ϕ is decreased if B is modified into $B - \lambda\nabla\phi(B)B$. This relative gradient rule is turned into a *stochastic* relative gradient rule by deleting the expectation operator in the relative gradient $\nabla\mathbf{E}f(B\mathbf{x}) = \mathbf{E}[\mathbf{f}'(\mathbf{y})\mathbf{y}^{\mathrm{T}}]$. This is then

$$B_{t+1} = B_t - \lambda_t \mathbf{f}'(\mathbf{y}_t)\mathbf{y}_t^{\mathrm{T}} B_t \tag{19}$$

for the stochastic minimization of $\mathbf{E}f(\mathbf{y})$, where λ_t is a sequence of positive scalars. The key point here is that (19) precisely is a serial update algorithm: Rules (19) and (4) are identical if we set $H(\mathbf{y}) = \mathbf{f}'(\mathbf{y})\mathbf{y}^{\mathrm{T}}$. Hence, for simple objective functions in the form $\phi(B) = \mathbf{E}f(\mathbf{y})$, the notion of (stochastic) relative gradient does yield the $H(\cdot)$ function, which defines a serial update algorithm. Note that the stochastic optimization of the same objective function

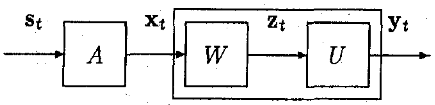


Fig. 3. Two-stage separation in batch processing.

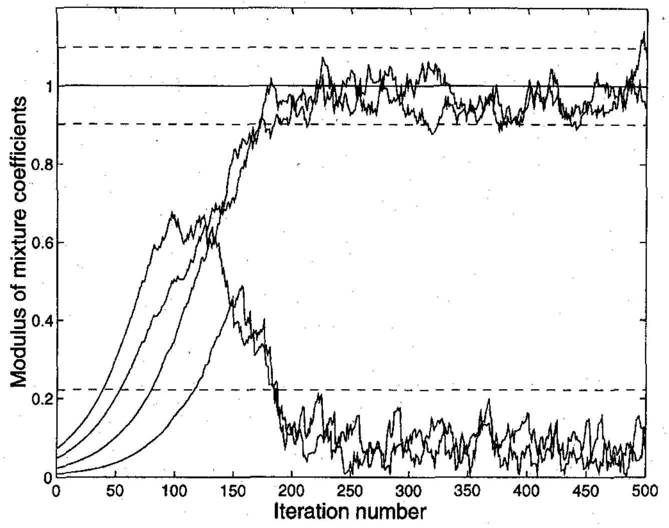


Fig. 4. Sample run. Convergence to 0 or 1 of the moduli of the coefficients of the global system B_tA. Fixed step size: $\lambda = 0.03$. Two QAM16 sources, cubic nonlinearities: $g_i(\mathbf{y}) = |y_i|^2 y_i$.

by a standard (nonrelative) gradient algorithm leads to the algorithm

$$B_{t+1} = B_t - \lambda_t \mathbf{f}'(\mathbf{y}_t)\mathbf{x}_t^{\mathrm{T}}$$

which is similar to (19) but does not enjoy uniform performance properties. The next section shows how the above results extend to find $H(\cdot)$ function solving orthogonally constrained optimization problems.

IV. Serial Updates for Orthogonal Contrasts

The contrast function ϕ_4 defined in (6) is in the form $\phi_4 = \mathbf{E}f(\mathbf{y})$ but must be optimized under the decorrelation constraint $R_y = \mathbf{E}\mathbf{y}\mathbf{y}^{\mathrm{T}} = I$. Batch procedures for optimizing contrast functions under this constraint have been described in [15], [13], and [27]; 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 intermediate vector $\mathbf{z}_t = W\mathbf{x}_t$, and the estimated source signal vector is $\mathbf{y}_t = U\mathbf{z}_t$ (see Fig. 3). By definition, W is a whitening matrix if its output is spatially white i.e.,

$$I = R_z \overset{\text{def}}{=} \mathbf{E}\big[\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}}\big] = WR_xW^{\mathrm{T}}. \tag{20}$$

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 $\mathbf{E}f(\mathbf{y}) = \mathbf{E}f(B\mathbf{x})$ over B under the constraint $R_y = I$ becomes that of minimizing $\mathbf{E}f(\mathbf{y}) = Ef(U\mathbf{z})$ over U under the constraint that U is orthogonal.

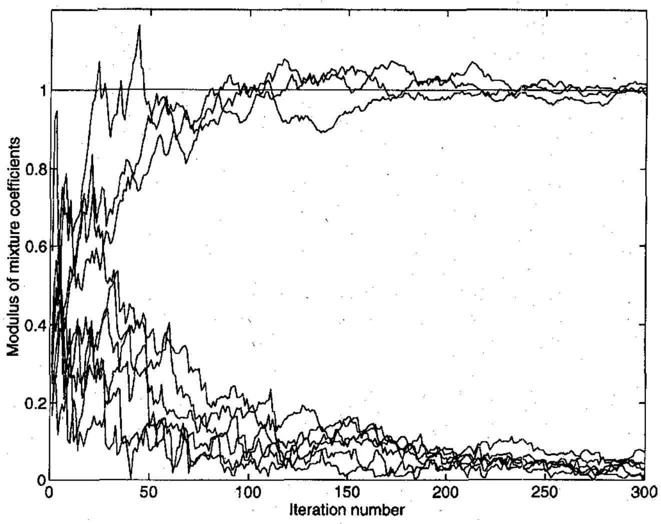


Fig. 5. Sample run. Convergence to 0 or 1 of the moduli of the coefficients of the global system B_tA. Three QAM16 sources. Decreasing step size: $\lambda_t = 2/t$.

How this program is completed in the adaptive context with serial updates is now described: 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 a separating matrix 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 may be obtained by minimizing a 'distance' between R_z and I. The Kullback–Leibler divergence [28] between two zero-mean normal distributions with covariance matrices equal to R_z and I, respectively, is

$$K(R_z) \overset{\text{def}}{=} \frac{1}{2}(\text{Trace}(R_z) - \log\det(R_z) - n). \tag{21}$$

The key property of this divergence measure is that $K(R_z) \geq 0$ with equality if and only if $R_z = I$. This may be proved by denoting $\mu_1, \ldots, \mu_n$ which are the eigenvalues of R_z. Then, $\text{Trace}(R_z) = \sum_{i=1,n}\mu_i$ and $\log\det(R_z) = \sum_{i=1,n}\log\mu_i$ so that $K(R_z) = 2^{-1}\sum_{i=1,n}\psi(\mu_i)$, where $\psi(\mu) = \mu - 1 - \log\mu$. This function is nonnegative because $\mu - 1 \geq \log\mu$ for any positive μ with equality only for $\mu = 1$. Hence, $K(R_z) \geq 0$ with equality i.f.f. $\mu_1 = \cdots = \mu_n = 1$, in which case, $R_z = I$ QED.

Thus, a whitening matrix is such that $K(R_z) = 0$; hence, it is a minimizer of

$$\phi_2(W) \overset{\text{def}}{=} K(WR_xW^{\mathrm{T}}). \tag{22}$$

The relative gradient of ϕ_2 is (see Appendix A)

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

A serial whitening algorithm is obtained, exactly as in (19), as

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

by deleting in the relative gradient descent the expectation operator in $\nabla\phi_2$. Interestingly enough, rule (24) can be shown to correspond to the first order (in λ) approximation of

Potter formula [29] 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 the contrast function (6) is minimized. However, the optimization is with respect to an orthogonal matrix U. We have seen that unconstrained minimization of such an objective leads to the updating rule (19). Of course, such a rule would *not* preserve the orthogonality of U, thus violating the orthogonality constraint. Note that if matrix U is orthogonal, i.e., $UU^{\mathrm{T}} = I$ and is modified into $U + \mathcal{E}U$, then

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

so that the orthogonality constraint is met at first order (in the sense that $(U+\mathcal{E}U)(U+\mathcal{E}U)^{\mathrm{T}} = I + o(\mathcal{E}))$ only if $\mathcal{E}^{\mathrm{T}} = -\mathcal{E}$, i.e., if $\mathcal{E}$ is skew symmetric. Thus, the steepest direction preserving the orthogonality of U is obtained by projecting the gradient onto the space of skew-symmetric matrices.

The orthogonal projection of $\nabla\phi_4$ onto the skew-symmetric matrix set is $(\nabla\phi_4 - \nabla\phi_4^{\mathrm{T}})/2$. We are thus led to consider the serial update obtained by skew symmetrizing the rule (19) into

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

Such an updating rule does not preserve orthogonality *exactly*, but only at first order in λ. The next section shows that this problem disappears when the whitening stage and the orthogonal stage are considered altogether.

Note that orthogonality could also be preserved by some parameterization of the orthogonal matrices (as product of Givens rotations for instance), but this solution cannot be considered here because it would result in the spoilage of the uniform performance property and because we ultimately want to get rid of the factorization of B into two distinct matrices W and U.

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 is updated according to (24), and U_t is updated according to (26).[2] This is

$$\begin{aligned} B_{t+1} &= U_{t+1}W_{t+1} \\ &= \left(U_t - \lambda_t\left[\mathbf{f}'(\mathbf{y}_t)\mathbf{y}_t^{\mathrm{T}} - \mathbf{y}_t\mathbf{f}'(\mathbf{y}_t)^{\mathrm{T}}\right]U_t\right) \\ &\quad \times \left(W_t - \lambda_t\left[\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}} - I\right]W_t\right) \\ &= \left(I - \lambda_t\left[\mathbf{f}'(\mathbf{y}_t)\mathbf{y}_t^{\mathrm{T}} - \mathbf{y}_t\mathbf{f}'(\mathbf{y}_t)^{\mathrm{T}}\right]\right)U_t \\ &\quad \times \left(I - \lambda_t\left[\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}} - I\right]\right)W_t \\ &= \left(I - \lambda_t\left[\mathbf{f}'(\mathbf{y}_t)\mathbf{y}_t^{\mathrm{T}} - \mathbf{y}_t\mathbf{f}'(\mathbf{y}_t)^{\mathrm{T}}\right]\right) \\ &\quad \times \left(I - \lambda_t\left[\mathbf{y}_t\mathbf{y}_t^{\mathrm{T}} - I\right]\right)U_tW_t \\ &= \left(I - \lambda_t\left[\mathbf{f}'(\mathbf{y}_t)\mathbf{y}_t^{\mathrm{T}} - \mathbf{y}_t\mathbf{f}'(\mathbf{y}_t)^{\mathrm{T}}\right.\right. \\ &\quad \left.\left. + \mathbf{y}_t\mathbf{y}_t^{\mathrm{T}} - I + O\left(\lambda_t^2\right)\right]\right)B_t \end{aligned} \qquad (27)$$

where we have used the identity $U_t(I - \lambda_t[\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}} - I]) = (I - \lambda_t[\mathbf{y}_t\mathbf{y}_t^{\mathrm{T}} - I])U_t$, which stems from $U_t^{\mathrm{T}}U_t = I$ and $\mathbf{y}_t = U_t\mathbf{z}_t$. Discarding the term of order λ_t^2, we finally obtain

$$B_{t+1} = B_t - \lambda_t H_4(\mathbf{y}_t)B_t \qquad (28)$$

where function $H_4(\cdot)$ appears to be

$$H_4(\mathbf{y}) = \mathbf{y}\mathbf{y}^{\mathrm{T}} - I + \mathbf{f}'(\mathbf{y})\mathbf{y}^{\mathrm{T}} - \mathbf{y}\mathbf{f}'(\mathbf{y})^{\mathrm{T}}. \qquad (29)$$

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

V. A Family of Equivariant Adaptive Algorithms

In the previous section, the notion of relative gradient applied to a fourth-order contrast function (6) provided us with a specific form (29) of function $H(\cdot)$. The source separation algorithms defined in this section and studied below improve on (29) by modifying it in two respects: Arbitrary nonlinear functions are considered for increased flexibility, and stabilization factors are introduced.

A. The EASI Algorithms

A 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 $H(\cdot) = H_4(\cdot)$ given by (29), this equation can be decomposed into symmetric and skew-symmetric parts, namely,

$$\mathrm{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}] = I \qquad (30)$$

$$\mathrm{E}[\mathbf{f}'(\mathbf{y})\mathbf{y}^{\mathrm{T}} - \mathbf{y}\mathbf{f}'(\mathbf{y})^{\mathrm{T}}] = 0. \qquad (31)$$

Condition (30) means that the output $\mathbf{y}$ is spatially white; it matches the normalization convention (5). This condition ensures *second-order* independence (i.e., decorrelation) of the separated signals. It is clearly verified in particular if $\mathbf{y} = B\mathbf{x}$, and B is a separating matrix. However, it is 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, something other than second-order conditions are required; these are provided by the skew-symmetric condition (31). If the components of $\mathbf{y}$ are mutually independent, then, for $i \neq j$, one has $\mathrm{E}[y_i f_j'(y_j)] = \mathrm{E}y_i \ \mathrm{E}f_j'(y_j) = 0$, where the first equality is by the independence assumption and the second equality by the zero mean assumption: $\mathrm{E}y_i = 0$. It follows that the off-diagonal entries of matrix $\mathrm{E}[\mathbf{y}\mathbf{f}'(\mathbf{y})^{\mathrm{T}}]$ are zero, and consequently, condition (31) is satisfied if B is a separating matrix.

We just proved that $\mathrm{E}H_4(\mathbf{y}) = 0$ whenever B is a separating matrix. In doing so, function $\mathbf{f}'(\cdot)$ was only assumed to operate component wise. Then, for n arbitrary nonlinear functions $g_1(\cdots), \ldots, g_n(\cdots)$, let us define

$$\mathbf{g}(\mathbf{y}) = [g_1(y_1), \ldots, g_n(y_n)]^{\mathrm{T}}, \qquad (32)$$

$$H_{\mathbf{g}}(\mathbf{y}) = \mathbf{y}\mathbf{y}^{\mathrm{T}} - I + \mathbf{g}(\mathbf{y})\mathbf{y}^{\mathrm{T}} - \mathbf{y}\mathbf{g}(\mathbf{y})^{\mathrm{T}}. \qquad (33)$$

The serial adaptation rule (4) with $H(\cdot) = H_{\mathbf{g}}(\cdot)$ still has any separating matrix as a stationary point since the stationarity condition $\mathrm{E}H_{\mathbf{g}}(\mathbf{y}) = 0$ is proved exactly as above.

[2] Note that there is no reason for considering different step sizes λ in these two rules (24) and (26) since a ratio different from 1 could be integrated in f. Optimal scaling of the nonlinear function f is determined in Section VI-C.

In summary, any component-wise nonlinear function g is associated with a corresponding EASI algorithm:

EASI Algorithms for Adaptive Source Separation

$$B_{t+1} = B_t - \lambda_t \left[\mathbf{y}_t\mathbf{y}_t^{\mathrm{T}} - I + \mathbf{g}(\mathbf{y}_t)\mathbf{y}_t^{\mathrm{T}} - \mathbf{y}_t\mathbf{g}(\mathbf{y}_t)^{\mathrm{T}}\right]B_t. \quad (34)$$

The remainder of this paper is devoted to analyzing the performance of EASI algorithms. In particular, we characterize nonlinear functions $g_1(\cdot), \ldots, g_n(\cdot)$ allowing stable separation. As usual, stability depends on the distributions of the sources.

B. Normalized Form of EASI Algorithms

Whenever fast convergence is required, speed can be increased by increasing adaptation steps. Large steps may cause explosive behavior and make the algorithm more sensitive to possible outliers: Some kind of stabilization is necessary. However, stabilization should not spoil the uniform performance property. Recall that uniform performance of serial updates has been established under quite general conditions, but it was implicitly assumed that the separating matrix could take any value. Thus, in order to preserve uniform performance, stabilization should *not* involve any constraint on the separating matrix itself, like clipping the diagonal entries or normalizing the rows. Hence, stabilization must be achieved by modifying $H_{\mathbf{g}}(\cdot)$ itself. We consider the following normalized form:

Normalized EASI Algorithms

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

which is very similar to the modification of the LMS algorithm into the 'normalized LMS.'

The stabilization solution offered by the form (35) is admittedly *ad hoc*, and other similar solutions could be considered. The mechanism behind it is simple: If B_tA is far from the identity matrix or if some outlying observation is received, then the output $\mathbf{y}_t$ is likely to be 'large' in the sense that $|\mathbf{y}_t|$ and/or $|\mathbf{y}_t^{\mathrm{T}}\mathbf{g}(\mathbf{y}_t)|$ can be much greater than λ_t^{-1}. In this case, the denominators in (35) prevent the update to take too large values. Actually, provided that $\lambda_t < 1/n$, the Frobenius norm of the term in square bracket in (35) is easily found to be upper bounded by $3/\lambda_t$ for any value of $\mathbf{y}$. On the contrary, $||H_{\mathbf{g}}(\mathbf{y})||_{\mathrm{Fro}}$ can be arbitrarily large (for large enough $|\mathbf{y}|$).

Besides protection against outliers, as discussed above, the normalized form of EASI offers the following advantages. First, it entails very little extra computation with respect to (33), and it does not introduce any additional parameter. Second, when the system is close to a stationary point, the covariance of $\mathbf{y}$ is close to the identity matrix. Thus, for small enough λ, the normalized version is expected to behave like the raw version (see an example in Fig. 6) for which a detailed performance analysis is available (Section VI). Finally, the normalized form has proved very satisfactory in numerous numerical experiments.

C. Discussion

Nonlinearities, Stability, and Permutations: The choice of the nonlinear function g is, of course, crucial to the performance of the algorithm. For any choice of g, a separating matrix is a stationary point, but the real issue is the *stability* of the separating matrices. A stability condition is established below (47) by an asymptotic analysis that also gives some clues as to how to choose and scale the nonlinear functions $g_1, \ldots, g_n$.

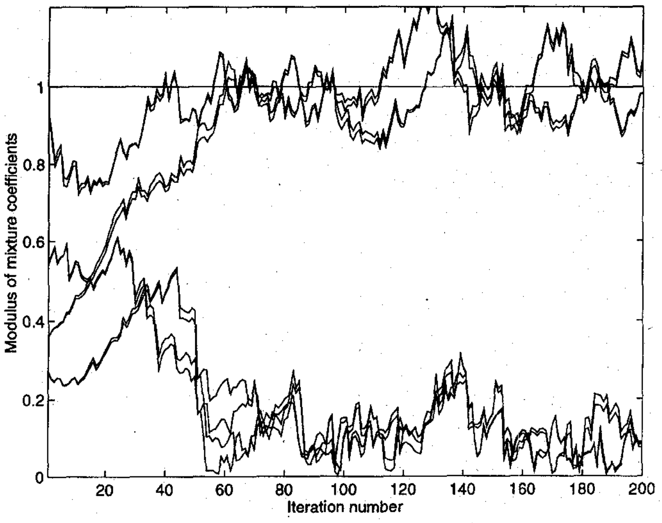


Fig. 6. Vertical axis: $20\log_{10}||C_t - I||_{\mathrm{Fro}}$. Lower panel: unbalanced nonlinearity. Convergence rate depends on the starting point. Upper panel: balanced nonlinearity. Isotropic convergence.

This analysis is conducted 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: We have shown above that uniform performance rigorously holds if model (1) is verified exactly. In particular, one can deal with arbitrarily ill-conditioned mixtures. This fact may appear paradoxical since the intrinsic hardness of most array processing problems usually depends on the conditioning of matrix A. However, this is not true in the specific case of model (1), which ignores any additive noise. Note that source separation remains statistically challenging even in the noiseless case unlike other array processing problems, like source bearing estimation for instance.[3]

Of course, some noise is always present in practice, and uniform performance can only be expected to hold in a more restricted sense. Intuition suggests that for high enough SNR, source separation performance is essentially unaffected by additive noise and that this 'high enough' SNR level actually depends on the conditioning of matrix A. This is in agreement with preliminary performance analysis results [26].

On the Scale Indetermination: Because of the scale indetermination 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

[3] In bearing estimation, matrix A is parameterized by source locations in such a way that the range space of A determines exactly the bearings and vice versa. In the noiseless case, this range space coincides exactly with the range of X_T for finite T: Deterministic identification is thus possible without statistical issue.

[2], [3], [9], or the rows of B_t are normalized [11]. We have chosen an alternate solution: Indeterminations are dealt with by requiring that the output signals have unit variance rather than by constraining the separating matrix. We recall that assuming an unconstrained separating matrix was necessary to establish the uniform performance property. However, requiring unit variance outputs offers another important benefit: Knowing in advance the range of the output signals allows the nonlinearities to be properly scaled. 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 that 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: The operating domain of the nonlinear functions becomes unpredictable.

VI. Performance Analysis

In this section, some quantities governing stability and performance are evaluated. Since theoretical results are mainly available in the limit of arbitrarily small step size, we use the form (33) of function $H(\cdot)$ rather than the normalized version of (35).

We informally recall some definitions and results (see [30]) about stochastic algorithms in the form

$$\theta_{t+1} = \theta_t - \lambda_t \psi(\theta_t, \mathrm{x}_t) \tag{36}$$

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

$$\Gamma \overset{\mathrm{def}}{=} \left. \frac{\partial \mathbf{E}\psi(\theta, \mathrm{x})}{\partial \theta} \right|_{\theta=\theta_\star} \tag{37}$$

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 Lyapunov equation

$$\Gamma \mathrm{Cov}(\theta_t) + \mathrm{Cov}(\theta_t)\Gamma^{\mathrm{T}} = \lambda P \tag{38}$$

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

$$P \overset{\mathrm{def}}{=} \mathrm{Cov}(\psi(\theta_\star, \mathrm{x})) = \mathbf{E}[\psi(\theta_\star, \mathrm{x})\psi^{\mathrm{T}}(\theta_\star, \mathrm{x})]. \tag{39}$$

Clearly, this result does not apply in full rigor to the source separation problem where, due to the permutation indetermination, 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 closed form solution of (38) is given below and is indeed found to predict with great accuracy the residual error of source separation observed in numerical simulations (see Table I).

TABLE I
Effect of Normalization

	$\lambda^{-1}\mathrm{ISI}_{12}$; QAM4 sources		
λ	Theor.	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
	$\lambda^{-1}\mathrm{ISI}_{12}$; QAM16 sources		
λ	Theor.	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

We recall that it is only necessary to study the dynamics of the global system C_t as given by (10). The above results apply to our algorithm by the identifications $\theta_t \to C_t$ and $\psi(\theta, \mathrm{x}) \to H_{\mathrm{g}}(C\mathrm{s})C$. It is also necessary to vectorize these matrices. The following convention turns out to be 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

$$\theta = [\underbrace{\ldots, c_{ij}, c_{ji}, \ldots}_{1 \leq i < j \leq n}, \underbrace{\ldots, c_{ii}, \ldots}_{1 \leq i \leq n}]^{\mathrm{T}} \tag{40}$$

and similarly for matrix $H_{\mathrm{g}}(C\mathrm{s})C$.

A. Stability of the Separating Matrices

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

$$\mathcal{H}(C) \overset{\mathrm{def}}{=} \mathbf{E}[H_{\mathrm{g}}(C\mathrm{s}_t)C]. \tag{41}$$

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

$$\mathcal{H}_{ii}(I + \mathcal{E}) = 2\mathcal{E}_{ii} + o(\mathcal{E}) \tag{42}$$

$$\begin{bmatrix} \mathcal{H}_{ij}(I + \mathcal{E}) \\ \mathcal{H}_{ji}(I + \mathcal{E}) \end{bmatrix} = DJ^{ij}D^{-1} \begin{bmatrix} \mathcal{E}_{ij} \\ \mathcal{E}_{ji} \end{bmatrix} + o(\mathcal{E}) \tag{43}$$

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

$$D \overset{\mathrm{def}}{=} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad J^{ij} \overset{\mathrm{def}}{=} \begin{bmatrix} 2 & 0 \\ \xi_i - \xi_j & \kappa_i + \kappa_j \end{bmatrix} \tag{44}$$

with the nonlinear moments of the source signals

$$\kappa_i \overset{\mathrm{def}}{=} \mathbf{E}[g_i'(s_i) - s_i g_i(s_i)] \tag{45}$$

$$\xi_i \overset{\mathrm{def}}{=} \mathbf{E}[g_i'(s_i) + s_i g_i(s_i)]. \tag{46}$$

The significant fact in (42) (holding for $1 \leq i \leq n$) and in (43) (holding for $1 \leq i < j \leq n$) is the pairwise decoupling. It means that, with the vectorization (40), 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.

Stability Condition: $\kappa_i + \kappa_j > 0 \quad \text{for } 1 \leq i < j \leq n.$ (47)

The stability condition for a separating matrix B such that BA is a permutation is similar. Indeed, if the source signal s_i is present at the $\sigma(i)$th output of B, then it undergoes the nonlinearity $g_{\sigma(i)}$. Hence, the stability of this separating B is again subject to (47), provided the moments κ_i are understood as $\mathbf{E}[g'_{\sigma(i)}(s_i) - s_i g_{\sigma(i)}(s_i)]$. 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 quasiidentity, 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 nonlinear moments κ_i deserve some comments. First note that if g_i is a cubic distortion $g_i(s_i) = s_i^3$, then $\kappa_i = 3 - \mathbf{E}|s_i|^4$ since $\mathbf{E}|s_i|^2 = 1$. This is simply the opposite of the fourth-order cumulant (or kurtosis) of s_i. The stability condition for cubic nonlinearities then is that the sum of the kurtosis of any two sources must be negative. Note that condition (47) actually is weaker than the requirement that all source signals have a negative kurtosis. In particular, stability condition (47) is verified if one source is Gaussian (in which case, its kurtosis is zero) and the other sources have negative kurtosis. In addtion, 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 nonlinear function g_i. This shows that stability condition (47) 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 except, possibly, one must be nonlinear to make a separating matrix stable.

B. Asymptotic Covariance and Rejection Rates

When the global system is $C = I + \mathcal{E}$, the ith estimated source signal (the ith output of C) is

$$\hat{s}_i = y_i = [(I + \mathcal{E})\mathbf{s}]_i = (1 + \mathcal{E}_{ii})s_i + \sum_{j \neq i} \mathcal{E}_{ij} s_j. \tag{48}$$

Since the signals are independent with unit variance and since $\mathcal{E}$ is of order $\sqrt{\lambda}$, (48) shows that the ratio of the variance of the (undesired) jth signal to the variance of the ith signal (of interest) is approximately equal to $|\mathcal{E}_{ij}|^2$. Hence, we are interested in computing pairwise rejection rates, which is defined as

$$\mathrm{ISI}_{ij} = \mathbf{E}|(C_t - I)_{ij}|^2 \quad 1 \leq i, j \leq n \tag{49}$$

which correspond to intersymbol interference (ISI) in the terminology of channel equalization.

If C_t is 'vectorized' in a n^2-dimensional parameter vector θ as in (40), these quantities are the diagonal elements of matrix $\mathrm{Cov}(\theta)$. Thanks to the regular structure of the EASI algorithms, the Lyapunov equation (38) can be solved in close form for $\mathrm{Cov}(\theta)$ for arbitrary $n, \mathbf{g}(\cdot)$ and signal distributions.

This general expression of ISI_{ij} is given in Appendix C in (83). For the sake of simplicity, this section only discusses the case of signals with identical distributions and of a single nonlinearity $g_i(\cdot) = g(\cdot)$ for $1 \leq i \leq n$. There is then only one extra moment to consider

$$\gamma \overset{\text{def}}{=} \mathbf{E}g^2(s)\mathbf{E}s^2 - \mathbf{E}^2[g(s)s] \tag{50}$$

where s is any of the s_i's. The rejection rates are (necessarily) identical. Equation (83) reduces to

$$\mathrm{ISI}_{ij} = \lambda\left(\frac{1}{4} + \frac{\gamma}{2\kappa}\right) \quad 1 \leq i \neq j \leq n. \tag{51}$$

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

$$\mathrm{ISI}_{ij} \geq \frac{\lambda}{4} \quad 1 \leq i \neq j \leq n \tag{52}$$

and this bound is reached when $s = \pm 1$ with equal probabilities, 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, again, discuss here only the simplest case, which is often encountered in practice, of sources with identical distributions. There is no reason then to use different nonlinearities: We take $g_i(\cdot) = g(\cdot)$ for $1 \leq i \leq n$, implying $\kappa_i = \kappa$ and $\gamma_i = \gamma$ for $1 \leq i \leq n$. These equalities are assumed throughout this section.

Local Convergence: The mean field $\mathcal{H}(C)$ has a very simple local structure when C is close to any quasiidentity attractor $C_\star$: Equations (42) and (43) combine into

$$\mathcal{H}(C_\star + \mathcal{E}) = 2\frac{\mathcal{E} + \mathcal{E}^{\mathrm{T}}}{2} + 2\kappa\frac{\mathcal{E} - \mathcal{E}^{\mathrm{T}}}{2} + o(\mathcal{E}) \tag{53}$$

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 2κ, respectively. When κ is known in advance or can be (even roughly) estimated, expression (53) suggests that we normalize the nonlinearity $g(\cdot)$ into $\tilde{g}(\cdot) = g(\cdot)/\kappa$ because then the nonlinear moment $\tilde{\kappa}$ associated with $\tilde{g}$ is $\tilde{\kappa} = 1$. With such a choice, the mean field in the neighborhood of an attractor becomes

$$\mathcal{H}(C_\star + \mathcal{E}) = 2\mathcal{E} + o(\mathcal{E}). \tag{54}$$

This is, in our opinion, a strong result because it means that any deviation $\mathcal{E}$ to a separator is pulled back to zero with a (mean) strength that is independent of the direction of the error. This very desirable property of isotropic convergence can only be achieved (in general) by resorting to second-order adaptive techniques, such as Newton-like algorithms.

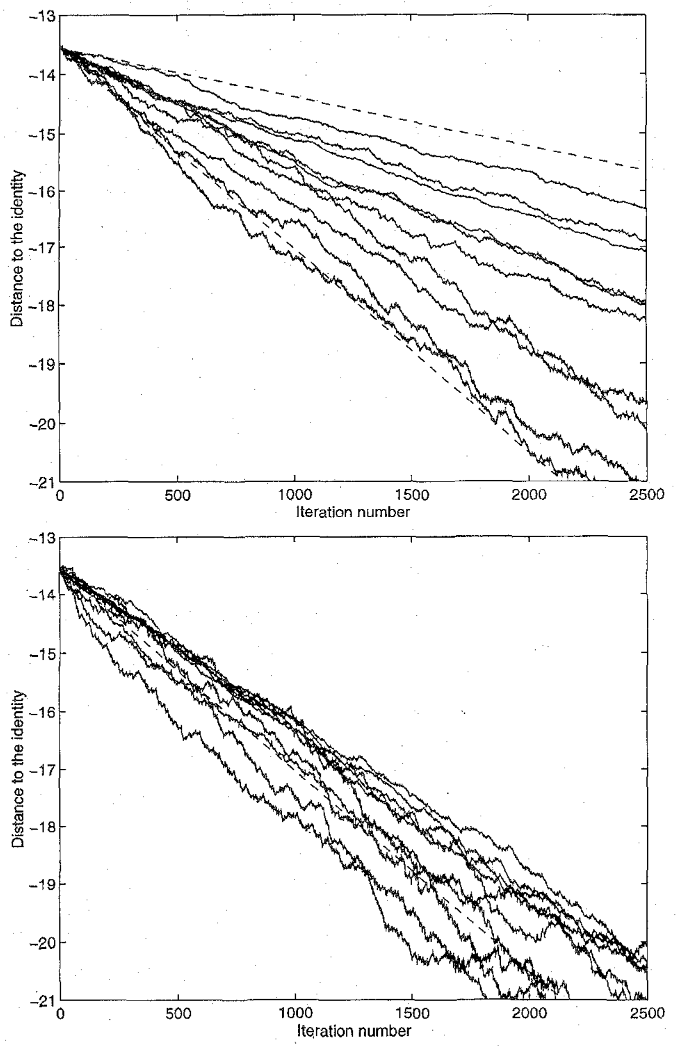


Fig. 7. Each row corresponds to a sampling time τ: From top to bottom, $\tau = 0, 30, 60, 90, 120, 150, 180$. Each column corresponds to one of the source signals. Each panel shows 200 estimated signal points in the portion $(-2, 2) \times (-2i, 2i)$ of the complex plane. For $\tau = 180$, the 'constant modulus feature' is approximately reached at both outputs, indicating a successful separation.

This benefit is seen to be obtained by our simple (first-order) gradient algorithm by simply adjusting the strength of the nonlinear function (see an example in Fig. 7).

Rejection Rates: The nonlinear function $g(\cdot)$ can be chosen to minimize the rejection rates under the constraint that its amplitude is fixed to ensure isotropic local convergence, which, as discussed above, is achieved for $\kappa = 1$. In view of (51), the optimal nonlinearity should minimize γ subject to $\kappa = 1$. This optimization problem is easily solved by the Lagrange multiplier method when the common probability distribution of the sources has a differentiable density $p(s)$ with respect to Lebesgue measure. The optimal nonlinearity is found to be

$$g_{\text{opt}}(s) = \frac{\psi(s)}{\mathbf{E}\psi^2(s) - 1} \quad \text{where } \psi(s) \overset{\text{def}}{=} -\frac{p'(s)}{p(s)}. \tag{55}$$

The resulting minimal rejection rate may be computed to be

$$\text{ISI}_{\min} = \lambda\left(\frac{1}{4} + \frac{1}{2(\mathbf{E}\psi^2(s) - 1)}\right). \tag{56}$$

It is interesting to note that the optimal nonlinear functions under the decorrelation constraint are proportional to those obtained without forcing this constraint (see the M.L. solution of [11]).

VII. 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 V-A on the stationary points carries over to the complex case with only one restriction: The diagonal terms of the skew-symmetric part of $\mathbf{E}H_{\mathrm{g}}(\mathrm{s})$ are not necessarily zero unless the scalar-to-scalar functions g_i are restricted to be phase-preserving, i.e., of the form

$$g_i(y_i) = y_i l_i(|y_i|^2) \quad 1 \leq i \leq n \tag{57}$$

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 have the following assumption:

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

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

$$\begin{bmatrix} \mathcal{H}_{ij}(I + \mathcal{E}) \\ \mathcal{H}_{ji}^*(I + \mathcal{E}) \end{bmatrix} = DJ^{ij}D^{-1} \begin{bmatrix} \mathcal{E}_{ij} \\ \mathcal{E}_{ji}^* \end{bmatrix} + o(\mathcal{E}) \tag{58}$$

where matrices D and J^{ij} are as in (44), but the nonlinear moments are slightly different

$$\kappa_i \overset{\text{def}}{=} \mathbf{E}[|s_i|^2 l_i'(|s_i|^2) + l_i(|s_i|^2) - |s_i|^2 l_i(|s_i|^2)], \tag{59}$$

$$\xi_i \overset{\text{def}}{=} \mathbf{E}[|s_i|^2 l_i'(|s_i|^2) + l_i(|s_i|^2) + |s_i|^2 l_i(|s_i|^2)]. \tag{60}$$

Stability condition (47) is then unchanged provided κ_i is defined according to (59). For cubic nonlinearities, i.e., for $l_i(s) = s$, one has $\kappa_i = 2 - \mathbf{E}|s_i|^4$ 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

$$\gamma_i \overset{\text{def}}{=} \mathbf{E}\left[|s_i|^2 l_i^2(|s_i|^2)\right] - [\mathbf{E}|s_i|^2 l_i(|s_i|^2)]^2 \tag{61}$$

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

which are the direct complex counterparts of those defined in (74). With these definitions, the rejections rates take the very same form, either in the i.i.d. case, as given by the simple formula (51) or in the general case, as given by the general expression (83).

VIII. 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 (except in Fig. 7).

Figs. 4–6 display trajectories of the modulus of the coefficients of the global system C_t. Hence, an experiment with n sources is illustrated by the n^2 trajectories of $|[C_t]_{ij}|$ as a function of t for all the pairs $1 \leq i, j \leq n$. Thus, a successful convergence is observed when n of these trajectories converge to 1 and $n^2 - n$ converge to 0.

Fast Convergence: Fast convergence is first illustrated by Fig. 4 for two i.i.d. QAM16 sources using the basic cubic nonlinearity $g_i(\mathbf{y}) = |y_i|^2 y_i$ for $1 \leq i \leq n$. The dashed lines represent ± two standard deviations computed from (51) and (77). Fig. 5 is similar except that three QAM16 sources are involved, and the step size is decreased according to the cooling scheme: $\lambda_t = 2/t$.

Effect of Normalization: The effect of normaliztion is investigated in Fig. 6. With the same QAM16 input, two serial algorithms are run with $\lambda = 0.01$: one with the normalized algorithm (35) and the other with the 'raw' algorithm (34). Both trajectories are displayed and show little discrepancy (see also Table I).

Isotropic Convergence: Isotropic convergence is illustrated by Fig. 7. It displays the evolution of a logarithmic distance of C_t to the identity, namely, $20 \log_{10} \|C_t - I\|_{\text{Fro}}$, 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 to make comparisons easier. Both panels are for cubic nonlinearities and uniformly distributed sources (this is the only experiment with real signals). Zero-mean uniformly distributed random variables have a normalized kurtosis equal to $-6/5$. We have seen that for $g(s) = s^3$, the moment κ is minus the kurtosis: $\kappa = 6/5$. According to discussion of Section VI-C, isotropic convergence is achieved by taking $g(s) = 5/6 \cdot s^3$ so that the corresponding moment is $\kappa = 1$. The resulting trajectories are displayed in the lower panel, where the straight dotted line corresponds to a distance varying as $\exp(-2\lambda t)$. The upper panel displays trajectories for $g(s) = 0.2s^3$. With this factor, the nonlinear moment takes the value $\kappa = 0.2 \times 6/5 = 0.24$. Thus, according to (53), the symmetric and skew-symmetric errors decay, respectively, as $\exp(-2\lambda t)$ and $\exp(-2 \cdot 0.24 \lambda t)$. These two functions are plotted as straight dotted lines in the upper panel. Various decay rates are observed in the upper panel, depending on the (random) proportion of symmetric and skew-symmetric errors in C_0. They are seen to be upper and lower bounded in accordance with theoretical predictions. The lower panel shows decay rates that are essentially independent of the initial condition C_0, evidencing the isotropic convergence obtained by a proper scaling of the nonlinear function.

Rejection Rates: Rejection rates predicted by (51) 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, $N_{\text{MC}} = 500$ 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 used to determine an experimental standard deviation. The table displays $\lambda^{-1}\text{ISI}_{12}$ and its empirical value plus and minus two standard deviations for both the normalized and non-normalized versions. There are no results presented for $\lambda = 0.1$ and QAM16 signals for the nonnormalized algorithms because a significant fraction of divergent trajectories have been observed. In all the other cases, representing 15×500 trajectories, no divergence was observed. It appears that asymptotic analysis correctly predicts the rejection rates for step sizes as large as $\lambda = 0.1$. Note that normalization does not affect much the empirical performance.

Application to Digital Communications: The performance of EASI is illustrated by running the algorithm on data generated from a digital communications testbed. The data set is the output of an eight-element linear array with a sensor spacing of 0.574 wavelength operating at 1.89 GHz. The complex baseband signals are digitized at 4.6 MHz, and the symbol rate is 1.15 MHz. Because the source signals have constant modulus,[4] the quality of separation is evidenced by displaying the separator outputs in the complex plane and checking that they do sit close to a circle.

In the following run, EASI is fed with the outputs of sensors 1 and 5; it is initialized with $B_0 = I_2$ and is run with a constant step size $\lambda = 0.01$. Fig. 7 shows the evolution of the separation. Because the mixing matrix is unknown, the evolution of the global system cannot be displayed here. To evidence convergence of B_t to a separating matrix, we have to use the data themselves. The following device is used: At time τ, the current value of B_τ is sampled and is applied to a batch of 200 samples $\{x_t \mid t = 1, 200\}$ of the array output. The first (resp. second) column shows, in the complex plane, the first (resp. second) coordinate of $B_\tau x_t$ for $t = 1, \ldots, 200$. Each row corresponds to a sampling time τ. We use $\tau = 0, 30, 60, 90, 120, 150, 180$. That a successful separation is achieved is seen by the fact that the points gather close to the unit circle.

IX. Summary and Conclusion

A class of equivariant adaptive algorithms for the blind separation of n independent sources was introduced. This class is defined by the 'serial updating' rule (4), and a particular serial algorithm in this class is determined by the specification of a vector-to-matrix mapping $H(\cdot)$ such that matrix $\mathbf{E}H(\mathbf{y}) = 0$ when vector $\mathbf{y}$ has independent components.

The very desirable property of 'uniform performance' is guaranteed by the simple device of serial updating. For adaptive algorithms, uniform performance 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, making it possible to tune and optimize the algorithm once for all mixtures.

[4] The GMSK modulation is used in this data set.

To obtain uniformly *good* performance, the (relative) gradient of an orthogonal contrast functions is computed, and the resulting form was then generalized into a family (33) of matrix-valued functions $H_{\mathrm{g}}(\cdot)$. A particular matrix $H_{\mathrm{g}}(\cdot)$ depends on a set of n arbitrary scalar functions $g_1, \ldots, g_n$. Performance of the algorithm then depends on an appropriate choice of these functions with respect to the distribution of the sources. This was investigated via an asymptotic analysis.

Stability conditions and rejection rates are given in close form for arbitrary $g_1, \ldots, g_n$. A striking result is that functions g_i can be scaled to obtain isotropic convergence; this Newton-like feature is obtained even though our algorithm basically is a stochastic gradient algorithm. Optimal nonlinearities, optimizing source rejection under the constraint of isotropic convergence, were also derived.

Numerical experiments confirm that the asymptotic analysis characterizes the algorithm with a very good accuracy, even for relatively large adaptation steps. Finally, a sample run on real array data (digital communications signals) was presented to illustrate the fast convergence of the algorithm.

The general theme of this paper is that when *matrices* are to be updated, specific rules may be considered that have no equivalent for a generic adaptive system with an unstructured vector of parameters. This specificity lies in the multiplicative nature of the source separation problem.

Appendix A
Relative Gradient for Whitening

To compute the relative gradient of ϕ_2, we recall that for a positive matrix $R, \log\det(R + \delta R) = \log\det(R) + \mathrm{Trace}\{R^{-1}\delta R\} + o(\delta R)$ so that the differential of function K at a positive R is

$$K(R+\mathcal{E}) = K(R) + \frac{1}{2}\mathrm{Trace}\{(I - R^{-1})\mathcal{E}\} + o(\mathcal{E}). \quad (63)$$

The first-order relative expansion of ϕ_2 follows

$$\begin{aligned}
&\phi_2(W + \mathcal{E}W)\\
&\quad = \phi_2((I+\mathcal{E})W)\\
&\quad = K((I+\mathcal{E})WR_xW^{\mathrm{T}}(I+\mathcal{E})^{\mathrm{T}})\\
&\quad = K((I+\mathcal{E})R_z(I+\mathcal{E})^{\mathrm{T}})\\
&\quad = K\big(R_z + \mathcal{E}R_z + R_z\mathcal{E}^{\mathrm{T}} + o(\mathcal{E})\big)\\
&\quad = K(R_z) + \frac{1}{2}\mathrm{Trace}\big\{\big(I - R_z^{-1}\big)\big(\mathcal{E}R_z + R_z\mathcal{E}^{\mathrm{T}}\big)\big\} + o(\mathcal{E})\\
&\quad = K(R_z) + \frac{1}{2}\mathrm{Trace}\{(R_z - I)(\mathcal{E} + \mathcal{E}^{\mathrm{T}})\} + o(\mathcal{E})\\
&\quad = K(R_z) + \mathrm{Trace}\{(R_z - I)\mathcal{E}\} + o(\mathcal{E})\\
&\quad = K(R_z) + \langle R_z - I \mid \mathcal{E}\rangle + o(\mathcal{E}).
\end{aligned}$$

Identifying the last expression with definition (13) yields expression (23).

Appendix B
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 definition (41) also reads

$$\mathcal{H}(I+\mathcal{E}) = \mathbf{E}[H(\mathrm{s} + \mathcal{E}\mathrm{s})(I+\mathcal{E})]. \quad (64)$$

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

$$\mathcal{H}(I+\mathcal{E}) = \mathbf{E}H(\mathrm{s}+\mathcal{E}\mathrm{s}) + o(\mathcal{E}). \quad (65)$$

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

$$\mathbf{E}[(\mathrm{s}+\mathcal{E}\mathrm{s})(\mathrm{s}+\mathcal{E}\mathrm{s})^{\mathrm{T}} - I] = \mathcal{E} + \mathcal{E}^{\mathrm{T}} + o(\mathcal{E}) \quad (66)$$

since our normalization convention is $\mathbf{E}[\mathrm{ss}^{\mathrm{T}}] = I$.

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

$$\begin{aligned}
y_ig_j(y_j) = s_ig_j(s_j) &+ \sum_a \mathcal{E}_{ia}s_ag_j(s_j)\\
&+ \sum_b \mathcal{E}_{jb}s_is_bg_j'(s_j) + o(\mathcal{E}). \quad (67)
\end{aligned}$$

There is no need to evaluate the terms for $i = j$ since they cancel after skew symmetrization. Focusing on terms with $i \neq j$, we next find that

$$\mathbf{E}s_ag_j(s_j) = \delta(j,a)\mathbf{E}s_jg_j(s_j) \quad (68)$$

$$\mathbf{E}s_is_bg_j'(s_j) = \delta(i,b)\mathbf{E}s_i^2\mathbf{E}g_j'(s_j) \quad \text{for } i \neq j \quad (69)$$

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

$$\mathbf{E}y_ig_j(y_j) = \mathcal{E}_{ij}\mathbf{E}s_jg_j(s_j) + \mathcal{E}_{ji}\mathbf{E}s_i^2\mathbf{E}g_j'(s_j) + o(\mathcal{E}). \quad (70)$$

Expectations (65), (66), and (70) then combine into

$$\begin{aligned}
\mathcal{H}_{ij}(I+\mathcal{E}) = &\,\mathcal{E}_{ij}\big(1 + \mathbf{E}s_j^2\mathbf{E}g_i'(s_i) - \mathbf{E}s_jg_j(s_j)\big)\\
&+ \mathcal{E}_{ji}\big(1 - \mathbf{E}s_i^2\mathbf{E}g_j'(s_j) + \mathbf{E}s_ig_i(s_i)\big) + o(\mathcal{E})
\end{aligned} \quad (71)$$

which, after symmetrization, yields (43).

Appendix C
Asymptotic Covariance

To solve (38), we must first evaluate matrix P. Using source independence, it is easily checked that most of the entries of $H_{\mathrm{g}}(\mathrm{s})$ are uncorrelated. The nonvanishing terms can be computed to be

$$\mathrm{Cov}(H_{ii}(\mathrm{s})) = \mathbf{E}s_i^4 - 1 \quad (72)$$

$$\mathrm{Cov}\left(\begin{bmatrix} H_{ij}(\mathrm{s}) \\ H_{ji}(\mathrm{s}) \end{bmatrix}\right) = DQ^{ij}D^{\mathrm{T}} \quad (73)$$

with the following definitions

$$Q^{ij} \overset{\mathrm{def}}{=} \begin{bmatrix} 1 & \mu_i - \mu_j \\ \mu_i - \mu_j & \gamma_i + \gamma_j + (\mu_i - \mu_j)^2 \end{bmatrix} \quad (74)$$

$$\gamma_i \overset{\mathrm{def}}{=} \mathbf{E}[g_i^2(s_i)] - [\mathbf{E}s_ig_i(s_i)]^2 \quad (75)$$

$$\mu_i \overset{\mathrm{def}}{=} \mathbf{E}[g_i(s_i)s_i]. \quad (76)$$

This is a pleasant finding since it means that P has the same block diagonal structure as Γ, allowing the Lyapunov equation (38) to be solved block wise.

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

$$\mathrm{Cov}(C_{ii}) = \lambda \frac{\mathbf{E}s_i^4 - 1}{4}. \tag{77}$$

The 2×2 block Lyapunov equation extracted from (38) for a pair $i \neq j$ is

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

where we set

$$R^{ij} \overset{\mathrm{def}}{=} \mathrm{Cov}\left(\begin{bmatrix} C_{ij}(\mathbf{s}) \\ C_{ji}(\mathbf{s}) \end{bmatrix}\right). \tag{79}$$

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

$$J^{ij}(D^{-1}R^{ij}D^{-\mathrm{T}}) + (D^{-1}R^{ij}D^{-\mathrm{T}})J^{ij\mathrm{T}} = \lambda Q^{ij}. \tag{80}$$

Matrix J^{ij} being lower triangular, the 2×2 Lyapunov equation may be solved explicitly. This purely algebraic task needs not be reported here. Only the following intermediate result is needed: If (x, y, t) is the solution of

$$\begin{bmatrix} a & 0 \\ c & b \end{bmatrix}\begin{bmatrix} x & t \\ t & y \end{bmatrix} + \begin{bmatrix} x & t \\ t & y \end{bmatrix}\begin{bmatrix} a & c \\ 0 & b \end{bmatrix} = \begin{bmatrix} \alpha & \gamma \\ \gamma & \beta \end{bmatrix} \tag{81}$$

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

$$x + y + 2t = \frac{\alpha}{2a} + \frac{\beta}{2b} + \frac{(2b - c)(2a\gamma - c\alpha)}{2ab(a + b)}. \tag{82}$$

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

$$\mathbf{E}|C_{ij}|^2 = \mathrm{Cov}(C_{ij}) = \lambda\left(\frac{1}{4} + \frac{1}{2}\frac{\gamma_i + \gamma_j}{\kappa_i + \kappa_j} + \beta_{ij}^{+} + \beta_{ij}^{-}\right) \tag{83}$$

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

$$\beta_{ij}^{+} \overset{\mathrm{def}}{=} \frac{2(\kappa_i + \kappa_j)(\mu_i - \mu_j)^2 + (\kappa_i - \kappa_j)^2}{4(\kappa_i + \kappa_j)(2 + \kappa_i + \kappa_j)} \tag{84}$$

$$\beta_{ij}^{-} \overset{\mathrm{def}}{=} \frac{(2\mu_i - \kappa_i) - (2\mu_j - \kappa_j)}{2(2 + \kappa_i + \kappa_j)}. \tag{85}$$

ACKNOWLEDGMENT

The authors are grateful to A. Carr and R. Arnott from ERA Technology for providing us with a real data set from their SCARF experiment.

REFERENCES

[1] L. Tong, R. Liu, V. Soon, and Y. Huang, "Indeterminacy and identifiability of blind identification," *IEEE Trans. Circuits Syst.*, vol. 38, pp. 499–509, May 1991.

[2] 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*, Nice, France, 1985, pp. 1017–1020.

[3] C. Jutten and J. Hérault, "Blind separation of sources: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, pp. 1–10, 1991.

[4] A. Dinç and Y. Bar-Ness, "Bootstrap: A fast blind adaptive signal separator," in *Proc. ICASSP*, vol. 2, 1992, pp. 325–328.

[5] J. Karhunen and J. Joutsensalo, "Representation and separation of signals using nonlinear PCA type learning," *Neural Networks*, vol. 7, no. 1, pp. 113–127, 1993.

[6] A. Cichocki and L. Moszczynski, "New learning alforithms for blind separation of sources," *Electron. Lett.*, vol. 28, pp. 1986–1987, 1992.

[7] A. Cichocki and R. Unbehauen, "New neural networks with on-line learning for blind identification and blind separation of sources," Preprint.

[8] N. Delfosse and P. Loubaton, "Adaptive separation of independent sources: A deflation approach," in *Proc. ICASSP*, vol. 4, 1994, pp. 41–44.

[9] E. Moreau and O. Macchi, "New self-adaptive algorithms for source separation based on contrast functions," in *Proc. IEEE Signal Processing Workshop Higher Order Statist.*, Lake Tahoe, 1993, pp. 215–219.

[10] A. J. Bell and T. Sejnowski, "Blind separation and blind deconvolution: An information-theoretic approach," in *Proc. ICASSP*, Detroit, 1995.

[11] D.-T. Pham, P. Garrat, and C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach," in *Proc. EUSIPCO*, 1992, pp. 771–774.

[12] A. Belouchrani and J.-F. Cardoso, "Maximum likelihood source separation for discrete sources," in *Proc. EUSIPCO*, Edinburgh, Sept. 1994, pp. 768–771.

[13] P. Comon, "Independent component analysis," in *Proc. Int. Workshop Higher Order Statist.*, Chamrousse, France, 1991, pp. 111–120.

[14] M. Gaeta and J.-L. Lacoume, "Source separation without a priori knowledge: The maximum likelihood solution," in *Proc. EUSIPCO*, 1990, pp. 621–624.

[15] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *Proc. Inst. Elec. Eng.*, pt. F, vol. 140, pp. 362–370, Dec. 1993.

[16] O. Shalvi and E. Weinstein, "New criteria for blind deconvolution of nonminimum phase systems (channels)," *IEEE Trans. Inform. Theory*, vol. 36, no. 2, pp. 312–321, 1990.

[17] L. Tong, Y. Inouye, and R. Liu, "Waveform preserving blind estimation of multiple independent sources," *IEEE Trans. Signal Processing*, vol. 41, pp. 2461–2470, July 1993.

[18] J.-F. Cardoso, "Iterative techniques for blind source separation using only fourth order cumulants," in *Proc. EUSIPCO*, 1992, pp. 739–742.

[19] ——, "Fourth-order cumulant structure forcing. Application to blind array processing," in *Proc. 6th IEEE SSAP Workshop*, Oct. 1992, pp. 136–139.

[20] K. Matsuoka, M. Ohya, and M. Kawamoto "A neural net for blind separation of nonstationary signals," *Neural Networks*, vol. 8, no. 3, pp. 411–419, 1995.

[21] L. Féty and J. P. Van Uffelen, "New methods for signal separation," in *Proc. 4th Int. Conf. HF Radio Syst. Techn.*, London, Inst. Elec. Eng., Apr. 1988, pp. 226–230.

[22] D. Pham and P. Garat, "Séparation aveugle de sources temporellement corrélées," in *Proc. GRETSI*, 1993, pp. 317–320.

[23] K. Abed Meraim, A. Belouchrani, J.-F. Cardoso, and É. Moulines, "Asymptotic performance of second order blind source separation," in *Proc. ICASSP*, Apr. 1994, vol. 4, pp. 277–280.

[24] S. V. Gerven and D. V. Compernolle, "On the use of decorrelation in scalar signal separation," in *Proc. ICASSP*, Adelaide, Australia, 1994.

[25] E. L. Lehmann, *Testing Statistical Hypothesis*. New York: Wiley, 1959.

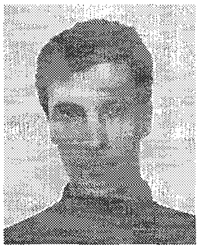
[26] J.-F. Cardoso, "On the performance of source separation algorithms," in *Proc. EUSIPCO*, Edinburgh, Sept. 1994, pp. 776–779.

[27] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, pp. 287–314, Apr. 1994, Special issue on *Higher-Order Statistics*.

[28] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[29] J. E. Potter, "New statistical formulas," Tech. Rep., Instrum. Lab., Mass. Inst. Technol., 1963.

[30] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*. New York: Springer Verlag, 1990.



Jean-François Cardoso (M'91) was born on March 1, 1958. He received the agrégation de physique degree from the École Normale Supérieure de Saint-Cloud in 1981 and the doctorat de physique degree from the University of Paris in 1984.

He currently is with the Centre National de la Recherche Scientifique (C.N.R.S.) and works in the 'Signal' department of École Nationale Supérieure des Télécommunications (E.N.S.T.). His research interests include statistical signal processing, with emphasis on (blind) array processing and performance analysis.

Dr. Cardoso is one of the coordinators of ISIS, the C.N.R.S. research group on signal and image processing, and he has been a member of the SSAP Technical Committee of the IEEE Signal Processing Society since 1995.

Beate Hvam Laheld was born in Oslo, Norway, in 1967. She graduated from the Norwegian Technical Institute (N.T.H.), Trondheim, Norway, in 1989. She received the Masters and Ph.D. degrees from École Nationale Supérieure des Télécommunications (E.N.S.T.), Paris, France, in signal and image processing, in 1991 and 1994, respectively.

From 1990 to 1994, she was with the 'Signal' Department of E.N.S.T. and worked on blind array processing problems, with emphasis on blind adaptive source separation. In 1994, she joined France Télécom Mobiles, where she is currently in charge of several optimization tasks in the GSM network.