# The invariant approach to source separation

Jean-François Cardoso
ENST / CNRS / GdR TdSI
46 rue Barrault, 75634 Paris, France. cardoso@sig.enst.fr
http://sig.enst.fr/~cardoso/stuff.html

ABSTRACT

The notion of *equivariance* is relevant to source separation because multiplication of mixed signals is equivalent to changing the unknown parameter (the mixing matrix) into another mixing matrix. Elaborating on this observation, a wide class of batch estimators of the mixing matrix is first shown to offer uniform performance: quality of separation does not depend on the hardness of the mixture. Equivariance is next extended to adaptive algorithms, by the device of 'serial updating'. Adaptive separators based on such a learning rule also exhibit uniform performance in a strong sense.

## I INTRODUCTION: SOURCE SEPARATION

Source separation, blind array processing, signal copy, independent component analysis, waveform preserving estimation...: these keywords refer to a signal model which is receiving increasing attention in both signal processing and neural network literature since the seminal paper [1]. This 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 this paper, we will consider the simplest case where any additive noise can be neglected and the number of mixtures is equal to the number of sources. The signal model then is that of a $n$-dimensional time series $\mathbf{x}_t$ in the form :

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

where $\mathbf{x}_t$ and $\mathbf{s}_t$ are $n \times 1$ vectors. and $A$ is a $n \times n$ matrix. The components of $\mathbf{s}_t$ are often termed 'source signals' and matrix $A$ is the 'mixing matrix'.

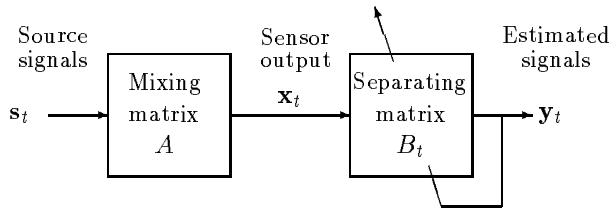
**Adaptive source separation** consists in updating an $n \times n$ matrix $B_t$ such that its output $\mathbf{y}_t$:

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

is as close as possible to the source signals, *i.e.*the *global system* $C_t$:

$$C_t \stackrel{\text{def}}{=} B_t A$$

should be driven to the identity matrix.




**Batch source separation** consists in computing an estimate $\widehat{A}$ of $A$ from of a batch $X_T$ of $T$ samples:

$$X_T \stackrel{\text{def}}{=} [\mathbf{x}_1, \ldots, \mathbf{x}_T].$$

The estimate $\widehat{A}$ returned by any particular estimator is a function of the observations. This may be denoted as

$$\widehat{A} = \mathcal{A}(X_T)$$

where the mapping $\mathcal{A}$ corresponds to a given estimator. Source signals $\mathbf{s}_t$ are then estimated as $\hat{\mathbf{s}}_t = \mathbf{y}_t = (\widehat{A})^{-1}\mathbf{x}_t$. As for adaptive algorithms, we define a global system $\widehat{C}$ by

$$\widehat{C} = (\widehat{A})^{-1} A$$

which should be as close as possible to the identity matrix, since $\hat{\mathbf{s}}_t = \widehat{C}\mathbf{s}_t$.

**Assumptions.** For the purpose of source separation, the following assumptions are made

*A1. Matrix A is invertible.*
*A2. Source signals are ergodic zero-mean processes.*
*A3. Source signals are statistically independent.*
*A4. Source signals 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 when the source signals originate from physically separated systems. We note that assumption 4 only is a *normalization convention* because the amplitude of each source signal can be incorporated into $A$. Assumptions 2, 3 and 4 imply that

$$R_s \stackrel{\text{def}}{=} \mathrm{E}\,[\,\mathbf{s}_t\mathbf{s}_t^{\mathrm{T}}] = I \tag{1}$$

Assumption 1 is expected to hold 'almost surely' in any physical situation. As a matter of fact, the most questionable assumption is the existence of $A$ itself *i.e.* the plausibility of observing instantaneous mixtures.

It is important to realize that without additional information, the outputs of a separating matrix cannot be ordered because the ordering of the source signals is itself immaterial (conventional): source signals can be at best recovered up to a permutation and a change of sign. To simplify exposition, it is assumed throughout that this indetermination can be fixed in one way or another, possibly using a priori information.

## II Source separation and equivariance

The notion of equivariance (see for instance Lehman [2]) arises in the following general context. The probability distribution of a random variable $V$ is assumed to belong to a parametric family $P_\theta$, $\theta \in \Omega$ where $\Omega$ is the parameter space. Let $g$ denote a transformation of the sample space. Assume that the transformed variable $gV$ is distributed according to $P_{\theta'}$ for some $\theta' \in \Omega$ whenever $V$ is distributed according to $P_\theta$. This defines a transformation $\bar{g}$ of the parameter space onto itself by setting $\bar{g}\theta = \theta'$. An equivariance principle is invoked when there is a whole group $G$ of transformations $g$ of the sample space. Loosely speaking, this principle states that inference procedures should be consistent with the group structure. In particular, *equivariant* estimators produce estimates which are transformed by $\bar{g}$ when the observations are transformed by $g$ for any $g \in G$.

This section stresses the equivariant structure of source separation which turns out to be a rewarding domain of application for equivariance ideas. It considers batch algorithms; algorithms are considered in section IV.

### A *Equivariance for source separation*

Equivariance, as briefly outlined above, applies straightforwardly to source separation. For any $n \times n$ invertible matrix $M$, denote $g_M$ the operation of multiplying the batch of observations $X_T$ by $M$:

$$g_M X_T \stackrel{\text{def}}{=} MX_T = [M\mathbf{x}_1, \ldots, M\mathbf{x}_T].$$

On the other hand, note that for a fixed distribution of the sources, the probability distribution of the observations depends only on the unknown mixing matrix $A$. Thus, we can identify $\theta$ and $A$ and denote $P_A$ the distribution of $X_T = AS_T$. Because $g_M X_T = MX_T = MAS_T = (MA)S_T$, it is clear that $g_M X_T$ is distributed according to $P_{MA}$. Thus, according to the definition of $\bar{g}$, we have $\bar{g}_M A = MA$.

We find that equivariance has a very simple algebraic expression in our context: the sample space is the set of $n \times T$ data matrices and the parameter space is the set of $n \times n$ invertible matrices. The group $G$ of transformations of interest is left multiplication by invertible $n \times n$ matrices. Since this group acts on the sample $X_T$ and on the parameter $A$ by the same left matrix multiplication, no additional notations are needed: we can identify $g$ and $\bar{g}$ and even forget about the whole group $G$ itself, since its action can be represented by matrix multiplication.

### B *Equivariant estimators and uniform performance*

We have seen that the relevant transformation group for source separation is left multiplication by invertible matrices. Thus, a particular estimator $\mathcal{A}$ of $A$ is said to be *equivariant* if it satisfies

$$\mathcal{A}(MX_T) = M\mathcal{A}(X_T) \tag{2}$$

for any invertible $n \times n$ matrix $M$.

Let $\mathcal{A}$ be an equivariant estimator. For a given realization $S_T = [\mathbf{s}_1, \ldots, \mathbf{s}_T]$ of the source signals, denote $C(S_T)$ the matrix

$$C(S_T) = \mathcal{A}(S_T)^{-1}.$$

Note that $C(S_T)^{-1} = \mathcal{A}(S_T)$ is the estimate of the mixing matrix when the data are $S_T$ *i.e.* when the source signals are *not* mixed. Thus, if $\mathcal{A}$ is a consistent estimator, matrix $\mathcal{A}(S_T)$ is close to the identity matrix for large enough $T$ and consequently so is $C(S_T)$.

Consider applying an equivariant algorithm $\mathcal{A}$ to the mixture $X_T = AS_T$. One has

$$\widehat{A} = \mathcal{A}(X_T) = \mathcal{A}(AS_T) = A\mathcal{A}(S_T) = AC(S_T)^{-1}.$$

Hence we find that the global estimated system is

$$\widehat{C} \stackrel{\text{def}}{=} (\widehat{A})^{-1}A = (AC(S_T)^{-1})^{-1}A = C(S_T).$$

We arrive at the simple but crucial result that the global system $\widehat{C}$ estimated by an equivariant algorithm does not depend on the mixing matrix but only on the particular realization $S_T$ of the source signals via function $C(S_T)$. The estimated source signals are estimated by an equivariant estimator $\mathcal{A}$

$$\widehat{\mathbf{s}}(t) = \widehat{C}\mathbf{s}_t = C(S_T)\mathbf{s}(t)$$

for a particular realization $S_T$. Thus, in terms of signal separation, the performance of an equivariant algorithm *does not depend at all on the mixing matrix.* We

call this property *uniform performance* of batch equivariant estimators. It can be given an even stronger sense in the case of adaptive algorithms (see sec. IV).

## III To be or not to be equivariant ?

So far, we have defined equivariant batch algorithms and shown that they were automatically enjoying uniform performance. The question remains to know if such algorithms exist and how they can be constructed. In this section, we show that several classes of equivariant algorithms do exist. Rather than establishing explicitely property (2), it is sometimes prefered to show the equivalent property that the estimates are such that $\widehat{C} = (\widehat{A})^{-1}A = C(S_T)$ for some function $C(S_T)$ depending on $S_T$ only.

### A Maximum likelihood

Although the ML estimator is often invariant under mild conditions, it is worthwhile to outline the mechanism of equivariance in the source separation context.

Assume that $\{\mathbf{s}_t\}$ is i.i.d. with a marginal probability density, denoted $p_s$. The log-likelihood $\log p(X_T|\tilde{A})$ that mixture $X_T$ has been produced by a mixing matrix $\tilde{A}$ is easily found to be

$$\log p(X_T|\tilde{A}) = \sum_{t=1,T} \log p_s(\tilde{A}^{-1}\mathbf{x}_T) - T\log|\det \tilde{A}|.$$

Define then the random function:

$$l(C, S_T) = \sum_{t=1,T} \log p_s(C\mathbf{s}_t) - T\log|\det C|$$

and denote $C_{ML}(S_T)$ its maximizer in the square matrix $C$:

$$C_{ML}(S_T) = \arg\max_C l(C, S_T)$$

It is readily seen that

$$\log p(X_T|\tilde{A}) = l(\tilde{A}^{-1}A, S_T) - T\log|\det A|.$$

The rightmost term not depending on $\tilde{A}$, the likelihood of the data is maximum for $\tilde{A} = \widehat{A}_{ML}$ verifying $\widehat{A}_{ML}^{-1}A = C_{ML}(S_T)$ by mere definition of $C_{ML}(S_T)$. Thus, we find that $\widehat{C} = C_{ML}(S_T)$, implying that the ML estimator for source separation is equivariant.

### B Contrast functions

The concept of contrast functions was introduced for source separation by Comon [3]. The general idea is to optimize some functional of the (empirical) distribution of $\mathbf{y} = B\mathbf{x}$. An appropriate functional is such that the optimum is reached at $\widehat{B}$ such that $\widehat{B}\mathbf{x}$ is 'close' to $\mathbf{s}$. The estimate of $A$ is $\widehat{A} = (\widehat{B})^{-1}$ or, more to the point, the estimated global system is $\widehat{C} = \widehat{B}A$.

We start here with a simple version. Let $\mathbf{s}$ be a random vector of independent components and let $f$ be a real function of a vector argument such that $\mathrm{E}f(C\mathbf{s})$ is minimized when $C$ is the identity matrix. Define

$$\phi_f(B) = \widehat{\mathrm{E}}f(B\mathbf{x}) = \widehat{\mathrm{E}}f(\mathbf{y}).$$

Here, as in the following, we use $\widehat{\mathrm{E}}$ to denote a sample mean, for instance:

$$\widehat{\mathrm{E}}f(\mathbf{y}) \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1,T} f(\mathbf{y}_t).$$

and any function of $\mathbf{y}$ implicitly is a function of $B$ via $\mathbf{y} = B\mathbf{x}$. Define then:

$$C_f(S_T) \stackrel{\text{def}}{=} \arg\min_C \widehat{\mathrm{E}}f(C\mathbf{s}).$$

By definition of $C_f(S_T)$, functional $\phi_f(B)$ is minimized at $\widehat{B}$ such that $\widehat{C} = \widehat{B}A = C_f(S_T)$. Since $C_f(S_T)$ depends only on $S_T$, minimizing of $\phi_f$ yields an equivariant estimate.

An instance of this type of contrast may be derived from the CMA (Constant Modulus Algorithm) which has been widely studied for blind equalization of communication channels [4]. Its extension to source separation, for binary (real) source signals is to consider

$$f_{CM}(\mathbf{y}) \stackrel{\text{def}}{=} \sum_{i=1,n} (y_i^2 - 1)^2.$$

More generally, one may think of designing contrast functions for source separation as above by summing (over the outputs $y_1, \ldots, y_n$ of a separating matrix) contrast functions designed for blind equalization or deconvolution. Unfortunately, this is not perfectly appropriate because nothing in such a sum of source-wise contrast functions prevents the same signal to be found twice at the outputs of the separator. This may be corrected by incorporating a decorrelation constraint. This is the topic of the next section.

### C Orthogonal contrast functions

Orthogonal contrast functions are —by definition— to be optimized under the constraint that the output of the separating matrix is empirically 'spatially white', *i.e.*

$$\widehat{\mathrm{E}}\mathbf{y}\mathbf{y}^{\mathrm{T}} = \frac{1}{T}Y_T Y_T^{\mathrm{T}} = I. \qquad (3)$$

Approximating the likelihood via Gram-Charlier expansion [5] or the mutual information between outputs [6] led to the following contrast function:

$$\phi_{CL}(B) \stackrel{\text{def}}{=} \sum_{i,j,k,l\neq i,i,i,i} |\widehat{\mathrm{Cum}}(y_i, y_i, y_i, y_i)|^2$$

to be minimized under constraint (3). Here, $\widehat{\mathrm{Cum}}$ denotes empirical cumulants.

Based on another approach (investigating the eigenstructure of the cumulant tensor), a similar orthogonal contrast is arrived at in [7]:

$$\phi_{CS}(B) \stackrel{\text{def}}{=} \sum_{i,k,l=1,n} |\widehat{\text{Cum}}(y_i, y_i, y_k, y_l)|^2$$

with the advantage that it can be efficiently optimized[1] while offering the same asymptotic performance as $\phi_{CL}$. When the source signals have kurtosis $\kappa_i = \text{Cum}(s_i, s_i, s_i, s_i)$ such that $\kappa_i + \kappa_j < 0$, then the following much simpler contrast

$$\phi_{M4}(B) \stackrel{\text{def}}{=} \widehat{\text{E}} \sum_{i=1,n} y_i^4$$

may be considered.

Other orthogonal contrast functions can be used for source separation. For any orthogonal contrast function $\phi$, such as $\phi_{CL}$, $\phi_{CS}$, $\phi_{M4}$, a matrix $C_\phi(S_T)$ is defined as

$$\begin{cases} C_\phi(S_T) = \arg\max_C \phi(CS_T) \\ \text{subject to } CS_T S_T^{\text{T}} C^{\text{T}} = T \end{cases}$$

and reasoning as in previous section, it is found that $\widehat{C} = \widehat{B}A = C_\phi(S_T)$, *i.e.* equivariance is granted. We stress that equivariance results from two facts: (i) contrasts are functions of the output $\mathbf{y}$ of the separator (ii) they are optimized under constraint (3) which also depends only on the output.

### D *Estimating equations*

We have considered so far maximum likelihood and contrast-based estimation where estimates are the optimizers of some functional. Let us then express the stationarity condition (cancelation of the gradient). For the ML estimate and simple contrasts like $\phi_f$ or $\phi_{M4}$, this occurs at points $\widehat{B}$ such that $\widehat{C} = \widehat{B}A$ verifies

$$\widehat{\text{E}} H(\widehat{C}\mathbf{s}) = 0 \qquad (4)$$

where $H(\cdot)$ is a vector-to-matrix mapping. For the ML estimator, mapping $H(\cdot)$ is easily computed to be

$$H_{ML}(\mathbf{y}) = \mathbf{l}(\mathbf{y})\mathbf{y}^T - I \qquad (5)$$

where vector $\mathbf{l}(\mathbf{y})$ is the gradient of $-\log p_s$, evaluated at $\mathbf{y}$. Optimizing $\phi_{M4}$ under constraint (3), some calculations show that the corresponding $H(\cdot)$ function is:

$$H_{M4}(\mathbf{y}) = \mathbf{y}\mathbf{y}^{\text{T}} - I + \mathbf{c}(\mathbf{y})\mathbf{y}^{\text{T}} - \mathbf{y}\mathbf{c}(\mathbf{y})^{\text{T}}$$

where the $i$-th component of $\mathbf{c}(\mathbf{y})$ is $y_i^3$. In both cases, $H(\cdot)$ is such that $\text{E}H(\mathbf{s}) = 0$. Indeed, the *estimating equation* $\widehat{\text{E}} H(\mathbf{y}) = 0$ (where $\mathbf{y} = \widehat{B}\mathbf{x} = \widehat{C}\mathbf{s}$) is a sample counterpart of $\text{E}H(\mathbf{s}) = 0$, where expectation is replaced by sample averaging and the 'true' sources are replaced by their estimates, namely: the output $\mathbf{y}$ of the separator.

As a matter of fact, it is not difficult to find vector-to-matrix functions $H(\cdot)$ such that $\text{E}H(\mathbf{s}) = 0$. For instance replacing $\mathbf{c}$ in $H_{M4}$ by any other component-wise non-linearity ensures that $\text{E}H(\mathbf{s}) = 0$. What is more difficult is to find $H$ such that resulting algorithms show good performance. In any case, whatever this performance may be, it is uniform in the mixing matrix $A$ since, proceeding as usual, we define $C_H(S_T)$ such that $\widehat{E}H(C_H(S_T)\mathbf{s}) = 0$. Then a solution of $\widehat{\text{E}} H(\mathbf{y}) = 0$ is such that the global estimated system is $\widehat{C} = C_H(S_T)$, not depending on $A$.

### E *Equivariant... not !*

In view of the previous examples, one may wonder if there exists any decent batch algorithm for source separation that would *not* be equivariant! Ironically enough, there is an apparently irrelevant technical detail that makes it easy to get into trouble with respect to equivariance.

Recall that there is an inherent indetermination in source separation: the amplitude of a particular source signal and the amplitude of the corresponding column of $A$ both affect the observations $X_T$ in the same way and thus cannot be distinguished. To fix this ambiguity, one may adopt the convention (or assumption) $A4$ that source signals have unit variance. Alternatively, one may for instance normalize $B$ by taking, as is often done, its diagonal elements equal to 1.

Constraining the separating matrix will generally lead to algorithms whose behavior depends strongly and unpredictably on the mixing matrix. Assume for instance that the problem is normalized by constraining the separating matrix $B$ to have its diagonal entries equal to 1. It is simple to find matrices $A$ such that $BA = \Lambda$ is a diagonal matrix only if $\Lambda$ has very large elements. For $n = 2$ for instance, take $A = \begin{bmatrix} \epsilon & 1 \\ 1 & \epsilon \end{bmatrix}$. Then $\Lambda = \lambda I$ with $\lambda = \epsilon - \epsilon^{-1}$. The result is an over-amplification of the output if $\epsilon$ is small and small variations of $\epsilon$ induce large variations of $\lambda$. This behavior must of course be avoided for the sake of robustness, especially for adaptive algorithms which are more difficult to tune.

## IV Equivariant adaptive source separation

This section addresses *adaptive* source separation. First, the idea of 'serial updating' is introduced as the adaptation rule consistent with equivariance. In particular it offers uniform performance in a strong sense, as shown and discussed in a second part. We introduce next specific adaptive source separation algorithms.

[1] Optimization of $\phi_{CS}$ is shown to be equivalent to joint diagonalization of cumulant matrices. A Matlab demo is available at http://sig.enst.fr/~cardoso/stuff.html or upon request from the author.

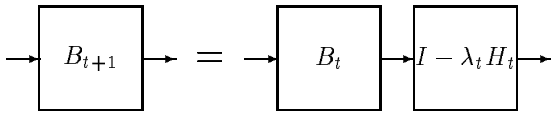
*A Serial update algorithms*

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

$$B_{t+1} = B_t - \lambda_t H(\mathbf{y}_t) B_t \tag{6}$$

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

The adaptation rule (6) is termed a 'serial update', because it reads equivalently $B_{t+1} = (I - \lambda_t H(\mathbf{y}_t))B_t$. This latter 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 $B_{t+1}$.




Serial update is multiplicative

Recall that the transformation group relevant to source separation is left matrix multiplication and note that system $B_t$ is serially updated in (6) by *left multiplication* by matrix $I - \lambda_t H(\mathbf{y}_t)$, depending only on the current output $\mathbf{y}_t$. In this sense, serial updating is consistent with equivariance.

The *stationary points* of serial update algorithms are these matrices $B_*$ such that the mean value of the update is zero when $B_t$ is kept at the fixed value $B_*$. The corresponding value of the global system is $C_* = B_* A$ and the stationarity condition then reads:

$$\mathrm{E} H(C_* \mathbf{s}) = 0$$

which is to be compared to (4), indeed.

*B Adaptive separation and uniform performance*

Consider the global mixing-unmixing system $C_t = B_t A$. Its evolution under the updating rule (6) is readily obtained by right multiplication of (6) by matrix $A$, yielding

$$C_{t+1} = C_t - \lambda_t H(C_t \mathbf{s}_t) C_t \tag{7}$$

where we used $\mathbf{y}_t = B\mathbf{x}_t = BA\mathbf{s}_t = C\mathbf{s}_t$. This is a trivial but remarkable result because it means that, under serial updating, the evolution law of the global system $C_t$ is independent of the mixing matrix $A$. Note that the argument parallels those of 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 equation (7), the subsequent trajectory $\{C_t | t > 1\}$ of the global system will be *strictly identical* to the trajectory that would be observed for another mixing matrix $A'$, provided the initial point is $B'_o = B_o A A'^{-1}$. This is pretty obvious since in both cases, the *global* system starts from the same initial condition $C_0$ and evolves according to (7) 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.

From this, it follows in particular that it is only needed to study the convergence of $C_t$ to the identity matrix under the stochastic rule (7) to completely characterize a serial source separation algorithm.

*C Relative gradient algorithms*

As in previous section, we have exhibited a general class of algorithms with uniform performance, before describing specific algorithms. In this section, we show that the notion of 'relative gradient' (see below) is instrumental in designing serial update adaptive algorithms, *i.e.* in choosing a specific function $H(\cdot)$.

The stationarity condition $\mathrm{E} H(C_* \mathbf{s}) = 0$ for adaptive algorithms is the counterpart of the estimating equation $\widehat{\mathrm{E}} H(\widehat{C}\mathbf{s}) = 0$ of batch algorithms. Such equations may be arrived at by expressing the stationarity of an objective function as seen in previous section. Not surprisingly, the most appropriate definition of gradient for this purpose turns out to be related to the invariance group of source separation, namely left matrix multiplication. It goes as follows.

Let $\phi(B)$ be a real differentiable function of a $n \times n$ matrix. The *relative gradient* of $\phi$ at $B$ is the matrix, denoted $\nabla\phi(B)$, verifying

$$\phi(B + \mathcal{E}B) = \phi(B) + < \nabla\phi(B) | \mathcal{E} > + o(\mathcal{E}) \tag{8}$$

for any square matrix $\mathcal{E}$. Here $< \cdot | \cdot >$ denotes the Euclidean scalar product of matrices, *i.e.* $< M | N >= \mathrm{Trace}(M^{\mathrm{T}} N)$.

This is indeed the gradient to use in a serial update: a relative gradient algorithm for minimizing $\phi$ consists in modifying $B$ into $B + \mathcal{E}B$ so that $\phi$ decreases. Take then $\mathcal{E} = -\lambda \nabla\phi(B)$: according to (8), the variation of $\phi$ is $< \nabla\phi | - \lambda \nabla\phi > + o(\lambda) = -\lambda ||\nabla\phi||^2 + o(\lambda)$ which is negative for small enough positive $\lambda$ (recall that $< M | M >= ||M||^2$, the squared Frobenius norm).

*D Stochastic relative gradient algorithms*

When the relative gradient of a contrast function $\phi(B)$ takes the form $\nabla\phi(B) = \mathrm{E} H(\mathbf{y})$ then an stochastic relative gradient (SRG) algorithm is obtained by deleting the expectation operator and updating $B$ into $B - H(\mathbf{y})B$. Indeed, this is the form of the serial update algorithm (6).

Note though that not any serial update algorithm is a stochastic relative gradient because mapping $H(\cdot)$ in (6) is not necessarily required to be the relative gradient of some objective function. Conversely not any stochastic relative algorithm is a serial update because we have requested (with uniform performance in view)

that $H$ should depend on $\mathbf{y}$ only. However, we will see below that this is generally the case, by considering two specific forms for function $H$, already met in section II.

**Likelihood.** Consider the likelihood function under the assumptions of section A. Its relative gradient (with opposite sign), identifying $\tilde{A}$ and $B^{-1}$, is $\widehat{E}H_{ML}(\mathbf{y})$ where $H_{ML}$ is given in (5). Thus, if the distribution of the sources is known, one obtains a maximum likelihood (ML) stochastic relative gradient (SRG) algorithm by using function $H_{ML}$ in (6). See the excellent paper [8] for a related batch ML algorithm. In particular, a smart approximation to the score function $\mathbf{l}(\mathbf{y})$ is proposed and studied therein.

**Orthogonal contrasts.** When orthogonal contrasts are used, function $H(\cdot)$ takes a special form: the symmetric part of $H(\mathbf{y})$ is $\mathbf{y}\mathbf{y}^T - I$. Indeed, the mean value of this part is zero when the output of the separator is white. We advocate the use of $H(\cdot)$ functions with the basic form

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

where $\mathbf{g}$ is some component-wise non-linear function. This family of algorithms is called EASI (equivariant adaptive separation via independence) even though, this is not the most general form of equivariant algorithms.

**Stabilization.** Most adaptive algorithms require to be stabilized. In the case of equivariant adaptive algorithms, stabilization procedure should not be based on clipping the entries of $B$ re-normalizing its rows; in fact, stabilization should *not* involve any action on $B$ itself, because this would spoil the uniform performance property. We rather suggest consider a modified form of $H$:

$$H_g(\mathbf{y}) = \frac{\mathbf{y}\mathbf{y}^{\mathrm{T}} - I}{1 + \lambda_t \ \mathbf{y}^{\mathrm{T}}\mathbf{y}} + \frac{\mathbf{g}(\mathbf{y})\mathbf{y}^{\mathrm{T}} - \mathbf{y}\mathbf{g}(\mathbf{y})^{\mathrm{T}}}{1 + \lambda_t \ |\mathbf{y}^{\mathrm{T}}\mathbf{g}(\mathbf{y})|}$$

which prevents $H(\cdot)$ to take too large values. This algorithm is studied in greater details (including asymptotic performance analysis) in [9, 10].

## V Discussion and Conclusion

We have seen that it exists a natural transformation group relevant to the source separation problem: left multiplication by invertible matrices. This equivariance structure entails a remarkable property: a large class of source separation algorithms perform uniformly well with respect to the mixture. This is a very desirable property, because it allows the performance of a particular algorithm to be predicted independently of the (un)observed mixture.

Key ingredients to the design of an equivariant batch algorithm are: *unconstrained* optimization of the separating matrix and optimization of objective functions (or resolution of estimating equations) depending only on the distribution of the outputs of the separator. In order to design equivariant *adaptive* algorithms, it is further necessary to adopt the serial update approach: the separating matrix is to updated by left multiplcation with a matrix close to the identity and depending on the outputs only. While this paper was in preparation, Pr Cichocki draw our attention to his algorithm [11] which is a serial updater and does achieve uniform performance.

A striking feature of uniform performance is that separation is as good for 'unmixed' signals as it is for arbitrarily bad mixtures. Of course, there is a trick here: when $A$ is ill conditioned, additive noise can no longer be neglected because it is amplified in the inversion of a poorly conditioned mixing matrix. This is main limit of the equivariant approach. This effect has been studied in [12].

## References


[1] 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), pp. 1017–1020, 1985.

[2] E. L. Lehmann, *Testing statistical hypothesis*. Wiley pub. in statistics, John Wiley, 1959.

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

[4] R. Gooch and J. Lundell, "The CM array: An adaptive beamformer for constant modulus signals," in *Proc. ICASSP*, 1986.

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

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

[7] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *IEE Proceedings-F*, vol. 140, pp. 362–370, Dec. 1993.

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

[9] B. Laheld and J.-F. Cardoso, "Adaptive source separation without prewhitening," in *Proc. EUSIPCO*, (Edinburgh), pp. 183–186, Sept. 1994.

[10] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation.." submitted to the IEEE Trans. on S.P., 1994.

[11] A. Cichocki, R. Unbehauen, and E. Rummert, "Robust learning algorithm for blind separation of signals," *Electronic letters*, vol. 30, no. 17, pp. 1386–87, 1994.

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