

Adaptive source separation with uniform performance

Beate Laheld and Jean-François Cardoso

Télécom Paris, 46 rue Barrault, 75634 Paris Cedex 13, France.
E-mail: cardoso@sig.enst.fr, Fax: (33) 1 45 88 79 35.

Abstract. This paper presents a family of adaptive algorithms for the blind separation of independent signals. Source separation consists in recovering a set of independent signals from some linear mixtures of them, the coefficients of the mixtures being unknown. In the noiseless case, the ‘hardness’ of the blind source separation problem does not depend on the mixing matrix (see the companion paper [1]). It is then reasonable to expect adaptive algorithms to exhibit convergence and stability properties that would also be independent of the mixing matrix. We show that this desirable *uniform performance* feature is simply achieved by considering ‘serial updating’ of the separating matrix. Next, generalizing from the gradient of a standard cumulant-based contrast function, we present a family of adaptive algorithms called ‘PFS’, based on the idea of serial updating. The stability condition and the theoretical asymptotic separation levels are given in closed form and, as expected, depend only on the distributions of the sources. Performance is also illustrated by some numerical experiments.

1. Blind source separation

Consider an array of m sensors receiving signals emitted by n statistically independent sources. The array output at time t is a $m \times 1$ vector $\mathbf{x}(t)$ modelled as

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

where the $m \times n$ matrix A is called the ‘mixing’ matrix and where the n source signals are collected in a $n \times 1$ vector denoted $\mathbf{s}(t)$. In the complex case, this is the familiar linear model used in narrow band array processing. Adaptive source separation consists in updating a $n \times m$ matrix $B(t)$, called the *separating matrix*, such that

$$\mathbf{y}(t) \stackrel{\text{def}}{=} B(t) \mathbf{x}(t) \quad (2)$$

is an estimate of the source signals, hence the terminology ‘source separation’.

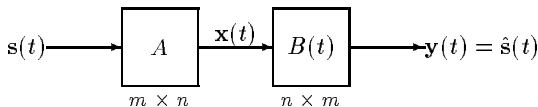


Figure 1. Blind source separation model

We emphasize that $B(t)$ should be updated *without resorting to any information about the spatial mixing matrix* A . This is in sharp contrast to ‘standard’ array processing and beamforming techniques where the columns of A or their dependence on the location of the sources is assumed to be known. Matrix A is supposed to be a fixed matrix with full column rank but no other assumptions are made. The crucial property source separation relies on is the mutual statistical independence of the source signals.

Note that under these assumptions, the $n \times n$ global system matrix

$$C(t) \stackrel{\text{def}}{=} B(t) A \quad (3)$$

may only be identified up to the product of a permutation and a diagonal matrix. This is because without a priori information on the amplitude of the source signals nor on matrix A , the scale of each source signal is essentially unobservable. The permutation indetermination stems from the fact that the labelling of the source signals is immaterial. Hence, without any loss of generality we take the convention that the source signals have unit variance, and the algorithms presented below are designed to yield unit variance outputs.

The following assumptions hold throughout the paper:

- H1:** Matrix A is full column rank, implying $n \leq m$.
- H2:** At each t , the components of $\mathbf{s}(t)$ are independent.
- H3:** Each component of $\mathbf{s}(t)$ is a stationary zero-mean process with non Gaussian marginal distribution.

The various techniques proposed to solve this spatial blind deconvolution problem differ in the way the source independence hypothesis is exploited. Block techniques based on 2nd- and 4th-order cumulants are proposed in [2, 3, 4, 5]. See also [6] for a (quasi)-likelihood approach. If the source signals are temporally correlated, separation may be based on second order statistics only [7, 8]. Adaptive algorithms may also be based either on optimization of cumulant-based contrast function [9, 10, 11] or on ‘estimating equations’ involving non-linear distortions of the output \mathbf{y} [9, 12, 13].

Outline. In this paper, we present a family of ‘serial’ adaptive algorithms for source separation. In section 2, we introduce the notion of ‘serial updating’. This is specific to the problem of updating a *matrix*. A key property of this approach is that it automatically yields algorithms whose performance *does not depend* on mixing matrix A . In section 3, we compute the serial updating rule which derives from a standard cumulant-based contrast function. This is then generalized to yield a family of serial source separa-

tion algorithms (section 4). The resulting performance is investigated by computing the asymptotic rejection rates and by numerical experiments in section 5.

For ease of exposition, we derive the main result in the real case (extension to the complex case is straightforward, as indicated in section 4). The following notational conventions hold throughout: scalars are denoted in lowercase, vectors in bold lowercase and matrices in uppercase. The superscript ‘t’ denotes transposition. The Euclidian scalar product between matrices M and N is $\langle M|N \rangle = \text{Trace}(NM^t)$.

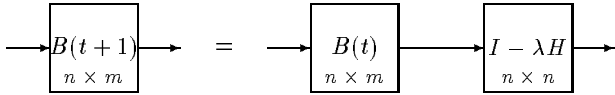
2. Serial update for uniform performance

2.1. Serial update

We consider adapting rules in a *serial form*:

$$B(t+1) = \{ I - \lambda H(\mathbf{y}(t)) \} B(t) \quad (4)$$

where matrix H is a function of the output \mathbf{y} only. The update (4) is called ‘serial’ since it may be interpreted as *chaining* a system close to the identity at the output of $B(t)$.



By right multiplication of (4) with the mixing matrix A , the evolution of the global system $C(t) = B(t)A$ is:

$$C(t+1) = \{ I - \lambda H(C(t)\mathbf{s}(t)) \} C(t). \quad (5)$$

This trivial mathematical property has an important practical consequence: equation (5) which describes the evolution of $C(t)$ depends only on the source signals. Hence the dynamics of the global system $C(t)$ do not depend on the mixing matrix.

This is a crucial feature since $\hat{\mathbf{s}}(t) = C(t)\mathbf{s}(t)$, i.e. the quality of the source separation, is itself determined by $C(t)$ only: some of its entries (one per row and per column) should converge to unit norm scalars while the others should be as small as possible. It must be also noted that *changing the mixing matrix A is equivalent to changing the initial point of the algorithm*.

Note finally that by the same argument, another pleasant practical consequence of the uniformity property is that, in the noiseless case, the study of the algorithm is exhausted by the study of its behaviour for $A = I$.

2.2. Estimating equations

The adaptive algorithm (4) may be seen as a stochastic approximation device to solve the equation:

$$EH(\mathbf{y}(t)) = 0, \quad (6)$$

since any full rank matrix B such that the distribution of $\mathbf{y}(t)$ satisfies eq. (6) is seen to be a stationary point of the algorithm (the *stability* is discussed below).

As an illustration and for later use, we first consider the case where matrix B should be updated such as to minimize an objective function $c(B)$ in the form:

$$c(B) \stackrel{\text{def}}{=} E\phi(\mathbf{y}) = E\phi(B\mathbf{x}), \quad (7)$$

where ϕ is differentiable function. Denoting $\phi'(\mathbf{y})$ the $m \times 1$ vector whose i -th component is $\frac{\partial \phi}{\partial y_i}(\mathbf{y})$, we have

$$\phi(\mathbf{y} + \delta\mathbf{y}) = \phi(\mathbf{y}) + \phi'(\mathbf{y})^t \delta\mathbf{y} + o(\delta\mathbf{y}), \quad (8)$$

and it is easily found that

$$\begin{aligned} c((I + \mathcal{E})B) &= c(B) + E\phi'(\mathbf{y})^t \mathcal{E}\mathbf{y} + o(\delta\mathcal{E}) \\ &= c(B) + \langle E\phi'(\mathbf{y})\mathbf{y}^t | \mathcal{E} \rangle + o(\delta\mathcal{E}). \end{aligned} \quad (9)$$

A gradient algorithm for minimizing $c(B)$ consists in updating B into $(I + \mathcal{E})B$, with \mathcal{E} proportional to $-\lambda E\phi'(\mathbf{y})\mathbf{y}^t$ where λ is a small positive number in order to ensure $c(B_{n+1}) = c((I + \mathcal{E})B_n) < c(B_n)$. The *stochastic* version of this algorithm amounts to dropping the expectation operator. This is just the updating rule (4) with $H(\mathbf{y}) = \phi'(\mathbf{y})\mathbf{y}^t$.

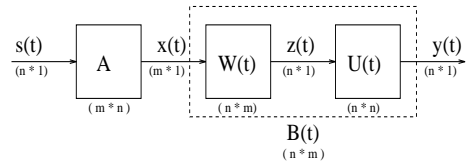
Regarding the source separation problem, we base our approach on contrast functions which are not in the form (7) so that we cannot implement directly the above result. It is nonetheless possible to exhibit a family of functions $H(\mathbf{y})$ such that separating matrices are stable stationary points of the algorithm (4) as shown next.

3. Serial updates for an orthogonal contrast function

As a first step in deriving general serial source separation algorithms, we derive a specific vector-to-matrix mapping $H(\mathbf{y})$ stemming from an ‘orthogonal’ contrast function. The following constrained optimization problem

$$\text{Minimize } \Phi \stackrel{\text{def}}{=} \sum_{i=1,n} E|y_i|^4 \quad \text{subject to } E\mathbf{y}\mathbf{y}^t = I \quad (10)$$

may be used to separate sources with negative kurtosis. It may be solved in a two step approach as in the figure below. First the $m \times 1$ observed vector $\mathbf{x}(t)$ is whitened



into an intermediate $n \times 1$ vector \mathbf{z} by means of a $n \times m$ whitening matrix W , i.e. $\mathbf{z} = W\mathbf{x}$ and $R_z = E\mathbf{z}\mathbf{z}^t = I$. In a second step, the spatially white vector \mathbf{z} is rotated by a unitary matrix U into the final output: $\mathbf{y} = U\mathbf{z}$. Matrix U is computed so as to minimize $\sum E|y_i|^4$, the constraint $E\mathbf{y}\mathbf{y}^t = I$ being satisfied by constraining U to remain unitary.

We show below how the two steps on W and on U can be implemented in a serial adaptive fashion and how they combine into a single adapting rule for $B(t) = U(t)W(t)$.

3.1. Whitening stage

Consider the following objective function

$$\Upsilon(W) \stackrel{\text{def}}{=} \text{Trace}(R_z) - \log \det(R_z) - n \quad (11)$$

which defines a ‘distance’ from R_z to the identity matrix and actually is a function of W since $\mathbf{z} = W\mathbf{x}$. This function admits the first order expansion

$$\Upsilon(W + \mathcal{E}W) = \Upsilon(W) + 2 \langle E\mathbf{z}\mathbf{z}^t - I | \mathcal{E} \rangle + o(\mathcal{E}). \quad (12)$$

Hence, even though the objective function Υ is not in the form of (7), the relative variation (12) of Υ is in the form of (9), suggesting the serial whitening algorithm:

$$W(t+1) = \{ I - \lambda(\mathbf{z}(t)\mathbf{z}(t)^t - I) \} W(t). \quad (13)$$

3.2. Orthogonal stage

The orthogonal stage attempts to optimize (10) which is in the form of (7) with $\phi(\mathbf{y}) = \sum_{i=1,n} |y_i|^4$ and Φ now a function of U since $\mathbf{y} = U\mathbf{z}$. Similar to (9), we have

$$\Phi(U + \mathcal{E}U) = \Phi(U) + \langle E\phi'(\mathbf{y})\mathbf{y}^t | \mathcal{E} \rangle + o(\mathcal{E}). \quad (14)$$

Since Φ must be minimized *under unitary constraint*, matrix U should not be updated into $(I + \mathcal{E})U$ with $\mathcal{E} = -\lambda\phi'(\mathbf{y})\mathbf{y}^t$, since this would destroy the orthogonality condition $UU^t = I$. We note that, if U is orthogonal,

$$(U + \mathcal{E}U)(U + \mathcal{E}U)^t = I + \mathcal{E} + \mathcal{E}^t + o(\mathcal{E}), \quad (15)$$

showing that orthogonality of $U + \mathcal{E}U$ is preserved at first order in \mathcal{E} whenever \mathcal{E} is skew-symmetric ($\mathcal{E} = -\mathcal{E}^t$). This suggests to align \mathcal{E} on the *skew-symmetric* part of $\phi'(\mathbf{y})\mathbf{y}^t$ and to update U as:

$$U(t+1) = \{ I - \lambda[\phi'(\mathbf{y}(t))\mathbf{y}(t)^t - \mathbf{y}(t)\phi'(\mathbf{y}(t))^t] \} U(t) \quad (16)$$

where λ is a small positive number. Of course, such an updating rule does not preserve *exactly* orthogonality but this problem disappears when the whitening stage and the orthogonal stage are considered altogether.

3.3. A unique adapting rule

Consider the matrix $B(t)$ obtained by concatenation : $B(t) = U(t)W(t)$. Updating $W(t)$ and $U(t)$ according to (13) and (16) and neglecting the terms in λ^2 results in updating $B(t)$ as in (4) with

$$H(\mathbf{y}) = \mathbf{y}\mathbf{y}^t - I + \phi'(\mathbf{y})\mathbf{y}^t - \mathbf{y}\phi'(\mathbf{y})^t \quad (17)$$

It is easily checked that if B solves the source separation problem (i.e., is such that BA is the identity matrix or any signed permutation), then $EH(\mathbf{y}(t)) = 0$ so that B is stationary point of (4).

4. A family of source separation algorithm

The result (17) of section 3 is generalized by noting that if $C = BA = I$ then $EH(\mathbf{y}(t)) = 0$ whenever ϕ' is replaced by any component-wise non-linear function \mathbf{g} . We thus obtain a family of adaptive source separation algorithms which we call PFS (Parameter free separators) to emphasize the fact that matrix B is updated without imposing any constraint on it. As hinted in section 2, this is one of the ingredients for uniform performance.

A specific algorithm then depends on the choice of some component-wise non-linear function \mathbf{g} . The generic algorithm is summarized by

$$g\text{-PFS} \begin{cases} B(t+1) &= \{ I - \lambda H(\mathbf{y}(t)) \} B(t) \\ H(\mathbf{y}) &= \mathbf{y}\mathbf{y}^t - I + \mathbf{g}(\mathbf{y})\mathbf{y}^t - \mathbf{y}\mathbf{g}(\mathbf{y})^t \\ (\mathbf{g}(\mathbf{y}))_i &= g_i(y_i) \end{cases} \quad (18)$$

The asymptotic ($\lambda \ll 1$) stability of the algorithm at point B_* such that $B_*A = I$ depends on the moments:

$$k_i^g \stackrel{\text{def}}{=} E s_i g_i(s_i) - E g_i'(s_i) E s_i^2 \quad i = 1, \dots, n \quad (19)$$

where $g_i'(s_i)$ is the derivative of g_i , the i -th component function of \mathbf{g} . The n^2 eigenvalues of the derivatives of the mean field are easily computed in the i.i.d. case. We find $n(n+1)/2$ eigenvalues equal to -2 and $n(n-1)/2$ equal to $k_i^g + k_j^g$ for each pair (i, j) of sources. Hence B_* is an attractor¹ iff

$$\forall i \neq j \quad k_i^g + k_j^g < 0. \quad (20)$$

If $g_i(x) = x^3$, these moments are the 4th-order cumulants of the source signals. Hence, for cubic non-linearities, stability requires only (20) which is weaker than the condition that all sources have negative kurtosis (this would be : $\forall i \quad k_i^g < 0$). In particular, one Gaussian source at most can be handled since for a normal distribution, $k_i^g = 0$ for any g .

In practice, algorithms are run with non vanishing adaptation steps; since $B(t)$ is adapted without constraint, explosive behaviours may be observed, especially if the non-linearities in \mathbf{g} are strongly increasing functions (like a cubic distortion for instance). Hence, it is recommended to implement a robustified version of PFS. An *ad hoc* stabilization is obtained by modifying $H(\mathbf{y})$ into

$$H(\mathbf{y}) = \frac{\mathbf{y}\mathbf{y}^t - I}{1 + \lambda |\mathbf{y}^t \mathbf{y}|} + \frac{\mathbf{g}(\mathbf{y})\mathbf{y}^t - \mathbf{y}\mathbf{g}(\mathbf{y})^t}{1 + \lambda |\mathbf{y}^t \mathbf{g}(\mathbf{y})|} \quad (21)$$

resulting in the so-called **SPFS** (stabilized PFS) family. This modification preserves the asymptotic stability of B_* and have proved very effective in numerical experiments.

The complex case. Extension to the complex case is straightforward when the source signals are circularly distributed. The algorithm is modified by i) understanding the transposition as transpose-conjugation and ii) using phase preserving non-linearities i.e. $g_i(\mathbf{y}) = f_i(|y_i|^2)y_i$ where each f_i is a real valued function. The stability conditions is like (20) but now depends on the moments :

$$k_i^f \stackrel{\text{def}}{=} E |s_i|^2 f_i(|s_i|^2) - E |s_i|^2 E f_i(|s_i|^2) - E |s_i|^2 E |s_i|^2 f_i'(|s_i|^2)$$

which, again, are the 4th-order cumulants in the complex circular case when g_i is cubic, i.e. when $f_i(x) = x$.

5. Performance

We quantify the performance in terms of ISI (intersymbol interference). Since our convention is that the source signals have unit variance, the residual energy of the j -th source signal in the estimate of the j -th source signal is $|(BA)_{ij}|^2$. If BA is close to the identity, we have $|(BA)_{ii}|^2 \approx 1$, so that the relative power of the j -to- i interference is just given by $|(BA)_{ij}|^2$.

5.1. Asymptotic performance

We have computed the mean j -to- i ISI measured as

$$\rho_{ij} \stackrel{\text{def}}{=} E |(BA)_{ij}|^2 \quad (22)$$

¹The result reported in [13] is incorrect.

after convergence in the limit $\lambda \ll 1$. In the real case with identical sources and identical non-linearities $g_i(s_i) = g(s_i)$, so that $k_i^g = k^g$, we get $\rho_{ij} = \rho$ for all $i \neq j$ with

$$\rho = \lambda \left[\frac{1}{4} + \frac{1}{2} \frac{E s_i^2 E g^2(s_i) - E^2 s_i g(s_i)}{E s_i^2 E g'(s_i) - E s_i g(s_i)} \right]. \quad (23)$$

Both the numerator and the denominator in the last term are positive by the Cauchy-Schwarz inequality and by the stability condition respectively. Hence, the performance is lower bounded by $\rho \geq \lambda/4$, regardless of the non-linearities in g . See [1] for a similar result on block processing algorithms. Note that this bound is reached if $|s_i(t)| = 1$ and g is an odd function.

5.2. Simulation results

The simulations show a good behaviour of the PFS and the SPFS algorithm. In particular, we did not observe spurious attractors, in the general case of n sources and monotonous non-linearities.

Figures 2 and 3 are for $m = 4$ sensors and $n = 3$ QAM4 modulated sources, $g_i(\mathbf{y}) = |y_i|^2 y_i$, $\lambda = 0.01$ and $\text{cond}(A) \approx 13$. Convergence is attained after about 300 samples in fig. 2 which displays the evolution of $|(B(t)A)_{ij}|$ for all i and j . Figure 3 shows the first 100 and the last 100 estimated source signals at each output.

Figures 4 and 5 show that *one* Gaussian source is allowed. Fives sources (2 QAM4, 1 QAM16, 1 PSK, 1 Gaussian) are recovered with $\lambda = 0.005$ and $\text{cond}(A) \approx 5.n$

Conclusion

Based on the idea of ‘serial updating’, a family of a adaptive source separation algorithms has been presented. They show, at low noise, a behavior which is independent of the mixing matrix. Stability conditions and the residual MSE are given in closed form, making it possible to adapt the non-linearities to the source distributions in order to gain isotropic convergence and unconditional stability.

References

- [1] J.-F. Cardoso, “On the performance of source separation algorithms,” *these proceedings*.
- [2] M. Gaeta and J.-L. Lacoume, “Source separation without a priori knowledge: the maximum likelihood solution,” in *Proc. EUSIPCO*, pp. 621–624, 1990.
- [3] P. Comon, “Independent component analysis,” in *Proc. Int. Workshop on Higher-Order Stat., Chamrousse, France*, pp. 111–120, 1991.
- [4] G. Giannakis and S. Shamsunder, “Modelling of non Gaussian array data using cumulants. DOA estimation with less sensors than sources,” in *Proc. Conf. Info. Sci. Syst., Baltimore*, 1991.
- [5] J.-F. Cardoso and A. Souloumiac, “Blind beamforming for non Gaussian signals,” *IEE Proceedings-F*, vol. 140, pp. 362–370, Dec. 1993.
- [6] D. T. Pham, P. Garrat, C. Jutten, “Separation of a mixture of independent sources through a ML approach,” in *Proc. EUSIPCO*, pp. 771–774, 1992.
- [7] L. Tong, R. Liu, “Blind estimation of correlated source signals,” in *Proc. Asilomar Conf.*, Nov. 1990.

- [8] K. Abed Meraim, A. Belouchrani, J.-F. Cardoso, and E. Moulines, “Asymptotic performance of second order blind source separation,” in *Proc. ICASSP*, vol. 4, pp. 277–280, Apr. 1994.
- [9] C. Jutten and J. Héroult, “Blind separation of sources: an adaptive algorithm based on neuromimetic architecture,” *Signal Processing*, vol. 24, pp. 1–10, 1991.
- [10] N. Delfosse and P. Loubaton, “Adaptive separation of independent sources: a deflation approach,” in *Proc. ICASSP*, 1994.
- [11] E. Moreau and O. Macchi, “A one stage self-adaptive algorithm for source separation,” in *Proc. ICASSP*, (Adelaide, Australia.), 1994.
- [12] A. Dinç and Y. Bar-Ness, “Bootstrap: A fast blind adaptive signal separator,” in *Proc. ICASSP*, vol. 2, pp. 325–328, 1992.
- [13] J.-F. Cardoso, A. Belouchrani, and B. Laheld, “A new composite criterion for adaptive and iterative blind source separation,” in *Proc. ICASSP*, vol. 4, pp. 273–276, Apr. 1994.

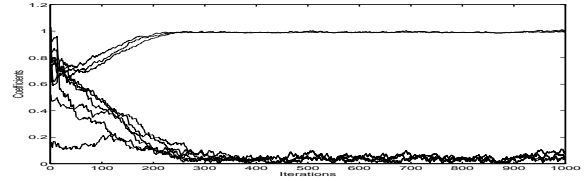


Fig. 2 : Convergence to 0 or 1 of $|(B(t)A)_{ij}|$

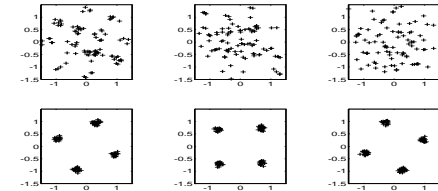


Fig. 3 : The first 100 and last 100 outputs $\mathbf{y}(t)$

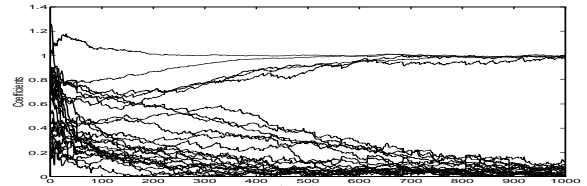


Fig. 4 : Convergence to 0 or 1 of $|(B(t)A)_{ij}|$

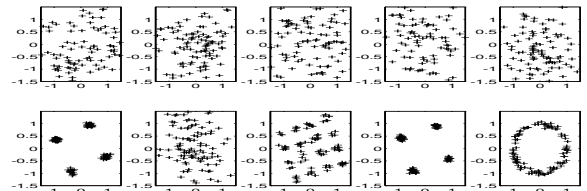


Fig. 5 : The first 100 and last 100 outputs $\mathbf{y}(t)$