Equivariant Adaptive Source Separation

Arindam Bose

***Abstract*—This report discusses a family of adaptive algorithms for the blind separation of independent signals. Source separation consists in recovering a set of independent signals when only some linear mixtures with unknown coefficients are observed. The 'hardness' of the blind source separation problem does not depend on the mixing matrix in particularly in the noiseless case [1]. It is reasonable to expect adaptive algorithms to exhibit convergence and stability properties that would also be independent of the mixing matrix. In this project we show that this uniform performance feature is simply achieved by 'serial updating' or 'iterative 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 Parameter Free Separators (PFS), based on the idea of serial updating. The stability condition and the theoretical asymptotic separation levels are given in closed form and, it depends only on the distributions of the sources using some numerical experiments.**

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

I. Introduction

BLIND source separation (BSS) is receiving some attention in signal processing literature [2]–[5]. The underlying model is that of n statistically independent signals whose m (possibly noisy) linear combinations are observed; the problem consists in recovering the original signals from their mixture. The 'blind' qualification refers to the coefficients of the mixture: no *a priori* information is assumed to be available about them. This feature makes the blind approach extremely versatile because it does not rely on modeling the underlying physical phenomena. In particular, it should be contrasted with standard narrow band array processing where a similar data model is considered but the mixture coefficients are assumed to depend in a known fashion on the location of the sources.

This report addresses the issue of adaptive source separation and considers the case where any additive noise can be neglected. 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 $\boldsymbol{x}_t$ modelled as

$$\boldsymbol{x}_t = A\boldsymbol{s}_t, \tag{1}$$

where the $m \times n$ matrix A is called the 'mixing' matrix and where the n source signals are collected in a $m \times 1$ vector denoted as $\boldsymbol{s}_t$. Adaptive source separation consists in updating a $n \times m$ matrix B_t, called the separating matrix (see Fig. I), such that

$$\boldsymbol{y}_t \overset{\text{def}}{=} B_t\boldsymbol{x}_t \tag{2}$$

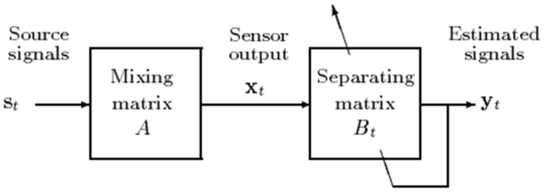


Fig. 1. Adapting a separating matrix

We emphasize that B_t should be updated without resorting to any information about the spatial mixing matrix A . This is in notable 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 *a priori*. Matrix A is assumed 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

$$\boldsymbol{C}_t \overset{\text{def}}{=} B_tA \tag{3}$$

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 in-observable. The permutation in-determination stems from the fact that the labeling 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. Ideally, an adaptive source separator should converge to a matrix B_* such that $B_*A = I$, or, equivalently, the global system C_t should converge to the $n \times n$ identity matrix I.

The following assumptions hold throughout.

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

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

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

Assumption 4. *The components of $\boldsymbol{s}_t$ 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

$$R_s \overset{\text{def}}{=} E[\boldsymbol{s}_t\boldsymbol{s}_t^T] = I. \tag{4}$$

A. Bose is a PhD student in the Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, IL 60607, USA, UIN: 665387232, e-mail: abose4@uic.edu

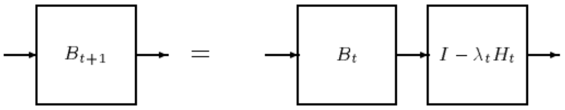


Fig. 2. Serial update of a matrix

II. Serial update for uniform performance

1) Serial update: Let us consider adapting rules in a *serial form* (See Fig. II-1):

$$B_{t+1} = \{I - \lambda H(\boldsymbol{y}_t)\}B_t \tag{5}$$

where matrix H is a function of the output $\boldsymbol{y}$ only and λ is a small positive number. The update (5) 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 (5) with the mixing matrix A, the evolution of the global system $C_t = B_t A$ is:

$$C_{t+1} = \{I - \lambda H(C_t \boldsymbol{s}_t)\}C_t \tag{6}$$

Equation (6) 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 $\hat{\boldsymbol{s}}_t = C_t \boldsymbol{s}_t$, 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 practical consequence of the uniformity property is that, in the noiseless case, the study of the algorithm is exhausted by the study of its behavior for $A = I$.

2) Estimating Equations: The adaptive algorithm (5) may be seen as a stochastic approximation device to solve the equation:

$$\mathrm{E}\{H(\boldsymbol{y}_t)\} = 0, \tag{7}$$

since any full rank matrix B such that the distribution of $\boldsymbol{y}(t)$ satisfies (7) 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:

$$c(B) \overset{\mathrm{def}}{=} \mathrm{E}\{\phi(\boldsymbol{y})\} = \mathrm{E}\{B\boldsymbol{x}\}, \tag{8}$$

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

$$\phi(\boldsymbol{y} + \partial \boldsymbol{y}) = \phi(\boldsymbol{y}) + \phi^{'}(\boldsymbol{y})^T \boldsymbol{y} + o(\partial \boldsymbol{y}), \tag{9}$$

and it is easily found that

$$\begin{aligned} c((I+\varepsilon)B) &= c(B) + \mathrm{E}\{\phi^{'}(\boldsymbol{y})^T \varepsilon \boldsymbol{y}\} + o(\partial\varepsilon) \qquad (10) \\ &= c(B) + < \mathrm{E}\{\phi^{'}(\boldsymbol{y})\boldsymbol{y}^T | \varepsilon\} > + o(\partial\varepsilon) \end{aligned}$$

A gradient algorithm for minimizing $c(B)$ consists in updating B into $(I+\varepsilon)B$, with ε proportional to $-\lambda \mathrm{E}\{\phi^{'}(\boldsymbol{y})\boldsymbol{y}^T\}$ where λ is a small positive number in order to ensure $c(B_{n+1}) = c((I+\varepsilon)B_n) < c(B_n)$. The stochastic version of this algorithm amounts to dropping the expectation operator. This is just the updating rule (5) with $H(\boldsymbol{y}) = \phi^{'}(\boldsymbol{y})\boldsymbol{y}^T$.

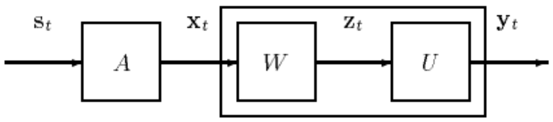


Fig. 3. Two-step separation in batch processing

Regarding the source separation problem, we base our approach on contrast functions which are not in the form (8) so that we cannot implement directly the above result. It is nonetheless possible to exhibit a family of functions $H(\boldsymbol{y})$ such that separating matrices are stable stationary points of the algorithm (5) as shown next.

III. 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 $H(\boldsymbol{y})$ stemming from an 'orthogonal' contrast function. The following constrained optimization problem

$$\begin{aligned} \min \qquad & \Phi \overset{\mathrm{def}}{=} \sum_{i=1}^{n} \mathrm{E}\{|y_i|^4\} \qquad (11) \\ s.t. \qquad & \mathrm{E}\{\boldsymbol{y}\boldsymbol{y}^T\} = I \end{aligned}$$

may be used to separate sources with negative kurtosis. It may be solved in a two step approach as in the Fig. III. First the $m \times 1$ observed vector $\boldsymbol{x}_t$ is whitened into an intermediate $n \times 1$ vector $\boldsymbol{z}$ by means of a $n \times m$ whitening matrix W, i.e. $\boldsymbol{z} = W\boldsymbol{x}$ and $R_z = \mathrm{E}\{\boldsymbol{z}\boldsymbol{z}^T\} = I$. In a second step, the spatially white vector $\boldsymbol{z}$ is rotated by a unitary matrix U into the final output: $\boldsymbol{y} = U\boldsymbol{z}$. Matrix U is computed so as to minimize $\sum_{i=1}^{n} \mathrm{E}\{|y_i|^4\}$, the constraint $\mathrm{E}\{\boldsymbol{y}\boldsymbol{y}^T\} = I$ 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 $B_t = U_t W_t$.

1) Whitening Stage: Let us consider the following objective function

$$\Upsilon(W) \overset{\mathrm{def}}{=} \text{Trace}\{R_z\} - \log\det(R_z) - n \tag{12}$$

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

$$\Upsilon(W + \varepsilon W) = \Upsilon(W) + 2 < \mathrm{E}\{\boldsymbol{z}\boldsymbol{z}^T\} - I | \varepsilon > + o(\varepsilon) \tag{13}$$

Hence, even though the objective function Υ is not in the form of (8), the relative variation (13) of Υ is in the form of (10), suggesting the serial whithening algorithm:

$$W_{t+1} = \{I - \lambda(\boldsymbol{z}_t \boldsymbol{z}^T - I)\}W_t \tag{14}$$

2) Orthogonal stage: The orthogonal stage attempts to optimize (11) which is in the form of (8) with $\phi(\boldsymbol{y}) = \sum_{i=1}^{n} |y_i|^4$ and Φ now a function of U since $\boldsymbol{y} = U\boldsymbol{z}$. Similar to (10), we have

$$\Phi(U + \varepsilon U) = \Phi(U) + < \mathrm{E}\{\phi^{'}(\boldsymbol{y})\boldsymbol{y}^T\}|\varepsilon > + o(\varepsilon) \tag{15}$$

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

$$(U + \varepsilon U)(U + \varepsilon U)^T = I + \varepsilon + \varepsilon^T + o(\varepsilon) \tag{16}$$

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

$$U_{t+1} = \{I - \lambda[\phi^{'}(\boldsymbol{y}_t)\boldsymbol{y}_t^T - \boldsymbol{y}_t\phi^{'}(\boldsymbol{y}_t)^T]\}U_t \tag{17}$$

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) 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 (14) and (17) and neglecting the terms in λ^2 results in updating B_t as in (5) with

$$H(\boldsymbol{y}) = \boldsymbol{y}\boldsymbol{y}^T - I + \phi^{'}(\boldsymbol{y})\boldsymbol{y}^T - \boldsymbol{y}\phi^{'}(\boldsymbol{y})^T \tag{18}$$

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 $\mathrm{E}\{H(\boldsymbol{y}(t))\} = 0$ so that B is stationary point of (5).

IV. A FAMILY OF SOURCE SEPARATION ALGORITHM

The result (18) of section III is generalized by noting that if $C = BA = I$ then $\mathrm{E}\{H(\boldsymbol{y}_t)\} = 0$ whenever $\phi^{'}$ is replaced by any component-wise non-linear function 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 II, this is one of the ingredients for uniform performance. A specific algorithm then depends on the choice of some component-wise non-linear function g. The generic algorithm is summarized by

$$g - \textbf{PFS} \left\{ \begin{array}{rcl} B_{t+1} & = & \{I - \lambda_t H(\boldsymbol{y}_t)\}B_t \\ H(\boldsymbol{y}) & = & \boldsymbol{y}\boldsymbol{y}^T - I + g(\boldsymbol{y})\boldsymbol{y}^T - \boldsymbol{y}g(\boldsymbol{y})^T \\ (g(\boldsymbol{y}))_i & = & g_i(\boldsymbol{y}_i) \end{array} \right. \tag{19}$$

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

$$k_i^g \stackrel{\mathrm{def}}{=} \mathrm{E}\{s_i g_i(s_i)\} - \mathrm{E}\{g_i^{'}(s_i)\}\mathrm{E}\{s_i^2\} \tag{20}$$
$$i = 1, \cdots, n$$

where $g_i^{'}(s_i)$ is the derivative of g_i, the i-th component function of 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 $k_i^g + k_g^j$ for each pair (i, j) of sources. Hence B_* is an attractor if and only if

$$k_i^g + k_g^j < 0 \qquad \forall i \neq j \tag{21}$$

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 (21) which is weaker than the condition that all sources have negative kurtosis (this would be : $k_i^g < 0,\ \forall i$). 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 behaviors may be observed, especially if the non-linearities in g are strongly increasing functions (like a cubic distortion for instance). Hence, it is recommended to implement a robustified verion of PFS. An *ad hoc* stabilization is obtained by modifying $H(\boldsymbol{y})$ into

$$H(\boldsymbol{y}) = \frac{\boldsymbol{y}\boldsymbol{y}^T - I}{1 + \lambda|\boldsymbol{y}^T\boldsymbol{y}|} + \frac{g(\boldsymbol{y})\boldsymbol{y}^T - \boldsymbol{y}g(\boldsymbol{y})^T}{1 + \lambda|\boldsymbol{y}^T g(\boldsymbol{y})|} \tag{22}$$

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.

V. NUMERICAL ANALYSIS

We analyze the performance of PFS 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 $|(BA)_{ij}|^2$.

1) Asymptotic performance: We have computed the mean j-to-i ISI measured as

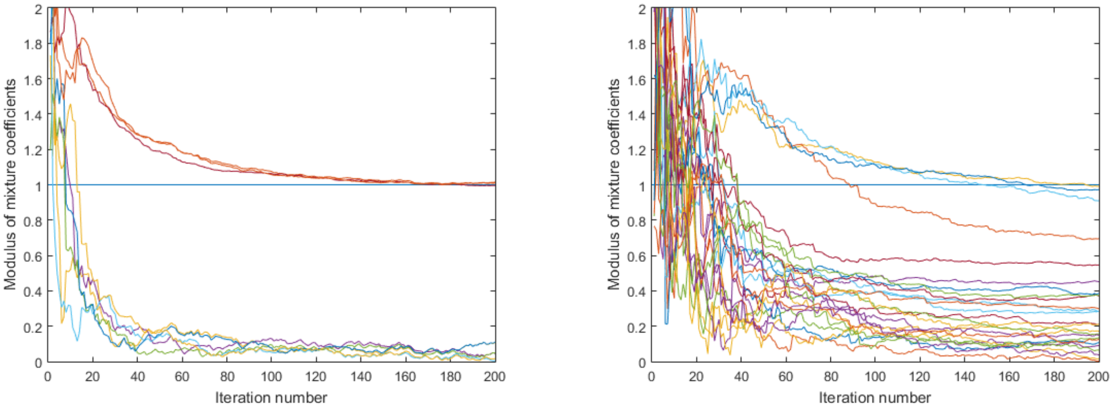
$$\rho_{ij} \stackrel{\mathrm{def}}{=} \mathrm{E}\{|(BA)_{ii}|^2\} \tag{23}$$

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

$$\rho = \lambda \left[\frac{1}{4} + \frac{1}{2} \frac{\mathrm{E}\{s_i^2\}\mathrm{E}\{g^2(s_i)\} - \mathrm{E}^2\{s_i g(s_i)\}}{\mathrm{E}\{s_i^2\}\mathrm{E}\{g'(s_i)\} - \mathrm{E}^2\{s_i g(s_i)\}} \right] \tag{24}$$

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. Note that this bound is reached if $|s_{it}| = 1$ and g is an odd function.

2) Simulation results: The simulations show a good behavior of the PFS alorithm. In particular, we did not observe spurious attractors, in the general case of n sources and monotonous non-linearities. Fig.4(a) is for $m = 4$ sensors and $n = 3$ QAM4 modulated sources, $g_i(\boldsymbol{y}) = |y_i|^2 y_i, \lambda = 0.01$. Convergence is attained after about 100 samples which displays the evolution of $|(B_t A)_{ij}|$ for all i and j. Fig. 4(b) shows that one Gaussian source is allowed. Fives sources (2 QAM4, 1 QAM16, 1 PSK, 1 Gaussian) are recovered with $\lambda = 0.005$.



(a)

(b)

Fig. 4. Convergence to 0 or 1 of $|(B_t A)_{i,j}|$ for (a) $m = 4$, $n = 3$, $\lambda = 0.01$, and (b) $m = 4$, $n = 5$, $\lambda = 0.005$.

VI. 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 orthogonal source separation algorithms," 1994.

[2] J. F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Transactions on Signal Processing*, vol. 44, no. 12, pp. 3017–3030, Dec 1996.

[3] J. F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct 1998.

[4] J. franois Cardoso, M. Martin, J. Delabrouille, M. Betoule, and G. Patanchon, "Component separation with flexible models: Application to multichannel astrophysical observations," in *Selected Topics in Signal Processing, IEEE Journal of*, 2008, p. 746.

[5] J. F. Cardoso, "Precision cosmology with the cosmic microwave background," *IEEE Signal Processing Magazine*, vol. 27, no. 1, pp. 55–66, Jan 2010.