SOURCE SEPARATION WITHOUT EXPLICIT DECORRELATION

Odile MACCHI [1] and *Eric MOREAU* [2]

[1] LSS, Supelec, Plateau de Moulon, F-91192 Gif-sur-Yvette, FRANCE
e-mail: macchi@lss.supelec.fr
[2] MS-GESSY, ISITV, Av. Pompidou, BP 56, F-83162 La Valette, FRANCE
e-mail: moreau@isitv.univ-tln.fr

ABSTRACT

The contrast approach has become classical for separating independent sources. It involves whitening (with decorrelation) as a pre-processing step. Here we propose a new contrast applicable to correlated signals, as long as they have unit power. The corresponding system involves output Automatic Gain Controls (AGC). An adaptive contrast maximization is proposed. Its achievements are shown to outperform the adaptive implementation of the classical pre-whitened contrast.

1 INTRODUCTION

In recent years, the problem of source separation has received an increasing interest because of the wide domain of potential applications, e.g. for telecommunication purposes and for image reconstruction. Source separation enters the general category of "signal separation" where several, say N, unknown, random (time) input signals $a_i(t)$ are jointly propagated inside a linear multiple input channel $\mathcal{F}$ with multiple outputs $x_j(t)$[1]. The problem is that each x_j is a mixture of several a_i and that the mixing effect of the channel $\mathcal{F}$ is usually unknown. Hence, the different signals a_i are unobserved and not easily separated. In this paper, we consider the general linear model

$$\boldsymbol{x} = \boldsymbol{F}\boldsymbol{a} + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{F}$ is a fixed deterministic matrix called "mixture", $\boldsymbol{a}$ is the vector of input "sources" a_i and $\boldsymbol{x}$ is the vector of outputs x_j, observed after the mixture. The additive noise $\boldsymbol{n}$, if any, is assumed to be zero-mean and independent of $\boldsymbol{a}$. For instance, in the telecommunication context, this model is suitable for the case of open atmosphere (wireless) radio communications with several transmitted interfering sources a_i and several receivers arranged in an antenna, and bringing multiple information x_j. Then $\boldsymbol{F}$ represents the propagation in the atmosphere. Another example is the image reconstruction context, where the input sources a_i are the light intensities at the various pixels of the true image, and the outputs x_j characterize the observed image pixels after the distorting and blurring process $\boldsymbol{F}$. The latter is caused by the electronics or by the optics of the camera. There are many other contexts where the model (1) is relevant.

Usually the problem is attacked in a supervised mode. There is a learning phase, where some "teacher" provides many examples of associated pairs $(\boldsymbol{a}, \boldsymbol{x})$. This permits the identification of the matrix $\boldsymbol{F}$ and of some kind of inverse for $\boldsymbol{F}$, denoted $\boldsymbol{H}$.

After this learning phase the separation task is solved by computing the vector

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} \tag{2}$$

which will be essentially equal to $\boldsymbol{a}$ when the noise is low and when identification is good.

In the so-called unsupervised or self-learning approaches[2], examples of pairs $(\boldsymbol{a}, \boldsymbol{x})$ are not available. Identification of $\boldsymbol{F}$ and $\boldsymbol{H}$ must be done with the sole knowledge of a sequence of observed $\boldsymbol{x}$, or possibly of its statistics: both the input $\boldsymbol{a}$ and the matrix $\boldsymbol{F}$ are unknown. It is of course impossible to solve this problem without any additional knowledge. In previous studies on source separation, the key assumption is always the statistical *joint independence of the N sources a_i*.

Without any loss of generality it can be assumed that the a_i have unit power:

$$\mathrm{E}[a_i^2] = 1 \tag{3}$$

2 CONTRAST FUNCTIONS

2.1 Contrast for whitened inputs

Contrast functions have been introduced in [1] to cancel linear mixing effects which affect independent sources. Let $\boldsymbol{C}$ be an $(N \times N)$ matrix, and assume that

$$\boldsymbol{y} = \boldsymbol{C}\boldsymbol{a}\ . \tag{4}$$

[1]Hereafter the time t is assumed discrete.

[2]These approaches are often called blind. We think that this denomination should be avoided because of its pejorative connotation.

Then it has been shown that the y_i are equal to the sources a_i up to N arbitrary non-zero gains and to an order permutation, if and only if (iff) some contrast function such as

$$\mathcal{J}_w(\boldsymbol{y}) = \sum_{i=1}^{N} |\mathrm{K}_{y_i y_i}| \tag{5}$$

is maximized by the matrix $\boldsymbol{C}$. In eq. (5), $\mathrm{K}_{y_i y_i}$ is the fourth order self-cumulant of the random variable y_i. Moreover $\boldsymbol{a}$ and thus $\boldsymbol{y}$ are taken zero-mean (with a slight loss of generality). Hence $\mathrm{K}_{y_i y_i} = \mathrm{E}[y_i^4] - 3(\mathrm{E}[y_i^2])^2$. It has also been shown that the maximum occurs iff

$$\boldsymbol{C} = \boldsymbol{D}\boldsymbol{P} \tag{6}$$

where $\boldsymbol{D}$ is an invertible diagonal matrix and $\boldsymbol{P}$ an arbitrary permutation matrix.[3] However this result is restricted to "white" vectors $\boldsymbol{y}$ in the sense that

$$\mathrm{E}[y_i y_j] = 0 \qquad \text{for} \quad i \neq j, \quad i,j = 1, \ldots, N \tag{7}$$

$$\mathrm{E}[y_i^2] = 1 \qquad \text{for} \quad i = 1, \ldots, N \ . \tag{8}$$

When the N sources are observed through N (possibly noisy) linear mixtures x_j, *i.e.*, when one observes a vector

$$\boldsymbol{x} = \boldsymbol{F}\boldsymbol{a} + \boldsymbol{n} \ , \tag{9}$$

this type of result allows to "separate" the N sources a_i by successively
(i) "pre-whitening" the vector $\boldsymbol{x}$, *i.e.* linearly transforming it into a vector $\boldsymbol{w}$ which exhibits properties (7) and (8),

$$\boldsymbol{w} = \boldsymbol{W}\boldsymbol{x} \tag{10}$$

(ii) then processing the vector $\boldsymbol{w}$ with the proper separation stage

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{w} \tag{11}$$

under the constraint that the matrix $\boldsymbol{H}$ is unitary ($\boldsymbol{H}^T = \boldsymbol{H}^{-1}$) so that the white character is retained for $\boldsymbol{y}$.

The corresponding system is called "pre-whitened separation" (PWS) for its adaptive implementation. In our implementation $\boldsymbol{W} = \boldsymbol{\Lambda}\boldsymbol{T}$ where $\boldsymbol{T}$ is a lower triangular matrix with unitary diagonal entries ensuring decorrelation (7), and the regular diagonal matrix $\boldsymbol{\Lambda}$ accounts for N AGC ensuring the normalizations (8), while the separation matrix $\boldsymbol{H}$ is achieved with Givens rotations [1].

In the present communication, we perform source separation directly with correlated vectors $\boldsymbol{y}$. This retains the N normalizing constraints (8) but cancels the $N(N-1)/2$ correlation constraints involved in eq. (7).

[3] *i.e.* P has one and only one nonzero entry (which is 1) per row and column.

2.2 Contrast for correlated inputs

For correlated signals, the contrast expression has to include cross moments. For instance, when $\boldsymbol{y}$ is "normalized" according to (8), and under the assumption, retained below, that the kurtosises of all the sources a_i have the same sign, denoted ε, the function

$$\mathcal{J}'_n(\boldsymbol{y}) = \sum_{i=1}^{N} |\mathrm{K}_{y_i y_i}| - \sum_{i,j=1;\ i<j}^{N} |\mathrm{K}_{y_i y_j}| \tag{12}$$

where $\mathrm{K}_{y_i y_j}$ is the fourth order cross cumulant, namely

$$\mathrm{K}_{y_i y_j} = \mathrm{E}[y_i^2 y_j^2] - 2(\mathrm{E}[y_i y_j])^2 - \mathrm{E}[y_i^2]\mathrm{E}[y_j^2]$$

is a contrast. Dropping the unnecessary constant values yields the new contrast

$$\mathcal{J}_n(\boldsymbol{y}) = \varepsilon \sum_{i=1}^{N} \mathrm{E}[y_i^4] - \sum_{i,j=1, i<j}^{N} \varepsilon\{\mathrm{E}[y_i^2 y_j^2] - 2(\mathrm{E}[y_i y_j])^2\} \ . \tag{13}$$

This property has been proved in [4], but it has not been checked by experiments, and no implementation is given to perform maximization of (13).

3 ADAPTIVE SOURCE SEPARATION

We provide a novel adaptive algorithm to separate the sources, based on the contrast $\mathcal{J}_n(\boldsymbol{y})$ and called NS for "normalized separation".

The set $\mathcal{Y}$ of definition for $\mathcal{J}_n(\boldsymbol{y})$ is the set of normalized vectors. As a result, the separation task involves a post-processing of the kind

$$\boldsymbol{y} = \boldsymbol{\Lambda}\boldsymbol{z} \tag{14}$$

where the diagonal matrix $\boldsymbol{\Lambda}$ ensures the normalizing condition (8) for the y_i. This can be done by placing an adaptive gain control (AGC) on each coordinate y_i. The proper separation stage, with matrix $\boldsymbol{H}$, comes first according to

$$\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x} \tag{15}$$

where the diagonal entries of matrix $\boldsymbol{H}$ can be constrained to unity without loss of generality.

It remains to estimate the separation matrix such that the contrast $\mathcal{J}_n(\boldsymbol{y})$ in (13) is maximum, or equivalently, such that the cost $c(\boldsymbol{H}) = -\mathcal{J}_n(\boldsymbol{y})$ is minimum. Our general strategy will be adaptive, based on gradient algorithms.

First a deterministic procedure is to reach the minimum of $c(\boldsymbol{H})$ by an iterative algorithm which updates $\boldsymbol{H}$ with the opposite gradient increment

$$\triangle \boldsymbol{h}_i = -\frac{\mu}{2} \nabla_{\boldsymbol{h}_i} c(\boldsymbol{H}) \quad i = 1, \ldots, N \tag{16}$$

In this equation $\boldsymbol{h}_i = (h_{i1}, \ldots, h_{iN})^T$ denotes the i-th row of the separation matrix $\boldsymbol{H}$, and μ is a small positive step-size. Accordingly, the optimum $\boldsymbol{H}$ will be found as the limit of the sequence

$$\boldsymbol{H}(n) = \boldsymbol{H}(n-1) - \frac{\mu}{2} \nabla_{\boldsymbol{H}}\, c(\boldsymbol{H})\big|_{\boldsymbol{H}=\boldsymbol{H}(n-1)} . \quad (17)$$

Very often, it is possible to express the cost $c(\boldsymbol{H})$ as an expectation according to

$$c(\boldsymbol{H}) = \mathrm{E}_{\boldsymbol{H}}[C(\boldsymbol{H}, \boldsymbol{x})] \quad (18)$$

where the notation $\mathrm{E}_{\boldsymbol{H}}$ indicates an expectation, conditioned by the value of $\boldsymbol{H}$: the expectation is performed with the probability distribution of the observed vector $\boldsymbol{x}$. Then

$$\nabla_{\boldsymbol{h}_i} c(\boldsymbol{H}) = \mathrm{E}_{\boldsymbol{H}}[\nabla_{\boldsymbol{h}_i} C(\boldsymbol{H}, \boldsymbol{x})] \quad (19)$$

The so-called "adaptive algorithm" is simpler. It consists in dropping the expectation involved in (19): the matrix $\boldsymbol{H}$ is updated each time a new observation $\boldsymbol{x}(n)$ is measured according to the time recursive algorithm

$$\boldsymbol{h}_i(n) = \boldsymbol{h}_i(n-1) - \frac{\mu}{2} \nabla_{\boldsymbol{h}_i} C(\boldsymbol{H}, \boldsymbol{x}(n))\big|_{\boldsymbol{H}=\boldsymbol{H}(n-1)} \quad (20)$$

where the increment is evaluated with $\boldsymbol{H}$ in the state $\boldsymbol{H}(n-1)$ which is available before the observation $\boldsymbol{x}(n)$ is gained. The matrix $\boldsymbol{H}(n)$ generated by (20) is, like $\boldsymbol{x}(n)$, a stochastic matrix. This algorithm enters the category of stochastic approximation, for which the stability and convergence investigations are, in general, difficult [3]. Yet, in many cases a few theoretical results can be obtained, while the algorithm turns out to be efficient in practice.

The diagonal matrix $\boldsymbol{\Lambda}$ in (14) will adaptively implement the N output gain controls λ_i in order to ensure (8)

$$\begin{aligned} y_i(n) &= \lambda_i(n-1) z_i(n) && (21)\\ \gamma_i(n) &= \gamma_i(n-1) + \mu_\gamma[1 - y_i^2(n)]\ , \mu_\gamma > 0 && (22)\\ \lambda_i(n) &= \sqrt{\gamma_i(n)} && (23) \end{aligned}$$

The second issue is the adaptation of the front separation matrix $\boldsymbol{H}$. Now according to (20)

$$C(\boldsymbol{H}, \boldsymbol{x}) = \sum_{i=1}^{N}(-\varepsilon y_i^4) + \sum_{i,j=1,\, i<j}^{N} \varepsilon(y_i^2 y_j^2 - 2 y_i y_j \mathrm{E}[y_i y_j]) \quad (24)$$

where $\boldsymbol{y}$ follows from $\boldsymbol{x}$ through (14) and (15). To simplify, consider only the case of $N = 2$, with the following parametrization

$$\boldsymbol{H} = \begin{pmatrix} 1 & h_1 \\ h_2 & 1 \end{pmatrix} \quad (25)$$

Clearly

$$\begin{aligned} -\frac{1}{2} \nabla_{\boldsymbol{H}}\, C(\boldsymbol{H}, \boldsymbol{x}) \quad = \quad & \sum_{i=1}^{2}(2\varepsilon y_i^3)\, \nabla_{\boldsymbol{H}}\, y_i - \frac{\varepsilon}{2} \nabla_{\boldsymbol{H}}\, \{y_1^2 y_2^2 \\ & -2 y_1 y_2 \mathrm{E}[y_1 y_2]\}\ . \qquad (26) \end{aligned}$$

According to (14), (15) and (25)

$$\frac{\partial y_2}{\partial h_1} = \frac{\partial y_1}{\partial h_2} = 0 \quad (27)$$

Moreover it is easy to check that the normalizing property (8) implies

$$\frac{\partial y_i}{\partial h_i} = \lambda_i(x_j - \lambda_i y_i(\mathrm{E}[x_1 x_2] + \mathrm{E}[x_j^2] h_i)) \quad (28)$$

where $(i, j) = (1, 2)$ or $(2, 1)$. Plugging (27) and (28) into (26) yields the adaptive gradient algorithm:

$$\begin{aligned} \Delta h_i \quad = \quad & \mu\lambda_i\{\varepsilon(2y_i^3 + y_j \mathrm{E}[y_1 y_2] - y_i y_j^2)\} \times \\ & (x_j - \lambda_i y_i(\mathrm{E}[x_1 x_2] + \mathrm{E}[x_j^2] h_i)) + \\ & \mu\varepsilon y_1 y_2 \lambda_i\{\mathrm{E}[x_1 x_2](\lambda_j h_j - \lambda_i \mathrm{E}[y_1 y_2]) + \\ & \mathrm{E}[x_j^2](\lambda_j - \lambda_i h_i \mathrm{E}[y_1 y_2])\} \qquad (29) \end{aligned}$$

where, again, $(i, j) = (1, 2)$ or $(2, 1)$. In these equations, the quantities are evaluated with the n-th observation $\boldsymbol{x}(n)$ for $\boldsymbol{x}$, and with the separating system $\boldsymbol{\Lambda H}$ in the state $\boldsymbol{\Lambda}(n-1)\boldsymbol{H}(n-1)$:

$$\boldsymbol{z}(n) = \boldsymbol{H}(n-1)\boldsymbol{x}(n)\ ; \quad \boldsymbol{y}(n) = \boldsymbol{\Lambda}(n-1)\boldsymbol{z}(n) \quad (30)$$

Finally, the expectation $\mathrm{E}[y_1 y_2]$ (and $\mathrm{E}[x_1^2]$, $\mathrm{E}[x_2^2]$, $\mathrm{E}[x_1 x_2]$) which appear in (29) can be estimated via the LMS-type algorithm

$$M(n) = M(n-1) + \mu_m[m(n) - M(n-1)] \quad (31)$$

where $m(n)$ is the n-th trial of $y_1 y_2$ (and of $x_1^2, x_2^2, x_1 x_2$), and μ_m is a small positive step-size.

The joint adaptive formulae (29) - (31), plus the AGC formulae for $\boldsymbol{\Lambda}(n)$ constitute an adaptive source separation system, which we call the "Normalized Separation" (NS) system.

4 COMPUTER SIMULATIONS

To investigate the separation performance of a method, we use the so-called "separation index" $s(\boldsymbol{C})$ [4] which depends on the overall matrix $\boldsymbol{C} = \boldsymbol{\Lambda H F}$ (mixture $\boldsymbol{F}$ followed by separation $\boldsymbol{\Lambda H}$) and is worth

$$\begin{aligned} s(\boldsymbol{C}) \quad = \quad & \frac{1}{2}\left[\sum_i \left(\sum_j \frac{c_{ij}^2}{\max_\ell c_{i\ell}^2} - 1\right)\right. \\ & \left. + \sum_j \left(\sum_i \frac{c_{ij}^2}{\max_\ell c_{\ell j}^2} - 1\right)\right] . \qquad (32) \end{aligned}$$

Clearly this positive index is zero iff $\boldsymbol{C}$ is of the kind (6) characteristic of separation.

Two mixtures are considered

$$\boldsymbol{F}_1 = \begin{pmatrix} 1 & 0.6 \\ 0.4 & 1 \end{pmatrix}\ , \qquad \boldsymbol{F}_2 = \begin{pmatrix} 1 & -0.6 \\ 0.4 & 1 \end{pmatrix}$$

and there is no noise. Two cases of sources are considered:

Case 1: The sources are binary. Fig. 1 (resp. Fig. 2) shows the plots of the separation index, averaged over 50 independent trials for the respective systems PWS and NS using mixture $\boldsymbol{F}_1$ (resp. $\boldsymbol{F}_2$). Clearly NS has better performance than PWS.

Case 2: The sources take the four values $\pm 1/\sqrt{5}$, $\pm 3/\sqrt{5}$ with equal probability. They are called 4-PAM. Fig. 3 (resp. Fig. 4) shows the plots of the separation index, averaged over 50 independent trials for the respective systems PWS and NS using mixture $\boldsymbol{F}_1$ (resp. $\boldsymbol{F}_2$). Clearly NS has again better performance than PWS.

In conclusion, a novel adaptive algorithm is proposed which does not require a pre-whitening stage and thus is applicable with correlated signals. Its performances are shown to outperform those of the classical two stages algorithm (with whitening) in the four presented cases.

References

[1] P. Comon, "Independent component analysis, a new concept?", *Signal Processing*, Vol. 36, pp 287-314, 1994.

[2] C. Jutten and J. Herault, "Blind separation of sources, Part I: An adaptative algorithm based on neuromimetic architecture", *Signal Processing*, Vol. 24, pp 1-10, 1991.

[3] O. Macchi, *Adaptive processing: the least mean square approach with applications in transmission*, Wiley, 1995.

[4] E. Moreau and O. Macchi, "High order contrasts for self-adaptive source separation", *International Journal of Adaptive Control and Signal Processing*, vol. 10, no. 1, pp 19-46, January 1996.

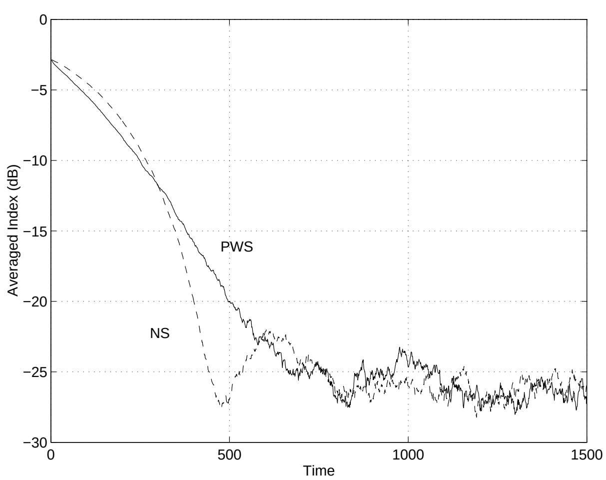

Figure 1: Case 1, averaged index using mixture $\boldsymbol{F}_1$.

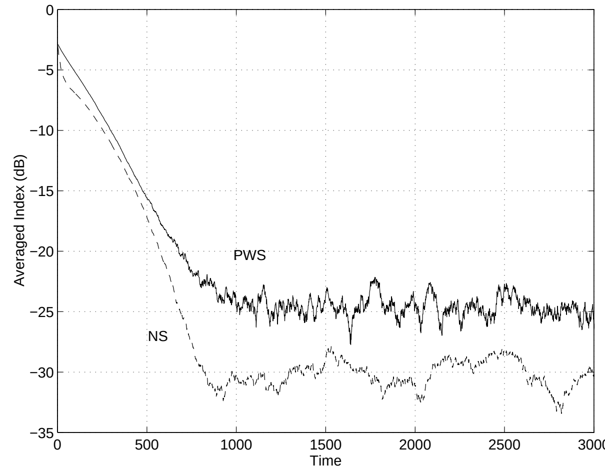

Figure 2: Case 1, averaged index using mixture $\boldsymbol{F}_2$.

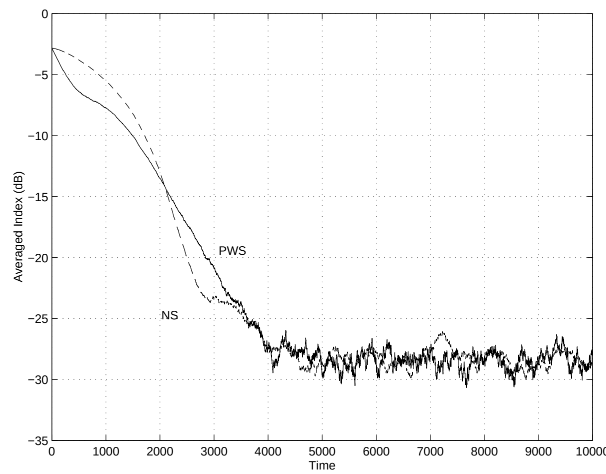

Figure 3: Case 2, averaged index using mixture $\boldsymbol{F}_1$.

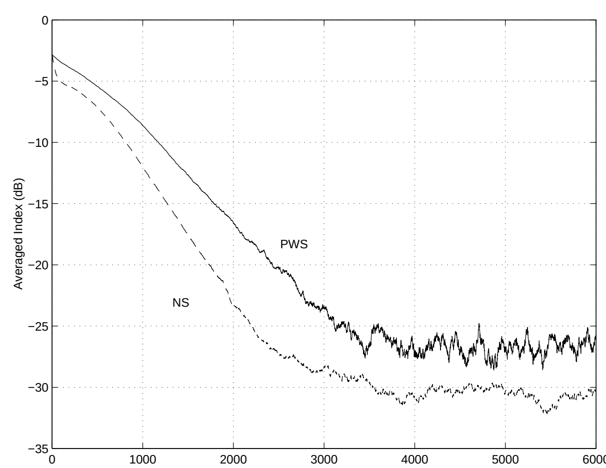

Figure 4: Case 2, averaged index using mixture $\boldsymbol{F}_2$.