NONLINEAR CONSTRAINED OPTIMIZATION USING LAGRANGIAN APPROACH FOR BLIND SOURCE SEPARATION

Benoit STOLL and *Eric MOREAU*

MS-GESSY, ISITV, Université de Toulon et du Var,
Av. G. Pompidou, BP 56, F-83162 La Valette du Var, FRANCE
e-mail: stoll@isitv.univ-tln.fr; moreau@isitv.univ-tln.fr

ABSTRACT

The paper deals with the blind source separation problem. We introduce two new adaptive algorithms based on the minimization of constrained contrast functions using a Lagrangian approach. The algorithms "only" require one stage for separation and the approach is general in the sense that it can be used with any contrasts working with normalized vectors. The computer simulation shows good performances in comparison to the EASI algorithm.

1 INTRODUCTION

We consider the source separation problem [3] which find numerous applications in diverse fields of engineering and applied sciences, e.g. data communications, seismic exploration, antenna processing, speech processing etc... It can be simply formulated as follows. Several linear mixture of different signals called sources are observed. We want to recover the unknown original sources without knowing the mixing system. Hence this must be realized from the only observations and this is the reason why this problem is often qualified as "blind" or "unsupervised".

Among the great number of approaches that have been proposed in the recent literature, we are primary concerned with high-order statistics inverse criteria based approaches. In this field, contrast functions constitute separation criteria in the sense that their maximization solve the source separation problem. Such contrasts have been first introduced in [2] and recently generalized in [7] and [8]. Some contrasts work with normalized vectors (or "white" vectors), e.g. [2], [7], [8], while in [5] one can find contrasts which do not require the normalization. Unfortunately, the optimization procedure of this last class of contrasts is very complex. Now the classical way for the maximization of the first class of contrast require two stages. The first stage consists in a normalization (whitening) of the observations and the second stage maximize the contrast over the set of unitary matrices.

The main purpose of this communication is to propose a one stage algorithm for the maximization of contrasts requiring normalization. For this task, the normalization is viewed as a constraint which is considered in the criterion thanks to a Lagrangian approach. It has to be noticed that in [4] a Lagrangian approach has also been considered but restricted to the estimation of an unitary matrix. In this latter case a first whitening stage is again required.

Because the parameter we are looking for is a matrix, the methods are derived using both a classical gradient and a relative gradient [1]. This leads to two algorithms whose performances are illustrated thanks to computer simulations in comparison to the EASI algorithm [1].

2 PROBLEM FORMULATION

The source separation problem consist in estimating a set of M independent signals from $N \geq M$ observed instantaneous mixtures of these signals.

We can express the classical linear memoriless mixture model as

$$\boldsymbol{x} = \boldsymbol{G}\boldsymbol{a} + \boldsymbol{b} \tag{1}$$

where $\boldsymbol{x}$ is the $(N,1)$ vector of observations, $\boldsymbol{a}$ the $(M,1)$ vector of sources, $\boldsymbol{b}$ the $(N,1)$ vector of additive noise and $\boldsymbol{G}$ the (N,M) invertible mixing matrix. In the following, we consider $M=N$ and $\boldsymbol{b}=0$.

The separation problem consist in estimating a separating matrix $\boldsymbol{H}$ such that the vector

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

restore the N input sources a_i. In the noiseless case, this is to identify an inverse $\boldsymbol{G}^{-1}$ of the mixing matrix. Ideally, considering the global system, the global matrix should be a unit matrix

$$\boldsymbol{S} = \boldsymbol{H}\boldsymbol{G} = \boldsymbol{I} \tag{3}$$

As sources are unobservable, there are some inherent indeterminacy in their estimation. That is, in general, we can not identify the order and the power of each components of the source vector $\boldsymbol{a}$. Hence $\boldsymbol{H}$ is called a separating matrix when the global matrix $\boldsymbol{S}$ reads

$$\boldsymbol{S} = \boldsymbol{D}\boldsymbol{P} \tag{4}$$

where $\boldsymbol{D}$ is an invertible constant diagonal matrix and $\boldsymbol{P}$ a permutation matrix.

For separation, the key hypothesis is the joint independence of the N sources and the non-zero character of some of their cumulants. Thus Gaussian sources are excluded. Without loss of generality, the sources can be assumed zero-mean with unit variance, *i.e.*

$$\mathrm{E}\left[\boldsymbol{a}\boldsymbol{a}^T\right] = I \tag{5}$$

where E stands for the mathematical expectation operator, the superscript T the transpose operator and $\boldsymbol{I}$ the (N,N) identity matrix.

3 RECALLS ON CONTRAST FUNCTION

Contrast functions constitute separation criteria in the sense that their maximization solve the separation problem, *i.e.* they are maximum if and only if the relation (4) holds for $\boldsymbol{S}$.

As originally defined [2], a contrast function has to be a symmetrical and scale invariant function to be maximized. According to this definition a first useful contrast was proposed in [2] for normalized vectors.

For real sources, it reads

$$I_2(\boldsymbol{y}) = \sum_{i=1}^{N} \left(\mathrm{C}_p[y_i]\right)^2 \tag{6}$$

where $\mathrm{C}_p[y_i]$ is the p-th order joint cumulant of y_i and p an integer greater or equal to 3.

It has also been shown that the minimization of $I_2(\boldsymbol{y})$ is equivalent to the minimization of the sum of squares of all cross-cumulants of the same order p.

Later in [5],[6], it was shown that squaring the cumulants in (6) is not necessary, hence, the proposed contrast reads

$$I_1'(\boldsymbol{y}) = \sum_{i=1}^{N} |\mathrm{C}_p[y_i]| \ . \tag{7}$$

Moreover, when $p=4$, one can still "simplify" to obtain the simplest contrast

$$I_1''(\boldsymbol{y}) = \varepsilon_4 \sum_{i=1}^{N} \mathrm{E}[y_i^4] \tag{8}$$

under the assumption that the fourth-order cumulants of all the sources a_i have the same sign denoted by ε_4.

Finally, in [7] extended forms of contrast functions are introduced based on some convex functions of the absolute values of joint cumulants, and in [8] non symmetrical contrasts are defined which allow to exhibit a novel wide class of contrast function whose maximization is proved to be a sufficient condition for separation. Two examples of non symmetrical contrasts are

$$J_1'(\boldsymbol{y}) = \sum_{i=1}^{N} \gamma_i |\mathrm{C}_p[y_i]| \tag{9}$$

and

$$J_1''(\boldsymbol{y}) = \varepsilon_4 \sum_{i=1}^{N} \gamma_i \mathrm{E}[y_i^4] \tag{10}$$

where $\gamma_1 \geq \cdots \geq \gamma_N > 0$ and assuming that the p-th order ($p=4$ for $J_1''(\boldsymbol{y})$) cumulants of sources satisfy one of the two following (non restrictive) conditions:

c1. $|\mathrm{C}_p[a_1]| \geq \cdots \geq |\mathrm{C}_p[a_N]| > 0$;

c2. $|\mathrm{C}_p[a_1]| \geq \cdots \geq |\mathrm{C}_p[a_{N-1}]| > |\mathrm{C}_p[a_N]| = 0$.

In particular this means that at most one of the cumulants $\mathrm{C}_p[a_i]$, $i \in \{1,\ldots,N\}$, is null.

It has to be noticed that the following derivation can be used for any contrast using white vectors and thus, in particular, the five hereabove contrasts.

4 SOURCE SEPARATION USING A LAGRANGIAN APPROACH

In order to achieve the separation, we intend to maximize a contrast, say $C(\boldsymbol{y})$, which works with normalized (white) vectors $\boldsymbol{y}$, *i.e.* vectors such that $\mathrm{E}[\boldsymbol{y}\boldsymbol{y}^T] = I$. We propose to realize the separation in one stage. Hence we consider the normalization of the output vector $\boldsymbol{y}$ as a constraint and the contrast based criterion reads

$$\max_{\boldsymbol{H}} C(\boldsymbol{y}) \quad \text{subject to} \quad \mathrm{E}[\boldsymbol{y}\boldsymbol{y}^T] = I \ .$$

It is now possible to write an unconstrained criterion as the maximization of the so-called Lagrangian function defined as

$$L(\boldsymbol{y}) = C(\boldsymbol{y}) + \mathrm{trace}[\boldsymbol{\Lambda}(\mathrm{E}[\boldsymbol{y}\boldsymbol{y}^T] - I)] \tag{11}$$

where $\boldsymbol{\Lambda}$ is a symmetrical matrix referred to as the Lagrange multiplier.

At the optimum, we have

$$\frac{\partial L}{\partial \boldsymbol{H}} = 0 \quad \text{and} \quad \frac{\partial L}{\partial \boldsymbol{\Lambda}} = 0 \ .$$

The first condition is a (first order) necessary condition for $\boldsymbol{H}$ to be a maximum of C under the constraint.

Using (2), $L(\cdot)$ can be written

$$L(\boldsymbol{H}) = C(\boldsymbol{H}) + \mathrm{trace}[\boldsymbol{\Lambda}(\boldsymbol{H}\mathrm{E}[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{H}^T - \boldsymbol{I})] \tag{12}$$

where we do, for simplicity, some abuse of notation writting $L(\cdot)$ as a function of $\boldsymbol{H}$. Hence

$$\frac{\partial L}{\partial \boldsymbol{H}} = \frac{\partial C}{\partial \boldsymbol{H}} + (\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^T)\boldsymbol{H}\mathrm{E}[\boldsymbol{x}\boldsymbol{x}^T]$$

and

$$\frac{\partial L}{\partial \boldsymbol{\Lambda}} = \boldsymbol{H}\mathrm{E}[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{H}^T - \boldsymbol{I} \ .$$

The optimum value of $\boldsymbol{H}$ and $\boldsymbol{\Lambda}$ satisfy

$$\frac{\partial C}{\partial \boldsymbol{H}} + (\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^T)\boldsymbol{H}\mathrm{E}[\boldsymbol{x}\boldsymbol{x}^T] = 0 \tag{13}$$

and

$$\boldsymbol{H}\mathrm{E}[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{H}^T = \boldsymbol{I} \ . \tag{14}$$

Right multiplying (13) by $\boldsymbol{H}^T$ and using (14) leads to

$$(\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^T) = -\frac{\partial C}{\partial \boldsymbol{H}}\boldsymbol{H}^T = -\nabla_{\boldsymbol{H}} C \tag{15}$$

where $\nabla_{\boldsymbol{H}} C$ is referred to as a " relative gradient" [1].

Transposing (15), we have

$$\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^T = -\left(\nabla_{\boldsymbol{H}} C\right)^T$$

and then

$$\nabla_{\boldsymbol{H}} C = (\nabla_{\boldsymbol{H}} C)^T \tag{16}$$

Thus an optimal solution of $L(\cdot)$ is solution to (16). Unfortunately (16) seems to be difficult to solve. Let us now consider two algorithms searching for a local minimum of $L(\cdot)$.

5 TWO LAGRANGIAN ALGORITHMS

The optimal values of $\boldsymbol{H}$ and $\boldsymbol{\Lambda}$ can not be derived from the above development. They have thus to be estimated.

Classically, the two equations of the Lagrange programming are given by

$$\triangle \boldsymbol{H} = \mu_n \frac{\partial L}{\partial \boldsymbol{H}} \tag{17}$$

$$\triangle \boldsymbol{\Lambda} = -\mu_l \frac{\partial L}{\partial \boldsymbol{\Lambda}} \tag{18}$$

The first equation corresponds to the maximization of L with respect to $\boldsymbol{H}$ ($\boldsymbol{\Lambda}$ "constant") while the second equation reach for a minimum of L with respect to $\boldsymbol{\Lambda}$ ($\boldsymbol{H}$ "constant") according to the "convex duality" theory. The algorithm based on (17) and (18) will be denoted by LAGC.

Now equation (17) corresponds to a classic gradient optimization scheme. Following some ideas of [1], we can also propose the use of a relative gradient as

$$\triangle \boldsymbol{H} = \mu_n' \nabla_{\boldsymbol{H}} L \tag{19}$$

and the algorithm based on (19) and (18) will be denoted by LAGR.

For simplicity, we only consider the use of the "simplest" contrast in (8) writing now

$$C(\boldsymbol{y}) = \varepsilon_4 \sum_{i=1}^{N} \mathrm{E}[y_i^4] \tag{20}$$

recalling $\varepsilon_4 = -1$ (resp. $+1$) when sources have negative (resp. positive) fourth-order cumulant. In that case, we easily have

$$\frac{\partial C}{\partial \boldsymbol{H}} = 4\varepsilon_4 \mathrm{E}[f(\boldsymbol{y})\boldsymbol{x}^T]$$

and using (2)

$$\nabla_{\boldsymbol{H}} C = 4\varepsilon_4 \mathrm{E}[f'(\boldsymbol{y})\boldsymbol{y}^T]$$

where $f(\cdot)$ is the component wise non linear cubic function.

In practice, a stochastic gradient algorithm is used by dropping one expectation operator in (17), (18), (19). Hence for the specific contrast in (20), the stochastic Lagrangian algorithm LAGC reads

$$\begin{aligned} \triangle \boldsymbol{H}_k &= \mu_n \left[4\varepsilon_4 f(\boldsymbol{y}_k) + (\boldsymbol{\Lambda}_{k-1} + \boldsymbol{\Lambda}_{k-1}^T)\boldsymbol{y}_k\right] \boldsymbol{x}_k^T \\ \triangle \boldsymbol{\Lambda}_k &= -\mu_l \left[\boldsymbol{y}_k \boldsymbol{y}_k^T - \boldsymbol{I}\right] \end{aligned} \tag{21}$$

where $\triangle \boldsymbol{H}_k = \boldsymbol{H}_k - \boldsymbol{H}_{k-1}$ and $\triangle \boldsymbol{\Lambda}_k = \boldsymbol{\Lambda}_k - \boldsymbol{\Lambda}_{k-1}$ and the stochastic algorithm LAGR reads

$$\triangle \boldsymbol{H}_k = \mu_n' \left[4\varepsilon_4 f(\boldsymbol{y}_k) + (\boldsymbol{\Lambda}_{k-1} + \boldsymbol{\Lambda}_{k-1}^T)\boldsymbol{y}_k\right] \boldsymbol{y}_k^T \boldsymbol{H}_{k-1}$$

together with (21).

6 COMPUTER SIMULATIONS

In order to illustrate the two hereabove algorithms we present computer simulations using the following mixing matrix

$$\boldsymbol{G} = \begin{pmatrix} 1 & 0.6 \\ 0.4 & 1 \end{pmatrix} . \tag{22}$$

The algorithm performance is evaluated thanks to the positive index [5]

$$\begin{aligned} i(\boldsymbol{S}) = \ & \frac{1}{2}\left[\sum_i \left(\sum_j \frac{|S_{ij}|^2}{\max\limits_{\ell}|S_{i\ell}|^2} - 1\right)\right. \\ & \left. + \sum_j \left(\sum_i \frac{|S_{ij}|^2}{\max\limits_{\ell}|S_{\ell j}|^2} - 1\right)\right] \end{aligned} \tag{23}$$

which equals zero when perfect separation is realized.

Two cases of sources are considered:

Case 1: The sources are binary. Using Monte Carlo runs, fig. 1 shows the mean of the index over 100 independent realizations and with respect to discrete time. Given the same convergence speed (defined here as the time for the algorithms to reach an index of -15dB), the figure clearly shows that the two novel algorithms work well in comparison to the EASI algorithm (the complexity being approximately the same).

Case 2: The sources take the four values $\pm 1/\sqrt{5}$, $\pm 3/\sqrt{5}$ with equal probability. Using Monte Carlo runs, fig. 2 shows the mean of the index over 100 independent realizations and with respect to discrete time. The figure shows that the LAGR algorithm leads to a better mean index after convergence in comparison to the EASI algorithm.

In conclusion, two novel adaptive stochastic gradient based algorithms are derived using a Lagrangian approach. The algorithms only require one stage for separation and the approach is general in the sense that it can be used with any contrasts working with normalized vectors. The computer simulation shows good performances in comparison to the EASI algorithm.

Acknowledgment: The second author would like to thank Dr. J.-F. Cardoso for fruitful discussions related to this work.

References

[1] J.F. Cardoso and B.H. Laheld, “ Equivariant Adaptive Source Separation ”, *IEEE Transactions on Signal Processing*, Vol. 44, no. 12, pp 3017-3030, December 1996.

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

[3] C. Jutten and J. Herault, “ Blind Separation of Sources: An adaptive algorithm based on neuromimetic architecture ” *Signal Processing*, Vol.24, pp 1-10, 1991.

[4] J. Karhunen and J. Joutsensalo, “Representation and separation of signal using nonlinear PCA Type learning”, *Neural Networks*, Vol.7, no.1, pp 113-127, 1994.

[5] 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.

[6] E. Moreau, “Criteria for complex sources separation”, in *proc. EUSIPCO’96*, Trieste, Italy, pp 931-934, September 1996.

[7] E. Moreau and J.-C. Pesquet, “Generalized contrasts for multichanel blind deconvolution of linear systems”, *IEEE Signal Processing Letters*, Vol.4, no. 6, pp 182-183, June 1997.

[8] E. Moreau and N. Thirion-Moreau, “Non symmetrical contrasts for source separation”, submitted to *IEEE Trans. Signal Processing*, December 1997.

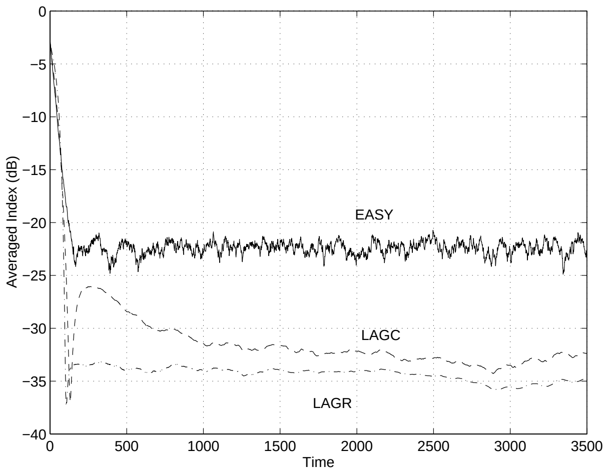

Figure 1: Performance of the two novel algorithms in comparison with EASI in case 1.

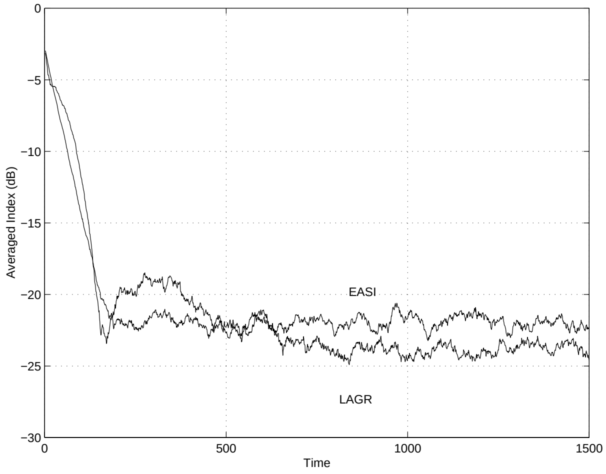

Figure 2: Performance of the LAGR algorithms in comparison with EASI in case 2.