# BLIND SOURCE SEPARATION IN POST NONLINEAR MIXTURES

*Sophie Achard, Dinh Tuan Pham*

Univ. of Grenoble
Laboratory of Modeling and Computation,
IMAG, C.N.R.S.
B.P. 53X, 38041 Grenoble Cedex, France
Sophie.Achard@imag.fr,
Dinh-Tuan.Pham@imag.fr

*Christian Jutten*

Laboratory of Images and Signals,
INPG
46 avenue Félix Viallet,
38031 Grenoble Cedex, France
Christian.Jutten@inpg.fr

## ABSTRACT

This work implements alternative algorithms to that of Taleb and Jutten for blind source separation in post nonlinear mixtures. We use the same mutual information criterion as them, but we exploit its invariance with respect to translation to express its relative gradient in terms of the derivatives of the nonlinear transformations. Then we develop algorithms based on these derivatives. In a semi-parametric approach, the latter are parametrized by piecewise constant functions. All algorithms require only the estimation of the score functions of the reconstructed sources. A new method for score function estimation is presented for this purpose.

## 1. INTRODUCTION

Blind source separation has been well studied in the case of linear mixtures [1, 2, 3, 4], but only begins to attract attention in the case of nonlinear mixtures. Yet in many situations, it is more realistic to assume a nonlinear rather than a linear mixture. The main difficulty in using this class of mixtures is that they are too broad and thus lead to the problem of identifiability. Indeed Darmois [5] has shown that there exist nonlinear mixtures of sources which preserve their independence. To avoid this difficulty, in this paper we shall follow Taleb and Jutten [6] by restricting the mixture to the smaller class of post nonlinear mixtures. In such mixture, the observed channels $X_1, \ldots, X_K$ are related to the sources $S_1, \ldots, S_K$ through the relations

$$X_i = f_i(\sum_{k=1}^{K} \mathbf{A}_{ik} S_k), \qquad i = 1, \ldots, K$$

where $\mathbf{A}_{ik}$ denotes the general element of the mixing matrix $\mathbf{A}$ and $f_1, \ldots, f_K$ are nonlinear functions. It is assumed here that there are the same number $K$ of sources and sensors and the matrix $\mathbf{A}$ is *invertible* and the functions $f_i$ are monotonous so that the sources can be recovered from the observations if one *knew* $\mathbf{A}$ and $f_1, \ldots, f_K$.

The goal of blind source separation is to recover the sources without any particular knowledge on their distributions and the mixing mecanism. The separation will be based solely on the mutual independence assumption of the sources. Specifically, one tries to find a matrix $\mathbf{B}$ and $K$ applications $g_1, \ldots, g_K$ so that the random variables

$$Y_i = \sum_{k=1}^{K} \mathbf{B}_{ik} Z_k, \quad i = 1, \ldots, K, \quad \text{where } Z_k = g_k(X_k), \tag{1}$$

which represent the reconstructed sources, are as independent as it is possible. Using the mutual information as the natural measure of independence, Taleb and Jutten [6] has obtained the criterion

$$C(B, g_1, \ldots, g_K) = \sum_{i=1}^{K} (H(Y_i) - H(Z_i)) - \log|\det \mathbf{B}|. \tag{2}$$

These authors have derived a separation algorithm based on the relative gradient of the criterion with respect to $\mathbf{B}$ and $g_1, \ldots, g_K$. Our approach stems from the remark that the above criterion, due to its invariance with respect to translation, is actually a function of the derivatives $g_1', \ldots, g_K'$ of $g_1, \ldots, g_K$ and not of these functions themselves. Thus we shall compute the gradient of the criterion with respect to these derivatives and obtain new algorithms by the gradient descent method. They are described in sections 2 and 3. Section 4 introduces a new method for the score function estimation. Our algorithms have the nice property that they can be expessed entirely in terms of the estimated score functions of the sources, hence these estimations play a crucial role. Section 5 discusses the numerical implementation of the algorithms and some simulation results are presented in section 6. Finally, some comments and discussions about our algorithms are made in section 7.

Thanks to the European Bliss Project for funding.

## 2. RELATIVE GRADIENT

We begin by computing the relative gradient of the criterion with respect to the derivatives $g_1', \ldots, g_K'$ of $g_1, \ldots, g_K$. The general result can then be specialized to the case where these functions are parametrized.

### 2.1. Nonparametric approach

To obtain the linear functional of the relative gradient of $C$, we express the first order Taylor expansion of

$$C(\mathbf{B} + \varepsilon\mathbf{B}, g_1 + \delta_1 \circ g_1, \ldots, g_K + \delta_K \circ g_K)$$

with respect to $\varepsilon, \delta_1', \ldots, \delta_K'$ around 0. We then deduce the relative gradient functional

$$\varepsilon, \delta_1', \ldots, \delta_K' \mapsto \sum_{1 \le i \ne k \le K} \varepsilon_{ik} E[\psi_{Y_i}(Y_i) Y_k] + \sum_{k=1}^{K} \int \left\{ E\Big[ \mathbb{1}_+(Z_k - z) \sum_{i=1}^{K} \psi_{Y_i}(Y_i) \mathbf{B}_{ik} \Big] - p_{Z_k}(z) \right\} \delta_k'(z) dz$$

where $\psi_{Y_i}$ is the score function of $Y_i$ (i.e. $\psi_{Y_i} = -(\log p_{Y_i})'$, $p_{Y_i}$ being the density of $Y_i$) and $\mathbb{1}_+$ is the indicator function of $[0, \infty)$ (i.e. $\mathbb{1}_+(x) = 1$ if $x \ge 0$, 0 else).

By setting this gradient to zero, we deduce the following estimating equations:

$$E[Y_j \psi_{Y_i}(Y_i)] = 0, \quad 1 \le i \ne j \le K$$
$$E\left[ \mathbb{1}_+(Z_k - z) \sum_{i=1}^{K} \psi_{Y_i}(Y_i) \mathbf{B}_{ik} \right] = p_{Z_k}(z), \quad 1 \le k \le K.$$

One can verify that these equations are indeed satisfied if the variables $Y_1, \ldots, Y_K$ are independent.

### 2.2. Semi-parametric approach

The above approach allows arbitrary values for the functions $g_k$ and thus could give them too much degree of freedom, ignoring the fact that they are generally smooth functions. It is thus of interest to consider a semi-parametric approach which restricts the degree of freedom of these functions but still allows a sufficiently rich nonlinear mixture model. Our approach consists of representing the functions $g_1, \ldots, g_K$ by continuous piecewise linear functions. The main reason, besides simplicity, is that the space of piecewise linear functions is stable with respect to composition. This will simplify considerably the calculation of the relative gradient, as the increments $\delta_1, \ldots, \delta_K$ in subsection 2.1 will be also piecewise linear. As we will show shortly, our general formula can be specialized to this case with ease.

Note that the function $g_k$ cannot and (need not) be defined outside the support $[\xi_{k,1}, \xi_{k,M}]$ of the distribution of $X_k$. Although $\xi_{k,1}$ and $\xi_{k,M}$ could theoretically be $-\infty$ and $+\infty$, in practice, one can be satisfied with them being the minimum and the maximum of the observed values of $X_k$, since there is no data outside $[\xi_{k,1}, \xi_{k,M}]$ and hence it is not necessary to estimate $g_k$ outside this interval. In this interval, let us denote by $\xi_{k,2}, \ldots, \xi_{k,M-1}$ the change points in slope of the function $g_k$, arranged in increasing order. As $g_k$ is linear in $[\xi_{k,m}, \xi_{k,m+1}]$, $m = 1, \ldots, M-1$, it is clear that its relative increment $\delta_k$ in subsection 2.1 is also linear in $[\zeta_{k,m}, \zeta_{k,m+1}]$, $m = 1, \ldots, M-1$, where $\zeta_{k,m} = g_k(\xi_{k,m})$, $m = 1, \ldots, M$. Therefore, the derivatives $\delta_k'$ of $\delta_k$ can be repesented as

$$\delta_k'(z) = \sum_{m=1}^{M-1} d_{k,m} \mathbb{1}_{[\zeta_{k,m}, \zeta_{k,m+1}[}(z)$$

Without loss of generality, we take $g_1, \ldots, g_K$ to be increasing, which is equivalent to the $\zeta_{k,m}$ being increasing. We will see in section 5.2 how this condition can be easily checked.

By replacing the above expression for $\delta_k'$ in the formula for relative gradient, we obtain,

- Relative gradient of $C$ according to $\mathbf{B}$:

$$\varepsilon \mapsto \sum_{1 \le i \ne k \le K} \varepsilon_{ik} E[\psi_{Y_i}(Y_i) Y_k].$$

- Relative gradient of $C$ according to $d_{k,1}, \ldots, d_{k,M-1}$:

$$d_{k,1}, \ldots, d_{k,M-1} \mapsto \sum_{m=1}^{M-1} \int_{\zeta_{k,m}}^{\zeta_{k,m+1}} E\left\{ \mathbb{1}_+(Z_k - z) \sum_{i=1}^{K} \psi_{Y_i}(Y_i) \mathbf{B}_{ik} \right\} dz - P(\zeta_{k,m} \le Z_k < \zeta_{k,m+1}) d_{k,m}, \qquad 1 \le k \le K.$$

## 3. GRADIENT DESCENT METHOD

The minimization of the criterion is achieved through the gradient descent method. The method consists of making successive small relative changes in $\mathbf{B}$ and $g_1, \ldots, g_K$ in the opposite direction of the relative gradient so as to decrease the criterion. Explicitly, one adjusts $\mathbf{B}$ and $g_1', \ldots, g_K'$ by the following iteration,

$$\begin{cases} \mathbf{B} & \mapsto & \mathbf{B} - \lambda \mathbf{D} \mathbf{B} \\ g_k' & \mapsto & g_k'(1 - \mu h_k' \circ g_k), \quad \le k \le K \end{cases}$$

where $\lambda$ and $\mu$ are two tuning parameters and $\mathbf{D}$ and $h_k'$ are obtained from the relative gradient.

### 3.1. Nonparametric approach

Here the matrix $\mathbf{D}$ has general element

$$\mathbf{D}_{ij} = \begin{cases} E[Y_j \psi_{Y_i}(Y_i)], & 1 \le i \ne j \le K, \\ 0, & \text{else} \end{cases} .$$

As for $h'_k$, $1 \leq k \leq K$,, there are two possibilities according to the adopted metric:

- with the normal metric (of the Hilbert space of square summable functions with respect to the Lebesgue measure):

$$h_k^{n'}(z) = E\Big[\mathbb{1}_+(Z_k - z)\sum_{i=1}^{K}\psi_{Y_i}(Y_i)\mathbf{B}_{ik}\Big] - p_{Z_k}(z).$$

- with the probabilistic metric (of the Hilbert space of square summable functions with respect to the probability measure):

$$h_k^{p'}(z) = \frac{1}{p_{Z_k}(z)}E\Big[\mathbb{1}_+(Z_k - z)\sum_{i=1}^{K}\psi_{Y_i}(Y_i)\mathbf{B}_{ik}\Big] - 1.$$

### 3.2. Semi-parametric approach

The matrix $\mathbf{D}$ is given by the same formula as above while $h'_k$ is given by

$$h'_k(z) = \sum_{m=1}^{M-1} d_{k,m}\mathbb{1}_{[\zeta_{k,m};\zeta_{k,m+1}[}(z)$$

where $d_{k,m}$, $m = 1, \ldots, M-1$, $k = 1, \ldots, K$, are given in two different ways according to the adopted metric:

- with the Lebesgue metric (associated to the scalar product $\langle d_{k,.}, d'_{k,.}\rangle = \sum d_{k,m}d'_{k,m}(\zeta_{k,m+1} - \zeta_{k,m})$):

$$d_{k,m} = \frac{1}{\zeta_k^{(m+1)} - \zeta_k^{(m)}}\Bigg\{P(\zeta_{k,m} \leq Z_k < \zeta_{k,m+1}) - \int_{\zeta_{k,m}}^{\zeta_{k,m+1}} E\left[\mathbb{1}_+(Z_k - z)\sum_{i=1}^{K}\psi_{Y_i}(Y_i)\mathbf{B}_{ik}\right]dz\Bigg\}.$$

- with the probabilistic metric (associated to the scalar product $\langle d_{k,.}, d'_{k,.}\rangle = \sum d_{k,m}d'_{k,m}P(\zeta_{k,m} \leq Z_k \leq \zeta_{k,m+1})$):

$$d_{k,m} = 1 - \frac{1}{P(\zeta_{k,m} \leq Z_k < \zeta_{k,m+1})} \times \int_{\zeta_{k,m}}^{\zeta_{k,m+1}} E\left[\mathbb{1}_+(Z_k - z)\sum_{i=1}^{K}\psi_{Y_i}(Y_i)\mathbf{B}_{ik}\right]dz.$$

## 4. SCORE FUNCTION ESTIMATION

In all our algorithms, only the score functions of the reconstructed sources $Y_1, \ldots, Y_K$ are involved. Different methods for score function estimation thus provide different algorithms. Note that unlike the linear mixture case where the score functions need not to be accurately estimated, since the estimation equations are still satisfied even with a wrong score function, in the present case, the accuracy of the score function estimation is crucial, since the estimating equations are *not satisfied when the score functions are wrong*. Several score function estimation methods are available: the kernel method (used by Taleb and Jutten [6]), the spline method [7, 8]. We present here a new method derived from an entropy estimator. It is suggested by the following expansion of the entropy

$$H(T + \Delta) - H(T) = E[\Delta\psi_T(T)] + \text{terms of higher order in } \Delta \quad (3)$$

where $T$ is a random variable, $\Delta$ is a small random increment and $\psi_T$ denotes the score function of $T$. Let $t_1, \ldots, t_N$ be a sample from $T$ and $\mathcal{H}(t_1, \ldots, t_N)$ an estimator of $H(T)$ based on this sample. The relation (3) suggests estimating $\psi_T$ at the sampling points $t_n$ by

$$\hat{\psi}_T(t_n) = N\frac{\partial\mathcal{H}(t_1, \ldots, t_N)}{\partial t_n}, \qquad 1 \leq n \leq N.$$

It is desirable that $\mathcal{H}$ be invariant with respect to translation and equi-variant with respect to scaling, that is $\mathcal{H}(t_1 + c, \ldots, t_N + c) = \mathcal{H}(t_1, \ldots, t_N)$ and $\mathcal{H}(ct_1, \ldots, ct_N) = \mathcal{H}(t_1, \ldots, t_N) + \log|c|$ for any real number $c$, since the entropy functional satisfies these properties. By differentiating the above equalities with respect to $c$ then putting $c = 0$ and $c = 1$, one gets

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_T(t_i) = 0 \quad (4)$$

and

$$\frac{1}{N}\sum_{i=1}^{N}t_i\hat{\psi}_T(t_i) = 1. \quad (5)$$

The above interesting properties of $\hat{\psi}_T$ mimic that of the true score function: $E[\psi_T(T)] = 0$ and $E[T\psi_T(T)] = 1$, which can be easily obtained by integration by part.

As an estimator of the entropy, one may take a discretization of its defining integral:

$$\mathcal{H}(t_1, \ldots, t_N) = -\sum_l [\log\hat{p}_T(\hat{\mu}_T + lb\hat{\sigma}_T)] \times \hat{p}_T(\hat{\mu}_T + lb\hat{\sigma}_T)b\hat{\sigma}_T + \log\hat{\sigma}_T$$

where $\hat{p}_T$ is an estimate of the density of $T$, $\hat{\mu}_T$ and $\hat{\sigma}_T$ are the sample mean and standard deviation of $T$, and $b$ represents the discretization step. Taking $\hat{p}_T$ to be a kernel estimator of the density of $T$ with the smoothing parameter proportional to its sample standard deviation, one can show that the estimator $\mathcal{H}$ possesses the desired translation invariance and scale equi-variance properties.

One can also use an empirical estimator of the mean, leading to

$$\mathcal{H}(t_1, \ldots, t_N) = -\frac{1}{N}\sum_{i=1}^{N}\log\hat{p}_T(t_i).$$

This estimator $\mathcal{H}$ again possesses the desired translation invariance and scale equi-variance properties, if one takes $\hat{p}_T$ to be a kernel estimator of the density of $T$ with a smoothing parameter proportional to its sample standard deviation. This second form of the entropy estimator, although appears simpler than the first, can actually be costlier computationally. The reason is that the first form involves only the estimated density at a *regular grid points* and computation of the kernel density over such a grid can be efficiently done through the *binning* technique. Further, the number of the grid points can usually be much less than the sample size $N$ without incurring any appreciable loss of accuracy.

Finally, the probability density function $p_{Z_k}$ and the probabilities $P(\zeta_{k,m} \leq Z_k < \zeta_{k,m+1})$, which are needed in the algorithm, can also be estimated via the estimated score function. Indeed, noting that the density $p_T$ of $T$ satisfies $p_T(t) = -E[\psi_T(T)\mathbb{1}_+(T-t)]$, we propose to estimate it by

$$\frac{1}{N}\sum_{n=1}^{N} \hat{\psi}_T(t_n)\mathbb{1}_+(t_n - t)$$

This expression can be seen to be constant in $t$ in each interval $]t^{(m)}, t^{(m+1)}[$, $m = 1, \ldots, N-1$, where $t^{(1)} \leq \cdots \leq t^{(N)}$ are the order statistics of the sample $t_1, \ldots, t_N$. By integration, one obtains the estimator

$$\hat{P}_T(t^{(m)} \leq T \leq t^{(n+1)}) = \frac{t^{(m)} - t^{(m+1)}}{N}\sum_{n=1}^{m} \hat{\psi}_T(t^{(n)}).$$

## 5. NUMERICAL IMPLEMENTATION

### 5.1. Algorithms

Let $x_{k,n}$, $n = 1, \ldots, N$, denote the observations recorded by the $k$-th sensor and put $z_{k,n} = g_k(x_{k,n})$. The algorithm developed by Taleb and Jutten [9] changes at each iteration the matrix $\mathbf{B}$ and the $z_{k,n}$. Based on the previous expression of the relative gradient, the proposed method adapts the matrix $\mathbf{B}$ and the differences $z_k^{(2)} - z_k^{(1)}, \ldots, z_k^{(N)} - z_k^{(N-1)}$ instead, where $z_k^{(m)} = g_k(x_k^{(m)})$ and $x_k^{(1)} \leq \ldots \leq x_k^{(N)}$ are the order statistics of $X_k$. The values of the $z_{k,n}$, and hence of the $y_{k,n}$ can be recovered by centering, for example.

The proposed algorithm, in the nonparametric approach with probabilistic metric, is described in the boxed text below, in which $\hat{P}_{k,m}$ is an estimation of

$$\begin{aligned} P_{k,m} &= P(x_k^{(m)} \leq X_k < x_k^{(m+1)}) \\ &= P(z_k^{(m)} \leq Z_k < z_k^{(m+1)}), \end{aligned}$$

$\{\pi_k(n)\}$ is the permutation defined by $z_k^{(m)} = z_{k,\pi_k(m)}$ and $\mu$ and $\lambda$ are two *"small"* tuning parameters. The algotrihms with the normal metric and in the semi-parametric approach are quite similar, so we don't detail them.

**Initialisations :**

$\mathbf{B} = I$

$Y = Z = X$

**loop :**

$\hat{\psi}_{Y_i}$ an estimation of $\psi_{Y_i}$,$i = 1, \ldots, K$.

$z_k^{(m+1)} - z_k^{(m)} \mapsto [z_k^{(m+1)} - z_k^{(m)}] \times$

$$\left[1 + \mu - \mu \frac{z_k^{(m+1)} - z_k^{(m)}}{\hat{P}_{k,m} N} \sum_{n=m+1}^{N} \sum_{i=1}^{K} \hat{\psi}_{Y_i}(y_{i,\pi_k(n)})\mathbf{B}_{ik}\right]$$

$k = 1, \ldots, K$, $n = 1, \ldots, N$

Normalization of $Z$.

$\mathbf{B} \mapsto \mathbf{B} - \lambda \hat{\mathbf{D}}\mathbf{B}$.

where $\hat{\mathbf{D}}_{ij} = \begin{cases} \frac{1}{N}\sum_{n=1}^{N} \hat{\psi}_{Y_i}(y_{i,n}) y_{j,n} & 1 \leq i \neq j \leq K, \\ 0 & \text{sinon.} \end{cases}$

$Y = \mathbf{B}Z$.

Normalization of $Y$.

**until convergence.**

### 5.2. Normalisations

It should be noted that there is a redundancy in the post nonlinear mixture model: the nonlinear transformations $f_1$, $\ldots$, $f_K$ can be multiplied by arbitrary constant factors and the corresponding columns of the matrix $\mathbf{A}$ divided by the same constants, without changing the mixtures. To avoid this ambiguity, which affects the convergence of the algorithm, one should normalize the functions $g_1$, $\ldots$, $g_K$ (or equivalently the rows of $\mathbf{B}$). Many different possibilities exist, the first which may come to mind is to require that the sample variance of $Z_k$ be 1 ($k = 1, \ldots, K$). But we find more convenient to require that $g_1, \ldots, g_K$ preserve the entropy, that is $\int \log|g_k'(x)| p_{X_k}(x)dx = 0$ ($k = 1, \ldots, K$), since our algorithm works with the derivatives and hence the last integral can be easily estimated. This still leaves the sign of $g_k'$ undetermined. For definiteness, we require that $g_k'$ be positive, which can be easily checked by looking at $z_k^{(m+1)} - z_k^{(m)} > 0$, $m = 1, \ldots, N-1$.

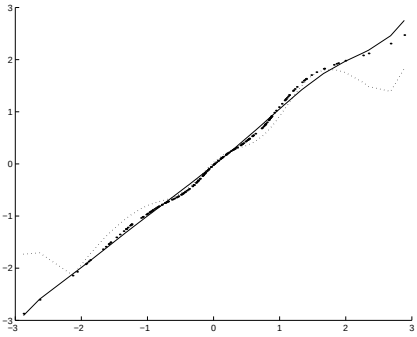
In blind source separation, it is well known that one can only recover the sources up to a scaling and a permutation. However, instead of normalizing $Y_1, \ldots, Y_K$ at each iteration step, we shall modify our algorithm to make it invariant with respect to scale change, by taking as $\hat{\mathbf{D}}$

$$\begin{bmatrix} \hat{D}_{ij} \\ \hat{D}_{ji} \end{bmatrix} = \begin{bmatrix} \hat{E}[\hat{\psi}_{Y_i}(Y_i)Y_j]/\widehat{\text{var}}(Y_j)\,\hat{E}[\hat{\psi}'_{Y_i}(Y_i)] \\ \hat{E}[\hat{\psi}_{Y_j}(Y_j)Y_i]/\widehat{\text{var}}(Y_i)\,\hat{E}[\hat{\psi}'_{Y_j}(Y_j)] \end{bmatrix}.$$

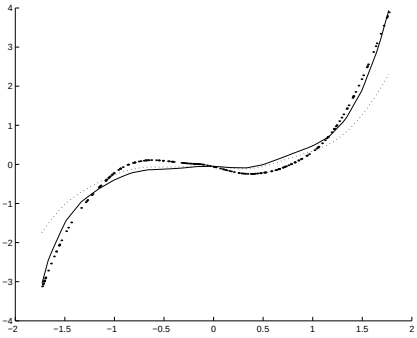
The denominators in the above formula are actually the diagonal terms of the Hessian of the criterion with respect to the linear transformations.

## 6. EXPERIMENTAL RESULTS

Figure 1 presents some comparison of the score function estimation by three different methods. One can see that the method with the splines and with the entropy discretization are rather similar. But the method with kernels is rather bad near the end points of the interval. This may be explained by the fact that the first two method possess the properties (4) and (5) while the kernel method does not.



Estimation of score function of a gaussian variable.



Estimation of score function of an uniform variable.

**Fig. 1**. Score function estimations with different methods: dotted line: kernel method, dashed line: spline method, solid line: method with discretization of entropy.

Next, we present in figure 2, the results of a simulation of the semi-parametric method with:

- a grid with 10 points among the observations.
- the probabilistic metric.
- the score function estimation with the method of the entropy discretization.
- $\mu = 0.2$, $\lambda = 0.5$.
- 32 iterations.
- $\mathbf{A} = \begin{pmatrix} 1 & 0.6 \\ 0.7 & 1 \end{pmatrix}$ and $f_1(x) = f_2(x) = \tanh(4x) + 0.1x$

Only the result of one simulation is presented here to save space. Others simulations performed have shown that the performances of the algorithms depend on the mixture components and the estimation of the score function. The two metrics also yield different effects on the convergence of the algorithms. The coefficients $\mu$ and $\lambda$ which control the gradient descent are fixed empirically.

## 7. DISCUSSION

A first comment regards the values of $h'_k$ at the end points of the interval $[z_k^{(1)}, z_k^{(N)}]$. We observ that these values are often quite large in comparison with the others. This may be attributed to the fact that there are too few data to achieve a good estimation of the score function near these points. To avoid instability of the algorithm caused by an unduly large value of $h'_k$ in usually a quite narrow region, we transform non linearly $h'_k$ into $\tanh(h'_k)$. When $h'_k$ is small this doesn't have any effect since then $\tanh(h'_k) \approx h'_k$, but when $h'_k$ is large, the hyperbolic tangente function reduces it to $\pm 1$. We find that this device can improve drastically the convergence of the algorithm.

Another comment regards the coefficients $\lambda$ and $\mu$ which control the gradient descent. It seems that keeping them fixed is not a good strategy. When one approaches the solution, these coefficents appear to be far too large, causing an oscilatory behaviour of the algorithm and may destroy its convergence. To avoid that, they must be drastically reduced with the consequence that the algorithm becomes extremely slow. A good strategy is to adapt $\lambda$ and $\mu$ such that the criterion is decreased at each iteration. Since the theoretical criterion $C$ is unknown, an estimator $\hat{C}$ of it must be used. But then, in order to ensure that $\hat{C}$ can be decreased, the relative gradient involved in the algorithm must be the gradient of $\hat{C}$ and not an estimated relative gradient of $C$. Naturally, one would estimate $\hat{C}$ by the same formula (2) but with $H(Y_k)$ and $H(Z_k)$ replaced by their estimators $\mathcal{H}(y_{k,1}, \ldots, y_{k,N})$ and $\mathcal{H}(z_{k,1}, \ldots, z_{k,N})$. Then it can be shown that the relative gradient of $\hat{C}$ is given by the same formula that for the relative gradient of $C$ in section 2.1, except that the score functions $\psi_{Y_k}$ and $\psi_{Z_k}$ are replaced by the score estimators of $Y_k$ and of $Z_k$ as defined in section 4 and that the expectation operator is replaced by a sample average. Simulations show that with this method, the algorithms are much more stable.

## 8. CONCLUSION

In this paper, we provide alternative methods for blind source separation in post nonlinear mixtures. Although we use the same mutual information as Taleb and Jutten, our algorithms differ in that they work with the derivatives of the nonlinear transformations and are based on the relative gradient of the criterion with respect to these derivatives. This approach can be extended easily to a semi-parametric setting in which the nonlinear transformations are represented by continuous piecewise linear finctions. The method has the nice property that it involves only certain estimated score functions and with an adequate choice of the latter, it amounts to minimizing some empirical criterion. This can be exploited to better control the convergence of the gradient descent. All the algorithms were implemented and tested in different situations.

## 9. REFERENCES


[1] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, 1995.

[2] D.-T. Pham, "Blind separation of instantaneous mixture of sources via an independent component analysis," *IEEE Transactions on Signal Processing*, vol. 44, no. 11, pp. 2768–2779, Nov. 1996.

[3] P. Comon, "Independent component analysis, a new concept ?," *Signal Processing*, vol. 3, no. 36, pp. 287–314, Apr. 1994.

[4] J.-F. Cardoso, "Blind signal separation : Statistical principles.," *Proceedings IEEE*, vol. 10, no. 86, pp. 2009–2025, Oct. 1998.

[5] G. Darmois, "Analyse gnrale des liaisons stochastiques," *Rev. Inst. Internat. Stat.*, vol. 21, pp. 2–8, 1953.

[6] A. Taleb and C. Jutten, "Sources separation in post-nonlinear mixtures," *IEEE Transactions on Signal Processing*, vol. 10, no. 47, pp. 2807–2820, Oct. 1999.

[7] D. D. Cox, "A penalty method for nonparametric estimation of the logarithmic derivative of density function," *Ann. Inst. Statist. Math.*, vol. 37, no. Part A, pp. 271–288, 1985.

[8] Pin T. Ng., "Smoothing spline score estimation," *S.I.A.M. J. Sci. Compt.*, vol. 15, no. 5, pp. 1003–1025, Sept 1994.

[9] A. Taleb, *Séparation de Sources dans les Mélanges Non Linéaires*, Ph.D. thesis, I.N.P.G. - Laboratoire L.I.S., 1999.

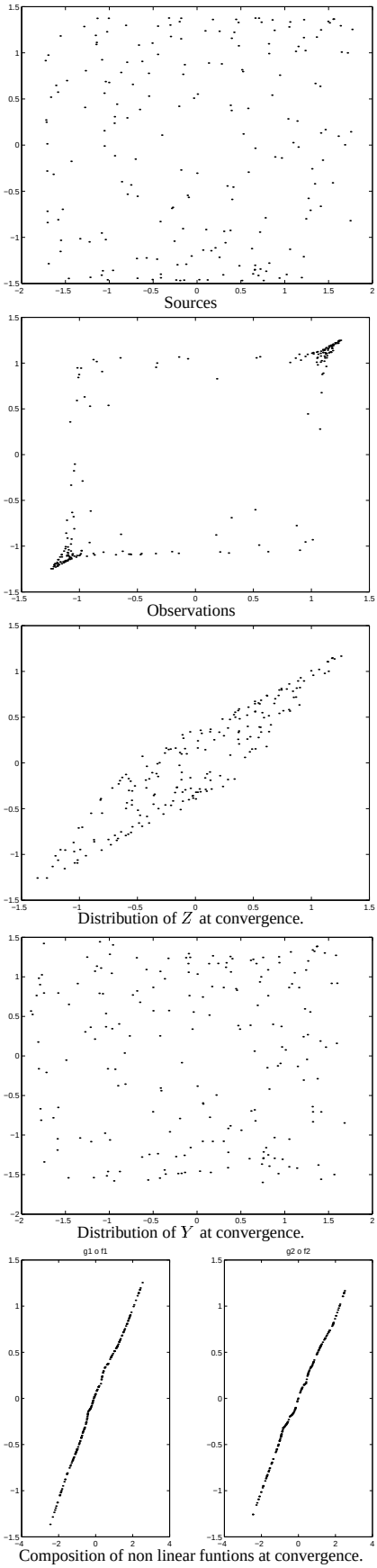




**Fig. 2**. Simulation