QUADRATIC DEPENDENCE MEASURE FOR NONLINEAR BLIND SOURCES SEPARATION

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 focuses on a quadratic dependence measure which can be used for blind source separation. After defining it, we show some links with other quadratic dependence measures used by Feuerverger and Rosenblatt. We develop a practical way for computing this measure, which leads us to a new solution for blind source separation in the case of nonlinear mixtures. It consists in first estimating the theoretical quadratic measure, then computing its relative gradient, finally minimizing it through a gradient descent method. Some examples illustrate our method in the post nonlinear mixtures.

1. INTRODUCTION

Blind source separation (BSS) consists in extracting independent sources from their mixtures without relying on specific assumptions about the mixture and the sources distribution *other than their independence*. Therefore most methods which have been proposed are based on minimizing some criterion related to independence. Such criterion often possesses the contrast property in the sense that it can be minimized if and only if the output of the separation system are mutually independent [1, 2]. In the context of linear mixtures, contrast functions can be constructed from cumulants [1] or even correlations if lagged correlations are included [3, 4]. This is possible because of the strong constraint of linearity of the mixture, since it is well known that the independence between a set of random variables cannot in general be inferred from the fact that some of their correlations and cumulants are zero. (One needs to consider all of them.) In the nonlinear mixtures problem, it is therefore of interest to consider dependence measures which completely characterizes independence, in the sense that the measure can be zero if and only if independence has been achieved. Of course, such measure can be of interest in the linear mixture context too.

The mutual information is a well known and widely used dependence measure. Its use in nonlinear BSS has been introduced in Taleb and Jutten [5, 6] and Babaie-Zadeh [7], among others. This measure is however difficult to estimate, as it involves the estimation of entropy which requires density estimation. This can cause severe difficulty for high dimensional data. Although it is possible to reduce a criterion based on mutual information to the one based only on the marginal entropies, this approach can lead to large bias due to bias in density estimation. For these reasons, it could be of interest to consider other dependence measures. Such a measure is considered in Murata [8] and in Eriksson, Kankainen and Koivunen [9], which can be traced back to Rosenblatt, [10] and another one in Jordan [11]. The latter measure does not seem easy to compute, while the former is nothing but a weighted sum of squared distances between the joint characteristic function and the product of the marginal characteristic functions, and is thus much simpler. This measure had been used by Feuerverger [12] to construct a test of dependence.

In this paper we study the above dependence measure, which we call quadratic dependence measure. We also make some extensions by introducing the use of scale factors (to make the measure scale invariant) and the use of general kernel functions. Further, we derive the gradient of our criterion in the general context of nonlinear BSS problem and we investigate the estimation of this gradient as well as the criterion itself. The criterion can thus be minimized through a gradient descent and one can control the step size of the algorithm to ensure that it decreases at each step. Then we focus on the particular case of the post nonlinear (PNL) mixture model. Some simulation results are given showing the good performance of the algorithm.

Thanks to the European Bliss Project for funding.

Section 2 defines the quadratic dependence measure and provides some interpretations. The next section applies this measure to the nonlinear BSS problem by providing a practical formula for the computation of an empirical version of the criterion and of its gradient. The special case of PNL is developed in greater detail in section 4 and some experimental results are provided in the last section.

2. THE QUADRATIC DEPENDENCE MEASURE

We first recall that a set of K random variables $Y_1, \dots, Y_K$ are mutually independent if and only if

$$E\left[\prod_{k=1}^{K} f_k(Y_k)\right] = \prod_{i=1}^{K} E[f_k(Y_k)],$$

for any set of K summable functions of a real variable f_1, $\dots$, f_K. In fact, taking f_i as the indicator functions of some measurable set, the above property is nothing but the usual definition of independence. That it also holds for arbitrary summable function is an easy extension. Thus we do not need to check the above equation for *all summable functions*, but only for a much smaller class of functions. A well known example is given by the class of complex exponential functions. Indeed, taking f_k to be of the form $f_k(x) = \exp(it_k x)$ for some real number t_k, the above equality means that the joint characteristic function of Y_1, $\dots, Y_K$ and the product of their marginal characteristic functions are equal. But it is well known that this property is a necessary and sufficient condition of mutual independence. In general there are many possibilities to construct a set of "test" functions $f_1, \dots, f_K$ which can characterize independence. A general method of construction is given in the following Lemma.

Lemma 2.1 *Let $\mathcal{K}$ be a summable kernel function with Fourier transform different from zero almost everywhere. Then the random variables $Y_1, \dots, Y_K$ are independent if and only if*

$$E\left[\prod_{i=1}^{K} \mathcal{K}(x_i - Y_i)\right] = \prod_{i=1}^{K} E[\mathcal{K}(x_i - Y_i)], \quad \forall x_1, \dots, x_K$$

To prove this result, one simply remarks that the Fourier transform of a convolution product of functions is a product of the Fourier transforms of the functions. Therefore the above equality is equivalent to

$$\prod_{k=1}^{K} \psi_{\mathcal{K}}(t_k)\psi_{Y_1,\dots,Y_k}(t_1, \dots, t_K) = \prod_{k=1}^{K} \psi_{\mathcal{K}}(t_k)\psi_{Y_k}(t_k)$$

for all $t_1, \dots, t_K$, where $\psi_{\mathcal{K}}$ denotes the Fourier transform of $\mathcal{K}$ and $\psi_{Y_1,\dots,Y_k}$ and ψ_{Y_k} denote the joint and marginal characteristic functions of $Y_1, \dots, Y_K$, respectively. This yields the desired result.

The above result leads to the definition of a quadratic measure of dependence, as follows

Definition 2.1 *Let $\mathcal{K}$ be a square summable kernel function with Fourier transform different from zero almost everywhere. For a set of K random variables $Y_1, \dots, Y_K$, we define the quadratic measure of their (mutual) dependence as*

$$Q(Y_1, \dots, Y_K) = \frac{1}{2}\int D_{\mathbf{Y}}(y_1, \dots, y_K)^2 dy_1 \dots dy_K.$$

where $\mathbf{Y} = [Y_1 \quad \cdots Y_K]^T$ *and*
$D_{\mathbf{Y}}(y_1, \dots, y_K) =$

$$E\left[\prod_{k=1}^{K} \mathcal{K}\left(y_k - \frac{Y_k}{\sigma_{Y_k}}\right)\right] - \prod_{k=1}^{K} E\left[\mathcal{K}\left(y_k - \frac{Y_k}{\sigma_{Y_k}}\right)\right]$$

σ_{Y_k} is a scale factor, that is a positive functional of the distribution of Y_k such that $\sigma_{\lambda Y_i} = |\lambda|\sigma_{Y_k}$, for all real constant λ.

It is worthwhile to note that in the above definition we assume that the kernel function is square summable (and not necessarily summable). We shall see later that this condition is sufficient for the integral in this definition to be well defined. Note also the presence of the scale factors σ_{Y_k} whose purpose is to make the measure scale invariant (multiplying the random variables Y_k by arbitrary constants does not change this measure).

From the above Lemma, the functional Q is clearly a dependence measure which completely characterizes independence: $Q(Y_1, \dots, Y_K) = 0$ if and only if $Y_1, \dots, Y_K$ are mutually independent. This measure may be related to those introduced by Feuerverger [12] or Rosenblatt [10].

A situation of interest is when the kernel $\mathcal{K}$ is an approximation to the Dirac function. Thus take $\mathcal{K}(u) = \tilde{\mathcal{K}}(u/h)/h$, where h is a (small) bandwidth parameter and $\tilde{\mathcal{K}}$ a density function. It is clear that $E[\prod_{i=1}^{K} \mathcal{K}(y_i - Y_i/\sigma_{Y_i})]$ and $\prod_{i=1}^{K} E[\mathcal{K}(y_i - Y_i/\sigma_{Y_i})]$ converge respectively to the joint density and the product of the marginal densities of the random variables $Y_1/\sigma_{Y_1}, \dots, Y_K/\sigma_{Y_K}$ when h tends to zero. Therefore, for small h, Q appears as the quadratic distance between an approximate smoothed version of the joint density function and the product of the marginal density functions, of the random variables $Y_1/\sigma_{Y_1}, \dots, Y_K/\sigma_{Y_K}$. But as it is shown above, Q is always a dependence measure whether h is small or not. In density estimation, the choice of the bandwidth parameter h must realize a compromise between bias and variance (which is not easy). Here we may choose h larger since we need not bother very much about bias, as we do not really need Q to be exactly a quadratic distance between the densities.

One can also express the measure Q in terms of the characteristic functions. After applying the Parseval formula (which states that the Fourier transform conserves the norm in L^2) and a change of integration variable :
$Q(Y_1, \dots, Y_K) =$

$$\frac{1}{2}\int \prod_{k=1}^{K} \left|\frac{\sigma_{Y_k}\psi_{\mathcal{K}}(\sigma_{Y_k}t_k)}{2\pi}\right|^2 |D_{\mathbf{Y}}^c(\mathbf{t})|^2 dt_1 \dots dt_K. \quad (1)$$

where $\mathbf{Y} = [Y_1 \quad \cdots Y_K]^T$ and

$$D^c_{\mathbf{Y}}(t_1, \ldots, t_K) = \psi_{\mathbf{Y}}(t_1, \ldots, t_K) - \prod_{k=1}^{K} \psi_{Y_k}(t_k) \qquad (2)$$

The measure (1) has been considered by Eriksson et al. [9] and Feuerverger [12], but only in the particular case where $\mathcal{K}$ is the Gaussian kernel and $\sigma_{Y_k} = 1$.

Equation (1) shows that Q is well defined as soon as the function $\psi_{\mathcal{K}}$ is square summable, since the characteristic functions are bounded. That justifies our previous assumption on $\mathcal{K}$ when we define Q.

Thus we have at our disposal a whole class of quadratic measures, depending on the choice of the kernel $\mathcal{K}$ and also on the bandwidth h if we choose the kernel to be a scaled kernel of the form $\tilde{\mathcal{K}}(\cdot/h)/h$. Let us stress that the kernel $\mathcal{K}$ does not need to be a density, and h does not need to be very small. Thus we have a lot of degrees of freedom in choosing them. Since we do not know how these choices will affect the performance of the method, we will have to choose them in an *ad hoc* manner.

3. BLIND SOURCE SEPARATION VIA THE QUADRATIC DEPENDENCE MEASURE

We now apply the quadratic dependence measure defined above to the nonlinear BSS problem. Here the observations $X_1, \ldots, X_K$ are related to the independent sources $S_1, \ldots, S_K$ through the relation,

$$\mathbf{X} = (X_1, \ldots, X_K)^T = f(\mathbf{S})$$

where $\mathbf{S} = (S_1, \ldots, S_K)^T$ and f is an invertible mapping. However, without further assumptions on the mapping f, it would be impossible to reconstruct the sources based on the sole hypothesis of their independence. Therefore one has to assume that the mapping f belongs to a restricted subclass of invertible mappings $\mathcal{F}$. For example, $\mathcal{F}$ can be the class of linear invertible maps, in which case we are in presence of linear mixtures. In the case of nonlinear mixtures, only the PNL mixtures will be investigated in detail.

Our goal is to find a mapping g in $\mathcal{F}^{-1}$, so that the components Y_k of $\mathbf{Y} = g(\mathbf{X})$ are as independent as possible, according to the quadratic measure Q. However Q is a theoretical measure which depends on the unknown distribution of the data. Therefore, two methods can be investigated :

1. replace Q by an estimate and then proceed to minimize the estimated criterion instead, which would (generally) require to compute its gradient.
2. proceed with the theoretical criterion, compute its gradient to obtain a system of equations and only then replace the unknown gradient by some estimator to get an estimating system of equations.

We prefer the first approach as it provides us a mean to control the decrease of the (empirical) criterion at each step of the algorithm. Thus it results in a more robust algorithm in terms of convergence properties.

In the sequel, the observed data will be denoted by $X_k(n)$, $n = 1, \ldots, N$, $k = 1, \ldots, K$, N being the sample size. The corresponding samples of Y_k are then $Y_k(n)$, the k-th components of $\mathbf{Y}(n) = g(\mathbf{X}(n))$.

3.1. Estimation of Q

Let us remark that the dependence measure Q involves only the expectation operator E. Thus a natural estimator of Q can be obtained by just replacing this operator with the sample average $\widehat{E}$, defined as $\widehat{E}\phi(\mathbf{X}) = \sum_{n=1}^{N} \phi(\mathbf{X}(n))/N$, where ϕ is any function of the data. Thus, an estimate of Q will be,

$$\widehat{Q}(Y_1, \ldots, Y_K) = \frac{1}{2} \int \{\widehat{D}_{\mathbf{Y}}(y_1, \ldots, y_K)\}^2 dy_1 \ldots dy_K.$$

where,

$$\widehat{D}_{\mathbf{Y}}(y_1, \ldots, y_K) = \widehat{E}\left[\prod_{k=1}^{K} \mathcal{K}\left(y_k - \frac{Y_k}{\widehat{\sigma}_{Y_k}}\right)\right] - \prod_{k=1}^{K} \widehat{E}\left[\mathcal{K}\left(y_k - \frac{Y_k}{\widehat{\sigma}_{Y_k}}\right)\right]$$

and $\widehat{\sigma}_{Y_k}$ is an empirical version of σ_{Y_k} depending only on the $Y_k(n)$.

However, the above expression of $\widehat{Q}$ is not suitable for computation because it requires multiple integration. Fortunately, these integrations can be avoided by the use of an alternative formula. Define a new kernel $\mathcal{K}_2$ as $\mathcal{K}_2(u) = \int \mathcal{K}(u+v)\mathcal{K}(v)dv$. Then, after expanding the square of $\widehat{D}_{\mathbf{Y}}$, and interchanging the summation and integration :

$$\widehat{Q}(Y_1, \ldots, Y_K) = \frac{1}{2}\left\{\widehat{E}\widehat{\pi}_{\mathbf{Y}}(\mathbf{Y}) + \prod_{k=1}^{K} \widehat{E}\widehat{\pi}_{Y_k}(Y_k) - 2\widehat{E}\prod_{k=1}^{K} \widehat{\pi}_{Y_k}(Y_k)\right\}$$

where

$$\begin{aligned} \widehat{\pi}_{\mathbf{Y}}(\mathbf{y}) &= \frac{1}{N}\sum_{n=1}^{N}\prod_{i=1}^{K} \mathcal{K}_2\left(\frac{y_i - Y_i(n)}{\widehat{\sigma}_{Y_i}}\right) \\ \widehat{\pi}_{Y_k}(y_k) &= \frac{1}{N}\sum_{n=1}^{N} \mathcal{K}_2\left(\frac{y_k - Y_k(n)}{\widehat{\sigma}_{Y_k}}\right). \end{aligned}$$

The first remark is that one can get a similar formula as above for the theoretical measure Q. We do not need it so we do not write it down. The second remark is that Q depends on $\mathcal{K}$ only indirectly through $\mathcal{K}_2$, therefore we can choose $\mathcal{K}_2$ directly without ever considering $\mathcal{K}$. But $\mathcal{K}_2$ cannot be arbitrary. From its definition, $\mathcal{K}_2$ must be choosen such that its Fourier transform is a positive summable even function, since its Fourier transform corresponds to $|\psi_{\mathcal{K}}|^2$ where $\psi_{\mathcal{K}}$ is the Fourier transform of a real square summable function.

Some possible choices for $\mathcal{K}_2$ are: ($\mathcal{K}_2^*$ denoting the Fourier transform of $\mathcal{K}_2$)

1. The Gaussian kernel :
   $\mathcal{K}_2(x) = e^{-x^2}$, $\mathcal{K}_2^*(t) = \sqrt{\pi}e^{-t^2/4}$
2. The square Cauchy kernel:
   $\mathcal{K}_2(x) = 1/(1+x^2)^2$, $\mathcal{K}_2^*(t) = \pi(|t|+1)e^{-|t|}$
3. The negative of the second derivative of the square Cauchy kernel :
   $\mathcal{K}_2(x) = (20x^2-4)/(1+x^2)^4$,
   $\mathcal{K}_2^*(t) = 4t^2\pi^3(|t|+1)e^{-|t|}$.

The first two kernels correspond to density kernels (after normalizing) and differ only by their tail behavior. But the last kernel does not and can be negative. One may note that the kernel $\mathcal{K}_2$ is related to Mercer kernels which are used especially in Support Vector Machine [13].

3.2. Gradient of the empirical criterion

To minimize the criterion $\widehat{Q}$ with respect to the separating transformation g, we need to compute its gradient. Actually, it is more convenient to work with relative gradient. Therefore we consider a “small” relative increment of g, namely $\Delta \circ g$ where Δ is a small transformation. This would induce a small increment $\Delta(\mathbf{Y})$ of $\mathbf{Y}$ and $\Delta_k(\mathbf{Y})$ of Y_k.

Since $\widehat{Q}$ depends only on the normalized variables $Y_k/\widehat{\sigma}_{Y_k}$, we first compute the variation of $Y_k/\widehat{\sigma}_{Y_k}$ and then apply the result to derive the change in this criterion.

The change of the normalized variable $Y_k/\widehat{\sigma}_{Y_k}$ induced by a small change $\Delta_k(\mathbf{Y})$ in Y_k is $\widehat{\Delta}_k^*(\mathbf{Y}(n))/\widehat{\sigma}_{Y_k} + o(\Delta_k)$ where

$$\widehat{\Delta}_k^*(y) = \Delta_k(y) - y_k\widehat{E}[\widehat{\varphi}_{Y_k}(Y_k)\Delta_k(\mathbf{Y})]$$

and

$$\widehat{\varphi}_{Y_k}(Y_k(n)) = N\frac{\partial \log \widehat{\sigma}_{Y_k}}{\partial Y_k(n)}(Y_k(n)).$$

In practice, $\widehat{\sigma}_{Y_k}$ is often the standard deviation of Y_k, then $\widehat{\varphi}_{Y_k}(y) = (y - \widehat{E}(Y_k))/\widehat{\sigma}_{Y_k}$. The change of $\widehat{Q}(Y_1,\dots,Y_K)$, corresponding to a small relative increment $\Delta \circ g$ of g, is therefore, up to the first order, $\widehat{E}[\sum_{k=1}^K \widehat{G}_k(\mathbf{Y})\widehat{\Delta}_k^*(\mathbf{Y})]$ where

$$\begin{aligned}\widehat{G}_k(\mathbf{y}) \quad = \quad & \widehat{\pi}'_{k,\mathbf{Y}}(\mathbf{y}) - \widehat{\pi}'_{Y_k}(y_k)\prod_{l\neq k}\widehat{\pi}_{Y_l}(y_l) + \\ & \widehat{\pi}'_{Y_k}(y_k)\prod_{l\neq k}\widehat{E}(\widehat{\pi}_{Y_l}(Y_l)) \\ & -\frac{1}{\widehat{\sigma}_{Y_k}}\widehat{E}\left[\mathcal{K}'_2\left(\frac{y_k - Y_k}{\widehat{\sigma}_{Y_k}}\right)\prod_{l\neq k}\widehat{E}(\widehat{\pi}_{Y_l}(Y_l))\right]\end{aligned}$$

with

$$\begin{aligned}\widehat{\pi}'_{k,\mathbf{Y}}(\mathbf{y}) = \frac{1}{N\widehat{\sigma}_{Y_k}}\sum_{n=1}^{N}\mathcal{K}'_2\left(\frac{y_k - Y_k(n)}{\widehat{\sigma}_{Y_k}}\right) \\ \prod_{l\neq k}\mathcal{K}_2\left(\frac{y_l - Y_l(n)}{\widehat{\sigma}_{Y_l}}\right),\end{aligned}$$

being the k-th component of the gradient of $\widehat{\pi}_{\mathbf{Y}}$ and

$$\widehat{\pi}'_{Y_k}(y_k) = \frac{1}{N\widehat{\sigma}_{Y_k}}\sum_{n=1}^{N}\mathcal{K}'_2\left(\frac{y_k - Y_k(n)}{\widehat{\sigma}_{Y_k}}\right)$$

being the derivative of $\widehat{\pi}_{Y_k}$.

Finally, the relative gradient of $\widehat{Q}$ is given by the linear maps :

$$\Delta_1,\dots,\Delta_K \rightarrow \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N}\widehat{G}_k^*[\mathbf{Y}(n)]\Delta_k(Y_k(n))$$

where $\widehat{G}_k^*(\mathbf{Y}(n)) = \widehat{G}_k[\mathbf{Y}(n)] - \widehat{E}[Y_k\widehat{G}_k(\mathbf{Y})]\widehat{\varphi}_{Y_k(n)}[Y_k(n)]$

Note that similar calculations can be applied to obtain the relative gradient of the theoretical measure Q. One gets the same formula except that $\widehat{\varphi}_{Y_k}$ and $\widehat{G}_k$ are replaced by their theoretical expression $\varphi_{Y_k} = (\log \sigma_{Y_k})'$ and

$$\begin{aligned}G_k(\mathbf{y}) \quad = \quad & \pi_{k,\mathbf{Y}}(y_1,\dots,y_k) - \pi'_{Y_k}(y_k)\prod_{l\neq k}\pi_{Y_l}(y_l) \\ & + \pi'_{Y_k}(y_k)\prod_{l\neq k}E[\pi_{Y_l}(Y_l)] \\ & - E\left[\frac{1}{\sigma_{Y_k}}\mathcal{K}'_2\left(\frac{y_k - Y_k}{\sigma_{Y_k}}\right)\prod_{l\neq k}\pi_{Y_l}(Y_l)\right].\end{aligned}$$

where $\pi_{\mathbf{Y}}(\mathbf{z}) = E[\prod_{l=1}^K \mathcal{K}_2((z_l - Y_l)/\sigma_{Y_l})]$, $\pi_{k,\mathbf{Y}}$ is its partial derivative with respect to the k-th component and $\pi_{Y_k}(z_k) = E[\mathcal{K}((z_k - Y_k)/\sigma_{Y_k})]$ and π'_{Y_k} is its derivative.

As $N \rightarrow \infty$, $\widehat{G}_k$ tends to G_k and it is clear that G_k vanishes when the variables $Y_1,\dots,Y_k$ are independent. Therefore the estimating equations, obtained by setting the relative gradient of $\widehat{Q}$ to zero, are satisfied in the limit.

Another remark concerns a property of the relative gradient arising from the invariance with respect to translation of Q and $\widehat{Q}$: one has $E(G_k(Y)) = 0$ and $\widehat{E}(\widehat{G}_k[\mathbf{Y}]) = 0$.

4. APPLICATION TO PNL MIXTURES

4.1. The problem

Let us focus on the PNL mixture, in which case the set $\mathcal{F}$ consists of all mappings f of the form

$$f(s_1,\dots,s_K) = \begin{bmatrix} f_1(\sum_{k=1}^K A_{1k}s_k) \\ \vdots \\ f_K(\sum_{k=1}^K A_{Kk}s_k)\end{bmatrix}$$

where A_{ik} are the elements of an invertible matrix $\mathbf{A}$ and $f_1, \ldots, f_K$ are monotonous functions.

The separating system consists of a mapping g in $\mathcal{F}^{-1}$ of the form

$$g(x_1, \ldots, x_K) = \begin{bmatrix} \sum_{k=1}^{K} B_{1k} g_k(x_k) \\ \vdots \\ \sum_{k=1}^{K} B_{Kk} g_k(x_k) \end{bmatrix}$$

where B_{ik} are the elements of an invertible matrix $\mathbf{B}$ and $g_1, \ldots, g_K$ are monotonous functions.

In a non parametric approach, only the assumption of the monotony of the nonlinear mapping g_k is made. But this is generally not enough since some smoothness must be imposed on these mappings, unless this is done implicitly through the estimating equations. In a parametric or semi-parametric approach, the mappings g_k are parametrized by a vector parameter with potentially large dimension in the latter case, hence they are inherently smoothed. In fact, in our semi-parametric approach, the mappings g_k are represented by piecewise linear continuous functions so that the vector parameter consists of the slope of these functions in each "piece".

In the sequel, we note $Z_k(n) = g_k(X_k(n))$ so that $Y_k(n) = \sum_{j=1}^{K} B_{kj} Z_j(n)$.

4.2. Non parametric approach

The particular form of $\mathcal{F}$ and $\mathcal{F}^{-1}$ leads us to express the small relative increment Δ of g in term of a small relative change $\varepsilon\mathbf{B}$ of $\mathbf{B}$ and $\delta_k \circ g_k$ of g_k where ε is a "small" matrix and δ_k are "small" functions. The expression of Δ in terms of ε and δ_k can be seen to be, up to the first order,

$$\Delta_i(\mathbf{y}) = \sum_{j=1}^{K} \left[\varepsilon_{ij} y_j + \mathbf{B}_{ij} \delta_j \left(\sum_{k=1}^{K} \mathbf{B}_{jk}^{-1} y_k \right) \right]$$

where Δ_i denotes the i-th component of (the vector) Δ and ε_{ij} are elements of the matrix ε.

We take the scale function σ_{Y_k} to be the standard deviation. The expression for the relative gradient can then be obtained, in this non parametric approach, as follows,

$$\begin{aligned}(\varepsilon, \delta_1, \ldots, \delta_K) \quad \rightarrow \quad & \sum_{1 \le j \ne k \le K} (\widehat{\Gamma}_{kj} - \widehat{\Gamma}_{kk} \widehat{\Sigma}_{kj} / \widehat{\Sigma}_{kk}) \varepsilon_{kj} \\ & + \sum_{j=1}^{K} \widehat{E} \left[\left\{ \sum_{k=1}^{K} \widehat{G}_k^*(\mathbf{Y}) \mathbf{B}_{kj} \right\} \delta_j \left(\sum_{l=1}^{K} \mathbf{B}_{jl}^{-1} Y_l \right) \right]\end{aligned}$$

where $\widehat{\Gamma}_{kj} = \widehat{E}[\widehat{G}_k(\mathbf{Y}) Y_j]$, $\widehat{\Sigma}_{kj} = \widehat{E}[(Y_k - \widehat{E}(Y_k))(Y_j - \widehat{E}(Y_j))]$ and $\widehat{G}_k^*(\mathbf{y}) = \widehat{G}_k(\mathbf{y}) - (\widehat{\Gamma}_{kk} / \widehat{\Sigma}_{kk})(y_k - \widehat{E}(Y_k))$

However, the property of the criterion $\widehat{Q}$ to be invariant with respect to translation shows that it depends only on the spacings between successive values of the ordered statistics $Z_j(1 : N) \le \ldots \le Z_j(N : N)$ of the sample $Z_j(1), \ldots, Z_j(N)$, assuming without loss of generality that g_j are increasing functions. Consequently, the gradient depends only on the changes in spacings: $\widetilde{\delta}_j(m) = \delta_j(Z_j(m : N)) - \delta_j(Z_j(m-1 : N))$, $m = 2, \ldots, N$. This remark leads to work only with the derivatives of the functions g_k which yields some constraints on the smoothness. Explicitly, with some calculations, the relative gradient can be rewritten as the linear maps

$$\delta_j \rightarrow \frac{1}{N} \sum_{n=2}^{N} \left\{ \sum_{m=n}^{N} \sum_{k=1}^{K} \widehat{G}_k^*(\mathbf{Y}(o_{j,m})) \mathbf{B}_{kj} \right\} \widetilde{\delta}_j(m)$$

4.3. Parametric and semi-parametric approach

These methods consist in parameterizing each g_k with a vector parameter θ_k. The expression of the relative gradient is now,

$$\delta_j \rightarrow \frac{1}{N} \sum_{n=2}^{N} \left\{ \sum_{m=n}^{N} \sum_{k=1}^{K} \widehat{G}_k^*(\mathbf{Y}(o_{j,m})) \mathbf{B}_{kj} \right\} \tilde{g}_{j,\theta_j}(n)$$

where $\tilde{g}_{j,\theta_j}(n) = \dot{g}_{j,\theta_j}(Z_j(n : N)) - \dot{g}_{j,\theta_j}(Z_j(n-1 : N))$ and $\dot{g}_{j,\theta_j} = g'_{j,\theta_j} \circ g_{j,\theta_j}^{-1}$, g'_{j,θ_j} being the ordinary gradient of g_{j,θ_j} with respect to θ_j.

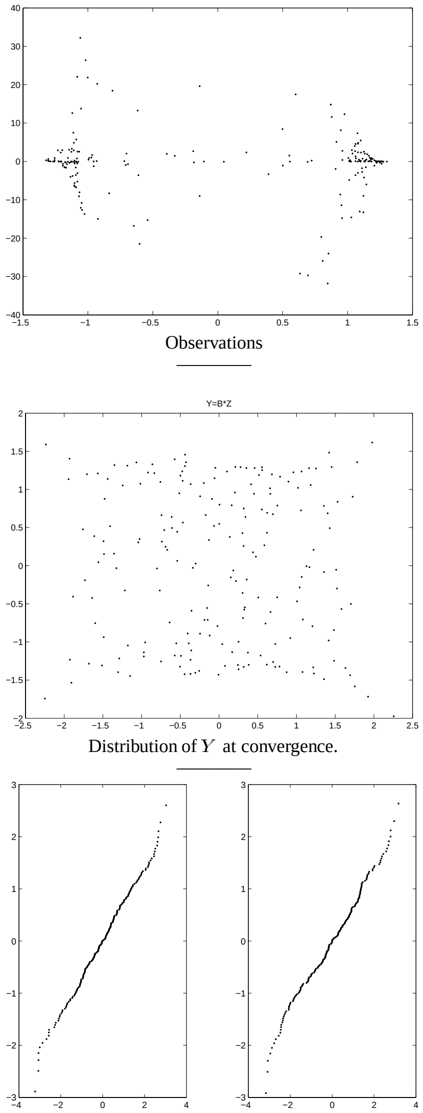
We have developed two different parameterizations. The first one consists in approximating the nonlinear transformations with piecewise linear functions, which yields the method called semi-parametric. The second one consists in using an approximation based on the quantile function. These two methods are not explained here due to lack of space. We present only an experimental result using them.

5. AN EXPERIMENTAL RESULT

We present on figure 1, the results of a simulation of the semi-parametric method with:

- a grid with 10 bins among the observations.
- $\mathcal{K}_2(x) = (4 - 20x^2)/(1 + x^2)^4$ (given at the end of part 3.1).
- $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$, $f_1(x) = \tanh(4x) + 0.1x$ and $f_2(x) = x^3 + 0.1x$

Here we present only the result of one simulation. We have also performed others simulations which have indicated that the performances of the algorithms depend on the linear mixture and that the choice of the kernel is very important. In particular, kernels with heavy tails seem to yield better results. Here, we present an example using a kernel, written above, with heavy tails which is not a density. In future works, we will attempt to relate some properties of these different kernels to the performance of the algorithms. The coefficients μ and λ which control the gradient descent in our experiment are fixed empirically : $\mu = 0.2$ and $\lambda = 1$. On figure 1, the top graph represents the distribution of the observations, and the middle one the distribution of the reconstructed sources. Finally, the bottom graph represents the compensation of the nonlinear functions $g_k \circ f_k$.



Observations

Distribution of Y at convergence.

Composition of non linear funtions at convergence.

Fig. 1. Simulation

6. CONCLUSION

In this paper, a quadratic measure is used for blind source separation. Its estimation based on the sample average and a computation trick leads to compute its gradient without requiring numerical integration. Then, its minimization is achieved through a gradient descent method. The algorithms has been implemented for PNL mixtures and tested in different situations. But we only present in this paper a simple simulation. Further perspectives of this work consist of choosing the shape of the kernels so as to improve the performances of the algorithms.

7. REFERENCES

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

[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] D.-T. Pham, “Blind separation of instantaneous mixture of sources via the gaussian mutual information criterion,” *Signal Processing*, vol. 81, pp. 855–881, 2001.

[4] A. Ziehe, M. Kawanabe, S. Harmeling, and K.R. Müller, “Separation of post-nonlinear mixtures using ace and temporal decorrelation,” *Proceedings ICA 2001 San Diego*, Dec. 2001.

[5] 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.

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

[7] M. Babaie-Zadeh, *On blind source separation in convolutive and nonlinear mixtures*, Ph.D. thesis, I.N.P.G. - Laboratoire L.I.S., 2002.

[8] N. Murata, “Properties of the empirical characteristic function and its application to testing for independence,” *Proceedings ICA 2001 San Diego*, Dec. 2001.

[9] J. Eriksson, A. Kankainen, and V. Koivunen, “Novel characteristic function based criteria for ica,” *Proceedings ICA 2001 San Diego*, Dec. 2001.

[10] M. Rosenblatt, “A quadratic measure of deviation of two-dimensional density estimates and a test of independence,” *The Annals of Statistics*, vol. 3, no. 1, pp. 1–14, 1975.

[11] F.R. Bach and M. I. Jordan, “Kernel independent component analysis,” *Journal of Machine Learning Research*, vol. 3, pp. 1–48, Jul. 2002.

[12] A. Feuerverger, “A consistent test for bivariate dependence,” *Internatinal Statistical Review*, vol. 61, no. 3, pp. 419–433, 1993.

[13] V. Vapnik, *The nature of statistical Learning Theory*, Springer, 1998.