Criteria based on mutual information minimization for blind source separation in post nonlinear mixtures

Sophie Achard, [a,1] Dinh-Tuan Pham, [a] Christian Jutten [b]

[a] *Univ. of Grenoble, IMAG, C.N.R.S.*
Laboratory of Modelling and Computation
B.P. 53X, 38041 Grenoble Cedex, France
sophie.achard@imag.fr, dinh-tuan.pham@imag.fr

[b] *Univ. of Grenoble, INPG, C.N.R.S.*
Laboratory of Images and Signals
46 avenue Félix Viallet
38031 Grenoble Cedex, France
christian.jutten@inpg.fr

Abstract

This work deals with the problem of blind source separation solved by minimization of mutual information. After having chosen a model for the mixture, we focus on two methods. One is based on the minimization of an estimation of I, the mutual information. The other one uses a minimization of an estimation of C, the mutual information after transforming all the joint entropy terms. We show the differences between these two approaches by studying statistical properties of the two estimators.

In this paper, we derive the bias of the estimators of the two criteria I and C. It is shown that under the hypothesis of independence, the estimator of I is asymptotically unbiased even if the bandwidth is kept fixed, whereas with a fixed bandwidth, the estimator of C is not asymptotically unbiased.

Further, the minimization is achieved by a relative gradient descent method and we show the differences between criteria I and C through the expression of their relative gradients.

Key words: Mutual information, biased and unbiased estimator, entropy, blind source separation, post nonlinear mixture.
PACS:

[1] present address:Brain Mappin Unit, University of Cambridge, Addenbrooke's Hospital, Cambridge, CB2 2QQ, UK, sa428@cam.ac.uk

1 Introduction

Clearly, the resolution of the problem of blind source separation (BSS) based on the sole hypothesis of the independence of sources requires a measure of dependence. For linear mixtures, several BSS methods are based on mutual information [16], maximum correlation [5], cumulants [10,7], characteristic functions [11], for most exhaustive references see the recent books [13,9] . The estimation of the mutual information however involves estimators of both the marginal and the joint entropies which in turn requires the estimations of marginal and joint densities. Especially, joint density estimation in a high dimensional space is difficult, because of the "curse of dimension". Usually, for overcoming this problem, the estimation of joint entropy and hence that of joint density, is avoided by expressing the joint entropy of the reconstructed sources as the sum of the observation joint entropy and of the expected Jacobian of the separating system (see equation 3). For a linear mixture, this trick leads to algorithms easy to implement, based on minimization of the mutual information [16,6].

However for post nonlinear (PNL) mixtures, the above method introduced some bias in estimating the reduced criterion and therefore, it might be preferable to consider a criterion based directly on the mutual information. The goal of this paper is to highlight these points. More precisely, we shall describe and compare two separation criteria. The first criterion is based on a reduction of the mutual information which avoids the estimation of the joint entropy. This method has been introduced in Taleb and Jutten [18] and adopted in [2]. The second one is based on the full expression of mutual information and was introduced recently in [4].

This paper is organized as follows: Section 2 explains the two separation criteria derived from the mutual information. We derive the bias of the main terms in section 3 and give a graphical illustration of the use of various bandwidths. In section 4, the comparison of the relative gradient of these estimators further highlights the difference between the two methods of blind source separation and we conclude in section 5.

2 Two empirical criteria for separation in PNL mixtures based on mutual information

2.1 The post nonlinear model

Let us first recall the definition of a post nonlinear mixture: the observed signals $X_1, \ldots, X_K$ 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 ik-th entry of the mixing matrix $\mathbf{A}$ and $f_1, \ldots, f_K$ are nonlinear functions. It is assumed that there is the same number K of sources and observations, 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 knows* $\mathbf{A}$ *and* $f_1, \ldots, f_K$.

The blind source separation problem consists in finding a matrix $\mathbf{B}$ and K applications $g_1, \ldots, g_K$ so that the random variables, $i = 1, \ldots, K$,

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

which represent the reconstructed sources, are independent. Indeed, it has been shown [3] that the independence of the output $Y_1, \ldots, Y_K$, implies $Y_i = \alpha_i S_{\sigma(i)}$ (where $\sigma(i)$ is a permutation over $\{1, 2, \ldots, K\}$ and $\alpha_1, \ldots, \alpha_K$ are scale factors), i.e. source separation is achieved with scale and permutation indeterminacies, as for linear mixtures. The mixture and separation structure are presented in Figure 1. In the sequel, we denote, $\mathbf{S} = (S_1, \ldots, S_K)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_K)^T$.

2.2 Theoretical independence criteria

As a measure of dependence, let us consider the mutual information of the random variables $Y_1, \ldots, Y_K$:

$$\begin{aligned} I(Y_1, \ldots, Y_K) &= -\int \log \left(\frac{p_{Y_1,\ldots,Y_K}(y_1, \ldots, y_K)}{\prod_{i=1}^{K} p_{Y_i}(y_i)} \right) p_{Y_1,\ldots,Y_K}(y_1, \ldots, y_K) dy_1 \ldots dy_K \\ &= \sum_{i=1}^{K} H(Y_i) - H(Y_1, \ldots, Y_K) \end{aligned} \tag{2}$$

where H denotes the entropy, $H(X) = -E[\log(p_X(X))]$, p_X is the density function of X.

As already shown in [15], the mutual information is always positive and is equal to zero if and only if the random variables $Y_1, \ldots, Y_K$ are independent. Thus, $I(Y_1, \ldots, Y_K)$ can be used as a criterion for blind source separation.

For a post non linear mixture, Taleb and Jutten [18] suggest to transform the above mutual information so as to keep only terms with marginal entropy. They obtain the reduced criterion:

$$C(Y_1, \ldots, Y_K) = \sum_{i=1}^{K} H(Y_i) - \sum_{i=1}^{K} H(Z_i) - \log|\det \mathbf{B}|. \tag{3}$$

Since the mutual information between $Z_1, \ldots, Z_K$ is equal to that between $X_1, \ldots, X_K$, it can be seen that,

$$I(Y_1, \ldots, Y_K) = C(Y_1, \ldots, Y_K) + I(X_1, \ldots, X_K). \tag{4}$$

As $I(X_1, \ldots, X_K)$ is a constant, the minimum of $C(Y_1, \ldots, Y_K)$ is the same as the one of $I(Y_1, \ldots, Y_K)$.

The above criteria $I(Y_1, \ldots, Y_K)$ and $C(Y_1, \ldots, Y_K)$ are theoretical criteria, in practice one has to estimate them. In the sequel, we shall consider such estimates and show that they lead to different algorithms.

In the following, let us denote $\mathbf{X}(1), \ldots, \mathbf{X}(N)$ a sample of $\mathbf{X} = (X_1, \ldots, X_K)^T$ of size N and for all $i = 1, \ldots, N$ and $k = 1, \ldots, K$, $Z_k(i) = g_k(X_k(i))$ and $Y_k(i) = \sum_{j=1}^{K} \mathbf{B}_{kj} Z_j(i)$. In addition, $\mathcal{K}$ denotes a kernel (a positive function whose integral is equal to one) and $\mathcal{K}_h(u) = h^{-1}\mathcal{K}(u/h)$, where h is a bandwidth.

2.3 Estimation of the reduced criterion C

Here, only the estimate of the marginal entropy is needed:

$$\widehat{C}(Y_1, \ldots, Y_K) = \sum_{i=1}^{K} \widehat{H}(Y_i) - \sum_{i=1}^{K} \widehat{H}(Z_i) - \log|\det \mathbf{B}|. \tag{5}$$

where marginal entropies and probability density functions (pdf) are estimated respectively by, for any sample $X(1), \ldots, X(N)$ of a random variable X:

$$\widehat{H}(X) = \frac{1}{N} \sum_{n=1}^{N} \log \widehat{p}_X(X(n)) \tag{6}$$

and

$$\widehat{p}_X(X(n)) = \frac{1}{N}\sum_{m=1}^{N} \mathcal{K}_h(X(n) - X(m)). \tag{7}$$

2.4 Estimation of mutual information I

In order to estimate the mutual information (2), we use the empirical mean and the kernel estimation of density functions:

Let $\mathbf{Y}(1), \ldots, \mathbf{Y}(N)$ be a sample of $\mathbf{Y}$

$$\widehat{I}(Y_1, \ldots, Y_K) = \sum_{i=1}^{K} \widehat{H}(Y_i) - \widehat{H}(\mathbf{Y}), \tag{8}$$

where marginal entropies and pdf are estimated according to equations (6) and (7), and joint entropy and pdf according to:

$$\widehat{H}(\mathbf{Y}) = \frac{1}{N}\sum_{n=1}^{N} \log \widehat{p}_{\mathbf{Y}}(\ \mathbf{Y}(n)) \tag{9}$$

and

$$\widehat{p}_{\mathbf{Y}}(\mathbf{Y}(n)) = \frac{1}{N}\sum_{m=1}^{N}\prod_{i=1}^{K}\ \mathcal{K}_h(Y_i(n) - Y_i(m)). \tag{10}$$

The choice of the kernel is discussed below.

2.5 Comments

(1) The estimator of the joint density (10) is expressed using a product of one dimensional kernel. This can be seen as a restriction since joint density does not factorize except if marginal variables are independent, but here this choice improves the bias of the estimator of mutual information (see below eq. 17).

(2) As the estimation of C requires only the estimation of marginal densities, the sample size does not need to be too large in order to get a small error. On the opposite, the estimation of I requires an estimation of joint density of $Y_1, \ldots, Y_K$ whose error is difficult to control because the sample size has to be huge when the number of sources is high. Moreover, we notice that the computational complexity of $\widehat{C}$ and $\widehat{I}$ is $O(N^2K)$, where N is the sample size and K is the number of sources. This cost thus grows quadratically with N. However, Pham [17] proposed an entropy estimator based on a discretization of integral, which allows to bring

down the computational cost to $O(NK)$ for $\widehat{C}$ and $O(N3^K)$ for $\widehat{I}$. The cost is then low for $\widehat{C}$ but grows exponentially with the dimension for $\widehat{I}$.

(3) The relation (4) shows that it is equivalent to look for the minimum of $I(Y_1, \ldots, Y_K)$ or of $C(Y_1, \ldots, Y_K)$. But the estimators $\widehat{I}$ and $\widehat{C}$ do not satisfy anymore the relation (4). Indeed, the kernel-density estimator does not satisfy the well-known relation between a density and a transformed density:

$$p_{g(X)}(y) = \frac{p_X(g^{-1}(y))}{|g'(g^{-1}(y))|}$$

where g is any continuously differentiable invertible function and X is any random vector admitting a density.

As a result, the minimum of $\widehat{C}$ does not correspond to the minimum of $\widehat{I}$.

A question worthwhile to consider is how to quantify the differences between these two methods based on these estimated criteria. This will be answered thanks to an asymptotic argument analysis.

3 Calculation of the bias

The results presented here are based on the work of Joe [14]. Let us mention also some related results by Hall and Morton [12]. In [14], the author proposed an estimator of the entropy and calculated the bias of this estimator. Here, we will combine and extend these results to obtain the bias of $\widehat{I}$ and $\widehat{C}$, written in terms of the entropy.

Assumptions **A** and **F** are assumed throughtout: (letters **A** and **F** refer to the corresponding assumptions in [14])

A. Tails of the density distributions: S is a bounded set such that the density distribution p is bounded below on it by a positive constant and

$$\int_S p\log(p) \approx \int_{\mathbb{R}^K} p\log(p).$$

As said in [14], this assumption claims that $\int_{\mathbb{R}^K} p\log(p) < \infty$ and can be approximated arbitrarily closely by $\int_S p\log(p)$ for a bounded set S. Thus, this is not a stringent restriction and in particular, all bounded densities satisfy this hypothesis. Therefore, for each expansion (see below), the bounded set S can be chosen such that $\int_{\mathbb{R}^K \setminus S} p\log(p)$ is neglegible in comparison with the terms of order N^{-1}:this term is thus contained in $o(N^{-1})$ in equations (13), (16), (18).

F. $\mathcal{K}$ is a symmetric univariate density satisfying $\int v^2\mathcal{K}(v)dv = 1$.

The other assumptions, more technical, are given by Joe [14, p. 685]. Then, two different calculations allow to write the bias of $\widehat{I}$ and $\widehat{C}$. Here, the development of the estimation of entropy given in [14, p. 692] is used.

In the sequel, we will use the following notations for the expectation of joint and marginal pdf estimates for a random vector $\mathbf{R} = (R_1, \ldots, R_n)^T$:

$$p_{\mathbf{R},h}(x) = E[\widehat{p}_{\mathbf{R}}(x)] = \int \prod_{i=1}^{K} \mathcal{K}_h(u_i) dF_{\mathbf{R}}(u) \tag{11}$$

$$p_{R_i,h}(x) = E[\widehat{p}_{R_i}(x)] = \int \mathcal{K}_h(u_i) dF_{R_i}(u_i) \tag{12}$$

where $dF_{\mathbf{R}}(u) = p_{\mathbf{R}}(u)du$ and $dF_{R_i}(u_i) = p_{R_i}(u_i)du_i$.

3.1 Bias of the reduced criterion $\widehat{C}$

Using the above notations (11), (12) and extending the result of Joe, the bias of the reduced criterion $\widehat{C}$ is:

$$\begin{aligned}
&E[\hat{C}(Y_1, \ldots, Y_K)] = \sum_{i=1}^{K} (E[\hat{H}(Y_i)] - E[\hat{H}(Z_i)]) - \log|det\mathbf{B}| \\
&= -\sum_{i=1}^{K} \int_S \log p_{Y_i,h}(x) dF_{Y_i}(x) + \sum_{i=1}^{K} \int_S \log p_{Z_i,h}(x) dF_{Z_i}(x) - \log|det\mathbf{B}| \\
&\quad + N^{-1} T^{AR}(h, K, \mathcal{K}, p_{Y_i}, p_{Z_i}) + o(N^{-1}) \\
&= -\sum_{i=1}^{K} \left(\int_S \log(p_{Y_i}(x)) dF_{Y_i}(x) - \int_S \log(p_{Z_i}(x)) dF_{Z_i}(x) \right) - \log|det\mathbf{B}| \\
&\quad - \sum_{i=1}^{K} \int_S (\log p_{Y_i,h}(x) - \log p_{Y_i}(x)) dF_{Y_i}(x) + \sum_{i=1}^{K} \int_S (\log p_{Z_i,h}(x) - \log p_{Z_i}(x)) dF_{Z_i}(x) \\
&\quad + N^{-1} T^{AR}(h, K, \mathcal{K}, p_{Y_i}, p_{Z_i}) + o(N^{-1}) \\
&= C(Y_1, \ldots, Y_K) - B^{AR}(h, K, \mathcal{K}, p_{Y_i}, p_{Z_i}) + N^{-1} T^{AR}(h, K, \mathcal{K}, p_{Y_i}, p_{Z_i}) + o(N^{-1})
\end{aligned} \tag{13}$$

where the exact expression of T^{AR} is written in appendix B, and

$$B^{AR}(h, K, \mathcal{K}, p_{Y_i}, p_{Z_i}) = \sum_{i=1}^{K} \int_S (\log p_{Y_i,h}(x) - \log p_{Y_i}(x)) dF_{Y_i}(x) - \sum_{i=1}^{K} \int_S (\log p_{Z_i,h}(x) - \log p_{Z_i}(x)) dF_{Z_i}(x). \tag{14}$$

As N goes to infinity, the last terms $N^{-1}T^{AR}(h, K, \mathcal{K}, p_{Y_i}, p_{Z_i}) + o(N^{-1})$ vanishes and $B^{AR}(h, K, \mathcal{K}, p_{Y_i}, p_{Z_i})$ (defined by formula (14)) represents the exact formula of the asymptotic bias of $\widehat{C}$ when N goes to infinity. In fact, B^{AR} is a term in h which does not depend on N. Let us give some comments about this term B^{AR},

Remark 1.

If all the densities are supposed three times differentiable with continuous derivatives (for expanding up to the second order), using the following expansion of $p_{Y_i,h}$, $p_{Y_i,h}(x) = p_{Y_i}(x) + 0.5h^2 p''_{Y_i}(x) + o(h^2)$, the bias is expressed as:

$$B^{AR}(h, K, \mathcal{K}, p_{Y_i}, p_{Z_i}) = 0.5h^2 \sum_{i=1}^{K} \int_S (p''_{Y_i}(x) - p''_{Z_i}(x))dx + o(h^2) \tag{15}$$

Thus the asymptotic bias B^{AR} goes to zero if and only if h tends to zero. Note that this term in h^2 may vanish if the density satisfy certain condition, but then there will be a term in h^3 again not depending on N.

Remark 2.

If the densities are not differentiable on all the space, B^{AR} does not vanish in general. The equation (15) suggests to study the following terms of order zero and one of the asymptotic expansion of $p_{Y_i,h}(x)$,

$$p_{Y_i,h}(x) - p_{Y_i}(x) = \int_S \mathcal{K}(u)(p_{Y_i}(x - uh) - p_{Y_i}(x))du$$
$$\frac{p_{Y_i,h}(x) - p_{Y_i}(x)}{h} = \int_S \mathcal{K}(u) \frac{p_{Y_i}(x - uh) - p_{Y_i}(x)}{h} du$$

Thus there exists no limit as soon as the density is not continuous or not differentiable in x, and it introduces a term which is not possible to control when h goes to zero.

3.2 Bias of mutual information $\widehat{I}$

The same idea can be applied to calculate the bias of $\widehat{I}$:

$$\begin{aligned} E[\widehat{I}(Y_1, \ldots, Y_K)] &= \sum_{i=1}^{K} E[\hat{H}(Y_i)] - E[\hat{H}(Y_1, \ldots, Y_N)] \\ &= \int_S \log\left(\frac{p_{\mathbf{Y},h}(x)}{\prod_{i=1}^{K} p_{Y_i,h}(x)}\right) dF_{\mathbf{Y}}(x) + N^{-1}T^{BR}(h, K, \mathcal{K}, p_{Y_i}, p_{\mathbf{Y}}) + o(N^{-1}) \end{aligned} \tag{16}$$

where the exact definition of T^{BR} is given in appendix A.

We note that, under the assumptions mentioned above,

$$\int \log\left(\frac{p_{\mathbf{Y},h}(x)}{\prod_{i=1}^{K} p_{Y_i,h}(x)}\right) dF_{\mathbf{Y}}(x) = 0 \text{ if } Y_1,\ldots,Y_K \text{ are independent.} \tag{17}$$

As a result, when $Y_1,\ldots,Y_K$ are **independent**, $\widehat{I}$ is asymptotically unbiased when N tends to infinity even if h is kept fixed.

In addition, by assumption on S,

$$I(Y_1,\ldots,Y_K) - \int_S \log\left(\frac{p_{\mathbf{Y}}(x)}{\prod_{i=1}^{K} p_{Y_i}(x)}\right) dF_{\mathbf{Y}}(x) = o(N^{-1}),$$

we obtain:

$$\begin{aligned} E[\widehat{I}(Y_1,\ldots,Y_K)] = I(Y_1,\ldots,Y_K) + \int_S \left\{\log\left(\frac{p_{\mathbf{Y},h}(x)}{\prod_{i=1}^{K} p_{Y_i,h}(x)}\right) - \log\left(\frac{p_{\mathbf{Y}}(x)}{\prod_{i=1}^{K} p_{Y_i}(x)}\right)\right\} p_{\mathbf{Y}}(x)dx \\ + N^{-1}T^{BR}(h,K,\mathcal{K},p_{Y_i},p_{\mathbf{Y}}) + o(N^{-1}) \end{aligned} \tag{18}$$

In this expression, the expression of the bias not depending on N is equal to,

$$B^{BR}(h,K,\mathcal{K},p_{Y_i},p_{\mathbf{Y}}) = \int_S \left\{\log\left(\frac{p_{\mathbf{Y},h}(x)}{\prod_{i=1}^{K} p_{Y_i,h}(x)}\right) - \log\left(\frac{p_{\mathbf{Y}}(x)}{\prod_{i=1}^{K} p_{Y_i}(x)}\right)\right\} dF_{\mathbf{Y}}(x)$$

Then, let us make a comment, if all the densities are supposed three times differentiable with continuous derivatives, the bias is expressed as:

$$B^{BR}(h,K,\mathcal{K},p_{Y_i},p_{\mathbf{Y}}) = 0.5h^2 \left\{\int_S \mathbf{tr}\; p''_{\mathbf{Y}}(x) - \sum_{i=1}^{K} \int_S p''_{Y_i}(x)dx\right\} + o(h^2)$$

where $p''_{\mathbf{Y}}$ is the Hessian matrix of $p_{\mathbf{Y}}$ and $\mathbf{tr}$ denotes the trace of a matrix.

When the variables $Y_1,\ldots,Y_K$ are **dependent**, the bias of $\widehat{I}$ will go to zero as N tends to infinity and h tends to zero.

Remark 3.

The difference between these two criteria $\widehat{I}$ and $\widehat{C}$ in the context of the PNL mixture is underlined by the terms in the bias due to the nonlinear terms in the mixture. Indeed, the bias of the reduced criterion $\widehat{C}$ is characterized by explicit terms coming from the nonlinear part of the mixture: $\sum_{i=1}^{K} \int_S (\log p_{Z_i,h}(x) - \log p_{Z_i}(x)) dF_{Z_i}(x)$, whereas the nonlinear terms of the mixture in the bias of mutual information $\widehat{I}$ are included in the calculation of the estimation of densities $p_{\mathbf{Y},h}$ and $p_{Y_i,h}$. Therefore, the nonlinear terms of the mixture have not the same influence in these two criteria.

3.3 Graphical illustrations of the bias for different density distributions

In order to illustrate the behaviour of the bias of the criteria $\widehat{I}$ and $\widehat{C}$ according to the bandwidth, we compute the two criteria for different size of the bandwidth and different density distributions. We do not intend to give a proof with these plots but just an illustration of our theoretical results obtained in section 3.1 and 3.2.

The mixture matrix $\mathbf{A}$ is orthogonal and we take only two sources, i.e. $K = 2$. The criteria are computed in the exact solution of the problem, i.e. Y_1, Y_2 are independent and Z_1, Z_2 satisfy $\mathbf{Z} = \mathbf{A}\mathbf{Y}$. We choose to represent uniform distribution which is not continuous and Gaussian distribution which is C^∞.

In Figure 2 and 3, the plots represent the computation of $\widehat{C} - C$ or $\widehat{I} - I$ for 99 samples of size N between 100 and 5000. Figure 2 illustrates the result that the bias of $\widehat{C}$ tends to zero only when h tends to zero and N tends to infinity, whereas the bias of $\widehat{I}$ tends to zero when N tends to infinity even if h is large. On Figure 3, we can see that with a C^∞ density distribution, the bias of $\widehat{C}$ tends to zero even if h is large. In particular, the bias in the estimation of C or I with Gaussian distribution will tend to zero even if h is large. But it still remains the particularity of Gaussian distributions that the independence is preserved even with non trivial mixtures. And, the critera can be minimum even if the mixture is not trivial.

3.4 Discussions

Both criteria $\widehat{C}$ and $\widehat{I}$ provide a solution to the problem of blind source separation, but their bias have different properties:

(1) If the variables $Y_1, \ldots, Y_K$ are independent, the main difference between the bias of $\widehat{C}$ and $\widehat{I}$ is its limit when N tends to infinity. Indeed, we notice that when N tends to infinity, the bias of $\widehat{C}$ tends to zero only if h tends to zero (with a sufficient low rate), while the bias of $\widehat{I}$ tends to zero when N tends to infinity even for fixed h. This suggest to use $\widehat{I}$ rather than $\widehat{C}$, so that the convergence does not depend on the choice of h. It also explains the efficiency of even simple histograms estimates [3] and the robustness concerning the choice of h in the kernel.

(2) It might happen that for some particular distributions of sources, the terms in h^2 in the asymptotic bias of $\widehat{C}$ (15) vanish but the bias does not vanish since there always remains terms in h^3, h^4... which does not depend on N. This could be of importance for specific applications. Let

us recall here that with a C^∞ density distribution, the bias of $\widehat{C}$ tends to zero even if h is large.

(3) Finally, to solve the problem of blind source separation, it is the behaviour at the point where the criterion is minimized. Further investigation is needed in order to compare this behaviour, by studying the gradient of these two criteria for example.

4 Comparison of the relative gradient:minimization of the criteria

A solution to the problem of blind source separation is obtained by minimizing the two criteria $\widehat{I}$ and $\widehat{C}$. It is possible to evaluate the exact gradient of the two criteria and to apply an exact gradient descent on $\widehat{I}$ and $\widehat{C}$. This approach is different from the methods developped by Taleb and Jutten [18], who apply a gradient descent on C, and Babaie-Zadeh [3], who uses I. Actually, they both use an estimation of the gradient calculated with theoretical expression of I and C. As they use only an estimation of the gradient, their methods do not consist in minimizing a known objective function but solving estimate equations, which require estimations of score functions, this will be delopped page 13.

Let us calculate the relative gradient [8] of the above estimate criteria whose the main idea consists in updating $\mathbf{B}$ and $g_1, \ldots, g_K$ according to:

$$\begin{cases} \mathbf{B} & \longleftarrow \mathbf{B} + \epsilon \mathbf{B} \\ g_i & \longleftarrow g_i + \delta_i \circ g_i \quad \text{for all } i = 1, \ldots, K \end{cases}$$

where ϵ and $\delta_1, \ldots, \delta_K$ denote the relative gradient of the linear and nonlinear part respectively. The following expression of the relative gradient are obtained in [1] just by applying Taylor expansion of both estimate criteria $\widehat{I}$ and $\widehat{C}$ in terms of ϵ and $\delta_1, \ldots, \delta_K$.

4.1 Relative gradient of the reduced criterion C

In the following, $\widehat{E}$ denotes the empirical mean, $\widehat{E}(\phi(X)) = \sum_{i=1}^{N} \phi(X_i)/N$, where $X_1, \ldots, X_N$ is a sample of X and ϕ is any function. The relative gradient of $\widehat{C}$ consists of two parts:

- relative gradient of linear part:

$$\varepsilon \mapsto \sum_{i=1}^{K} \sum_{k \neq i, k=1}^{K} \varepsilon_{ik} \widehat{E}[\widehat{\psi}_{Y_i}(Y_i) Y_k] \tag{19}$$

- relative gradient of non linear part:

$$\text{For all } k,\ 1 \leq k \leq K,\ \delta_k \mapsto \widehat{E}\left\{\delta_k(Z_k)\left[\sum_{i=1}^{K} \mathbf{B}_{ik}\widehat{\psi}_{Y_i}(Y_i) - \widehat{\psi}_{Z_k}(Z_k)\right]\right\} \tag{20}$$

where
- $\widehat{\psi}_{Y_i}(Y_i(j)) = N\partial^2_{ij}\widehat{H}(Y_i)$
- $\widehat{\psi}_{Z_i}(Z_i(j)) = N\partial^2_{ij}\widehat{H}(Z_i)$

and ∂^2_{ij} denotes the derivative with respect to $Y_i(j)$ (or $Z_i(j)$).

4.2 Relative gradient of mutual information I

The relative gradient of $\widehat{I}$ consists of two parts, too:

- relative gradient of linear part:

$$\varepsilon \mapsto \sum_{i=1}^{K} \sum_{k\neq i, k=1}^{K} \varepsilon_{ik}\widehat{E}\{Y_i\widehat{\beta}_k(\mathbf{Y})\} \tag{21}$$

- relative gradient of non linear part:

$$\delta_1, \ldots, \delta_K \mapsto \sum_{k=1}^{K} \widehat{E}\left\{\delta_k(Z_k)\sum_{i=1}^{K} \widehat{\beta}_i(\mathbf{Y})B_{ik}\right\} \tag{22}$$

where, $\widehat{\beta}_k(\mathbf{Y}(j)) = \widehat{\psi}_{Y_k}(Y_k(j)) - \widehat{\phi}_k(\mathbf{Y}(j))$, for all $k = 1, \ldots, K$ and

- $\widehat{\phi}_i(\mathbf{Y}(j)) = N\partial^2_{ij}\widehat{H}(\mathbf{Y})$
- $\widehat{\psi}_{Y_i}(Y_i(j)) = N\partial^2_{ij}\widehat{H}(Y_i)$

and ∂^2_{ij} denotes the derivative with respect to $Y_i(j)$.

4.3 Comments

(1) To compare these two relative gradients, let us subtract the relative gradient of linear part and non linear part, respectively:
   - difference of relative gradients of linear part for $i \neq k$:

$$\widehat{E}\{Y_i\widehat{\beta}_k(\mathbf{Y})\} - \widehat{E}[\widehat{\psi}_{Y_i}(Y_i)Y_k] = -\widehat{E}\{Y_i\widehat{\phi}_k(\mathbf{Y})\} \tag{23}$$

   - difference of relative gradients of nonlinear part:

$$\widehat{E}\left\{\delta_k(Z_k)\sum_{i=1}^{K}\widehat{\beta}_i(\mathbf{Y})B_{ik}\right\} - \widehat{E}\left\{\delta_k(Z_k)\left[\sum_{i=1}^{K}\mathbf{B}_{ik}\widehat{\psi}_{Y_i}(Y_i) - \widehat{\psi}_{Z_k}(Z_k)\right]\right\}$$
$$= \widehat{E}\left\{\delta_k(Z_k)\left[\sum_{i=1}^{K}\widehat{\phi}_i(\mathbf{Y})B_{ik} - \widehat{\psi}_{Z_k}(Z_k)\right]\right\} \tag{24}$$

Using the definition of score functions and integrations by part, it can be seen that when the estimate functions are replaced by the exact functions, expressions (23) and (24) are equal to zero. Thus, the use of estimate leads to the difference between the two relative gradients and then in the calculation of minima.

(2) The main difference between the approach used by Taleb [18] and Babaie-Zadeh [3] comes from the use of estimations. In addition, there exist two different approaches for solving the problem of minimizing an independence criterion such as I or C:

strategy differentiate first. (used in [18] and [3])

It consists in calculating the exact relative gradient ∇I or ∇C of I or C respectively, and then take an estimation of it to solve the estimating equations $\widehat{\nabla I} = 0$ or $\widehat{\nabla C} = 0$

strategy estimate first. (used in the present paper)

Our approach consists in first estimating the independence criteria I and C, denoted $\widehat{I}$ and $\widehat{C}$ respectively. The solution of BSS problem is reached by minimizing $\widehat{I}$ or $\widehat{C}$, using their exact relative gradients. In this strategy, the use of calculating the bias of $\widehat{I}$ and $\widehat{C}$ is obvious.

These two strategies are not equivalent, but the first one is not a straightforward characterization of independence and would require further investigations.

5 Conclusion

The use of mutual information to solve the problem of blind source separation, leads to implement two different methods. One consists in estimating mutual information without any assumptions on the mixture. The other consists in estimating a simple expression of mutual information without the terms with joint entropy. A statistical study of the bias shows that when the variables are independent, the estimator of the full expression of mutual information is asymptotically unbiased whereas, the estimator of the reduced criterion is biased when the bandwidth is fixed and does not tend to zero. The comparison of the relative gradients shows that these two criteria will lead to two different algorithms for blind source separation. As the solution of blind source separation is the minimum of the estimator, an interesting further development is to study more precisely the difference between the minima of these two criteria using the expression of the relative gradients.

A Expression of the bias of the estimator of mutual information I

All these calculations are deduced by extending the work of Joe [14].

$$
\begin{aligned}
T^{BR}(h,K,\mathcal{K},p_{Y_i},p_{\mathbf{Y}}) = 0.5(K-1) - \Bigg\{ & h^{-1}[0.5K_{02}-\mathcal{K}(0)]\sum_{i=1}^{K}\int_S p_{Y_i}(x)(p_{Y_i,h}(x))^{-1}dx \\
& - h^{-K}[0.5K_2-\mathcal{K}(0)^K]\int_S p_{\mathbf{Y}}(x)(p_{\mathbf{Y},h}(x))^{-1}dx \Bigg\} \\
+0.5\Bigg\{ & h^{-1}K_{02}\sum_{i=1}^{K}\int_S (p^*_{Y_i,h}(x)-p_{Y_i}(x))p_{Y_i}(x)(p_{Y_i,h}(x))^{-1})^{-2}dx \\
& -h^{-K}K_2\int_S (p^*_{\mathbf{Y},h}(x)-p_{\mathbf{Y}}(x))(p_{\mathbf{Y},h}(x))^{-2}dx \Bigg\}
\end{aligned}
$$

where $K_2 = K_{02}^K$, $K_{02} = \int \mathcal{K}^2(v)dv$, $l(u) = \mathcal{K}(u)/K_{02}$ and $p^*_{\mathbf{Y},h}(x) = h^{-K}\int\prod_{i=1}^K l((x_i - y_i)/h)p_{\mathbf{Y}}(\mathbf{y})d\mathbf{y}$ and $p^*_{Y_i,h}(x) = h^{-1}\int l((x-y)/h)p_{Y_i}(y)dy$.

B Expression of the bias of the estimator of the reduced criterion C

All these calculations are deduced by extending the work of Joe [14], , and use the same notations than in Appendix A.

$$
\begin{aligned}
T^{AR}(h,K,\mathcal{K},p_{Y_i},p_{Z_i}) = {} & h^{-1}[0.5K_{02}-\mathcal{K}(0)]\sum_{i=1}^{K}\int_S \left\{p_{Y_i}(x)(p_{Y_i,h}(x))^{-1} - \int_S p_{Z_i}(x)(p_{Z_i,h}(x))^{-1}\right\}dx \\
& +0.5h^{-1}K_{02}\sum_{i=1}^{K}\int_S (p^*_{Y_i,h}(x)-p_{Y_i}(x))p_{Y_i}(x)(p_{Y_i,h}(x))^{-2}dx \\
& -0.5h^{-1}K_{02}\sum_{i=1}^{K}\int_S (p^*_{Z_i,h}(x)-p_{Z_i}(x))p_{Z_i}(x)(p_{Z_i,h}(x))^{-2}dx
\end{aligned}
$$

References

[1] S. Achard. *Mesures de dépendance pour la séparation aveugle de sources, application aux mélanges post non linéaires.* PhD thesis, Université Joseph Fourier, Grenoble, 2003. .

[2] S. Achard, D.T. Pham, and C. Jutten. Blind source separation in post nonlinear mixtures. In *Proc. Int. Workshop on Independent Component Analysis and*

Blind Signal Separation, ICA2001, pages 295–300, San Diego, California, Dec. 2001. .

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

[4] M. Babaie-Zadeh, C. Jutten, and K. Nayebi. Minimization-projection (MP) approach for blind souces separation in different mixing models. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA2003*, pages 1083–1088, Nara, Japan, Apr. 2003.

[5] F.R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, Jul. 2002.

[6] J.F. Cardoso. Blind signal separation : Statistical principles. *Proceedings IEEE*, 86(10):2009–2025, Oct. 1998.

[7] J.F. Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11:157–192, 1999.

[8] J.F. Cardoso and B.H. Laheld. Equivariant adaptative source separation. *IEEE Transactions on Signa Processing*, 44(12):3017–3029, Dec. 1996.

[9] A. Cichoky and S.-I. Amari. *Adaptative blind signal and image processing. Learning algorithms and applications.* John Willey and Sons, 2002.

[10] P. Comon. Independent component analysis, a new concept ? *Signal Processing*, 3(36):287–314, Apr. 1994.

[11] J. Eriksson, A. Kankainen, and V. Koivunen. Novel characteristic function based criteria for ICA. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA2001*, pages 108–113, San Diego, California, Dec. 2001.

[12] P. Hall and S. C. Morton. On the estimation of entropy. *Ann. Inst. Statist. Math.*, 45(1):69–88, 1993.

[13] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* New York: John Wiley & Sons, 2001.

[14] H. Joe. Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Statist. Math.*, 41(4):683–697, 1989.

[15] S. Kullback. *Information theory and statistics.* John Wiley & Sons, 1959.

[16] D.-T. Pham. Blind separation of instantaneous mixture of sources via an independent component analysis. *IEEE Transactions on Signal Processing*, 44(11):2768–2779, Nov. 1996.

[17] D.-T. Pham. Flexible parametrisation of postnonlinear mixture model in blind source separation. *IEEE Signal Processing Letters*, to appear.

[18] A. Taleb and C. Jutten. Sources separation in post-nonlinear mixtures. *IEEE Transactions on Signal Processing*, 10(47):2807–2820, Oct. 1999.

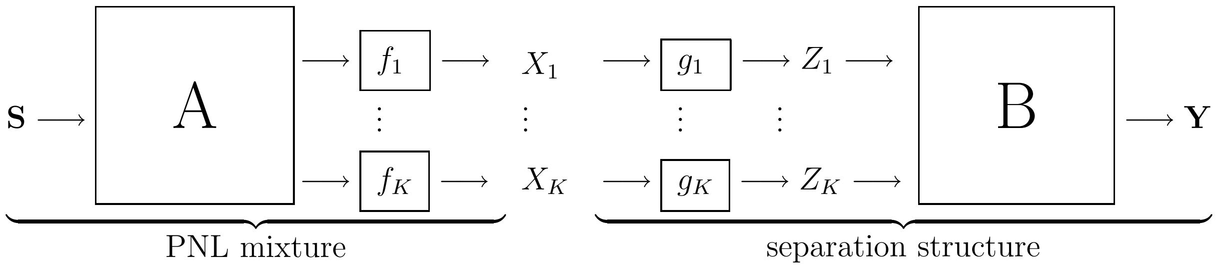

Fig. 1. Mixture and separation structure of a PNL mixture.

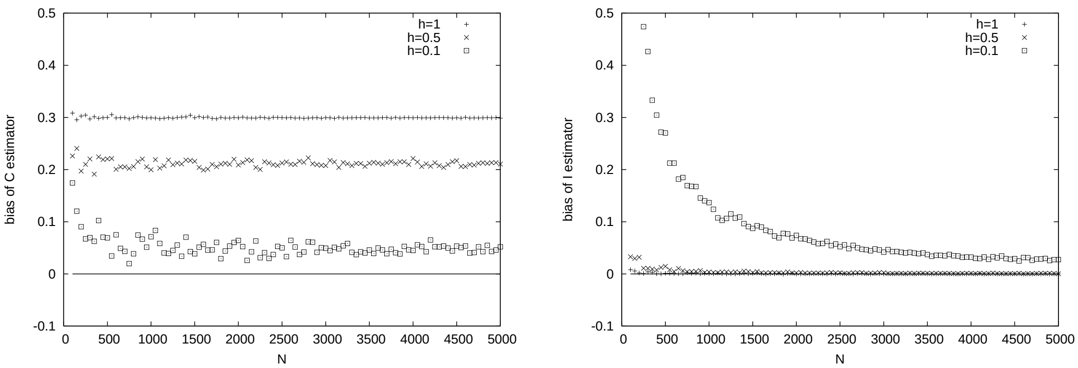

Fig. 2. Representation of the bias of $\widehat{C}$ and $\widehat{I}$ computed for various bandwidths and samples of uniform distribution, $C = -0.3$ and $I = 0$

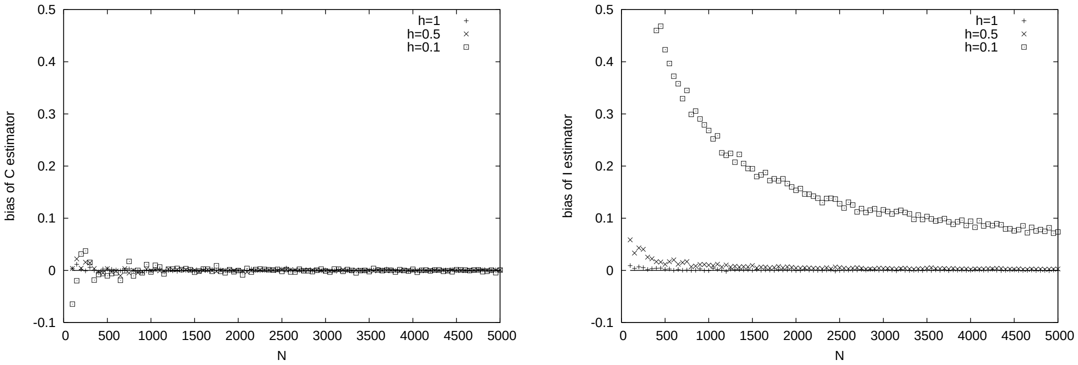

Fig. 3. Representation of the bias of $\widehat{C}$ and $\widehat{I}$ computed for various bandwidths and samples of Gaussian distribution, $C = 0$ and $I = 0$