# MUTUAL INFORMATION AND QUADRATIC DEPENDENCE FOR POST NONLINEAR BLIND SOURCE SEPARATION

*Sophie Achard, Christian Jutten, and Dinh-Tuan Pham*

University of Cambridge
Brain Mapping Unit
Cambridge CB3 2EB

Polytechnic Institute of Grenoble
Laboratory of Images and Signals
38031 Grenoble Cedex

CNRS
Laboratory of Modelling and Computation
38041 Grenoble Cedex

tel: 0044 1223 769 664, fax: 0044 1223 764 675, sa428@cam.ac.uk
http://www-bmu.psychiatry.cam.ac.uk/people/sa428/

**ABSTRACT**

This work focuses on solving the problem of Blind Source Separation (BSS) using Independent Component Analysis (ICA) method. Since ICA methods require a dependence measure, we will investigate the use of mutual information and quadratic dependence. Mutual information has already often been used for solving BSS problem, but difficulties occur in order to carry out an asymptotic study. In contrast, the quadratic dependence was introduced recently and has already been used for independence tests. Finally, the difficulty of solving the BSS problem is illustrated through examples of the shape of the objective-functions.

## 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 outputs of the separation system are mutually independent [7, 11]. In the context of linear mixtures, contrast functions can be constructed from cumulants [7] or even correlations if lagged correlations are included [13]. 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 [12] and Babaie-Zadeh [5], 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 Achard *et al.* [4] called the quadratic dependence measure. Thanks to simple computation and the possibility to carry out an asymptotic study, the quadratic dependence allows us to better characterize the behaviour of the estimators and the solution of the BSS problem.

In this paper, we compare the two dependence measures, the mutual information and the quadratic dependence. First from a statistical point of view, we show the good properties of the quadratic dependence in contrast of the difficulty to carry out an asymptotic study for the mutual information. Then, some examples of the shape of the objective-functions in the context of a linear mixture and nonlinear mixture are described, which show the difficulty to apply a minimization method due to the estimation problem but also to the complexity of the mixture.

Section 2 shows the statistical properties of the estimators of the mutual information and quadratic dependence. This results in a detailled comparison between the two dependence measures in terms of asymptotic behaviour of their estimators. Section 3 exhibits examples of the landscape of the objective-functions which have to be minimized. Especially, we note an important increasing in complexity for solving the problem of BSS in the context of a post nonlinear mixture.

**The post nonlinear model**

Let us 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, $i = 1, \ldots, K$, $Y_i = \sum_{k=1}^{K} \mathbf{B}_{ik} Z_k$, where $Z_k = g_k(X_k)$, which represent the reconstructed sources, are independent. Indeed, it has been shown [5, 2] 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.

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 $Y_k(i)=\sum_{j=1}^K \mathbf{B}_{kj}Z_j(i)$.

## 2. DEPENDENCE MEASURES

### 2.1 Mutual Information

As a measure of dependence, let us consider the mutual information of the random variables $Y_1,\ldots,Y_K$: $I(Y_1,\ldots,Y_K)=\sum_{i=1}^K H(Y_i)-H(Y_1,\ldots,Y_K)$, where $H$ denotes the entropy, $H(X)=-E[\log(p_X(X))]$, $p_X$ is the density function of $X$.

As already shown in [10], 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.

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 1). For a linear mixture, this trick leads to algorithms easy to implement, based on minimization of the mutual information [11, 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. For a post nonlinear mixture, Taleb and Jutten [12] suggest to transform the above mutual information so as to keep only terms with marginal entropy. They obtain the reduced criterion:

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

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,

$$I(Y_1,\ldots,Y_K)=C(Y_1,\ldots,Y_K)+I(X_1,\ldots,X_K). \quad (2)$$

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. For the estimation of these two criteria, we will simply use an estimation of entropy defined as:

$$\widehat{H}(\mathbf{Y})=\frac{1}{N}\sum_{n=1}^N \log\widehat{p}_{\mathbf{Y}}(\mathbf{Y}(n)) \quad (3)$$

and $\widehat{p}_{\mathbf{Y}}$ is a kernel estimation of density.

But, as shown in [3], the estimators $\widehat{I}$ and $\widehat{C}$ do not satisfy anymore the relation (2). 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 $\widehat{I}$. This difference leads to different minimization algorithms.

We notice also, see [3] for a proof, that 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 [5] and the robustness concerning the choice of $h$ in the kernel.

### 2.2 Quadratic dependence

#### *2.2.1 Definition*

Let us first recall the definition of the quadratic dependence as defined in [4].

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

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

*where*

$$\pi_{\mathbf{Y}}(\mathbf{y}) = E\left[\prod_{i=1}^K \mathscr{K}\left(\frac{y_i-Y_i(n)}{\widehat{\sigma}_{Y_i}}\right)\right]$$

$$\pi_{Y_k}(y_k) = E\left[\mathscr{K}\left(\frac{y_k-Y_k(n)}{\widehat{\sigma}_{Y_k}}\right)\right].$$

*and $\sigma_{Y_k}$ 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 $\lambda$.*

This dependence measure is called a quadratic dependence because it can be written in terms of an integral of the square difference between the joint and marginal characteristic functions, weighted by the Fourier transform of the kernel.

Thus we have at our disposal a whole class of quadratic measures, depending on the choice of the kernel $\mathscr{K}$ and also on the bandwidth $h$ if we choose the kernel to be a scaled kernel of the form $\mathscr{K}(\cdot/h)/h$. Let us stress that the kernel $\mathscr{K}$ 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. Due to the large degree of freedom in the choice of the kernel, the estimation of the quadratic dependence will be more robust in terms of the choice of the kernel and the bandwidth.

#### *2.2.2 Estimation*

As 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 $\phi$ is any function of the data.

*2.2.3 Asymptotic properties*

• Law under the hypothesis of independence (denoted $H_0$):
This result is due to Kankainen [9]. The estimator $N\widehat{Q}$ follows a law of $\gamma\chi^2(\beta)$ where $\gamma$ and $\beta$ are defined as, $\gamma = V_1/2E_1$ and $\beta = 2E_1^2/V_1$, where $E_1$ is the mean of $\widehat{Q}$ under $H_0$,and $V_1$ is the variance of $\widehat{Q}$ under $H_0$.
• Law under the hypothesis of dependence (denoted $H_1$):
The derivation of the law of the estimator of the quadratic dependence comes form results about U-statistics, [8, 9]. $\sqrt{N}(\widehat{Q}-Q)$ follows asymptotically a normal law with 0 mean and $\sigma^2$ variance, where $\sigma^2$ is,

$$\sigma^2 = \Sigma_{11} - 4\Sigma_{12} + 2\Sigma_{13} - 4\Sigma_{23} + 4\Sigma_{22} + \Sigma_{33}$$

with $\Sigma$ the variance-covariance matrix of the corresponding U-statistics dependent on $\mathscr{K}$ and $h$. Due to the lack of space, the reader is invited to refer to [1] for the exact formulas of the variances under each hypotheses.

These results allow us to propose a solution of the choice of the optimal bandwidth given a particular kernel. In the sequel, we will focus on two different kernels, the Gaussian kernel: $\mathscr{K}(x) = e^{-x^2}$ and the second derivative of the square Cauchy kernel: $\mathscr{K}(x) = -(20x^2-4)/(1+x^2)^4$ Figure 1 (a) illustrates the behaviour of the size of the confidence intervals in terms of the bandwitdh with two different kernels. $x$ is the solution of the equation: $P(-x \leq \widehat{Q}-Q \leq x) = 0.95$. Clearly, we observe that it is worthwhile to use a large bandwidth in order to get a very small variance. But with a large bandwidth, the power of the independence test can be very low, as described by the following figure 1 (b). In figure 1 (b), for the computation of $p$, we first compute $q_\alpha$ such that $P_{H_0}(\widehat{Q} > q_\alpha) = \alpha$, with $\alpha = 0.95$ and then, $p = 1 - P_{H_1}(\widehat{Q} < q_\alpha)$.

In conclusion, we have to choose the bandwith in order to have a small variability but also in order to keep a high power for the independence test. The possibility to construct an independence test is also very interesting in the sense that we will be able to control the gradient descent method in the minimization process. Indeed, this allows us to recognize a local minimum from a global minimum, and to propose an efficient criteria to control the convergence of the algorithm.

## 3. ILLUSTRATIONS

In this section, our objective is to illustrate what kind of difficulties may appear in the research of the minimum of the objective-functions like the quadratic dependence or the mutual information.

### 3.1 Linear mixtures

Figure 2 represents the landscape of the quadratic dependence in a simple example of a linear mixture with two sources $\mathbf{S} = (S_1, S_2)^T$: $\mathbf{X} = \mathbf{AS}$, with $\mathbf{A}$ a rotation matrix of angle $\pi/8$. Then the separation structure is defined by: $Y_1 = X_1 + aX_2$, $Y_2 = bX_1 + X_2$. In this context, it is possible to show that the initialization point of the minimization does not affect the convergence of the algorithm [1].

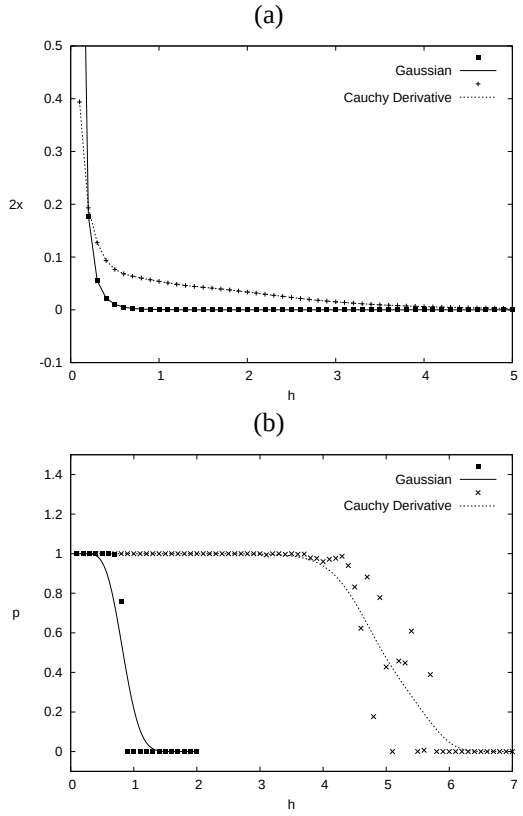


Figure 1: (a) Size of the confidence intervals and (b) power of the independence test in terms of the bandwidth using two different kernels

### 3.2 Post nonlinear mixtures

In the case of a post nonlinear mixture, some difficulties appear because of the existence of some local minima and the performance of the estimation. The landscape of the quadratic dependence given in figure 3 shows the complexity of the minimization in a simple example of a post nonlinear mixture with two sources $\mathbf{S} = (S_1, S_2)^T$, a rotation matrix $\mathbf{A}$ with angle $\pi/8$ and only one nonlinear function, $f_{1,\lambda}(x) = \frac{sign(x)}{2\lambda}(-1+\sqrt{1+4\lambda|x|})$, $\lambda = 3$. Then the separation structure is defined by, a rotation matrix $\mathbf{B}$ with angle $\theta$ and two functions $g_{1,\lambda}(x) = x + \lambda x|x|$ *and* $g_2(x) = x$. In this simple example, we observe that the choice of the initialization point for the minimization can be crucial, because of the existence of local minima.

Finally, figure 4 is a representation, in a logarithm scale, of figure 3 around the global minima, using a small bandwidth. The small oscillations observed on figure 4 are close to the global true minimum. This shows how the choice of the bandwidth can improve the accuracy of the method.

## 4. CONCLUSION

The study of different dependence measures is crucial in the improvments of ICA methods, especially to solve the BSS problem for nonlinear mixtures. In contrast to the mutual information, the quadratic dependence measure is easy to implement even for nonlinear mixtures. And due to the possibility to carry out an asymptotic study, it is possible to propose efficient control for the minimization algorithm. Further works will consist in characterizing precisely the behaviour

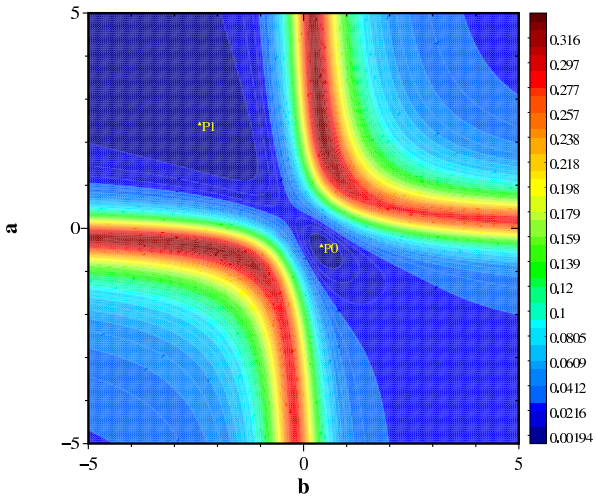


Figure 2: Representation of the estimator of the quadratic dependence measure with a Gaussian kernel and $h = 0.5$ in terms of $a$ and $b$. P0 and P1 denote the global minima.

of the estimation of the solution of the BSS problem in terms of the parameters used in the definition of the quadratic dependence.

## 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 and C. Jutten. Identifiability of post non linear mixtures. *IEEE Signal Processing letters*, in press, 2005.

[3] S. Achard, C. Jutten, and D.T. Pham. Criteria for blind source separation in post non linear mixture with mutual information. *signal processing, special issue on Information Theoretic Signal Processing*, in press, 2005.

[4] S. Achard, D.T. Pham, and C. Jutten. Quadratic dependence measure for non linear blind souces separation. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA2003*, pages 263–268, Nara, Japan, Apr. 2003. .

[5] M. Babaie-Zadeh, C. Jutten, and K. Nayebi. A geometric approach for separating post non-linear mixtures. In *Proc. Int. Workshop EUSIPCO 2002*, pages Vol. II, 11–14, Toulouse, France, sept. 2002.

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

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

[8] W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.*, 19:293–325, 1948.

[9] A. Kankainen. *Consistent testing of total independence based on empirical characteristic functions*. PhD thesis, University of Jyväskylä, 1995.

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

[11] 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.

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

[13] A. Ziehe, M. Kawanabe, S. Harmeling, and K.R. Müller. Separation of post-nonlinear mixtures using ace and temporal decorrelation. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA2001*, pages 433–438, San Diego, California, Dec. 2001.


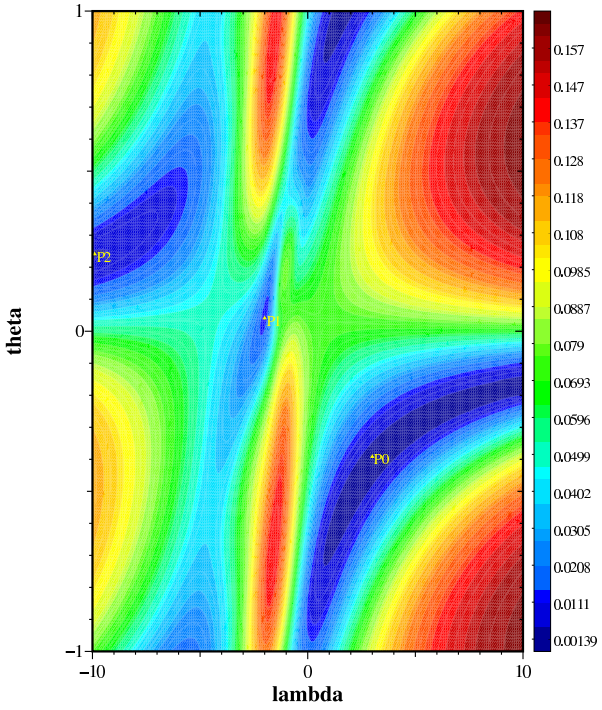


Figure 3: Representation of the estimator of the quadratic dependence measure with a derivative square Cauchy kernel and $h = 3$ in terms of $\theta$ and $\lambda$. P0 denotes the global minimum, P1 and P2 denote local minima.

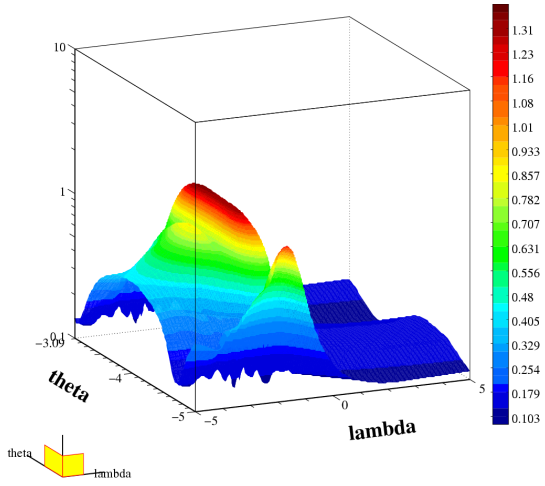


Figure 4: Representation of the logarithm of the estimator of the quadratic dependence measure with a derivative square Cauchy kernel and $h = 0.5$ in terms of $\theta$ and $\lambda$