SOURCE SEPARATION: FROM DUSK TILL DAWN

Christian JUTTEN[*]

INPG-LIS,
46, av. Félix Viallet,
38031 Grenoble Cedex,
France
Christian.Jutten@inpg.fr

Anisse TALEB[†]

ATRI and School of ECE,
Curtin University of Technology,
GPO Box U1987, Perth WA 6845,
Australia
anisse@atri.curtin.edu.au

ABSTRACT

The first part of this paper is concerned by the history of source separation. It include our comments and those of a few other researchers on the development of this new research field. The second part is focused on recent developments of the separation in nonlinear mixtures.

1. INTRODUCTION

First papers on source separation, and the genesis of the concept itself, can be traced back to early 80's and consists of the work of Ans, Hérault and Jutten [27, 3, 28], for modelling the biological problem of motion coding, and perhaps independently by Bar-Ness *et al.* in communications [5]. Although these results have had a weak impact in the neural networks community, they got an increasing interest in the signal processing community since 1989, especially in France and Europe.

Many signal processing conferences and more recently neural networks conferences, reserved sessions devoted to the problem of source separation: for instance, the French conference GRETSI since 1993, and many other international conferences, NOLTA 95 (Las Vegas, USA), ISCASS 96 (Atlanta, USA), EUSIPCO 96 (Trieste, Italia), NIPS 96 post-workshop (Denver, USA), ESANN'97 (Bruges, Belgium), *et cetera.* Finally, the first international workshop on blind source separation (BSS) and independent component analysis (ICA), ICA'99, brought together 130 researchers during one week in Aussois in the French Alps. Many BSS and ICA papers have been published in various journals, especially Signal Processing, Neural Networks, Neural Computation and IEEE Transactions on Signal Processing.

First papers on BSS concerned instantaneous or memoryless mixtures, but since 1991, source separation in convolutive mixtures have raised a greater interest. Issues in the separation of nonlinear mixtures remain almost unaddressed until very recently.

Initially, we wanted to focus this paper on source separation in nonlinear mixtures. However, the collegues from the HUT insisted on the importance of an historical recall, especially for those of you who are new in the field. The final idea was then to write an informal part on the "history" of BSS and ICA from the genesis, followed by the recent developments concerning nonlinear mixtures. The title of the paper may recall that kind of movies where anything can happen, but stay confident we tried to make the paper as easy to read as possible.

2. HISTORICAL COMMENTS OF BSS AND ICA

This part is a recall of the initial biological problem, followed by few comments explaining the development of BSS and ICA in the 80's context. We hope that this attempt will be neither too partial nor deadly boring. For a state-of-the-art, we recommend the papers of Cardoso [12] or of Karhunen [34].

2.1. A biological problem

The problem came to light in our group, in 1982, during an informal discussion with neuroscientists working

[*] Christian Jutten is professor in the Institut des Sciences et Techniques de Grenoble (ISTG) of the Université Joseph Fourier (UJF).

[†] Anisse Taleb is currently jointly affiliated with the Australian Telecommunications Research Institute (ATRI) and the School of Electrical and Computer Engineering at Curtin University of Technology.

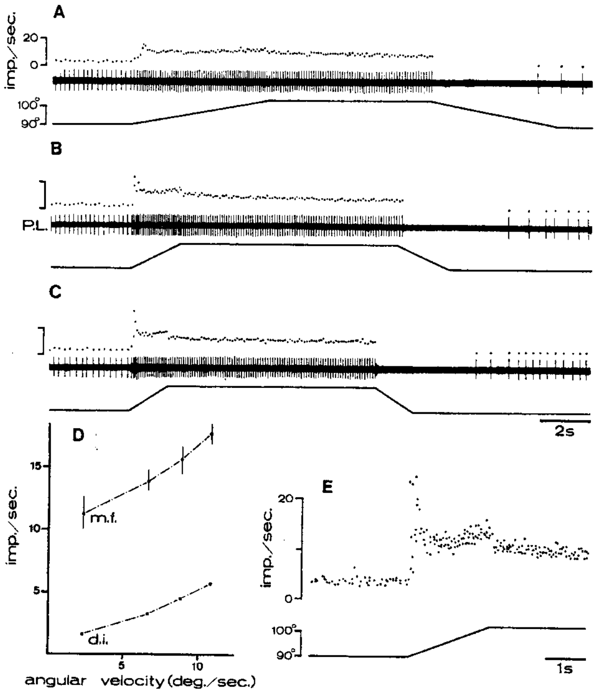

Figure 1: A, B and C) Primary ending responses to a joint movement imposed at 3 different constant velocities D) Frequency versus angular velocity E) Superimposition or three responses obtained with the same angular velocity. Reprinted from Roll [48]

on motion coding by proprioceptive fibers.
The joint motion is due to muscle contraction, each muscle spindle being controlled by the nervous system *via* a motor ending. Conversely, the contraction is measured (and transmitted to the central nervous system) on each muscle spindle by two types of (afferent) sensory endings located in the tendons, and called primary and secondary endings. One can easily observe the sensory responses for a passive joint movement imposed at constant velocity. Generally, the stability and the reproducibility of the responses are estimated by the frequencygram method, obtained by superposing several sensory ending responses corresponding to the repetition of the same imposed movement (Fig. 1 and 2). In the frequencygram, time is in abscissa and a spike occuring at time t is represented by one dot at time t, ordinate of which is equal to the instantaneous frequency, *i.e.* the inverse of the time interval between the spike and the previous one. Thus, with this representation, the response means the instantaneous frequency. Observations noted by Roll *et al.* [48] are the following:

- at constant joint position, the response of both primary and secondary endings is constant. The instantaneous frequency increases as the muscle is stretched *i.e.* as the joint is maintained at another angle. The ratio (frequency/muscle stretching) is similar, on the average, on both endings.
- conversely, at constant velocity, the responses of both endings differ. The response of primary endings is characterized by an initial burst (see Fig. 1), and then, during the movement, a uniform step associated to the uniform velocity motion, is superimposed to the response associated to the joint position. For the same angular motion (90 to 100), Fig. 1 shows that the step amplitude increases with the motion velocity. Secondary endings have similar responses, but with a low-pass transient behavior: no initial burst, the frequency decreases slowly as the motion finishes . Nevertheless, during the motion, the frequency increases with the velocity, but slower, on the average, than the instantaneous frequency recorded on primary endings.

Then, curiously, although there are two types of afferent endings, the messages of primary and secondary endings are mixtures of information on joint position and joint velocity.

We summarized these observations with the following simplified model. If one except the transient behavior of primary endings, denoting $p(t)$ the joint angular position, $v(t) = dp(t)/dt$ the joint angular velocity, $f_1(t)$ and $f_2(t)$ the instantaneous frequencies of one primary ending and one secondary ending, respectively, we proposed:

$$\begin{aligned} f_1(t) &= a_{11}v(t) + a_{12}p(t) \\ f_2(t) &= a_{21}v(t) + a_{22}p(t) \end{aligned}$$

with $a_{ij} > 0$ and $a_{11} > a_{21}$, but where $v(t)$, $p(t)$ as well as $a_{ij}, i, j = \{1, 2\}$ are unknown.

It seemed impossible to recover joint position and velocity from only the frequency sequence. However, as remarked by Mc Closkey [40] in 1978: *"Clearly, if spindle discharges are to be usefull for kinesthetic sensations, the central nervous system must be able to distinguish which part of the activity is attribuable to muscle stretch and which part is caused by fusimotor activity"*. Moreover, we were convinced that, even with imposed motions, even with close eyes, joint position and joint velocity could be separated by the central nervous system, although mixed in the primary and secondary ending responses.

Denoting $\boldsymbol{x}(t) = (f_1(t), f_2(t))^T$, where T stands for vector transposition, $\boldsymbol{s}(t) = (v(t), p(t))^T$, and $\boldsymbol{A}$ the

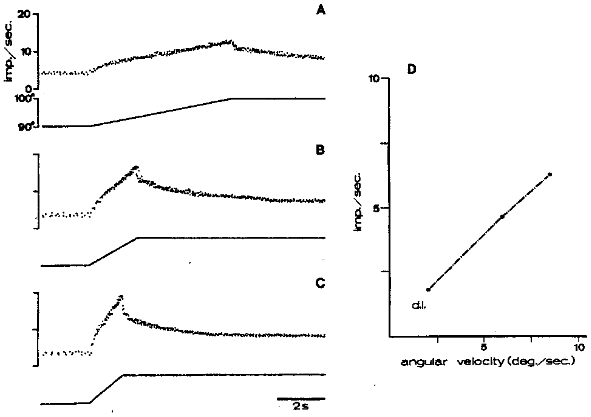

Figure 2: A, B and C) Secondary ending responses to a joint movement imposed at 3 different constant velocities D) Frequency versus angular velocity. Reprinted from Roll [48]

matrix with entries a_{ij}, we get the classic model of instantaneous mixtures

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t), \qquad (1)$$

In the following subsection, we sketch the difficulties for presenting the problem in 80's, and which explains the slow development of BSS up to 90's.

2.2. Contextual difficulties

2.2.1. Independence

The first problem concerns statistical independence of $v(t)$ and $p(t)$. In fact, due to the relation $v(t) = dp(t)/dt$, people often argued that the two quantities are dependent! In 1983-84, most peoples in signal processing or neural networks communities, are not very familiar with statistics. We tried to explain qualitatively that $v(t)$ and $p(t)$ are statistically independent, since knowing $v(t)$ gives no information on $p(t)$, and conversely. In other words, the distribution of $p(t)$ does not depend on the velocity value v, and conversely. Afterwards, to overcome this difficulty, we presented the problem, out of the biological context, but I remembered that there also were confusions for many peoples between statistical independence and linear independence (necessary between the observations for insuring existence of the mixture matrix inverse).

2.2.2. Second-order versus high-order statistics

This difficulty was also related to independence versus non correlation. Although the difference was emphasized in undergraduate probability lectures, in early 80's, researchers (except perhaps statisticians) assumed basically Gaussian models of signals and noises, and consequently merged non correlation and independence. Explaining this difference becomes easier with the rising interest for high order statistics (HOS), which started at about the same time: remember that the first workshop on HOS has been held in 1989 at Vail (Colorado, USA). Moreover, for approximating statistical independence, the first (heuristic) algorithm was based on the cancellation of several high order moments, but was not clearly derived from the cancellation of a unique independence criterion.

2.2.3. Separable or not?

In 1983, Bienvenu and Kopp [9] had proved that the eigenvectors of the spectral matrix span the signal subspace but cannot provide the sources. This last result is based on algebraic arguments: the number of equations was less than the number of parameters. The source separation problem was then considered as impossible to solve by signal processing researchers. When we presented a poster [28] in GRETSI'85, many peoples have then been intrigued by the result, but their comments were more or less incredulous: *"No, it is impossible"* or *"It is not impossible that it can work"*. Later, in 1987, J.-L. Lacoume, although he thought (because of [9]) separation was impossible, used high order statistics (4-*th* order cumulants) for expressing the independence assumption. His idea was to get more equations (as Nikias and others did for other problems) than using only second order statistics, which could perhaps overcome the above result [9]. And it worked [36]! He remarked later relationships between source separation and statistical independence, especially in developing a Maximum Likelihood approach for source separation, in which densities were approximated using a Gram-Charlier expansion [26].

2.2.4. Neural network context

In 1985 and 1986, we presented this work in neural networks conferences Cognitiva'85 and Snowbird 86. These communications raised probably curiosity of a few researchers. However, let us remember that, in Cognitiva'85, LeCun *et al.* [37] published a new learning scheme for multi-layer neural networks, which became very popular under the name 'backpropagation algorithm'. During Snowbird'86, neural networks researchers were more excited by Hopfield models, Kohonen self-organizing maps, multi-layer perceptrons (MLP) and backpropagation. I remember that, during this conference, Terry Sejnowski gave a very nice NETtalk demo, illustrating MLP applications. Recently, he told me that his interest for BSS and ICA began during this meeting: *"Because I did not understand why your network model could get the results that it did, new students in my lab often were offered that question as a research problem. Shaolin Li, a Chinese postdoc, made some progress by combining beamforming with H-J [38]. This project was started around 1991."*

2.2.5. The Terminology

One should remark that the concepts 'blind source separation' and 'independent component analysis' did not appear immediately. As an example, the first paper (published in 1984, in French, in in *Comptes Rendus de l'Académie des Sciences de Paris* [27], by J. Hérault and B. Ans) as well as the GRETSI'85 communication have so long and complex titles. Here is the translation: *Detection of primary quantities in a composite message with a neural computing architecture in unsupervised learning*! Then, in 1986, we spoke about the 'source discrimination' and finally, we used ICA and BSS since 1987.

2.3. Researcher mobilization

We believe that the success of the Signal Processing papers [32, 20, 50] has been due to a surprising performance with respect to the simplicity of the model and of the algorithm. Half a day was sufficient to write and test the algorithm. It has surprising performance, even if it depends on the mixture hardness and suffers from the lack of theory ... Many reasons to interest many people.

2.3.1. A few pioneers

During GRETSI'85, we actually raised great interest of L.Kopp (Thomson-Sintra company) who hired P.Comon for working on this problem in 1988, among others in antenna array processing. Comon tells: *"Thomson obtained a contract with the army, but nobody wanted to address that problem because it seemed so strange and unwonted. For the same reasons, I felt very attracted, and discovered after several weeks the central role played by cross cumulants. I could propose later an analytical solution in the noiseless case in the presence of two sources, and an iterative way to separate more sources [15]. I defined only later, in 1990, an information theoretic framework that led me to the formal definition of ICA and to contrast functions, allowing the computation of a solution in the presence of noise with unknown distribution [17]. The works of Darmois and Gassiat helped me in understanding the problem more deeply (but I unfortunately discovered the nice paper of Donoho on scalar deconvolution only two years later). I also noticed that the optimization was to be performed in the orthogonal group, and proposed a Jacobi-like algorithm. The full paper appeared in a special issue in 1994 [18]. I also tackled multichannel convolutive mixtures [16], but proposed rather late a class of contrasts without being able to establish the link to entropy criteria [19]. Now I am more interested in underdetermined mixtures (fewer sensors than sources).*

In September 1987, J.-F. Cardoso visited our lab, and we showed him a source separation demo with a hardware demonstrator: a purely analog device based on operational amplifiers, transistors, built in 1985, and which is able to separate, in real time, two audio sources mixed by potentiometers. Immediately, P.Comon and J.-F. Cardoso became enthousiastic about source separation, they met together later, in Vail (Colorado) in 1989. Cardoso explains: *"I became interested in source separation in 1987 after hearing Yann LeCun defend his thesis on back propagation: neural networks were arriving on our radar screens and I decided to look for applications of neural networks to signal processing (SP). A quick search led me to the first papers of Jutten and Hérault and I saw that the ICA model not only was relevant to sensor array processing –a hard core topic in SP– but was also bringing a fresh approach to it: the possibility of blind array processing. I started hacking fourth-order cumulant tensors in the spring of 1988 and in 1989 I had a paper [11] at the first IEEE SP High Order Statistics workshop in Vail, where I met Comon. Since doing blind array processing was crazy enough, I thought it would be safer to propose a simple algebraic method in which all the tensorial insanity would be concealed (tensors can be 'contracted' into innocuous matrices). [...] with JADE [14], we came out with a method which optimizes a contrast, but does so using a fast joint diagonalization algorithm. With the thesis of Beate Laheld, I started investigating on-line separation techniques and we were delighted to find that adaptive separation could be made simpler, faster and more efficient with 'relative gradient' algorithms [13] which offer a uniform performance. The underlying concept here is the so-called 'equivariant' nature of the ICA model. To me, other important dates are linked to Pham and Amari. In a beautiful paper [45], Pham showed us an efficient quasi-maximum likelihood solution which not only elucidates the role of non-linearity in source separation but also offers a simple technique to find good non-linear functions. [...] Another important moment for me has been when Amari joined the ICA band. I was fascinated by his 'information geometry' and was overjoyed that he brought in his insights and encouraged me to explore the information geometry of ICA."*

2.3.2. French and European supports

BSS and ICA took advantage of the researcher interactivity inside the French signal processing community. Since 1967, the biennal conference GRETSI bring together signal (and now image) processing researchers, now about six hundred people. Moreover, in 1989, a research group, funding by the French National Center for Research (CNRS) and Ministry of Research, has been created, for organizing scientific (informal) meetings in various topics (working groups). One of this working group, focused on HOS and then on BSS and ICA, has been supervised by J.-F. Cardoso from 1990 to 1997, who organized about 3 technical meetings per year, with on the average 8 talks and 30 attendees. This working group is still active, supervised now by E. Moreau. Finally, the Working Group ATHOS (Advance Topics in High Order Statistics) coordinated by P. Comon and funded by the European Community, contributed to promote BSS and ICA in the signal processing community. It is surprising that American researchers addressed the problem very late. We believe, and we have had some proofs in review reports of papers we submitted, that they considered the problem was simply a special case of blind multichannel equalization with trivial 0-order Moving Average filters. Of course, it is wrong, since in BSS and ICA the sources are not imposed to be i.i.d., but we often forgot to emphasize on this point.

2.3.3. From neural PCA to ICA

Independently, E. Oja, J. Karhunen *et al.* came to ICA and BSS by extending PCA neural networks which have been popular at the end of 80's. In a recent email, Karhunen explains: *"However, we knew that PCA can be realized more efficiently and accurately using standard numerical software, because the problem is linear. Therefore, we wanted to study nonlinear generalizations of PCA neural networks and learning rules. In those problems, there usually does not exist any such efficient conventional solution. We were also looking for nonlinear generalizations of PCA which could achieve something more than standard PCA. We developed several nonlinear neural extensions of linear PCA from different starting points. These developments are summarized in my two journal papers published in Neural Networks in 1994 and 1995... However, a problem with these extensions was that we had not at that time any convincing applications showing that nonlinear PCA is really useful and can provide something more than standard PCA. Independent component analysis is an extension of linear PCA, where uncorrelatedness assumptions are replaced by the stronger independence assumptions while relaxing the requirement of mutually orthogonal basis vectors. I was interested in that, especially after seeing your 1991 papers published in Signal Processing."*

2.3.4. From neural coding to ICA

The well known contribution of T. Bell and T. Sejnowsky [7] proposed some links between neural networks and entropy. However, the ideas which guided T. Bell were closer of theoretical biology. As said Terry Sejnowsky: *"Tony's motivation was based on a deep intuition that nature has used optimization principles to evolve nervous systems that need to self-organize channels in dendrites (his PhD thesis in Computer Science in Belgium) and in organizing the representations in the visual system. Although his 1995 paper in Neural Computation gets more citations, his 1997 paper in Vision Research is closer to his heart"* [8]. In the same spirit, J.-P. Nadal and N. Parga [41], from reflexions on information theory and the concept of sparse neural coding, introduced by Barlow at beginning of 60's [4], very early did interesting, although unrecognized, contributions [41]. Nadal explains: *"In 1994, together with Nestor Parga (Dep. de Fisica Teorica, UAM, Madrid), I was working on information theoretic approaches to sensory coding. At that time there were in the literature papers modelling, e.g., the early visual system, making use of two different "basic principles" : redundancy reduction, as proposed by the biologist H. Barlow in the 60's, and information maximization (Infomax). Predictions for, e.g., the receptive fields of ganglion cells, were quite similar. What we did with Nestor Parga [41] was to show that, under some conditions, Infomax leads to redundancy reduction. While writting the paper we realized that, in terms of signal processing, redundancy reduction is equivalent to ICA, and thus that we had demonstrated that Infomax is a proper cost function for performing ICA/BSS. This is said in the conclusion of our 94 paper. One year latter (1995) Bell and Sejnowski proposed their algorithm precisely based on Infomax"*.

2.3.5. RIKEN contributions

Last years, RIKEN institute in Japan, especially the groups of Amari and Cichocki in Wako-shi near Tokyo, have been very active in the field of BSS and ICA. Moreover, many researchers, coming from the whole world, have been invited or hired in their research groups for working on these problems. A. Cichocki writes: *"I have started close and fruitful collaboration with Professor Amari and also other researchers [...] from April 1995 when I joined Frontier Research Program Riken,* *JAPAN and I would like to mention that I have learned a lot form Professor Amari and His ideas.*

Before this fruitful collaboration I have started to study BSS/ICA since 1991 after reading several of your influential papers, including your Doctor of Science Thesis and works of your Ph.D students. When I was in Germany in 1992-1994, at University Erlangen Nuremberg we have published several brief papers (Electronics Letters 1992/94 IEEE Transaction on Circuits and Systems) and also in our book (in April 1993) we presented neural network approach to BSS: Neural Networks for Optimization and Signal Processing by A. Cichocki and R. Unbehauen (J. Wiley 1993 pp.461-471)."

And Shun-Ichi Amari adds : *'I knew the Jutten-Hérault idea of source separation in late eighties, and had interest in, but did not do any work on that subject. It was in 1994 when Cichocki visited us and emphasized the importance of the subject that I had again interest. He showed me a number of papers in 1995, one of which is Bell-Sejnowski paper. I was impressed by that one, and thought that I could study more general mathematical aspects.*

One of the results is the idea of natural gradient, which I we proposed in our 1995 NIPS paper (Amari-Cichocki-Yang, appeared in Proc. NIPS, 1996). The algoritm itself was proposed by Cichocki earlier, and also by Cardoso in 1995. But ours has a rigorous foundation based on the Lie group invariance and Riemannian metric derived therefrom.

From that on, I have carried out intensive research on this interesting subject, in particular, its mathematical foundations, in collaboration with Cichocki and many others.

One is its foundation from the point of semiparametric statistical models and information geometry [2]. [...] I also have studied efficiency and super-efficiency of algorithms [1]. There are a number of other ideas, but it is too much to state all of them".

2.3.6. From statistics to ICA

Clearly, independence is related to probability and statistics, and statisticians can bring a lot on source separation. For instance, it appears that factorial analysis, intensively studied in statistics in 50's, is another way to formalize ICA, especially the separability problem, and that results were available for many years. The Darmois's results [22] have been brought to light by P. Comon in 1991 [17], and more recently, researchers used a few theorems published in the statistics book [33] published in 1973! We finish by an anecdote, which shows that probably it would have been possible to go faster. In 1990, for administrative reasons, a PhD student had to register in statistics instead signal processing post-graduate courses in Grenoble. But he wanted absolutely to work on source separation, and D.T. Pham accepted to supervise him. J. Hérault and myself gave a short informal talk to D.T. Pham on source separation. Three days after, he sent us a 5 or 6-page note in which he sketched the Maximum Likelihood solution and emphasized on the relevance of score functions. Finally, the nonlinear functions used in our first algorithm correspond to a heuristic choice (fortunately robust), optimal for particular distributions. Since this date, he bring nice contributions [45, 46, 44] to the problem.

3. SOURCE SEPARATION IN NONLINEAR MIXTURES

When linear models fail, nonlinear models, because of their better approximation capabilities, appear to be powerfull tools for modeling practical situations. Examples of actual nonlinear systems include digital satellite and microwave channels which are composed of a linear filter followed by a memoryless nonlinear travelling wave tube amplifier [25], magnetic recording channel, *et cetera.* Such systems are, therefore, of great theoretical and practical interest.

While linear source separation in both the instantaneous and the convolutive case has been intensively studied, extension and generalisation to nonlinear models has only been done in a very sparse way. It is amusing to notice that this is not specific to source separation but seems to be a quite general observation.

3.1. Brief state-of-the-art

One may think that the interest in the separation of nonlinear mixtures is recent, but this is not true. The first reference to nonlinear mixtures is, at our best knowledge, by Jutten [31] in 1987. He used soft nonlinear mixtures in order to assess the robustness and the performance of the HJ algorithm. The mixtures he used was those obtained in stereoscopic vision and those obtained by a spherical coordinates transformation. He showed experimentally that the algorithm converges to the linear approximation of the mixture when this one is not too hard. Later, Burel [10] has proposed a neural network based solution for known nonlinearity depending on unknown parameters, however the cost function he used was very complex and leads to a very complex algorithm.

Pajunen *et al.* [42] have addressed the problem using self-organizing maps. This approach, although simple and very attractive, requires a huge number of neurons for good accuracy, and is restricted to sources having probability density functions with bounded supports, and the more important is that it tends to modify the sources distribution and somehow have uniformally distributed outputs. To overcome this difficullty, Pajunen *et al.*[42] use a modified GTM (generative topographic mapping) who is inspired itself by the SOM. The modification consists in forcing the output to have a specified, *a priori* known, source distribution.

Deco and Brauer [23] have also addressed the problem, considering a volume conservation condition on the nonlinear transforms. This constraint leads to very restrictive transforms. We also notice the contribution of Yang *et al.* [58] who proposed algorithms for special nonlinear mixtures, similar to post-nonlinear mixtures which will be discussed *infra*, in which the nonlinearity is not componentwise and whose inverse is supposed to be approximated by a two-layer perceptron.

Most of the works cited above, even if it was based on a good intuition, lacked theoritical justification. Ideally, solid theory should be behind practice. It is well known that, unlike the case of linear systems, prior knowledge of the model is necessary for nonlinear system identification [49]. So let us define the generic model of nonlinear mixtures.

3.2. Generic model

In this paper we are only concerned by the instantaneous nonlinear mixtures (yes things are quite complex in this case!). The problem of source separation in general nonlinear instantaneous mixtures consists in retrieving unobserved sources $\boldsymbol{s}(t) = (s_1(t), \ldots, s_n(t))^T$, by only observing a nonlinear mixture $\boldsymbol{x}(t)$:

$$\boldsymbol{x}(t) = \mathcal{F}(\boldsymbol{s}(t)) \qquad (2)$$

where $\mathcal{F}$ is an unknown nonlinear one-to-one mapping. This is done by constructing a nonlinear transform $\mathcal{G}$ (separation structure) in order to isolate in each component of the output vector $\boldsymbol{y}(t) = \mathcal{G}(\boldsymbol{e}(t))$ the image of one source. In other words, each component of the output vector $\boldsymbol{y}(t)$ must depend on only one component of the sources vector $\boldsymbol{s}(t)$. In the general case, this image (dependency) can be a nonlinear function of the source. In summary, the goal is to obtain:

$$y_i(t) = k_i(s_{\sigma(i)}(t)), \ i = 1, 2, \ldots, n \qquad (3)$$

where σ is a permutation on $\{1, 2, \ldots, n\}$, and k_i is a function which represents a probable residual distortion. It is clear that, without additional assumptions on the sources, the problem is ill posed. A natural assumption consists in supposing that the sources are statistically independent, which means that the joint density of the sources factorizes as the product of their marginal densities:

$$p(s_1, s_2, \ldots, s_n) = p(s_1)p(s_2)\ldots p(s_n) \qquad (4)$$

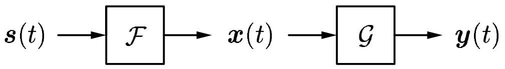

Figure 3: Instantaneous nonlinear mixture and its separation structure

Given this assumption, the role of the separation structure is to transform the observations vector $\boldsymbol{e}(t)$ to a vector $\boldsymbol{y}(t)$ with independent components.

Is the statistical independence assumption sufficient to obtain a separation in the sense of equation (3) and at what indeterminacies? The next section is concerned by this question, but we can already affirm that the separability will be closely related to the structural constraints on the separation structure and the indeterminacies will characterize the function k_i of (3).

In the following, the dependency over time will be ignored since the model is instantaneous and only spatial properties are used, moreover we assume that all the signals are real valued.

3.3. Separability

From the previous discussion, the separation structure provides:

$$\boldsymbol{y} = \mathcal{G} \circ \mathcal{F}(\boldsymbol{s}) = \mathcal{H}(\boldsymbol{s}) \qquad (5)$$

Provided the independence assumption, separability consists in determining the form of the transformations $\mathcal{H}$ which leaves the components of $\boldsymbol{s}$ independent.

There is a strong relation between the objective of source separation, as defined by equation (3), and the statistical independence assumption. This comes from the notion of trivial transformations.

A one-to-one mapping $\mathcal{H}$ is called *trivial*, if it transforms any random random vector $\boldsymbol{s}$ with independent components into a random vector with independent components. The set of trivial transformations will be denoted by $\mathfrak{Z}$.

Trivial transformations are then transformations conserving the independence property of any random vector. One can easily show that a one-to-one mapping $\mathcal{H}$ is trivial if and only if it writes as:

$$\mathcal{H}_i(u_1, u_2, \ldots, u_n) = h_i(u_{\sigma(i)}), \ i = 1, 2, \ldots, n \qquad (6)$$

where h_i are arbitrary functions and σ is any permutation over $\{1, 2, \ldots, n\}$.

This result establishes a link between the independence assumption and the objective of source separation. In fact, it becomes clear that the source separation objective is, using the independence assumption, to impose that the global transformation $\mathcal{H} = \mathcal{G} \circ \mathcal{F}$ is trivial.

However this is not possible without imposing additional structural constraints on $\mathcal{H}$, as we shall see in the next section.

3.3.1. General results from factor analysis

In the general case, *i.e.* the transformation $\mathcal{H}$ has no particular form, a well known statistical result shows that the independence conservation constraint is not strong enough to insure the separability in the sense of equation (3). This result has been established, early in the 50's, by Darmois [21][1] where he used a simple constructive method for decomposing any random vector as a non trivial mapping of independent variables.

This result is negative, in the sense that it show the existence of non trivial transformations $\mathcal{H}$ which *"mix"* the variables while conserving their statistical independence. Hence, for general nonlinear transformations and without constraints on the transformations model, source separation is simply *impossible* by only using the statistical independence.

In the conclusion of [21], Darmois clearly states: *"These properties,[...], clarify the general problem of factor analysis by showing the large indeterminacies it presents as soon as one leaves the field, already very wide, of linear diagrams."*

If you are still not convinced, here is a nice and simple example:

$$p_{XY}(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right), \quad (x, y) \in \mathbb{R}^2 \tag{7}$$

and consider the following nonlinear transform:

$$\left\{ \begin{array}{l} X = r\cos\theta \\ Y = r\sin\theta \end{array} \right. \tag{8}$$

with $r \in \mathbb{R}^+$ and $\theta \in [0, 2\pi[$. This transform has a full rank Jacobian matrix provided that $r \neq 0$:

$$\boldsymbol{J} = \left[\begin{array}{cc} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{array} \right], \tag{9}$$

and the joint pdf of R and Θ is then:

$$p_{R,\Theta}(r, \theta) = \left\{ \begin{array}{ll} \frac{r}{2\pi} e^{-r^2/2} & (r, \theta) \in \mathbb{R}^+ \times [0, 2\pi[\\ 0 & \text{otherwise} \end{array} \right. \tag{10}$$

Relation (10) shows that the two random variables, R and Θ are statistically independent. Other examples can be found in the litterature (see for example Lukacs [39]) or can be easily constructed.

3.3.2. Specific model

The previous negative result is due to the fact that we assume no constraints on the transformation $\mathcal{H}$. Constraining the transformation $\mathcal{H}$ in a certain set of transformation $\mathfrak{Q}$ can reduce these strong indeterminacies.

To characterize the indeterminacies for a specific model $\mathfrak{Q}$, one must solve the independence conservation equation which writes as:

$$\forall E \in \mathfrak{M}_n,$$
$$\int_E dF_{s_1} dF_{s_2} \cdots dF_{s_n} = \int_{\mathcal{H}(E)} dF_{y_1} dF_{y_2} \cdots dF_{y_n} \tag{11}$$

where $\mathfrak{M}_n$ is a σ-algebra on $\mathbb{R}^n$. Let $\mathfrak{P}$ denote the set

$$\mathfrak{P} = \{(F_{s_1}, F_{s_2}, \ldots, F_{s_n}), \text{ such that } \exists \mathcal{H} \in \mathfrak{Q} \backslash (\mathfrak{Z} \cap \mathfrak{Q}) : \\ \mathcal{H}(\boldsymbol{s}) \text{ has independent components}\} \tag{12}$$

This set contains all sources distributions for which there exists a non trivial transformation $\mathcal{H}$ belonging to the model $\mathfrak{Q}$ and conserving the independence of the $\boldsymbol{s}$ components.

An ideal model will be such that $\mathfrak{P}$ is empty and $\mathfrak{Z} \cap \mathfrak{Q}$ contains the identity as a unique element. However, in general this is not fulfilled. We then say that source separation is possible when the sources distribution is in $\bar{\mathfrak{P}}$, the complement of $\mathfrak{P}$, sources are then restored up to a trivial transformation belonging to $\mathfrak{Z} \cap \mathfrak{Q}$.

3.3.3. Example: Linear models

In the case of linear models, the transformation $\mathcal{F}$ is linear and can be represented by an $n \times n$ matrix $\boldsymbol{A}$, the observed signals write then as $\boldsymbol{e} = \boldsymbol{As}$. Source separation consists then in finding a matrix $\boldsymbol{B}$ such that $\boldsymbol{y} = \boldsymbol{Be} = \boldsymbol{Hs}$ has independent components.

The set of linear trivial transformations $\mathfrak{Z} \cap \mathfrak{Q}$ is the set of matrices equal to the product of a permutation and a diagonal matrix. By the Darmois-Skitovich theorem [33], the set $\mathfrak{P}$ contains the distributions having at least two Gaussian components. We then conclude that source separation is possible whenever we have at most one Gaussian source. Sources are then restored up to a permutation and a diagonal matrix.

3.4. Separation of PNL mixtures

A postnonlinear model (PNL) consists in observing :

$$x_i(t) = f_i(\sum_{j=1}^{n} a_{ij} s_j(t)), \; i = 1, \ldots, n, \tag{13}$$

Figure 4 shows what this model looks like. One can see that this model is a cascade of a linear mixture and a

[1] This paper can hardly be obtained and, at our knowledge, it is available at The British Library.

componentwise nonlinearity, *i.e.* acts on each output independently from the others. The nonlinear functions (distortions) f_i are supposed invertible. Besides its theoritical interest, this model, belonging to the L-ZMNL[2] family, sticks perfectly for a lot of real world applications. For instance, such models can be found in sensors arrays [43], satellite and microwave communications [47], and in a lot of biological systems [35].

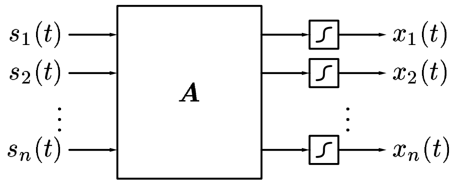

Figure 4: Postnonlinear mixture

As discussed before, the most important thing when dealing with nonlinear mixtures is the separability problem. First, we must think about the separation structure $\mathcal{G}$ which has as constraints:

1. **Can invert the mixing system** in the sense of equation (3): this constraint is quite obvious because that's what we want!

2. **Be as simple as possible**: In fact we want to reduce, in case we are successful, the residual distortions k_i which are the blind spot of the independence assumption.

By defining these two constraints, we have no other choice that selecting for $\mathcal{G}$ the mirror structure of $\mathcal{F}$ (Fig. 5). The total transformation $\boldsymbol{y}(t) = \mathcal{H}(\boldsymbol{s}(t)) = \mathcal{G} \circ \mathcal{F}(\boldsymbol{s}(t))$ can then be written as:

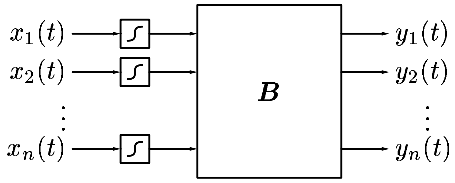

Figure 5: Separation structure

$$\begin{aligned} y_i(t) &= \sum_{j=1}^{n} b_{ij} g_j(x_j(t)) \\ &= \sum_{j=1}^{n} b_{ij} h_j(\sum_{k=1}^{n} a_{jk} s_k(t)), \; i = 1, \dots, n, \end{aligned} \tag{14}$$

where $h_j = g_j \circ f_j$ for $j = 1, 2, \dots, n$. In this case, the set $\mathfrak{Q}$ consists in a L-ZMNL-L transformation.

One needs to know what are the trivial transformations belonging to the model, *i. e.* characterize the set $\mathfrak{Z} \cap \mathfrak{Q}$. This comes by solving the following system of functional equations:

$$\sum_{j=1}^{n} b_{ij} h_j(\sum_{k=1}^{n} a_{jk} u_k) = k_i(u_{\sigma(i)}) \; i = 1, \dots, n, \tag{15}$$

where $\boldsymbol{u} \in \mathbb{R}^n$. It can easily be shown [55] that the solutions of this functionnal equation are linear functions when the matrix $\boldsymbol{A}$ has at least two non zero elements per row or per column. In this case the matrix $\boldsymbol{A}$ is called mixing enough.

The second, and the more difficult, step consists in determining the set $\bar{\mathfrak{P}}$ of distributions for which $\mathcal{H}$ will necessarily belongs to $\mathfrak{Z} \cap \mathfrak{Q}$. This, unfortunately, has not yet been fully established. What is shown is that, under some conditions, distributions which vanish at some points can be separable [55, 51].

The study of the separability of PNL mixtures gives birth to a new problem in the factorization theory of probability distributions [33], it consists in characterizing the distribution of the random variables X, Y which admitt the following two representations:

$$\begin{aligned} X &= f(\sum_{i=1}^{n} a_i x_i) = \sum_{i=1}^{p} \alpha_i y_i \\ Y &= g(\sum_{i=1}^{n} b_i x_i) = \sum_{i=1}^{p} \beta_i y_i \end{aligned} \tag{16}$$

where both $\{x_1, x_2, \dots, x_n\}$ and $\{y_1, y_2, \dots, y_p\}$ are independent, f and g bijective. We conjecture that if X and Y are not independent then f and g are necessarily linear in their domain of definition (support of $\sum_{i=1}^{n} a_i x_i$ and $\sum_{i=1}^{n} b_i x_i$).

We first introduced the PNL model in [53], where MLPs are used to invert the nonlinear functions f_i. We used the maximum likelihood as a cost function, where we replaced the unknown sources pdfs by a Gram-Charlier expansion up to the forth order. The obtained results were quite good when the nonlinearities were not too strong. In the case of strong nonlinearities, the

[2] L stands for Linear and ZMNL stands for Zero-Memory Non-Linearity.

inversion of the nonlinearities was poor. We suspected the approximation of the pdfs and in fact we were right. By replacing this approximation by an online estimation of the score functions developed in [52] we got good results even in bad situations [55] (see for instance Fig. 6 and Fig. 7). Moreover, contrary to linear mixtures, separation performance in nonlinear mixtures depends on the estimation performance of score functions [51].

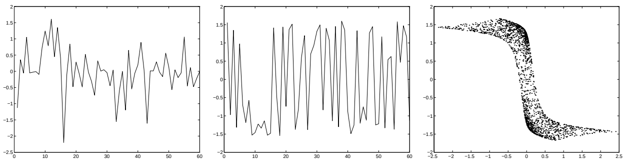

Figure 6: Observed mixture signals (left, center), and their joint distribution (right)

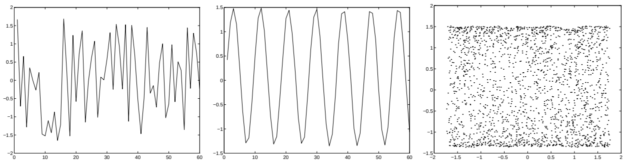

Figure 7: Algorithm outputs (left, center), and their joint distribution (right)

Finally, in ICA'99, we proposed a nonparametric algorithm [54], the idea was to get rid of any parametrisation of the functions g_i. The algorithm performes well, and it is able to invert with equal performance either hard saturations and hard cubic-like distortions[3].

3.5. Extensions

In our work, we focused on PNL mixtures as a first step towards understanding nonlinear mixtures and the new problems they opened. Other nonlinear mixtures can be considered. For instance, cascade of PNL mixtures which model successive nonlinear signal amplification and transmission. Convolutive PNL mixtures, in which the mixing matrix consists of filters, and which can model low-quality microphones. We can also consider Wiener nonlinear systems which are the counterpart of PNL models in the time-domain (Fig. 8) and also the cascade of Wiener systems.

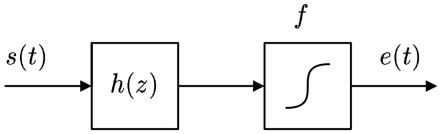

Figure 8: Wiener system

The class of Wiener systems is not only another nice and mathematically attracting model, but also a model found in various areas, such as biology: study of the visual system [24], relation between the muscle length and tension [29], industry: description of a distillation plant [6], sociology and psychology. See also [30] and the references therein. Despite its interest, at our knowledge, no blind procedure exists for the inversion of such systems.

Wiener systems are very similar to PNL models, in fact when the input $s(t)$ is an iid process, which means that the samples are independent, one can consider that we have an infinite number of independent sources which are linearly mixed (filter h) and then distorded by the ZMNL f. An efficient algorithm for the inversion of such systems is proposed in [56] and [57]. However, the separability result (in infinite dimension) is only conjectured.

In our opinion, the problem is somehow related to channel coding. In fact, if we consider that the nonlinearity f represents the channel, and the filtered signal $h * s(t)$ a convolutive channel coding of $s(t)$. Then, it seems that this coding is sufficient to completely identify the channel and to blindly compensate its effect. Is the redundacy introduced by the filter sufficient? What is the minimum required redundacy? These questions have presently no answer. The funnier thing is that this h-convolutive coding is done, in almost all situations, by mother nature itself. This remark also holds for PNL models, where the coding is done by the mixing matrix $\boldsymbol{A}$.

4. CONCLUSION

In this paper, we recalled the genesis of source separation in our group as well as in a few others. We are conscious that this part does only consider a few developments of BSS and ICA. Especially, advances for convolutive mixtures (due to Yellin and Weinstein, Nguyen Thi *et al.*, Broman and Lindgren, Loubaton *et al.*, *et cetera*) have not been touched, despite their importance in communications. The second part pointed out recent results, mainly developed in Taleb's Ph.D [51], on nonlinear mixtures. This opens large perspectives for investigating the problem of ICA and BSS from a more general point of view: other separable nonlinear mixtures, both convolutive and nonlinear mixtures, ICA for image and data analysis, *etc.* and for applying the methods in many domains: communications, medicine,

[3] An online demonstration of the algorithm and Matlab code can be found at `http://www.atri.curtin.edu.au/csp/anisse/ss-demo/` or at `http://helio.inpg.fr/weblis1/personnes/jutten/index.html`

etc. Let us conclude with J.-F. Cardoso: *"It is amazing to see how much attention the deceptively simple model of ICA as attracted over more than 10 years and to realize that the field is still open to investigation"*.

5. REFERENCES

[1] S.I. Amari. Superefficiency in blind source separation. *IEEE Trans. on SP*, 47(4):936–944, April 1999.

[2] S.I. Amari and Cardoso J.-F. Blind source separation - semiparametric statistical approach. *IEEE Trans. on SP*, 45(11):2692–2700, November 1997.

[3] B. Ans, J. Hérault, and C. Jutten. Adaptive neural architectures: Detection of primitives. In *Proceedings of COGNITIVA'85*, pages 593–597, Paris, France, 4-7 June 1985.

[4] H.B. Barlow. Possible principles underlying the transformation of sensory messages. In W. Rosenblith, editor, *Sensory Communication*, page 217. MIT Press, 1961.

[5] J.W. Barness, Y. Carlin and M.L. Steinberger. Bootstrapping adaptative interference cancelers: some practical limitations. In *Proc. The Globecom. Conference*, pages 1251–1255, 1982.

[6] R. Bars, I. Bèzi, B. Pilipàr, and B. Ojhelyi. Nonlinear and long range control of a distillation pilot plant. In *Identification and Syst. Parameter Estimation; Preprints 9th IFAC/IFORS Symp.*, pages 848–853, Budapest (Hungary), July 1990.

[7] T. Bell and T. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Comutation*, 7(6):1004–1034, 1995.

[8] T. Bell and T. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[9] G. Bienvenu and L. Kopp. Optimality of high-resolution array processing using the eigensystem approach. *IEEE Trans. on ASSP*, 31(5):1235–1248, 1983.

[10] G. Burel. Blind separation of sources: a nonlinear neural algorithm. *Neural Networks*, 5:937–947, 1992.

[11] J.-F. Cardoso. Blind identification of independent signals. In *Workshop on Higher-Order Spectral Analysis*, Vail (CO), USA, June 1989.

[12] J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings IEEE*, 9:2009–2025, 1998.

[13] J.-F. Cardoso and B. Laheld. An information-maximization approach to blind separation and blind deconvolution. *IEEE Trans. on SP*, 44:3017–3030, 1996.

[14] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140:362–370, December 1993.

[15] P. Comon. Separation of stochastic processes. In *Proc. Workshop on Higher-Order Spectral Analysis*, pages 174–179, Vail, Colorado, June 28-30 1989. IEEE-ONR-NSF.

[16] P. Comon. Analyse en composantes indépendantes et identification aveugle. *Traitement du signal*, 7(5):435–450, 1990.

[17] P. Comon. Independent component analysis. In J.-L. Lacoume, M. A. Lagunas, and C. L. Nikias, editors, *International Workshop on High Order Statistics*, pages 111–120, Chamrousse, France, July 1991.

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

[19] P Comon. Contrasts for multichannel blind deconvolution. *IEEE Signal Processing Letters*, 3(7):209–211, 1996.

[20] P. Comon, C. Jutten, and J. Hérault. Blind separation of sources, Part II: Statment problem. *Signal Processing*, 24(1):11–20, 1991.

[21] G. Darmois. Analyse des liaisons de probabilité. In *Proceedings Int. Stat. Conferences 1947*, volume III A, page 231, Washington (D.C.), 1951.

[22] G. Darmois. Analyse générale des liaisons stochastiques. *Rev. Inst. Intern. Stat.*, 21:2–8, 1953.

[23] G. Deco and W. Brauer. Nonlinear higher-order statistical decorrelation by volume-conserving architectures. *Neural Networks*, 8:525–535, 1995.

[24] A. C. den Brinker. A comparison of results from parameter estimations of impulse responses of the transient visual system. *Biol. Cybern.*, 61:139–151, 1989.

[25] K. Feher. *Digital Communications–Satellite/Earth Station Engineering*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[26] M. Gaeta and J.-L. Lacoume. Estimateurs du maximum de vraisemblance étendus à la séparation de sources non gaussiennes. *Traitement du Signal*, 7(5):419–434, 1990.

[27] J. Hérault and B. Ans. Circuits neuronaux à synapses modifiables : décodage de messages composites par apprentissage non supervisé. *C.R. de l'Académie des Sciences*, 299(III-13):525–528, 1984.

[28] J. Hérault, C. Jutten, and B. Ans. Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. In *Actes du Xeme colloque GRETSI*, pages 1017–1022, Nice, France, 20-24 Mai 1985.

[29] I. W. Hunter. Frog muscle fiber dynamic stiffness determined using nonlinear system identification techniques. *Biophys. J.*, 49:81a, 1985.

[30] I. W. Hunter and M. J. Korenberg. The identification of nonlinear biological systems: Wiener and Hamerstein cascade models. *Biol. Cybern.*, 55:135–144, 1985.

[31] C. Jutten. *Calcul neuromimétique et traitement du signal : analyse en composantes indépendantes.* Thèse d'état ès sciences physiques, UJF-INP Grenoble, 1987.

[32] C. Jutten and J. Hérault. Blind separation of sources, Part I: an adaptive algorithm based on a neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

[33] A.M. Kagan, Y.V. Linnik, and C.R. Rao. *Characterization Problems in Mathematics Statistics.* John Wiley & Sons, 1973.

[34] J. Karhunen. Neural approaches to independent component analysis and source separation. In *ESANN'96, European Symposium on Artificial Neural Networks*, pages 249–266, Bruges, Belgium, April 1996.

[35] M.J. Korenberg and I.W. Hunter. The identification of nonlinear biological systems: Lnl cascade models. *Biological Cybernetics*, 43(12):125–134, December 1995.

[36] J.-L. Lacoume and P. Ruiz. Sources identification: a solution based on cumulants. In *IEEE ASSP Workshop*, Mineapolis, USA, August 1988.

[37] Y. Le Cun. A learning scheme for assymetric threshold network. In *Proceedings of COGNITICA'85*, pages 599–604, Paris, France, 1985.

[38] S. Li and T. Sejnowski. Adaptive separation of mixed broadband sound sources with delays by a beamforming hérault-jutten network. *IEEE Trans. on Oceanic Engineering*, 20:73–79, 1995.

[39] E. Lukacs. A characterization of the gamma distribution. *Ann. Math. Statist.*, (26):319–324, 1955.

[40] D.I. Mc Closkey. Kinesthetic sensibility. *Physiol. Reviews*, 58(4):763–820, 1978.

[41] J.-P. Nadal and N. Parga. Nonlinear neurons in the low-noisy limit: a factorial code maximizes information transfer. *Network*, 5:565–581, 1994.

[42] P. Pajunen, A. Hyvarinen, and J. Karhunen. Non linear source separation by self-organizing maps. In *ICONIP 96*, Hong-Kong, September 1996.

[43] A. Parashiv-Ionescu, C. Jutten, A.M. Ionescu, A. Chovet, and A. Rusu. High performance magnetic field smart sensor arrays with source separation. In *MSM 98*, pages 666–671, Santa Clara, USA, April 1998.

[44] D. T. Pham. Blind separation of instantaneous mixture of sources based on order statistics. *IEEE Trans. on SP*, 48:363–375, 2000.

[45] D. T. Pham and Ph. Garat. Blind separation of mixture of independent sources through a quasimaximum likelihood approach. *IEEE Trans. on SP*, 45:1712–1725, 1997.

[46] D.T. Pham. Mutual information approach to blind separation of stationary sources. In *Proceedings of ICA'99*, pages 215–220, Aussois, France, January 1999.

[47] S. Prakriya and D. Hatzinakos. Blind identification of lti-zmnl-lti nonlinear channel models. *IEEE trans. S.P.*, 43(12):3007–3013, December 1995.

[48] J.-P. Roll. *Contribution à la proprioception musculaire, à la perception et au contrôle du mouvement chez l'homme.* PhD thesis, Université d'Aix-Marseille 1, 1981.

[49] M. Schetzen. Nonlinear system modeling based on the Wiener theory. *Proc. IEEE*, 69:1557–1573, December 1981.

[50] E. Sorouchyari. Blind separation of sources, Part III: Stability analysis. *Signal Processing*, 24(1):21–29, 1991.

[51] A. Taleb. *Séparation de sources dans des mélanges non-linéaires.* PhD thesis, Inpg, 1999. (In French).

[52] A. Taleb and C. Jutten. Entropy optimization, application to blind source separation. In *ICANN 97*, pages 529–534, Lausanne (Switzerland), October 1997.

[53] A. Taleb and C. Jutten. Non-linear source separation: the post-non-linear mixtures. In *ESANN'97*, pages 279–284, Bruges, Belgium, 1997.

[54] A. Taleb and C. Jutten. Batch algorithm for source separation in postnonlinear mixtures. In *ICA 99*, pages 155–160, Aussois (France), January 1999.

[55] A. Taleb and C. Jutten. Source separation in post nonlinear mixtures. *IEEE Tr. on SP*, 47(10):2807–2820, 1999.

[56] A. Taleb, J. Sole i Casals, and C. Jutten. Blind inversion of wiener systems. In *Proceedings of IWANN'99*, Alicante, Spain, 1999.

[57] A. Taleb, J. Sole i Casals, and C. Jutten. Quasi-nonparametric blind inversion of wiener systems. *IEEE Tr. on SP*, 1999. Submitted.

[58] H. H. Yang, S. Amari, and A. Cichocki. Information-theoritic approach to blind separation of sources in nonlinear mixture. *Signal Processing*, 64(3):291–300, 1998.