Redundancy Reduction and Independent Component Analysis: Conditions on Cumulants and Adaptive Approaches

Jean-Pierre Nadal
Laboratoire de Physique Statistique de l'E.N.S.[*]
Ecole Normale Supérieure
24, rue Lhomond, F-75231 Paris Cedex 05, France

Nestor Parga
Departamento de Física Teórica
Universidad Autónoma de Madrid
Cantoblanco, 28049 Madrid, Spain

Abstract

In the context of both sensory coding and signal processing, building *factorized codes* has been shown to be an efficient strategy. In a wide variety of situations, the signal to be processed is a linear mixture of statistically independent sources. Building a factorized code is then equivalent to performing blind source separation. Thanks to the linear structure of the data, this can be done, in the language of signal processing, by finding an appropriate linear filter, or equivalently, in the language of neural modeling, by using a simple feedforward neural network.

In this paper we discuss several aspects of the source separation problem. We give simple conditions on the network output which, if satisfied, guarantee that source separation has been obtained. Then we study adaptive approaches, in particular those based on redundancy reduction and maximisation of mutual information. We show how the resulting updating rules are related to the BCM theory of synaptic plasticity. Eventually we shortly discuss extensions to the case of non linear mixtures. In all the paper we take care to put into perspective our work with other studies on source separation and redundancy reduction. In particular we review algebraic solutions, pointing out to their simplicity but also their common drawbacks.

Submitted to *Neural Computation*

[*]Laboratoire associé au C.N.R.S. (U.R.A. 1306), à l'ENS, et aux Universités Paris VI et Paris VII.

Introduction

In the recent years many studies have been devoted to the study of sensory coding following a general framework initiated by H. Barlow more than thirty years ago [6]. The general idea is to define a cost function based on the properties one thinks a neural code should satisfy. Then, given a neural architecture with a simpler enough neuron model, one derives the parameters of the network (synaptic efficacies, transfer functions,...) which minimizes this cost function. A great deal of work has been done on cost functions based on information theoretic criteria [9, 30, 3]. The result for the receptive fields will crucially depend on the statistical properties of the signal to be processed (visual scenes, olfactory stimuli,...). Several cost functions have been proposed. One is based on the original idea of Barlow [6], which is that *redundancy reduction* should be performed. This leads to the notion of *factorial code* [9, 34]: in the pool of neurons defining the neural code (the "output" neurons), each neuron should code for features statistically independent of the features encoded by the other neurons. Redundancy reduction has been explored with a lot of details in the case of visual processing, with a linear neural network model [3]. A different proposal has been promotted by Linsker under the name of *infomax* principle [30]: one asks the network to simply maximize the *mutual information* between the input (the signal received onto the receptors) and the output (the neural code). Again, most studies have been performed for linear networks in the context of visual processing [30, 39, 17]. The predictions of these two strategies, redundancy reduction and maximization of mutual information, appeared to be very similar. In fact, we have shown [32] that, in the low synaptic noise limit with non linear outputs, infomax implies redundancy reduction, when optimization is done over both the synaptic efficacies and the non linear transfer functions. As a result, the optimal neural representation is a factorial code, for both cost functions in the limit of low processing noise. When noise is present, it can easily be shown [3] that pure redundancy reduction is not a meaningful strategy. Essentialy, an efficient code is one which gives the independent features, but add some redundancy to compensate for the noise. This can be studied in detail for a linear neural network [17], where one can find, for instance, the number of meaningful principal components at a given noise level.

In the case of signal processing and data analysis, however, the processing noise can indeed be neglected. It is then interesting to search for algorithms able to perform redundancy reduction, and to determine the neural architectures the best adapted to a given type of signal. Of particular interest is the case where a factorial code can be found by a feedforward network with no hidden layer. This implies that there exists a linear combination of the input that produces a factorial code. This arises when the input (the signal to be processed) itself is a linear mixture of statistically independent sources. Then finding the synaptic efficacies leading to a factorial code is equivalent, in the signal processing language, to finding the linear filter performing *blind source separation* (**BSS**). After the pioneering works of Bar-Ness [11] and of Herault and Jutten [26], interest for source separation has considerably increased during the recent years [15, 14, 31]. It is interesting that the link between signal processing and sensory coding is present in this field from the beginning: Jutten and Herault proposed a now well-known heuristic for performing BSS, using a neuromimetic architecture for analysing a signal coming from muscles. Conversely, experience gained with the study of this algorithm allowed for new approaches in the study of the olfactory system [22, 35]. On the neuro-psychological side, it is involved in the *coktail party effect* (one is able to focus on one speaker, separating his speech from all the ambiant noise). In signal processing BSS can be used for making *cancellers*, that is systems for substracting the noise from the signal [11]. In fact BSS has a wide variety of applications in data processing (see, e.g., [19]). In speech and sound processing, there is an additional complication: the signal is not simply a linear superposition of sources, but also a convolution [26]. In this paper, we will however not consider the case of blind source deconvolution.

Although working on linear mixtures of sources, researchers soon realized that finding independent components might be a useful strategy for any data processing. The term *Independent Component Analysis* (**ICA**) was then promoted as a generalization of BSS to non Gaussian data, and as an alternative to the standard principal component analysis (**PCA**) [26, 15]. To one of his papers on BSS Comon has given the title: "Independent Component Analysis: a new concept ?" [15]. To this question one may answer *no*, since this concept has been already proposed by Barlow at the begining of the sixties, under the name of factorial coding - and even before by Attneave [4], although in a less precise way, under the name of "economy of coding". Still, it is quite interesting that, many years after Barlow suggested that factorial coding could be a major strategy used by Nature for sensory coding, engineers arrived at the same conclusion in the context of signal analysis. In the same vain, it has been shown [15] that there is a particularly well chosen cost function for BSS, and this special cost function is nothing but the *redundancy reduction* cost function mentioned above in the context of sensory coding. While this convergence in the definition of relevant concepts may appear today not surprising, one should remember that this has not always been the case. In particular there are stimulating discussions by Barlow [6] where he is *opposing* the Nature's and engineer's approaches to coding.

In the case of BSS framework, as we said previously one is working under the hypothesis that linear processing is able to produce a factorial code. The main goal in BSS is then to find efficient algorithms in order to compute the filter (or synaptic efficacies) that leads to source separation. Besides heuristic algorithms such as the Herault-Jutten (HJ) algorithm [26], there are adaptive algorithms based on appropriate cost functions [15, 13, 12, 18], and algebraic solutions [14, 31, 37] in which the filter is algebraically derived from a particular set of measurements on the data.

In this paper we present new results on BSS. We deal with different mathematical aspects of BSS, keeping in mind the possible application to sensory systems modeling, and the general framework of factorial coding and redundancy reduction. Because there is a widespread literature on this subject that belongs to both sensory coding and signal processing, we found necessary to put into perspective our contribution with previous works on source separation and redundancy reduction. In fact we will use a general framework, namely the reduction to the search for an orthogonal matrix (as explained in section 1), which will appear convenient for both deriving the new results and presenting related works - some times in a simpler way as compared to their original formulation. First, in section 2, we give necessary and sufficient conditions for the network to be a solution of the ICA. These conditions are that a given small set of cumulants have to be set to zero. The number of cumulants which enter in any of these conditions is smaller than what is usually found in the literature. Next, in section 3, we present a short review of known algebraic solutions, making use of the same framework. In the next two sections we deal with adaptive approaches. In section 4, after reminding of the relationship between maximizing mutual information and minimizing redundancy (hence building a factorial code), we discuss several possible approaches based on these information theoretic concepts. We consider with more details one of them, and show how the resulting updating rules are related to the BCM theory of synaptic plasticity [8]. In section 5, we discuss shortly adaptive algorithms from cost functions based on cumulants, these costs being built from the conditions derived in section 2. Eventually in section 6 we also shortly discuss possible extensions to non linear processing. Perspectives are given in the Conclusion.

1 Linear mixtures of independent sources

1.1 Statement of the problem

Here and in the rest of the paper as well, we will consider the case where the factorisation problem has *a priori* an exact solution using a linear filter or a feedforward neural network with no hidden units (we will however comment, whenever appropriate, on the possible extensions to cases where a multilayer network would be needed). This means that we are assuming the input data to be a *linear mixture of independent sources*. More precisely, we assume that at each time t the observed data $\mathbf{S}(t)$ is a N dimensional vector given by

$$S_j(t) = \sum_{a=1}^{N} M_{j,a}\sigma_a(t), \ \ j = 1, ..., N \tag{1}$$

(in vector form $\mathbf{S} = \mathbf{M}\boldsymbol{\sigma}$) where the σ_a are N independent random variables, of unknown probability distributions, and $\mathbf{M}$ is an unknown, constant, $N \times N$ matrix, called the *mixture matrix*. By hypothesis, all the source cumulants are diagonal, in particular the two point correlation at equal time $\mathbf{K}^0$:

$$K^0_{a,b} \equiv < \sigma_a(t)\sigma_b(t) >_c = \delta_{a,b}K^0_a \tag{2}$$

where $\delta_{a,b}$ is the Kronecker symbol. Without loss of generality, one can always assume that the sources have zero average:

$$< \sigma_a >= 0, \ \ a = 1, ..., N \tag{3}$$

(otherwise one has to estimate the average of each input, and substract it from that input).

Performing source-separation (or equivalently factorizing the output distribution) means finding the network "couplings" or "synaptic efficacies" $\mathbf{J}$ (the linear filter $\mathbf{J}$

in the language of signal processing), such that the N-dimensional (filter, or network) ouput $\mathbf{h}$

$$h_i(t) = \sum_{j=1}^{N} J_{i,j} S_j(t), \;\; i = 1, ..., N \tag{4}$$

gives a reconstruction of the sources: ideally, one would like to have $\mathbf{J} = \mathbf{M}^{-1}$. However, as it is well known and clear from the above equations, one can recover the sources only up to an arbitrary permutation, and up to a sign-scaling factor for each source: nothing tells us which output should be equal to which source; one cannot distinguish $\mathbf{M}\boldsymbol{\sigma}$ from $\mathbf{M'}\boldsymbol{\sigma'}$ with $\mathbf{M'} = \mathbf{M}\boldsymbol{\Lambda}$ and $\boldsymbol{\sigma'} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\sigma}$, where $\boldsymbol{\Lambda}$ is any diagonal matrix having non zero diagonal elements. In particular, one can fix the absolute value of the arbitrary diagonal matrix by asking for $\mathbf{J}$ to be the inverse of $\mathbf{M}\mathbf{K}^{0\,\frac{1}{2}}$ - which is equivalent to state that all we can expect is to estimate normalized cumulants of the sources, such as the k-order normalized cumulants ζ_k^a:

$$\zeta_k^a \equiv \frac{< \sigma_a^k >_c}{< \sigma_a^2 >_c^{k/2}}, \;\; a = 1, ..., N \tag{5}$$

To conclude, any solution $\mathbf{J}$ that one may find will thus be equal to the inverse of $\mathbf{M}\mathbf{K}^{0\,\frac{1}{2}}$ up to what we will call a "sign-permutation", that is the product of a permutation by a diagonal matrix with only ± 1 diagonal elements.

As it is usually done in the study of source separation, we have assumed that the number of sources is known (we have N observations, e.g. N captors, for N independent sources), and we assume $\mathbf{M}$ to be invertible. The difficulty comes from the fact that the statistics of the sources are not known, the mixture matrix is not known and is not necessarily (and in general it is not) an orthogonal matrix.

1.2 Reducing the problem to finding an orthogonal matrix

An elementary, but extremely useful remark, first made by Comon [15] and Cardoso [14], is that the search for a solution $\mathbf{J}$ can be reduced to the search for an orthogonal matrix. Indeed, the $\mathbf{J}$ we are looking for has in particular to diagonalize the two point connected correlation of the inputs; as it is well known, the diagonalization of a symmetric positive matrix defines a matrix only up to an arbitrary orthogonal matrix. Hence, performing first *whitening* will leave us with an orthogonal matrix, which has to be determined from, e.g., higher cumulants. Because we will make an extensive use of this fact, and in order to introduce our notation, we give below a detailed derivation.

So let us write that $\mathbf{J}$ sets the variance output to the $N \times N$ identity matrix $\mathbf{1}_N$:

$$< \mathbf{h}\mathbf{h}^T >_c = \mathbf{1}_N \tag{6}$$

This reads, in term of the correlation $\boldsymbol{C_0}$ between the input data,

$$\boldsymbol{C_0} \equiv < \mathbf{S}\mathbf{S}^T >_c = \mathbf{M}\mathbf{K}^0\mathbf{M}^T \tag{7}$$

as

$$\mathbf{J}\boldsymbol{C_0}\mathbf{J}^T = \mathbf{1}_N \tag{8}$$

One should keep in mind that $\boldsymbol{C_0}$ is what one can measure, and the r.h.s. of equation (7) is its expression in function of the unknowns. Every solution of this equation (8) can be writen as

$$\mathbf{J} = \Omega\, \boldsymbol{C_0}^{-1/2} \tag{9}$$

where Ω is an orthogonal matrix:

$$\Omega\, \Omega^T = \mathbf{1}_N \tag{10}$$

The inverse of the square root of $\boldsymbol{C_0}$ is defined from the principal component analysis of the real positive matrix $\boldsymbol{C_0}$. If we denote by $\boldsymbol{J^0}$ the matrix whose rows are the orthonormal eigenvectors of $\boldsymbol{C_0}$, and by $\boldsymbol{\Lambda}_0$ the diagonal matrix of the associated eigenvalues, one has

$$\begin{aligned} \boldsymbol{J^0}\boldsymbol{C}_0\boldsymbol{J^{0T}} &= \boldsymbol{\Lambda}_0 \\ \boldsymbol{J^0}\boldsymbol{J^{0T}} &= \mathbf{1}_N \end{aligned} \tag{11}$$

and

$$\boldsymbol{C_0}^{-1/2} = \boldsymbol{J^{0T}}\boldsymbol{\Lambda}_0^{-\frac{1}{2}}\boldsymbol{J^0} \tag{12}$$

Now suppose that one first projects the input data onto the normalised principal components, that is one computes $\mathbf{h}^0$ defined by:

$$\mathbf{h}^0 = \boldsymbol{\Lambda}_0^{-\frac{1}{2}}\boldsymbol{J^0}\mathbf{S}, \tag{13}$$

Then, taking $\mathbf{J}$ as in (9) and replacing $\boldsymbol{C_0}^{-1/2}$ by its expression (12), $\mathbf{h}$ can be written as

$$\mathbf{h} = \Omega\boldsymbol{J^{0T}}\boldsymbol{\Lambda}_0^{-\frac{1}{2}}\boldsymbol{J^0}\mathbf{S} \;=\; \Omega\boldsymbol{J^{0T}}\mathbf{h}^0 \tag{14}$$

Since $\mathcal{O} \equiv \Omega\boldsymbol{J^{0T}}$ is again an orthogonal matrix, finding $\mathbf{J}$ means finding an orthogonal matrix $\mathcal{O}$ such that the output

$$\mathbf{h} = \mathcal{O}\mathbf{h}^0 \tag{15}$$

is a vector of N mutually independent components.

This means equivalently that, going from $\mathbf{S}$ to $\mathbf{h}^0$, leads to the problem of separating an *orthogonal* mixture. Let us check that $\mathbf{h}^0$ is indeed an orthogonal mixture of the sources. First we write, in term of the unknowns, that the rows of $\boldsymbol{J^0}$ are the eigenvectors of $\boldsymbol{C_0}$:

$$\boldsymbol{J^0}\mathbf{M}\mathbf{K}^0\mathbf{M}^T\boldsymbol{J^{0T}} = \boldsymbol{\Lambda}_0 \tag{16}$$

Then this implies that $\boldsymbol{\Lambda}_0^{-\frac{1}{2}}\boldsymbol{J^0}\mathbf{M}\mathbf{K}^{0\,\frac{1}{2}}$ is an (unknown) orthogonal matrix which, for later convenience, we define as the transposed of some orthogonal matrix $\mathcal{O}^0$:

$$\begin{aligned} \boldsymbol{\Lambda}_0^{-\frac{1}{2}}\boldsymbol{J^0}\mathbf{M}\mathbf{K}^{0\,\frac{1}{2}} &= \mathcal{O}^{0T} \\ \mathcal{O}^0\mathcal{O}^{0T} &= \mathbf{1}_N \end{aligned} \tag{17}$$

Hence, the projected data $\mathbf{h}^0$ as defined in (13) can be expressed, in function of the unknown sources, as an orthogonal mixture:

$$\mathbf{h}^0(t) = \mathcal{O}^{0T} \boldsymbol{K}^{\mathbf{0}\ -1/2} \boldsymbol{\sigma}(t) \tag{18}$$

Eventually, from (18) and (15), one can also write $\mathbf{h}$ in terms of the sources as

$$\mathbf{h} = \mathbf{X}\ \boldsymbol{K}^{\mathbf{0}\ -1/2} \boldsymbol{\sigma} \tag{19}$$

where $\mathbf{X}$ is the orthogonal matrix $\mathcal{O}\mathcal{O}^{0T}$. Clearly, the desired solution is, up to a sign-permutation, $\mathcal{O}^0$ for $\mathcal{O}$ or equivalently $\mathbf{1}_N$ for $\mathbf{X}$.

We conclude this section with some additional remarks.

1. In reducing the source separation problem to the search for an orthogonal matrix, we have seen that one can choose different, although equivalent, formulations. Each of them may be specifically useful depending on the particular choosen approach. In particular, in the next section 2, where we will determine the family of couplings $\mathbf{J}$ solution of a given set of equations, it will be convenient to use the formulation (19) taking $\mathbf{X}$ as unknown; next, for the algebraic approach in section 3, we will use only the quantities related to the principal components, that is $\mathbf{h}^0$ and $\mathcal{O}^0$ with the basic equation (18); and in section 5 we will consider adaptive algorithms for computing either $\mathbf{J}$ itself or $\mathcal{O}$ as defined in (15).

2. An important practical fact is that, here and in the following, averages need not be considered as ensemble averages: they can be taken as empirical time averages as well (e.g. a time average performed over some large enough time window T). In fact, no stationarity or ergodicity asumptions will be required. All what is required implicitely is to have, for the sources, null cross-cumulants when these cumulants are defined from the chosen averaging procedure. Furthermore, this has to be true only for the cumulants which one has to compute in a given approach (see section 3.1 for details).

3. In many sensory coding and data analysis problems, PCA is a natural thing to do (clearly for close to Gaussian data, but also for much more complex cases, e.g. in the analysis of time series). In all cases, one can go beyond the strict PCA, using the freedom in the choice of the orthogonal matrix Ω. An important example is in the application of redundancy reduction to the modeling of the first stages in the visual system [30, 3, 27]. There, after whitening (in fact an extension of whitening which takes noise processing into account), this freedom is used to satisfy as much as possible some additional constraints, such as locality of receptive fields and invariance by dilatation. As a result, one finds [27] a block-diagonal orthogonal matrix which leads to a multiscale representation of the visual scenes.

2 Criteria based on cross-cumulants of output activities

The main purpose of this section is to present a new necessary and sufficient condition for $\mathbf{J}$ to be a solution of the ICA. This condition is that a given set of cumulants, associated to correlations at equal time, are set to zero. For comparison we will also give other conditions based on correlations at equal time and on time correlations which are related to known algebraic solutions. Eventually we will comment on the relationship and differences with other criteria proposed in the literature.

2.1 Condition on a set of non symmetric higher cumulants

The claim is that for $\mathbf{J}$ to be a solution, it is sufficient (and of course necessary) that $\mathbf{J}$ performs whitening and sets altogether to zero a given set of cross-cumulants of some given order k, the number of which being only of order N^2. More precisely, we have the following theorem:

Theorem 1 *Let k be an integer at least equal to 3 for which the k-cumulants of the sources, $\zeta_k^a, a = 1, ..., N$, are not identically null; then*

- *(i) if at most one of these k-cumulants is null, $\mathbf{J}$ is equal to the inverse of $\mathbf{M}$ (up to a sign-permutation and a rescaling as explained above), if and only if one has:*

  for every i, i',

$$\begin{cases} < h_i h_{i'} >_c = \delta_{i,i'} \\ < h_i^{(k-1)} h_{i'} >_c = 0 \; for \; i \neq i'. \end{cases} \tag{20}$$

  where $\mathbf{h}$ is the output vector as defined in (4).

- *(ii) If only $1 \leq L \leq N-2$ k-order cumulants are nonzero, then any solution $\mathbf{J}$ of (20) is the product of a sign-permutation by a matrix which separates the L sources having non zero k- cumulants, and such that the restriction of $\mathbf{J}\ \mathbf{M}\mathbf{K}^{0\ \frac{1}{2}}$ to the space of the $N-L$ other sources is still an arbitrary $(N-L) \times (N-L)$ orthogonal matrix.*

One should note that the second line in (20) means that the matrix $< h_i^{(k-1)} h_{i'} >_c$ is diagonal, but that the values of the diagonal terms, to be named below Δ_i, need not be known in advance. The detailed proof is given in the Appendix. Here we give only the sketch of it, in the case where all the k-cumulants are non zero. First, solving for the diagonalization of the 2nd order cumulant, one uses the representation in term of orthogonal matrices as in section 1. As a result, we have seen that one can write $\mathbf{h}$ as $\mathbf{h} = \mathbf{X}\ \boldsymbol{K}^{\mathbf{0}\ -1/2} \boldsymbol{\sigma}$, where $\mathbf{X}$ is the unknown orthogonal matrix. Replacing $\mathbf{h}$ by this expression in the higher order cumulants, one then uses several times the fact that $\mathbf{X}$ is orthogonal. In particular, because we are considering for each pair i, i' a cumulant which is linear in $h_{i'}$, one can obtain an equation for each $X_{i,a}$ separately. As a result, for each pair i, a, either $X_{i,a}$ is zero or factorizes into the product of two terms, one depending only on i, and one depending only on a (namely the inverse of the cumulant ζ_k^a). From the condition that $\mathbf{X}\mathbf{X}^T$ has zero off-diagonal elements, one then gets that $\mathbf{X}$ has on each row and each column a unique non zero element, and this implies that $\mathbf{X}$ is a sign-permutation.

Some remarks are in order:

1. In the case of $N - L \leq N$ null cumulants, for a typical solution L outputs will then appear as still correlated with one another, yet uncorrelated with the other $N - L$ outputs. One has then to try another value of k, but only for the subspace of yet unseparated sources.

2. As one would expect, the conditions of application of the theorem are not satisfied, for any k, in the case of Gaussian sources - for which all the cumulants of order higher than 2 are zero, and nothing more than whitening can be done.

2.2 Conditions related to algebraic solutions

In the approach above, one has $N^2 + N$ unknowns (the mixture matrix elements and the k-cumulants of the sources), and we use more equations, namely $\frac{N(N+1)}{2} + N^2$ equations. Coming back to our result, the counting argument suggests that one should be able to separate the sources with less conditions, that is exactly $N^2 + N$. One possibility would have been that diagonalizing the second order cumulants and the *symmetric* part of the k-order cumulants in (20) would do. This is not the case: taking $N = 2$, one can easily check that in doing so unwanted solutions appear. However, looking at the derivation of the above theorem one sees that it results from the fact that we are using cumulants linear in at least one of the h_i. One has thus to try cumulants which are symmetric in i, i' and linear in both h_i and $h_{i'}$. This appears to be easy, and one gets another family of sufficient conditions with exactly $N^2 + N$ conditions for $N^2 + N$ unknowns. An additional advantage is that, because of the linearity in $h_i, h_{i'}$, one can solve the equations algebraically: in fact these conditions appear to be related to (or to the extansion of) known algebraic solutions - which we will then review in section 3.

There is however a price to pay: the restriction that we had in Theorem 1, namely that of having non zero k-order statistics for the sources, is replaced here by the much more restricting condition of having *different* cumulants, as stated below.

2.2.1 Conditions on symmetric higher cumulants

We first consider conditions based on correlations at equal time.

Theorem 2

- *(i) Let all the cumulants ζ_4^a, as defined in (5) for $k = 4$, be two by two different, then $\mathbf{J}$ is equal to the inverse of $\mathbf{M}$ (up to a sign-permutation and a rescaling as explained above), if and only if one has:*

for every i, i',

$$\begin{cases} < h_i h_{i'} >_c = \delta_{i,i'} \\ < h_i \; \sum_{i''=1}^{N} (h_{i''})^2 \; h_{i'} >_c = 0 \; for \; i \neq i'. \end{cases} \tag{21}$$

where $\mathbf{h} = \{h_i, i = 1, ..., N\}$ is the output vector as defined in (4).

- *(ii) If one or several sets of sources have a comon value for ζ_4^a (but different from one set to the other), then any $\mathbf{J}$ solution of (21) is the product of a sign-permutation by a matrix which separates the sources having a distinct 4-cumulant, and separate the sets globally - the restriction of $\mathbf{J} \; \mathbf{MK}^{0\,\frac{1}{2}}$ to the subspace of a given set is still an arbitrary orthogonal matrix.*

The elementary proof is short enough to be given here. As before, we first solve the diagonalization of the two point correlation, and express $\mathbf{h}$ in term of the unknown orthogonal matrix $\mathbf{X}$. The higher order cumulants in (21) can then be expressed as

$$< h_i \sum_{i''=1}^{N} (h_{i''})^2 \; h_{i'} >_c = \sum_a X_{i,a} \; \{\sum_{i"} (X_{i",a})^2\} \; \zeta_4^a \; X_{i',a} \tag{22}$$

But since $\mathbf{X}$ is orthogonal, $\sum_{i"} (X_{i",a})^2 = 1$, and the above equation (22) reduces to

$$< h_i \sum_{i''=1}^{N} (h_{i''})^2 \; h_{i'} >_c = \sum_a X_{i,a} \; \zeta_4^a \; X_{i',a} \tag{23}$$

What we want is to impose

$$< h_i \sum_{i''=1}^{N} (h_{i''})^2 \; h_{i'} >_c = \Delta_i \delta_{i,i'} \tag{24}$$

where the Δ_i are yet undetermined constants. Comparing the two equations (23) and (24), one sees that $\mathbf{X}$ gives the eigen-decomposition of a matrix, which is in fact already diagonal. This implies that if all the eigenvalues, that is the ζ_4^a, are distinct, $\mathbf{X}$ is a sign-permutation. Otherwise $\mathbf{X}$ is, up to a sign-permutation, the identity matrix on the subspace corresponding to the distinct eigenvalues, and an arbitrary $\ell \times \ell$ orthogonal matrix on each subspace associated to an eigenvalue of degeneracy ℓ. The algebraic solution associated to the above theorem is the one proposed by Cardoso [14] (see section 3.3).

Finally, we note that in the above statement one may replace $\sum_i h_i^2$ by any $Q(\mathbf{h})$ being a scalar depending polynomially on $\mathbf{h}$ (and by extension any analytical function of $\mathbf{h}$), such that, for *any* orthogonal matrix $\mathbf{X}$, denoting its rows by $\mathbf{X}_a, \; a = 1, ..., N$, the N numbers $< Q(\sigma_a \mathbf{X}_a) \sigma_a^2 >, \; a = 1, ...N$ are distinct (here every σ_a should be understood as the normalised source $\frac{\sigma_a}{\sqrt{<\sigma_a^2>_c}}$). In practice, it may not be easy to prove that such a condition holds, even for the simplest cases, say $Q(\mathbf{h}) = \sum_i h_i^k$ with k even at least equal to 4.

2.2.2 Conditions on time correlations

We consider now conditions related to algebraic solutions based on time correlations [20, 38, 5, 31]. These are extremely simple. If the sources present time correlations, namely the 2-point correlation matrix $\mathbf{K}(\tau)$ for some time delay $\tau > 0$,

$$K(\tau)_{a,b} \equiv < \sigma_a(t)\ \sigma_b(t-\tau) >_c \tag{25}$$

has non zero diagonal elements:

$$K(\tau)_{a,b} = \delta_{a,b}\ K_a(\tau) \tag{26}$$

then one can state:

Theorem 3

- *(i) Let all the cumulants $K_a(\tau)$, as defined in (26) be two by two different, then* $\mathbf{J}$ *is equal to the inverse of* $\mathbf{M}$ *(up to a sign-permutation and a rescaling as explained above), if and only if one has:*

  for every i, i',

$$\begin{cases} < h_i(t) h_{i'}(t) >_c = \delta_{i,i'} \\ < h_i(t)\ h_{i'}(t-\tau) >_c = 0 \ \textit{for}\ i \neq i' \end{cases} \tag{27}$$

  where $\mathbf{h} = \{h_i, i = 1, ..., N\}$ is the output vector as defined in (4).

- *(ii) If one or several sets of sources have a comon value for $K_a(\tau)$ (but different from one set to the other), then any* $\mathbf{J}$ *solution of (27) is the product of a sign-permutation by a matrix which separates the sources having a distinct $K_a(\tau)$-cumulant, and separate the sets globally - the restriction of* $\mathbf{J}\ \mathbf{M}\mathbf{K}^{0\,\frac{1}{2}}$ *to the subspace of a given set is still an arbitrary orthogonal matrix.*

The proof is essentially the same as the one of theorem 2: one writes

$$< h_i(t)\ h_{i'}(t-\tau) >_c = \sum_a X_{i,a}\ K_a(\tau)\ X_{i',a} \tag{28}$$

What we want is to impose

$$< h_i(t)\ h_{i'}(t-\tau) >_c = \Delta_i \delta_{i,i'} \tag{29}$$

where the Δ_i are yet undetermined constants. These two equations (28) and (29) play the same role as (23) and (24) in the case of theorem 2, and the rest of the proof follows in the same way.

In practical situation, averages are conveniently computed as time averages over a time window of some size T. In general, it could happen that the two point correlation $K_a(\tau)$ is a function of the particular time window $[t-T, t]$ considered. However if, on that time window, the sources are sufficiently independent ($K_a(\tau)$ is diagonal), since the mixture matrix does not depend on time, the solution $\mathbf{J}$ obtained from the data collected on that single time window is an exact, time independent, solution.

2.3 Comparison with other criteria

To conclude this section, we comment shortly on other criteria, namely those proposed by Comon [15]. This author has shown that independent component analysis is obtained from maximizing the sum of the square of k-order cumulants, for any chosen $k > 2$. That is, it is sufficient to find an (absolute) minimum of

$$\mathcal{C} \equiv \; - \sum_{i=1}^{N} < (h_i)^k >_c^2 \tag{30}$$

for any given $k > 2$, after whitening has been performed. This is a very nice result since it involves only N cumulants. However, one should note that defining a cost function is not exactly the same as giving explicit conditions onto the cumulants (even though one can easily derive a cost function from these conditions, as we will do in section 5). In fact, the minimization of a cost function such as (30) does involve implicitly the computation of order N^2 cumulants, as it should (see the counting argument at the begining of this subsection). This can be seen in two ways. First, since the value of the k-cumulants at the minimum is not known, the minimization of $\mathcal{C}$ given in (30) is not equivalent to a set of equations for these cumulants; then, if one wants to perform a gradient descent, one has to take the derivative of the cost function with respect to the couplings $\mathbf{J}$, and this will generate cross cumulants. It is the resulting fixed point equations which then matters in the counting argument (we will come back to the algorithmic aspects in section 5). Second, Comon showed also that the minimization of (30) is equivalent to setting to zero *all* the non diagonal cumulants of the same order k (still in addition to whitening):

$$< h_{i1} h_{i2} ... h_{ik} >_c = \; 0, \text{for every set of } k \text{ non identical indices.} \tag{31}$$

Hence, what we have obtained is that only a small subset of these cumulants have to be considered.

3 Algebraic solutions

In this section we will present three family of algebraic solutions. This short revue on algebraic solutions is intended to point out to their common advantages and drawback, to insist onto their simplicity when formulated within the framework described in section 1, and to relate them to the results of section 2.

3.1 Using time delayed correlations

In the case where each source σ_a shows time correlations, it has been shown [20, 38, 5, 31] that there is a simple algebraic solution using only second-order cumulants. More precisely, let us assume that the 2-point correlation matrix $\mathbf{K}(\tau)$ for some time delay $\tau > 0$,

$$K(\tau)_{a,b} \equiv < \sigma_a(t) \; \sigma_b(t - \tau) >_c \tag{32}$$

has non zero diagonal elements:

$$K(\tau)_{a,b} = \delta_{a,b}\ K_a(\tau). \tag{33}$$

It follows [31] that source separation is obtained by asking for the rows of $\mathbf{J}$ to be the left-eigenvectors of the following *non symmetric* matrix

$$\mathbf{C}(\tau)\ \boldsymbol{C_0}^{-1} \tag{34}$$

where $\mathbf{C}(\tau)$ is the 2nd order cumulant of the inputs at time delay τ:

$$\mathbf{C}(\tau) =< \mathbf{S}(t)\ \mathbf{S}^T(t-\tau) >_c . \tag{35}$$

What we propose now is to simplify further by working with a symmetric matrix. This could be done by muliplying the above matrix (34) by $\boldsymbol{C_0}^{-1/2}$ on the left and $\boldsymbol{C_0}^{1/2}$ on the right, and then to relate the desired solution $\mathbf{J}$ to the eigenvectors of the symmetric matrix. An equivalent, but more natural, way to do so is to use the reduction to the search for an orthogonal matrix presented in section 1. Indeed, let us first perform whitening (note that this implies the resolution of the eigen-problem for $\boldsymbol{C_0}$, a task of the same complexity as the computation of its inverse which is needed in (34)). We then compute the correlations at time delay τ of the projected inputs $\mathbf{h}^0$ (as given by (13)), that is the matrix $< \mathbf{h}^0(t)\ \mathbf{h}^{0T}(t-\tau) >_c$. Using the expression (18) of $\mathbf{h}^0$ as a function of the sources, one sees that this matrix is in fact symmetric, and given by:

$$< \mathbf{h}^0(t)\ \mathbf{h}^{0T}(t-\tau) >_c = \mathcal{O}^{0T}\ \boldsymbol{K}^{\mathbf{0}\ -1/2}\ \boldsymbol{K(\tau)}\ \boldsymbol{K}^{\mathbf{0}\ -1/2}\ \mathcal{O}^0 \tag{36}$$

This shows that the desired orthogonal matrix is obtained from solving the eigen decomposition of this correlation matrix (36).

Remark: in the present section averages might be conveniently time averages, such as $< A(t) >= \int dt'\ A(t-t')\ \exp(-t'/T)$. In that case, τ has to be small compared to T in such a way that e.g. $\mathbf{K}^0$ is the same when averaging on $t' < t$ and on $t' < t - \tau$.

3.2 Using correlations at equal time

Quite recently Shraiman [37] showed how to reduce the problem of source separation to the diagonalisation of a certain symmetric matrix $\mathcal{D}$ built upon third order cumulants. We refer the reader to [37] for the elegant derivation of this result. Here we will show how to derive this matrix from the approach introduced in the previous sections 1.2 and 2. We will do so by working with the k-order cumulants, giving thus a generalization of Shraiman's work to any k at least equal to 3.

We start with the expression of the data projected onto the principal components as in (18):

$$\mathbf{h}^0(t) = \mathcal{O}^{0T} \boldsymbol{K}^{\mathbf{0}\ -1/2} \boldsymbol{\sigma}(t) \tag{37}$$

One computes the k-order statistics $< h^0_{i_1} h^0_{i_2} ... h^0_{i_k} >_c$. From (37), these cumulants have the following expression in term of sources cumulants:

$$< h^0_{i_1} h^0_{i_2} ... h^0_{i_k} >_c = \sum_{a=1}^{N} \mathcal{O}^0_{a,i_1} ... \mathcal{O}^0_{a,i_k} \zeta^a_k \tag{38}$$

where the ζ_k^a are the normalised cumulants as defined in (5). Now, one multiplies two such cumulants having $k-1$ identical indixes, and one sums over all the possible values of these indixes. One sees that this will produce the contraction of $k-1$ matrices $\mathcal{O}^0$ with their transposed, leading to just a $\mathbf{1}_N$, and only two terms $\mathcal{O}^0$ will remain. Explicitly, we consider the symmetric matrix, to be referred to as the k-Shraiman matrix,

$$\mathcal{D}_{i,i'} = \sum_{i_1,i_2,\ldots,i_{k-1}} < h_i^0 h_{i_1}^0 \ldots h_{i_{k-1}}^0 >_c < h_{i'}^0 h_{i_1}^0 \ldots h_{i_{k-1}}^0 >_c \tag{39}$$

which, from (38), is equal to

$$\mathcal{D}_{i,i'} = \sum_{a=1}^{N} \mathcal{O}_{a,i}^0 \mathcal{O}_{a,i'}^0 (\zeta_k^a)^2 \tag{40}$$

The above formula is nothing but an eigen-decomposition of the matrix $\mathcal{D}$. This shows that the rows of $\mathcal{O}^0$ are the eigenvectors of the k-Shraiman matrix, the eigenvalues being $(\zeta_k^a)^2$ for $a = 1, ..., N$. A solution of the source separation problem is thus given by the diagonalisation of one k-Shraiman matrix (e.g. taking $k = 3$ or $k = 4$).

3.3 A simple solution using fourth order cumulants

We now consider the solution based on fourth order cumulants [14], which is directly related to the results obtained in section 2.2.1. Let us consider the following cumulants of the input data projected onto the principal components, $\mathbf{h}^0$:

$$\mathbf{C}_{4\,i,i'} \equiv < h_i^0 \sum_{i''=1}^{N} h_{i''}^{02} h_{i'}^0 >_c \tag{41}$$

In term of the orthogonal matrix $\mathcal{O}^0$ and of the cumulants ζ_4^a it reads

$$(\boldsymbol{C_4})_{i,i'} = \sum_{a=1}^{N} \mathcal{O}_{a,i}^0 \zeta_4^a \sum_{i''=1}^{N} (\mathcal{O}_{a,i''}^0)^2 \; \mathcal{O}_{a,i'}^0 \tag{42}$$

Since $\mathcal{O}^0$ is orthogonal, this reduces to the equation

$$\boldsymbol{C_4} = \mathcal{O}^0 \; \boldsymbol{K_4} \; \mathcal{O}^{0T} \tag{43}$$

where $\boldsymbol{K_4}$ is the diagonal matrix of the 4th order cumulants,

$$(\boldsymbol{K_4})_{a,b} = \delta_{a,b} \zeta_4^a \tag{44}$$

This shows that $\mathcal{O}^0$ can be found by solving for the eigen decomposition of the cumulant $\boldsymbol{C_4}$.

We conclude with some remarks which apply to any of the algebraic solutions considered in this section. First, they all suffer from the same drawback, which is that sources having the same statistics at the orders under consideration will not be separated. Moreover numerical instability may occur if these statistics are different but very close to one another. Second, one can see that all these solutions are based on the same two facts: (i) the diagonalization of a real positive symmetric matrix leaves an arbitrary orthogonal matrix, which in turn can be used to diagonalize another symmetric matrix; (ii) but because one is dealing with a linear mixture, applying this to two well chosen correlation matrices is precisely enough to solve the BSS problem (in particular we have seen that working with two symmetric matrices provides at least as many equations as unknowns).

4 Cost functions derived from information theory

We switch now to the study of adaptive algorithms. To do so, one has first to define proper cost functions. Whereas in the next section we will consider cost functions based on cumulants, in the present section we will consider the particular costs derived from information theory. In both cases we will take advantage of the results obtained in section 2. We will see that an important outcome is the derivation of updating rules for the synaptic efficacies closely related to the Bienenstock, Cooper and Munro (**BCM**) theory of cortical plasticity [8].

4.1 From infomax to redundancy reduction

Our starting point is the main result obtained in [32], namely that maximization of the mutual information between the input data and the neural code leads to redundancy reduction, hence to source separation for both linear and non linear mixtures. To be more specific, we first give a short derivation of that fact (for more details see [32]). We consider a network with N inputs and p outputs, and *nonlinear transfer functions* $f_i, i = 1, ..., p$. Hence the output $\vec{V}$ is given by a gain control after some (linear or non linear) processing:

$$V_i(t) = f_i(h_i(t)),\ \ i = 1, ..., p \tag{45}$$

In the simplest case (in particular in the context of **BSS**), $\vec{h}$ is given by the linear combination of the inputs:

$$h_i(t) = \sum_{j=1}^{N} J_{i,j} S_j(t),\ \ i = 1, ..., p \tag{46}$$

However here the $h_i(t)$ can be as well *any* deterministic (hence not necessarily linear) functions of the inputs $\mathbf{S}(t)$ (we will in particular make use of this fact in section 6). The *mutual information* $\mathcal{I}$ between the input and the output is given by [10]:

$$\mathcal{I} \ \equiv \int d^pV \ d^NS P(\vec{V}, \vec{S}) \ \log \frac{P(\vec{V}, \vec{S})}{Q(\vec{V}) \ P(\vec{S})}. \tag{47}$$

This quantity is well defined only if noise processing is taken into account (e.g. resolution noise). Under the hypothesis of additive output noise (so that $V_i = f_i(h_i)+$ noise), the mutual information can be written as

$$\mathcal{I} = H(Q) - H_\nu \tag{48}$$

where $H(Q)$ is the differential entropy of the output distribution $Q = Q(\vec{V})$,

$$H(Q) = -\int d^pV\ Q(\vec{V})\ \log Q(\vec{V}), \tag{49}$$

and H_ν the differential entropy of the noise probability distribution. In the limit of vanishing noise, the constant H_ν diverges, but $H(Q)$ has a finite limit which is the entropy of the deterministic output as given by (45). Maximizing the mutual information is then equivalent to maximizing the output entropy (49). In the r.h.s. of (49), one can make the change of variable $\vec{V} \to \vec{h}$, using

$$\prod_{i=1}^{p} dV_i Q(\vec{V}) = \prod_{i=1}^{p} dh_i \Psi(\vec{h}) \tag{50}$$

and

$$dV_i = f_i'(h_i)dh_i, \quad i = 1, ..., p. \tag{51}$$

This gives

$$H(Q) = -\int d\vec{h}\Psi(\vec{h})\ \ln \frac{\Psi(\vec{h})}{\prod_{i=1}^{p} f_i'(h_i)} \tag{52}$$

This implies that $H(Q)$, hence $\mathcal{I}$, is maximal when $\Psi(\vec{h})$ factorizes,

$$\Psi(\vec{h}) = \prod_{i=1}^{p}\ \Psi_i(h_i), \tag{53}$$

and at the same time for each output neuron the transfer function f_i has its derivative equal to the corresponding marginal probability distribution:

$$f_i'(h_i) = \Psi_i(h_i), \quad i = 1, ..., p. \tag{54}$$

As a result, infomax implies redundancy reduction. The optimal neural representation is a factorial code - provided it exists.

4.2 The specific case of BSS

Let us now come back to the **BSS** problem for which the $\vec{h}$ are taken as linear combinations of the inputs. By hypothesis, the N-dimensional input is a linear mixture of N independent sources. Then for $p \leq N$ synaptic efficacies can indeed be found such that $\Psi(\vec{h})$ factorizes. For $p = N$, factorization is equivalent to **BSS**. For $p < N$ there will be several different solutions. More precisely, in a factorial code the sources will be clustered in p disjoint sets, such that each output i codes for the cluster i composed of $p_i \geq 1$ sources ($p \leq \sum_{i=1}^{p} p_i \leq N$): the field h_i, although independent of the $h_{i'}$ for $i' \neq i$, is still a mixture of p_i sources. In the following we consider only $p = N$.

One should note that the factorial code is obtained by the network processing *before* applying the nonlinear function at each output neuron. From the algorithmic aspect, as suggested in [32] this gives us two possible approaches:

- One is to optimize globally, that is to maximize the mutual information over *both* the synaptic efficacies and the transfer functions. In that case, infomax is used in order to perform ICA - the nonlinear transfer functions being there just to enforce factorization.

- Another possibility is to *first* find the synaptic efficacies leading to a factorial code, and *then* compute the optimal transfer functions (which depend on the statistical properties of the stimuli). In that case, one may say that it is ICA which is used in order to build the network which maximizes information transfer. Still, if one considers that the transfer functions are chosen at each instant of time according to (54), the mutual information becomes precisely equal to minus the *redundancy* cost function $\mathcal{R}$:

  $$\mathcal{R} \equiv \int d\vec{h} \Psi(\vec{h}) \; \ln \frac{\Psi(\vec{h})}{\prod_{i=1}^{p} \Psi_i(h_i)} \tag{55}$$

  In the context of blind source separation the relevance of the redundancy cost function has been recognized by Comon [15] (one should mention also a work by Burel [13] where a different but related cost function is considered).

Remark on the terminology: the quantity (55), called here, and in the literature related to sensory coding, the *redundancy*, is called in the signal processing literature (in particular in [15]) the *mutual information*, as a short name for the mutual information between the joint probability distribution and the factorial distribution of the outputs. But this mutual information, that is the redundancy (55), should not be mistaken for the mutual information (47) that we introduced above, which is defined in the usual way, that is between the input and the output of a processing channel [10].

Although it is appealing to work with either the mutual information or the redundancy, doing so may not be easy. It is convenient to rewrite the output entropy, making in (52) the change of variable $\vec{h} \to \vec{S}$ as done in [12]. Since the input entropy $H(P)$ is a constant (it does not depend on the couplings $\mathbf{J}$), the quantitiy which has to be maximized is

$$\mathcal{E} = < \ln \mid \mathbf{J} \mid > \; + \sum_i < \log c_i(h_i) > \tag{56}$$

where $\mid \mathbf{J} \mid$ is the absolute value of the determinant of the coupling matrix $\mathbf{J}$, and c_i is either f_i' if one considers the mutual information, or Ψ_i if one considers the redundancy (the mutual information for the optimal transfer function at a given $\mathbf{J}$). In the first case, one has to find an algorithm for searching for the optimal transfer functions; in the second case, one has to estimate the marginal distributions. In the two following subsections we consider practical approaches.

4.3 Working with a given family of transfer functions

A practical way to search for the best f_i' is to restrict the search to an a priori chosen family of transfer functions. We have seen that, at the end of the optimization process, the transfer functions will be related to the probability distributions of the independent sources. Since precisely these distributions are not known - and cannot be estimated without first performing source separation! -, the choice of a proper parametrized family may be a problem. Still, any prior knowledge on the sources and any reasonable assumption may be used to limit the search to a family of functions controlled by a small number of parameters. A very simple strategy has been tried by Bell and Sejnowski [12] who obtained very promising results on some specific applications. Their numerical simulations suggest that one can take transfer functions with a simple behaviour (that is, e.g., with one peak in the derivative when the data show only one peak in their distribution), and to optimise just the gain and the threshold in each transfer function - which means fitting the location of the peak and its height.

4.4 Cumulant expansion of the marginal distributions

When working with the redundancy, one would like to avoid to have to estimate the marginal distributions from histograms, since this would take a lot of time. One may parametrize the marginal probability distributions, and adapt the parameters at the same time one is adapting the couplings: this is exactly the same as working with the mutual information with a parameterized family of transfer functions. Another possibility, already considered in [15] is to replace each marginal by a simple probability distribution with the same first cumulants as the ones of the actual distribution. Very recently this approach has been used in [1]. We consider here this expansion with a slight different point of view than in [15], in order to relate this approach to the result of section 2.

The general idea is the following. We know that if we had Gaussian distributions, every required computation would be easy. Now, if N is large, each field h_i is a sum of a large number of random variables, so that before adaptation (that is with arbitrary synaptic efficacies), the marginal distribution for h_i is a Gaussian. However, through adaptation, each h_i becomes proportional to one source σ_α - whose distribution is in general not a Gaussian, and not necessarily close to Gaussian. Still, there is another, and stronger, motivation for considering such an approximation. Indeed, the result of section 2, which is that conditions on a limited set of cumulants are sufficient in order to obtain factorization, strongly suggests to replace the unknown distribution with a simple distribution having the same first cumulants up to some given order.

So let us consider the systematic close-to-Gaussian cumulant expansion of $\Psi_i(h_i)$ [2]. At first non trivial order, it is given by

$$\Psi_i(h_i) \approx \Psi_i^1(h_i) \equiv \Psi^0(h_i)\,[1 + \lambda_i^{(3)}\,\frac{h_i(h_i^2-3)}{6}], \tag{57}$$

where $\Psi^0(h_i)$ is the normal distribution

$$\Psi^0(h_i) \equiv \frac{1}{\sqrt{2\pi}}\,\exp(-\frac{h_i^2}{2}), \tag{58}$$

and $\lambda_i^{(3)}$ is the third (true) cumulant of h_i:

$$\lambda_i^{(3)} \equiv < h_i^3 >_c . \tag{59}$$

In the above expression (57) we have taken into account that, as explained in section 1, one can always take

$$< h_i >= 0. \tag{60}$$

and

$$< h_i^2 >_c = 1. \tag{61}$$

Replacing $\Psi_i(h_i)$ by $\Psi_i^1(h_i)$ in the cost function (56), and expanding the logarithm $\ln[1 + \lambda_i^{(3)} \frac{h_i(h_i^2-3)}{6}]$, the quantity to be maximized is then, up to a constant,

$$\mathcal{E} =< \ln \mid \mathbf{J} \mid > + \frac{1}{6}\sum_{i=1}^{N}[\lambda_i^{(3)}]^2 \tag{62}$$

Since optimization has to be done under the constraints (61), we add Lagrange multiplyers:

$$\mathcal{E}(\rho) = \mathcal{E} - \sum_{i=1}^{N} \frac{1}{2}\rho_i(< h_i^2 >_c -1) \tag{63}$$

Taking into account (60), one then obtains the updating equation for a given synaptic efficacy J_{ij}:

$$\begin{aligned} \Delta J_{ij} &\propto - \frac{d\mathcal{E}(\rho)}{dJ_{ij}} \\ \frac{d\mathcal{E}(\rho)}{dJ_{ij}} &= - J_{ij}^{T\ -1} - < h_i^3 >_c< (h_i^2 - 1)S_j > + \rho_i < h_i S_j > \end{aligned} \tag{64}$$

We now consider the fixed point equation, that is $\Delta J_{ij} = 0$. Multiplying by $J_{i'j}$ and summing over j, it reads:

$$\delta_{ii'} = \rho_i < h_i h_{i'} >_c - < h_i^3 >_c< h_i^2 h_{i'} >_c \tag{65}$$

together with $< h_i^2 >_c= 1$ for every i. The parameters ρ_i are obtained by writing the fix point equation (65) at $i = i'$, that is

$$\rho_i = 1 + < h_i^3 >_c^2 . \tag{66}$$

Note that, in particular, $\rho_i > 0$ for all i.

It follows from the result (20) of section 2, that the exact, desired, solutions are particular solutions of the fixed point equation (65), giving a particular absolute minimum of the cost function with the close-to- Gaussian approximation. However, there is no guarantee that no other local minimum exists: there could be solutions for which (65) is satisfied with non-diagonal matrices $< h_i h_{i'} >_c$ and $< h_i^2 h_{i'} >_c$.

4.5 Link with the BCM theory of synaptic plasticity

Let us now consider a possible stochastic implementation of the gradient descent (64). Since there are products of averages, it is not possible to have a simple stochastic version, where the variation of J_{ij} would depend on the instanteneous activities only. Still, by removing one of the averages in (64) one gets the following updating rule:

$$\Delta J_{ij} \;\; = \;\; \epsilon \; \{ \; J^{T\; -1}_{ij} \; - \; \rho_i \; h_i S_j + \; < h_i^3 >_c h_i^2 S_j \; \} \tag{67}$$

where ϵ is a parameter controling the rate of variation of the synaptic efficacies. The parameters ρ_i can be taken at each time according to the fixed point equation (66). It is quite interesting to compare the above updating equation (67) with the BCM theory of synaptic plasticity [8]. In the latter, a qualitative synaptic modification rule was proposed in order to account for experimental data on neural cell development in the early visual system. This BCM rule can be seen as a non linear variant of the Hebbian covariance rule. One of its possible implementations reads, in our notation:

$$\Delta J_{ij} = \epsilon \; \gamma_i \; \{ \, - \, h_i S_j \; + \; \Theta_i h_i^2 S_j \; \} \tag{68}$$

where γ_i and Θ_i are parameters possibly depending on the current statistics of the cell activities. The particular choices $\Theta_i =< h_i^2 >$, $\gamma_i = 1$ or Θ_i^{-1} have been studied with some detail [25, 29]. The two main features of the BCM rule are: (i) there is a synaptic increase or decrease depending on the post synaptic activity relatively to some threshold Θ_i which, itself, varies according to the cell mean activity level; (ii) at low activity there is no synaptic modification. Since ρ_i is positive, the rule we derived above is quite similar to (68), with a threshold Θ_i equal to $\frac{<h_i^3>_c}{\rho_i}$. The main difference is in the constant (that is activity independent) term $J^{T\; -1}_{ij}$. This term plays a crucial role: it couples the N neural cells. One should note that in fact the BCM rule has been mostly studied for a single cell, and only crude studies of its possible extension to several cells have been performed [36]. One should note also that in our formulae we have always assumed zero mean activity ($< S_j >= 0$, hence also $< h_i >= 0$); if this were not the case, the corresponding averages have to be subtracted ($S_j \rightarrow S_j - < S_j >$ for every j, $h_i \rightarrow h_i - < h_i >$ for every i).

Finally we note that, if the third cumulants are zero, one has to make the expansion up to fourth order. The corresponding derivation and the conclusions are similar: apart from numerical factors, essentially the square of the third cumulant in the cost is replaced by the square of the fourth cumulant, and one gets again a plasticity rule similar to (67), that is with the same qualitative behaviour as the BCM rule.

5 Adaptive algorithms from cost functions based on cumulants

5.1 A gradient descent based on Theorem 1

Among the algorithms using correlations at equal time, only algebraic solutions (discussed in section 3) and the recently proposed *deflation algorithm* [18], which extracts the independent components one by one, guarantee to find them in a rather simple and efficient way. All other approaches suffer from the same problem: empirical updating rules based on high moments, as the Herault-Jutten algorithm, and gradient methods based on some cost function (most often a combination of cumulants), may have unwanted fixed points (see [15, 18] and ref. therein).

We point out that, whatever the algorithm one is working with, the conditions derived in section 2 can be used in order to check whether a correct solution has been found. Clearly, one can also define cost functions from these conditions, by just taking the sum of the square of every cumulant which has to be set to zero. We thus have a cost function for which, in addition to having only good solutions as absolute minima, the value of the cost at an absolute minimum is known: it is zero.

For illustrative purpose, we consider with more detail a gradient descent for a particular choice of cost based on our Theorem 1 in section 2. Specifically, we ask for the diagonalization of the two point correlation and of the third order cumulants $< \ h_i \ h_{i'}^2 \ >_c$. Here again we use the reduction to the search for an orthogonal transformation, as explained in section 1. We thus consider the optimisation of the orthogonal matrix. The cost is then defined by

$$\mathcal{E} \ = \ \frac{1}{2} \sum_{i \neq i'} < \ h_i \ h_{i'}^2 \ >_c^2 \ - \frac{1}{2} \, Tr[\rho(\mathcal{O}\mathcal{O}^T - \mathbf{1}_N)] \tag{69}$$

where ρ is a symmetric matrix of Lagrange multipliers, and h_i has to be written in term of $\mathcal{O}$ (see (15) section 1.2):

$$h_i = \sum_{\alpha=1}^{N} \mathcal{O}_{i,\alpha} \ h_\alpha^0 \tag{70}$$

the $\mathbf{h}_i^0$ being the projections of the inputs onto the principal components, as given by (13).

Taking the derivative of $\mathcal{E}$ with respect to $\mathcal{O}_{i,\alpha}$, one gets:

$$\frac{d\mathcal{E}}{d\mathcal{O}_{i,\alpha}} \ = \ \sum_{i'(\neq i)} \Big[< \ h_i \ h_{i'}^2 \ >_c < \ h_\alpha^0 h_{i'}^2 \ >_c \ + 2 \ < \ h_{i'} \ h_i^2 \ >_c < \ h_{i'} h_i h_\alpha^0 \ >_c \Big] \\ - \sum_{i'} \rho_{i,i'} \ \mathcal{O}_{i',\alpha} \tag{71}$$

One can either adapt ρ according to $\frac{d\rho}{dt} \ = \ -\varepsilon \frac{d\mathcal{E}}{d\rho}$, or choose ρ imposing at each time $\mathcal{O}\mathcal{O}^T = \mathbf{1}_N$, which we do here. The equation for ρ is obtained by writing the orthonogonality condition for $\mathcal{O}$, $(\mathcal{O} + d\mathcal{O})(\mathcal{O}^T + d\mathcal{O}^T) = \mathbf{1}_N$, that is:

$$\mathcal{O} \ d\mathcal{O}^T \ + \ d\mathcal{O} \ \mathcal{O}^T \ = \ 0. \tag{72}$$

Paying attention to the fact that ρ is symmetric, one gets

$$\rho_{i,i'} \ = \ \frac{1}{2} \sum_{k(\neq i')} \Big[< \ h_i \ h_k^2 \ >_c < \ h_{i'} h_k^2 \ >_c \ + 2 \ < \ h_k \ h_i^2 \ >_c < \ h_k h_i h_{i'} \ >_c \Big] \\ + \frac{1}{2} \sum_{k(\neq i)} \Big[< \ h_i \ h_k^2 \ >_c < \ h_{i'} h_k^2 \ >_c \ + 2 \ < \ h_k \ h_{i'}^2 \ >_c < \ h_k h_i h_{i'} \ >_c \Big] \tag{73}$$

Replacing in (71) ρ by its expression (73), multiplying both sides of (71) by $\mathcal{O}_{i',\alpha}$ for some i' and summing over α, and using (70), one gets the rather simple equations for the projections of the variations of $\mathcal{O}$ onto the N vectors $\mathcal{O}_i$:

$$\sum_\alpha \mathcal{O}_{i',\alpha} \frac{d\mathcal{O}_{i,\alpha}}{dt} = -\varepsilon \sum_\alpha \mathcal{O}_{i',\alpha} \frac{d\mathcal{E}}{d\mathcal{O}_{i,\alpha}} \equiv -\varepsilon\, \eta_{i,i'} \tag{74}$$

where:

$$\begin{aligned} \eta_{i,i'} = {} & \frac{3}{2}\Big[< h_{i'}^3 >_c < h_i\, h_{i'}^2 >_c - < h_i^3 >_c < h_{i'}\, h_i^2 >_c \Big] \\ & + \sum_k < h_k h_i h_{i'} >_c \Big[< h_k\, h_i^2 >_c - < h_k\, h_{i'}^2 >_c \Big] \end{aligned} \tag{75}$$

From the above expression, one can easily write the (less simple) updating equations for either $\mathcal{O}$ (multiplying by $\mathcal{O}_{i',\alpha}$ and summing over i'), or $\mathbf{J}$ (multiplying by $\left[\mathcal{O}\mathbf{\Lambda}_0^{-\frac{1}{2}}\mathcal{O}^0\right]_{i',j}$ and suming over i').

Although we derived the updating equation from a global cost function, one may also derive an adaptative version. One possibility is to use the approach considered in the preceding section: in each term containing a product of averages, one average $< . >$ is replaced by the instantaneous value. The remaining average is computed as a time average on a moving time-window. An alternative approach is to replace each average by a time average, taking different time constants in order to obtain an estimate of the product of averages -and not the average of the product.

5.2 From feedforward to lateral connections

We conclude with a general remark concerning the choice of the architecture. In all the above derivations, we worked with a feedforward network, with no lateral connections. As it is well known, one may prefer to work with adaptable lateral connections, as it is the case in the Herault-Jutten algorithm [26]. One can in fact perform any given linear processing with either one or the other architecture - it is only the algorithmic implementation which might be simpler with a given architecture. Let us consider here this equivalence. A standard way to use a network with lateral connections is the one in [26]. One has a unique link from each input S_i to the output unit i, and lateral connections $\mathcal{L}$ between output units. The dynamics of the postsynaptic potentials u_i of the output cells is given by

$$\frac{du_i}{dt} = -\sum_{i'} \mathcal{L}_{i,i'}\, u_{i'} + S_i \tag{76}$$

If $\mathcal{L}$ has positive eigenvalues, then the dynamics converge to a fixed point $\mathbf{h}$ given by $\mathcal{L}\,\mathbf{h} = \mathbf{S}$. As a result, after convergence, the network gives the same linear processing as the one with feedforward connections $\mathbf{J}$ given by

$$\mathcal{L} = \mathbf{J}^{-1}. \tag{77}$$

In the particular case considered above, one gets the updating rule for $\mathcal{L}_{j,i'} = \mathbf{J}^{-1}_{j,i'}$ by multiplying (74) by $\left[\mathcal{O}\mathbf{\Lambda}_0^{\frac{1}{2}}\mathcal{O}^0\right]_{i,j}$ and summing over i.

For a given adaptive algorithm derived from the minimization of a cost function, one can thus work with either the feedforward or the lateral connections. It is clear that, in general, updating rules will look different whether they are for the lateral or feedforward connections. However, it is worth mentioning that, if we consider the updating rule after whitening (which is to assume that a first network is performing PCA, providing the h^0_α as input to the next layer), then the updating rule for the feedforward and the lateral connections are essentially the same. Indeed, the feedforward coupling matrix allowing to go from $\mathbf{h}^0$ to $\mathbf{h}$ is an orthogonal transformation $\mathcal{O}$; hence the associated lateral network has as couplings the inverse of that orthogonal transformation, that is its transposed, $\mathcal{O}^T$.

6 Possible extensions to non linear processing

There has been already works proposing redundancy reduction criteria for non linear data processing [32, 23, 33], and for defining unsupervised algorithms in the context of automatic data clustering [21]. Here we just point out that all the criteria and cost functions discussed in the present paper may be applied to the *output* layer of a multilayer network for performing Independent Component Analysis on nonlinear data. Indeed, if a mulitlayer network is able to perform ICA, this implies that, in the layer preceeding the output, the data representation is a linear mixture. The main questions are then how many layers are required in order to find such a linear representation, and is it always possible to do so ?

Assuming that there exists at least one (possibly non linear) transformation of the data leading to a set of independent features, we suggest two lines of research. The first one is based on general results on function approximation. It is known that a network with one hidden layer with sufficiently many units is able to approximate a given function with any desired accuracy (provided sufficiently many examples are available) [16, 24, 7]. Then there exists a network with one hidden layer and N outputs such that the ith output unit gives an approximation of the particular function which extracts the ith independent component from the data. Hence we know that it should be enough to take a network with one hidden layer.

Another approach is suggested by the study of the infomax-redundancy reduction criteria [32]. It is easy to see that the cost function (56) has a straightforward generalization to a multilayer network where every layer has the same number N of units. Indeed, if one calls $\mathbf{J}_k$ the couplings in the kth layer, and $c_{ki}(h_{ki})$ the derivative of the transfer function (or the marginal distribution, see section 4.2) of the ith neuron in the kth layer, the mutual information between the input and the output of the multilayer network can be written as

$$\mathcal{E}_L = \sum_k < \ln \mid \mathbf{J}_k \mid > \; + \sum_k \sum_i < \log c_{ki}(h_{ki}) > \tag{78}$$

Hence the cost $\mathcal{E}_L$ is a sum of terms, each one tending to impose factorisation in a given layer. This allows an easy implementation of a gradient descent algorithm. Moreover, this additive structure of the cost suggests a constructive approach: one may start with one layer; if factorisation is not obtained, one can add a second layer, and so on (one should note however that the couplings of a given layer have to be readapted each time one adds a new layer).

7 Conclusion

In the present paper we have presented several new results on Blind Source Separation. Focusing on the mathematical aspects of the problem, we obtained several necessary and sufficient conditions which, if fulfilled, guarantee that separation has been performed. These conditions are on a limited set of cross-cumulants, and can be used either for defining an appropriate cost function, or just in order to check, when using any BSS algorithm, that a correct solution has been reached. Next we showed how algebraic solutions can be easily understood, and for some of them generalized, within the framework of the reduction to the search for an orthogonal matrix.

Eventually we discussed adaptive approaches, the main focus being on cost functions based on information theoretic criteria. In particular, we have shown that the resulting updating rule appears to be, in a loose sense, Hebbian and more precisely quite similar to the type proposed by Bienenstock, Cooper and Munro in order to account for experimental data on the developement of the early visual system [8]. We also showed how some cost functions could be conveniently used for non linear processing, that is for, say, a multilayer network.

In all cases, we payed attention to relate our work to other similar approaches. We showed how the reduction to the search for an orthogonal transformation is a convenient tool for analysing the BSS problem, and finding new solutions. This, of course, does not mean that one cannot perform BSS without whitening, and indeed there are interesting approaches to BSS in which whitening is not required [28].

Acknowledgements

This work has been partly supported by the French-Spanish program "Picasso", the E.U. grant CHRX-CT92-0063, the Universidad Autónoma de Madrid, the Université Paris VI and Ecole Normale Supérieure. NP and JPN thank, respectively, the Laboratoire de Physique Statistique (ENS) and the Departamento de Física Teórica (UAM) for their hospitality. We thank B. Shraiman for communicating his results prior to publication. We thank P. Del Giudice and A. Campa for useful discussions.

Appendix: proof of theorem 1

Let us consider a matrix $\mathbf{J}$ for which (20) is true. First we use the fact that $\mathbf{J}$ diagonalize the two point correlation. Hence, with the notation and results of section 1, we have to determine the family of orthogonal matrices $\mathbf{X}$ such that, when

$$\mathbf{h} = \mathbf{X}\boldsymbol{K}^{\mathbf{0}\ -1/2}\boldsymbol{\sigma} \tag{1}$$

the k-order cumulants in (20) are zero. Using the above expression of $\mathbf{h}$, we have for any k-order cumulant in (20):

$$< h_i^{(k-1)} h_{i'} >_c = \sum_{a=1}^{N} (X_{i,a})^{(k-1)} X_{i',a} \zeta_k^a \tag{2}$$

where the ζ_k^a are the normalized k-cumulants

$$\zeta_k^a = \frac{< \sigma_a^k >_c}{< \sigma_a^2 >_c^{k/2}}, \quad a = 1, ..., N \tag{3}$$

We first consider the case when for every a, ζ_k^a is not zero.

Since we want the k-order correlations to be zero whenever $i \neq i'$, we can write

$$\Delta_i \delta_{i,i'} = \sum_{a=1}^{N} (X_{i,a})^{(k-1)} X_{i',a} \zeta_k^a \tag{4}$$

for some yet undetermined constants Δ_i. Using the fact that $\mathbf{X}$ is orthogonal, we multiply both sides of this equation by $X_{i',a} = X^T_{a,i'}$ for some a, and sum over i'. This gives

$$\Delta_i X_{i,a} = (X_{i,a})^{(k-1)} \zeta_k^a \tag{5}$$

There are now two possibilities, for each pair (i, a): either $X_{i,a} = 0$, or $X_{i,a}$ is non zero (and then Δ_i as well), and we can write

$$X_{i,a}^{(k-2)} = \varepsilon_{i,a} \frac{\Delta_i}{\zeta_k^a} \tag{6}$$

where $\varepsilon_{i,a}$ is 1 or 0. We now use the fact that $\mathbf{X}$ is orthogonal; first for each i the sum over a of the $X_{i,a}^2$ is one, hence for at least one a $\varepsilon_{i,a}$ is nonzero - and it follows also that for every i Δ_i is non zero. Secondly, for every pair $i \neq i'$, $\sum_a X_{i,a} X_{i',a} = 0$. Then, from (6), we have

$$\sum_a \varepsilon_{i,a} \varepsilon_{i',a} (\zeta_k^a)^{\frac{2}{k-2}} = 0 \tag{7}$$

The l.h.s. is a sum of positive terms, hence each of them has to be zero. It follows that for every a either $\varepsilon_{i,a}$ or $\varepsilon_{i',a}$ is zero (or both). The argument can be repeated exchanging the roles of the indices i and a, so that it is also true that for each pair $a \neq a'$, for each i either $\varepsilon_{i,a}$ or $\varepsilon_{i,a'}$ is zero (or both). Hence $\mathbf{X}$ is a matrix with, on each row and on each column, a single nonzero element, which, necessarily, is then ± 1: $\mathbf{X}$ is what we called a sign-permutation, and this completes the proof of part

(i) of the theorem for the case where the k-order cumulants are non zero for every source.

Now we consider the case where only $L < N$ k-order cumulants are nonzero (and this will include the case of only one source with zero cumulant). Without loss of generality, we will assume that they are the first L sources ($a = 1, ..., L$). We have to reconsider the preceding argument. It started with equation (4) in which, now, the r.h.s can be considered as a sum over $a = 1, ..., L$. It follows that a particular family of solutions is the set of block-diagonal matrices $\mathbf{X}$, with a $L \times L$ block being a sign-permutation matrix, followed by a $(N-L) \times (N-L)$ block being an arbitrary $(N-L) \times (N-L)$ orthogonal matrix. We show now that these are, up to a global permutation, the only solutions.

Since the k-cumulants are zero for $a > L$, we have now the following possibilities for each i:

1. $\Delta_i = 0$ and $X_{i,a} = 0, a = 1, ..., L$;

2. $\Delta_i \neq 0$, $X_{i,a} = 0, a > L$, and for $a = 1, ..., L$ (6) is valid with at least one $\varepsilon_{i,a}$ non zero.

By applying an appropriate permutation, we can assume that it is the first ℓ indices, $i = 1, ..., \ell$, for which $\Delta_i \neq 0$. Hence $\mathbf{X}$ has a non zero upper-left $\ell \times L$ block, $\mathbf{X}^1$, which satisfies all the equation derived previously (in particular $\mathbf{X}^1\mathbf{X}^{1T} = \mathbf{1}_N$, but with $i = 1, ..., \ell$ and $a = 1, ..., L$; and a non zero lower-right $(N-\ell) \times (N-L)$ block, $\mathbf{X}^2$, for which the only constraint is $\mathbf{X}^2\mathbf{X}^{2T} = \mathbf{X}^{2T}\mathbf{X}^2 = \mathbf{1}_N$. It follows from the discussion of the case with non zero cumulants that $\mathbf{X}^1$ has a single non zero element per line and per column, which implies that this is a square matrix: $\ell = N$, and this completes the proof. In addition, if only one source has zero k-order cumulant, then the $\mathbf{X}^2$ matrix is just a 1×1 matrix, whose unique element is thus ± 1. Hence, in that case, it is in fact all the N sources which are separated.

References

[1] S. I. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*. MIT Press, 1996, in press.

[2] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Dover, 1972.

[3] J. J. Atick. Could information theory provide an ecological theory of sensory processing. *NETWORK*, 3:213–251, 1992.

[4] F. Attneave. Informational aspects of visual perception. *Psychological Review*, 61:183–193, 1954.

[5] A. Belouchrani and K. A. M. Séparation aveugle au second ordre de sources corrélées. In *GRETSI'93*, Juan-Les-Pins, 1993.

[6] H. B. Barlow. Possible principles underlying the transformation of sensory messages. In W. Rosenblith, editor, *Sensory Communication*, page 217. M.I.T. Press, Cambridge MA, 1961.

[7] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. I. T.*, 39, 1993.

[8] E. Bienenstock, L. Cooper, and P.W. Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal Neurosciences*, 2:32–48, 1982.

[9] H. B. Barlow, T. P. Kaushal, and G. J. Mitchison. Finding minimum entropy codes. *Neural Comp.*, 1:412–423, 1989.

[10] R. E. Blahut. *Principles and Practice of Information Theory*. Addison-Wesley, Cambridge MA, 1988.

[11] Y. Bar-Ness. Bootstrapping adaptive interference cancelers: some practical limitations. In *The Globecom Conf.*, pages 1251–1255, paper F3.7, Miami, November 1982.

[12] A. Bell and T. Sejnowski. An information-maximisation approach to blind separation and blind deconvolution. *Neural Comp.*, 7:1129–1159, 1995.

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

[14] J.-F. Cardoso. Source separation using higher-order moments. In *Proc. Internat. Conf. Acoust. Speech Signal Process.-89*, pages 2109–2112, Glasgow, 1989.

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

[16] G. Cybenko. Approximations by superpositions of a sigmoidal function. *Math. Contr. Signals Syst.*, 2:303–314, 1989.

[17] P. Del Giudice, A. Campa, N. Parga, and J.-P. Nadal. Maximization of mutual information in a linear noisy network: a detailed study. *NETWORK*, 6:449–468, 1995.

[18] N. Delfosse and Ph. Loubaton. Adaptive blind separation of independent sources: a deflation approach. *Signal Processing*, 45:59–83, 1995.

[19] Y. Deville and L. Andry Application of blind source separation techniques to multi-tag contact less identification systems. Proceedings of NOLTA'95, pp 73-78, Las Vegas 1995

[20] L. Féty. Méthodes de traitement d'antenne adaptée aux radio-communications. In *PhD Thesis*, ENST Paris, 1988.

[21] G. E. Hinton, P. Dayan, B J Frey, and R M Neal. The wake-sleep algorithm for unsupervised neural networks. *Science*, 268:1158–1160, 1995.

[22] J.J. Hopfield. Olfactory computation and object perception. *Proc. Natl. Acad. Sci. USA*, 88:6462–6466, 1991.

[23] M. Haft, M. Schlang, and G. Deco. Information theory and local learning rules in a self-organizing network of ising spins. *Phys. Rev. E*, 52:2860–2871, 1995.

[24] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1991.

[25] N. Intrator and L. Cooper. bjective function forulation of the bcm theory of visual cortical plasticity: statistical connections, stability conditions. *Neural Networks*, 5:3–17, 1992.

[26] C. Jutten and J. Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[27] Z. Li and J. J. Atick. Efficient stereo coding in the multiscale representation. *Neural Comp.*, 1993, to appear.

[28] B. Laheld and J.-F. Cardoso. Adaptive source separation without prewhitening. In *Proc. EUSIPCO*, pages 183–186, Edinburgh, 1994.

[29] Law and L. Cooper. Formation of receptive fields in realistic visual environments according to the bcm theory. *Proc. Natl. Acad. Sci. USA*, 91:7797–7801, 1994.

[30] R. Linsker. Self-organization in a perceptual network. *Computer*, 21:105–17, 1988.

[31] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72:3634–3637, 1994.

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

[33] L. C. Parra. Symplectic nonlinear component analysis. *preprint*, 1996.

[34] A. N. Redlich. Redundancy reduction as a strategy for unsupervised learning. *Neural Comp.*, 5:289–304, 1993.

[35] J.-P. Rospars and J.-C. Fort. Coding of odor quality: roles of convergence and inhibition. *NETWORK*, 5:121–145, 1994.

[36] C.L. Scofield and L. Cooper. Development and properties of neural networks. *Contemporary Physics*, 26:125–145, 1985.

[37] B. Shraiman. *private communication, and: technical memomorandum: AT&T Bell Labs TM-11111- 930811-36*, 1993.

[38] L. Tong, V. Soo, R. Liu, and Y. Huang. Amuse: a new blind identification algorithm. In *Proc. ISCAS*, New Orleans, USA, 1990.

[39] J.H. van Hateren. Theoretical predictions of spatiotemporal receptive fields of fly lmcs, and experimental validation. *J. Comp. Physiology A*, 171:157–170, 1992.