

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

Jean-Pierre Nadal

*Laboratoire de Physique Statistique, Ecole Normale Supérieure, Paris, France*

Nestor Parga

*Departamento de Física Teórica, Universidad Autónoma de Madrid, Canto Blanco, Madrid, Spain*

**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 article, we discuss several aspects of the source separation problem. We give simple conditions on the network output that, if satisfied, guarantee that source separation has been obtained. Then we study adaptive approaches, in particular those based on redundancy reduction and maximization of mutual information. We show how the resulting updating rules are related to the BCM theory of synaptic plasticity. Eventually we briefly discuss extensions to the case of nonlinear mixtures. Throughout this article, 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 their simplicity but also their drawbacks.**

### 1 Introduction ---

Recently many studies have been devoted to the study of sensory coding following a general framework initiated by H. Barlow more than thirty years ago (Barlow, 1961). 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 simple enough neuron model, one derives the parameters of the network (synaptic efficacies, transfer functions, etc.) that minimize this cost function. A great deal of work has been done on cost functions based on information-theoretic criteria (Barlow, Kaushal, & Mitchison, 1989; Linsker, 1988; Atick, 1992). The result for the receptive fields will crucially

depend on the statistical properties of the signal to be processed (visual scenes, olfactory stimuli, etc.). Several cost functions have been proposed. One is based on Barlow's original idea (Barlow, 1961), which is that redundancy reduction should be performed. This leads to the notion of factorial code (Barlow et al., 1989; Redlich, 1993): 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 in detail in the case of visual processing, with a linear neural network model (Atick, 1992). A different proposal has been promoted by Linsker (1988) under the name of the *infomax* principle: one asks the network to 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 (Linsker, 1988; van Hateren, 1992; Del Giudice, Campa, Parga, & Nadal, 1995). The predictions of these two strategies, redundancy reduction and maximization of mutual information, appeared to be very similar. In fact, we have shown (Nadal & Parga, 1994) that in the low synaptic noise limit with nonlinear outputs, infomax implies redundancy reduction, when optimization is done over both the synaptic efficacies and the nonlinear 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 (Atick, 1992) that pure redundancy reduction is not a meaningful strategy. Essentially, an efficient code is one that gives the independent features but add some redundancy to compensate for the noise. This can be studied in detail for a linear neural network (Del Giudice et al., 1995), 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 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). Since the pioneering work of Bar-Ness (1982) and Jutten and Herault (1991), interest in source separation has increased considerably (Comon, 1994; Cardoso, 1989; Molgedey & Schuster, 1994). The link between signal processing and sensory coding has been present in this field from the beginning. Jutten and Herault proposed a now well-known heuristic for performing BSS using a neuromimetic architecture for analyzing a signal coming from muscles. Experience gained with the study of this algorithm led to new approaches to the study of the olfactory system

(Hopfield, 1991; Rospars & Fort, 1994). On the neuropsychological side, it is involved in the “cocktail party effect” (being able to focus on a single speaker, separating his or her speech from all the ambient noise). In signal processing, BSS can be used for making cancellers, that is, systems for subtracting the noise from the signal (Bar-Ness, 1982). In fact, BSS has a wide variety of applications in data processing (see, e.g., Deville & Andry, 1995). In speech and sound processing, there is an additional complication: the signal is not simply a linear superposition of sources but also a convolution (Jutten & Herault, 1991). In this article, however, we will not consider the case of blind source deconvolution.

While working on linear mixtures of sources, researchers soon realized that finding independent components might be a useful strategy for data processing. The term independent component analysis (ICA) was then promoted as a generalization of BSS to nongaussian data and as an alternative to the standard principal component analysis (PCA) (Jutten & Herault, 1991; Comon, 1994). To one of his papers on BSS, Comon (1994) has given the title: “Independent Component Analysis: A New Concept?” The answer to this question is no, since this concept was proposed by Barlow at the beginning of the 1960s, under the name of factorial coding. In fact, even earlier, Atneave (1954) had addressed it, although in a less precise way, under the name of economy of coding. Still, 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 vein, it has been shown (Comon, 1994) 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. This convergence in the definition of relevant concepts may not seem surprising, but this was not always the case. In particular, Barlow (1961) put forth some stimulating discussions opposing nature’s and engineers’ approaches to coding.

In the case of the BSS framework, the hypothesis is 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 (Jutten & Herault, 1991), there are adaptive and gradient descent algorithms based on appropriate cost functions (Bar-Ness, 1982; Burel, 1992; Comon, 1994; Laheld & Cardoso, 1994; Bell & Sejnowski, 1995; Delfosse & Loubaton, 1995; Amari et al., 1996), and algebraic solutions (Féty, 1988; Cardoso, 1989; Tong, Soo, Liu, & Huang, 1990; Molgedey & Schuster, 1994; Shraiman, 1993) in which the filter is algebraically derived from a particular set of measurements on the data.

In this article 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 it necessary to put our contribution into perspective with previous works on source separation and redundancy reduction. In fact we will use a general framework—the reduction to the search for an orthogonal matrix (explained in section 2)—for both deriving the new results and presenting related work, sometimes in a simpler way as compared to the original formulation.

In section 3, 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 that enter in any of these conditions is smaller than what is usually found in the literature. In section 4, 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 5, after a reminder 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 one of them in more detail and show how the resulting updating rules are related to the BCM theory of synaptic plasticity (Bienenstock, Cooper, & Munro, 1982). In section 6, we briefly discuss adaptive algorithms from cost functions based on cumulants, these costs being built from the conditions derived in section 3. In section 7 we also briefly discuss possible extensions to nonlinear processing. Perspectives are given in section 8.

## 2 Linear Mixtures of Independent Sources

**2.1 Statement of the Problem.** Here and in the rest of the article as well, we consider the case where the factorization problem has a priori an exact solution using a linear filter or a feedforward neural network with no hidden units (we will 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  $\mathbf{S}(t)$  are an  $N$ -dimensional vector given by

$$S_j(t) = \sum_{a=1}^N M_{j,a} \sigma_a(t), \quad j = 1, \dots, N, \quad (2.1)$$

(in vector form  $\mathbf{S} = \mathbf{M}\boldsymbol{\sigma}$ ) where the  $\sigma_a$  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  $\mathbf{K}^0$ :

$$\mathbf{K}_{a,b}^0 \equiv \langle \sigma_a(t) \sigma_b(t) \rangle_c = \delta_{a,b} \mathbf{K}_a^0, \quad (2.2)$$

where  $\delta_{a,b}$  is the Kronecker symbol. Here and throughout this article  $\langle z_1 z_2, \dots, z_k \rangle_c$  denotes the cumulant of the  $k$  random variables  $z_1, \dots, z_k$ .

Without loss of generality, one can always assume that the sources have zero average:

$$\langle \sigma_a \rangle = 0, \quad a = 1, \dots, N \quad (2.3)$$

(otherwise one has to estimate the average of each input and subtract 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) output  $\mathbf{h}$ ,

$$h_i(t) = \sum_{j=1}^N J_{i,j} S_j(t), \quad i = 1, \dots, N, \quad (2.4)$$

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 nonzero 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 stating that all we can expect is to estimate normalized cumulants of the sources, such as the  $k$ -order normalized cumulants  $\zeta_k^a$ :

$$\zeta_k^a \equiv \frac{\langle \sigma_a^k \rangle_c}{\langle \sigma_a^2 \rangle_c^{k/2}}, \quad a = 1, \dots, N. \quad (2.5)$$

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  $\pm 1$  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.

**2.2 Reducing the Problem to Finding an Orthogonal Matrix.** An elementary but extremely useful remark, first made by Comon (1994) and Cardoso (1989), 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 to

diagonalize the two-point connected correlation of the inputs. As 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 higher cumulants. Because we will make extensive use of this fact, and in order to introduce our notation, we give a detailed derivation.

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

$$\langle \mathbf{h}\mathbf{h}^T \rangle_c = \mathbf{1}_N. \quad (2.6)$$

This reads, in term of the correlation  $C_0$  between the input data,

$$C_0 \equiv \langle \mathbf{S}\mathbf{S}^T \rangle_c = \mathbf{M}\mathbf{K}^0\mathbf{M}^T \quad (2.7)$$

as

$$\mathbf{J}C_0\mathbf{J}^T = \mathbf{1}_N. \quad (2.8)$$

Keep in mind that  $C_0$  is what one can measure, and the right-hand side of equation 2.7 is its expression in function of the unknowns. Every solution of equation 2.8 can be written as

$$\mathbf{J} = \Omega C_0^{-1/2}, \quad (2.9)$$

where  $\Omega$  is an orthogonal matrix:

$$\Omega \Omega^T = \mathbf{1}_N. \quad (2.10)$$

The inverse of the square root of  $C_0$  is defined from the PCA of the real positive matrix  $C_0$ . If we denote by  $\mathbf{J}^0$  the matrix whose rows are the orthonormal eigenvectors of  $C_0$ , and by  $\Lambda_0$  the diagonal matrix of the associated eigenvalues, one has

$$\begin{aligned} \mathbf{J}^0 C_0 \mathbf{J}^{0T} &= \Lambda_0 \\ \mathbf{J}^0 \mathbf{J}^{0T} &= \mathbf{1}_N \end{aligned} \quad (2.11)$$

and

$$C_0^{-1/2} = \mathbf{J}^{0T} \Lambda_0^{-\frac{1}{2}} \mathbf{J}^0. \quad (2.12)$$

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

$$\mathbf{h}^0 = \Lambda_0^{-\frac{1}{2}} \mathbf{J}^0 \mathbf{S}. \quad (2.13)$$

Then, taking  $\mathbf{J}$  as in equation 2.9 and replacing  $C_0^{-1/2}$  by its expression (see equation 2.12),  $\mathbf{h}$  can be written as

$$\mathbf{h} = \Omega \mathbf{J}^{0T} \Lambda_0^{-\frac{1}{2}} \mathbf{J}^0 \mathbf{S} = \Omega \mathbf{J}^{0T} \mathbf{h}^0. \quad (2.14)$$

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

$$\mathbf{h} = \mathcal{O} \mathbf{h}^0 \quad (2.15)$$

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 terms of the unknowns, that the rows of  $\mathbf{J}^0$  are the eigenvectors of  $C_0$ :

$$\mathbf{J}^0 \mathbf{M} \mathbf{K}^0 \mathbf{M}^T \mathbf{J}^{0T} = \Lambda_0. \quad (2.16)$$

This implies that  $\Lambda_0^{-\frac{1}{2}} \mathbf{J}^0 \mathbf{M} \mathbf{K}^0 \frac{1}{2}$  is an (unknown) orthogonal matrix, which, for later convenience, we define as the transposed of some orthogonal matrix  $\mathcal{O}^0$ :

$$\begin{aligned} \Lambda_0^{-\frac{1}{2}} \mathbf{J}^0 \mathbf{M} \mathbf{K}^0 \frac{1}{2} &= \mathcal{O}^{0T} \\ \mathcal{O}^0 \mathcal{O}^{0T} &= \mathbf{1}_N. \end{aligned} \quad (2.17)$$

Hence, the projected data  $\mathbf{h}^0$  as defined in equation 2.13 can be expressed, in function of the unknown sources, as an orthogonal mixture:

$$\mathbf{h}^0(t) = \mathcal{O}^{0T} \mathbf{K}^{0-1/2} \boldsymbol{\sigma}(t). \quad (2.18)$$

Eventually, from equations 2.15 and 2.18, one can also write  $\mathbf{h}$  in terms of the sources as

$$\mathbf{h} = \mathbf{X} \mathbf{K}^{0-1/2} \boldsymbol{\sigma}, \quad (2.19)$$

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  $\mathbf{X}$ .

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 may be specifically useful depending on the particular approach. In section 3, where we will determine the family of couplings  $\mathbf{J}$

solution of a given set of equations, it will be convenient to use equation 2.19, taking  $\mathbf{X}$  as unknown. For the algebraic approach in section 4, we will use only the quantities related to the principal components, that is,  $\mathbf{h}^0$  and  $\mathcal{O}^0$  with the basic equation 2.18. And in section 6 we will consider adaptive algorithms for computing either  $\mathbf{J}$  itself or  $\mathcal{O}$  as defined in equation 2.15.

Throughout this article, we will mainly consider the ideal situation where one would have access to the exact values of the cumulants. Nevertheless, we can make some general remarks about the practical cases, where cumulants are computed empirically (e.g., as time averages performed over some large enough time window  $T$ ). In practice, all that is required implicitly is to have for the sources small enough cross-cumulants when these cumulants are defined from the chosen averaging procedure. This has to be true only for the cumulants needed to compute in a given approach. We will not address the problem of the accuracy of the solution as a function of the quality of the estimation of cumulants.

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, as 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  $\Omega$ . An important example is in the application of redundancy reduction to the modeling of the first stages in the visual system (Linsker, 1988; Atick, 1992; Li & Atick, 1993). There, after whitening (in fact, an extension of whitening that 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 (Li & Atick, 1994) a block-diagonal orthogonal matrix, which leads to a multiscale representation of the visual scenes.

### 3 Criteria Based on Cross-Cumulants of Output Activities

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

**3.1 Condition on a Set of Nonsymmetric Higher Cumulants.** The claim is that for  $\mathbf{J}$  to be a solution, it is sufficient (and 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 odd integer at least equal to 3 for which the  $k$ -cumulants of the sources,  $\zeta_k^a$ ,  $a = 1, \dots, N$ , 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  $i, i'$ ,*

$$\begin{cases} \langle h_i h_{i'} \rangle_c = \delta_{i,i'} \\ \langle h_i^{(k-1)} h_{i'} \rangle_c = 0 \text{ for } i \neq i'. \end{cases} \quad (3.1)$$

where  $\mathbf{h}$  is the output vector as defined in equation 2.4.

- (ii) *If only  $1 \leq L \leq N - 2$   $k$ -order cumulants are nonzero, then any solution  $\mathbf{J}$  of equation 3.1 is the product of a sign permutation by a matrix that separates the  $L$  sources having nonzero  $k$ -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.*

Note that the second line in equation 3.1 means that the matrix  $\langle h_i^{(k-1)} h_{i'} \rangle_c$  is diagonal but the values of the diagonal terms, to be named below  $\Delta_i$ , need not be known in advance. The detailed proof is given in appendix A. Here we give only a sketch of it, in the case where all the  $k$ -cumulants are nonzero. First, solving for the diagonalization of the second-order cumulant, one uses the representation in terms of orthogonal matrices as in section 2. As a result, we have seen that one can write  $\mathbf{h}$  as  $\mathbf{h} = \mathbf{X} \mathbf{K}^{0 -1/2} \boldsymbol{\sigma}$ , 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 that 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  $\zeta_k^a$ ). 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 nonzero element, and this implies that  $\mathbf{X}$  is a sign permutation.

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 then has to try another value of  $k$ , but only for the subspace of yet unseparated sources.

As expected, 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.

For  $k$  even, one can easily find an example showing that the conditions (see equation 3.1) are not sufficient (see appendix A).

An interesting application of this theorem concerns the adaptive algorithm in Jutten and Herault (1991). In its simplest version, this algorithm aims at setting to zero the two-point correlation and the cross-cumulants  $\langle h_i^2 h_{i'} \rangle_c$  for  $i \neq i'$ . If the algorithm does reach that particular fixed point, theorem 1 asserts that full source separation has been obtained.

It is not difficult to find other families of cumulants of a given order  $k$  for which a similar theorem will hold. We illustrate this by giving an analogous result for a set of cumulants involving more indices.

**Theorem 2.** *Let  $k$  and  $m$  be two integers with  $m$  at least equal to 2 and  $k$  strictly greater than  $m$ , for which the  $k$ -cumulants of the sources,  $\zeta_k^a$ ,  $a = 1, \dots, N$ , 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  $i, i', i''$ ,

$$\begin{cases} \langle h_i h_{i'} \rangle_c = \delta_{i,i'} \\ \langle h_i^{(k-m)} h_{i'}^{m-1} h_{i''} \rangle_c = 0 \text{ for at least two nonidentical indices,} \end{cases} \quad (3.2)$$

where  $\mathbf{h}$  is the output vector as defined in equation 2.4.

- (ii) *If only  $1 \leq L \leq N - 2$   $k$ -order cumulants are nonzero, then any solution  $\mathbf{J}$  of equation 3.2 is the product of a sign permutation by a matrix that separates the  $L$  sources having nonzero  $k$ -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.*

The proof is given in appendix B. In theorem 2, the cases  $k = m$  and  $m = 1$  are excluded so that equation 3.2 never reduces to the conditions in equation 3.1 of theorem 1. In fact, these conditions are a subset of the conditions in equation 3.2, corresponding to the cases  $i = i'$ . However, theorem 2 is not exhausted by theorem 1: no condition on the parity of  $k$  (the order of the cumulants) is required for the second theorem. Theorem 2 shows that only a subset of all the  $k$ -order cumulants with three indices is enough in order to obtain separation. The case  $m = 2$  is related to the joint diagonalization approach of Cardoso and Souloumiac (1993), which we will discuss in section 4, where algebraic solutions are presented.

**3.2 Conditions Related to Algebraic Solutions.** In the above approach, 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 in the case of Theorem 1. Coming back to our result, the counting argument suggests that one should be able to separate the sources with fewer 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 equation 3.1 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 Theorems 1 and 2, 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 that 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 extension of) known algebraic solutions, which we review in section 4.

There is a price to pay for working with the minimal number of equations: the restriction that we had in Theorems 1 and 2, namely, that of having nonzero  $k$ -order statistics for the sources is replaced here by the much more restricting condition of having different cumulants, as stated below.

**3.2.1 Conditions on Symmetric Higher Cumulants.** We first consider conditions based on correlations at equal time.

**Theorem 3.** *Let all the cumulants  $\zeta_4^a$ , as defined in equation 2.5 for  $k = 4$  be different from one another; 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  $i, i'$ ,

$$\begin{cases} \langle h_i h_{i'} \rangle_c = \delta_{i,i'} \\ \langle h_i \sum_{i'=1}^N (h_{i'})^2 h_{i'} \rangle_c = 0 \text{ for } i \neq i'. \end{cases} \quad (3.3)$$

where  $\mathbf{h} = \{h_i, i = 1, \dots, N\}$  is the output vector as defined in equation 2.4.

If one or several sets of sources have a common value for  $\zeta_4^a$  (but different from one set to the other), then any  $\mathbf{J}$  solution of equation 3.3 is the product of a sign permutation by a matrix that separates the sources having a distinct 4-cumulant, and separates 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 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 equation 3.3 can then be expressed as

$$\left\langle h_i \sum_{i'=1}^N (h_{i'})^2 h_{i'} \right\rangle_c = \sum_a X_{i,a} \left\{ \sum_{i'} (X_{i',a})^2 \right\} \zeta_4^a X_{i',a}. \quad (3.4)$$

But since  $\mathbf{X}$  is orthogonal,  $\sum_i (X_{i,a})^2 = 1$ , and equation 3.4 reduces to

$$\left\langle \mathbf{h}_i \sum_{i'=1}^N (h_{i'})^2 \mathbf{h}_{i'} \right\rangle_c = \sum_a X_{i,a} \zeta_4^a X_{i',a}. \quad (3.5)$$

What we want is to impose

$$\left\langle \mathbf{h}_i \sum_{i'=1}^N (h_{i'})^2 \mathbf{h}_{i'} \right\rangle_c = \Delta_i \delta_{i,i'} \quad (3.6)$$

where the  $\Delta_i$  are yet indeterminate constants. Comparing equations 3.5 and 3.6, one sees that  $\mathbf{X}$  gives the eigendecomposition of a matrix, which is in fact already diagonal. This implies that if all the eigenvalues, that is the  $\zeta_4^a$ , 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 with an eigenvalue of degeneracy  $\ell$ . The algebraic solution associated with Theorem 3 was proposed by Cardoso (1989) (see section 4.3).

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  $\mathbf{X}_a$ ,  $a = 1, \dots, N$ , the  $N$  numbers  $\langle Q(\sigma_a \mathbf{X}_a) \sigma_a^2 \rangle_c$ ,  $a = 1, \dots, N$  are distinct (here every  $\sigma_a$  should be understood as the normalized source  $\frac{\sigma_a}{\sqrt{\langle \sigma_a^2 \rangle_c}}$ ). 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.

**3.2.2 Conditions on Time Correlations.** We consider now conditions related to algebraic solutions based on time correlations (Féty, 1988; Tong et al., 1990; Belouchrani, 1993; Molgedey & Schuster, 1994). These are extremely simple. If the sources present time correlations, namely, the two-point correlation matrix  $\mathbf{K}(\tau)$  for some time delay  $\tau > 0$ ,

$$\mathbf{K}(\tau)_{a,b} \equiv \langle \sigma_a(t) \sigma_b(t - \tau) \rangle_c \quad (3.7)$$

has nonzero diagonal elements:

$$\mathbf{K}(\tau)_{a,b} = \delta_{a,b} K_a(\tau), \quad (3.8)$$

then one can state:

**Theorem 4.** *Let all the cumulants  $K_a(\tau)$ , as defined in equation 3.8 be different from one another; 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  $i, i'$ ,

$$\begin{cases} \langle h_i(t) h_{i'}(t) \rangle_c = \delta_{ii'} \\ \langle h_i(t) h_{i'}(t - \tau) \rangle_c = 0 \text{ for } i \neq i' \end{cases} \quad (3.9)$$

where  $\mathbf{h} = \{h_i, i = 1, \dots, N\}$  is the output vector as defined in equation 2.4.

If one or several sets of sources have a common value for  $K_a(\tau)$  (but different from one set to the other), then any  $\mathbf{J}$  solution of equation 3.9 is the product of a sign permutation by a matrix that separates the sources having a distinct  $K_a(\tau)$ -cumulant, and separate the sets globally. The restriction of  $\mathbf{J} \mathbf{M} \mathbf{K}^{0.5}$  to the subspace of a given set is still an arbitrary orthogonal matrix.

The proof is essentially the same as the one of Theorem 3. One writes

$$\langle h_i(t) h_{i'}(t - \tau) \rangle_c = \sum_a X_{i,a} K_a(\tau) X_{i',a}. \quad (3.10)$$

What we want is to impose

$$\langle h_i(t) h_{i'}(t - \tau) \rangle_c = \Delta_i \delta_{ii'} \quad (3.11)$$

where the  $\Delta_i$  are yet indeterminate constants. Equations 3.10 and 3.11 play the same role as equations 3.5 and 3.6 in the case of Theorem 3, and the rest of the proof follows in the same way.

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.

**3.3 Comparison with Other Criteria.** To conclude this section, we comment briefly on other criteria, namely, those proposed by Comon (1994). This author has shown that ICA 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

$$\mathcal{C} \equiv - \sum_{i=1}^N \langle (h_i)^k \rangle_c^2 \quad (3.12)$$

for any given  $k > 2$ , after whitening has been performed. This is a nice result since it involves only  $N$  cumulants. However, defining a cost function is not exactly the same as giving explicit conditions to the cumulants (although

one can easily derive a cost function from these conditions, as we will do in section 6). In fact, the minimization of a cost function such as equation 3.12 does involve implicitly the computation of order  $N^2$  cumulants, as it should (see the counting argument at the beginning 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 equation 3.12 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 that then matter in the counting argument (we will come back to the algorithmic aspects in section 6). Second, Comon showed also that the minimization of equation 3.12 is equivalent to setting to zero *all* the nondiagonal cumulants of the same order  $k$  (still in addition to whitening):

$$\langle h_{i_1} h_{i_2}, \dots, h_{i_k} \rangle_c = 0, \text{ for every set of } k \text{ nonidentical indices.} \quad (3.13)$$

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

#### 4 Algebraic Solutions

In this section we present four families of algebraic solutions. We point out their advantages and drawbacks insisting on their simplicity when formulated within the framework described in section 2 and relating them to the results of section 3.

**4.1 Using Time-Delayed Correlations.** In the case where each source  $\sigma_a$  shows time correlations, it has been shown (Féty, 1988; Tong et al., 1990; Belouchrani, 1993; Molgedey & Schuster, 1994) that there is a simple algebraic solution using only second-order cumulants. More precisely, let us assume that the two-point correlation matrix  $\mathbf{K}(\tau)$  for some time delay  $\tau > 0$ ,

$$K(\tau)_{a,b} \equiv \langle \sigma_a(t) \sigma_b(t - \tau) \rangle_c \quad (4.1)$$

has nonzero diagonal elements:

$$K(\tau)_{a,b} = \delta_{a,b} K_a(\tau). \quad (4.2)$$

It follows (Molgedey & Schuster, 1994) that source separation is obtained by asking for the rows of  $\mathbf{J}$  to be the left eigenvectors of the following non-symmetric matrix

$$\mathbf{C}(\tau) \mathbf{C}_0^{-1}, \quad (4.3)$$

where  $C(\tau)$  is the second-order cumulant of the inputs at time delay  $\tau$ :

$$C(\tau) = \langle \mathbf{S}(t) \mathbf{S}^T(t - \tau) \rangle_c. \tag{4.4}$$

The equivalent, but more natural, approach of Féty (1988) and Tong et al. (1990) is to work with a symmetric matrix, making use of the reduction to the search for an orthogonal matrix presented in section 2. Indeed, let us first perform whitening (note that this implies the resolution of the eigenproblem for  $C_0$ , a task of the same complexity as the computation of its inverse, which is needed in equation 4.3). We then compute the correlations at time delay  $\tau$  of the projected inputs  $\mathbf{h}^0$  (as given by equation 2.13), that is the matrix  $\langle \mathbf{h}^0(t) \mathbf{h}^{0T}(t - \tau) \rangle_c$ . Using the expression equation 2.18 of  $\mathbf{h}^0$  as a function of the sources, one sees that this matrix is in fact symmetric and given by:

$$\langle \mathbf{h}^0(t) \mathbf{h}^{0T}(t - \tau) \rangle_c = \mathcal{O}^{0T} \mathbf{K}^{0-1/2} \mathbf{K}(\tau) \mathbf{K}^{0-1/2} \mathcal{O}^0. \tag{4.5}$$

This shows that the desired orthogonal matrix is obtained from solving the eigendecomposition of this correlation matrix (see equation 4.5).

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

**4.2 Using Correlations at Equal Time.** Shraiman (1993) has shown how to reduce the problem of source separation to the diagonalization of a certain symmetric matrix  $\mathcal{D}$  built on third-order cumulants. We refer readers to Shraiman (1993) for the elegant derivation of this result. Here we will show how to derive this matrix from the approach introduced in sections 2.2 and 3. We will do so by working with the  $k$ -order cumulants, giving 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 equation 2.18:

$$\mathbf{h}^0(t) = \mathcal{O}^{0T} \mathbf{K}^{0-1/2} \boldsymbol{\sigma}(t). \tag{4.6}$$

One computes the  $k$ -order statistics  $\langle h_{i_1}^0 h_{i_2}^0, \dots, h_{i_k}^0 \rangle_c$ . From equation 4.6, these cumulants have the following expression in term of source cumulants:

$$\langle h_{i_1}^0 h_{i_2}^0, \dots, h_{i_k}^0 \rangle_c = \sum_{a=1}^N \mathcal{O}_{a,i_1}^0, \dots, \mathcal{O}_{a,i_k}^0 \zeta_k^a, \tag{4.7}$$

where the  $\zeta_k^a$  are the normalized cumulants as defined in equation 2.5. Now, one multiplies two such cumulants having  $k - 1$  identical indices and sums

over all the possible values of these indices. This will produce the contraction of  $k-1$  matrices  $\mathcal{O}^0$  with their transposed, leading to  $\mathbf{1}_N$ , and only two terms  $\mathcal{O}^0$  will remain. Explicitly, we consider the symmetric matrix, to be referred to as the  $k$ -Shraiman matrix,

$$\mathcal{D}_{i,\tilde{i}} = \sum_{i_1, i_2, \dots, i_{k-1}} \langle h_i^0 h_{i_1}^0, \dots, h_{i_{k-1}}^0 \rangle_c \langle h_{\tilde{i}}^0 h_{i_1}^0, \dots, h_{i_{k-1}}^0 \rangle_c, \quad (4.8)$$

which, from equation 4.7, is equal to

$$\mathcal{D}_{i,\tilde{i}} = \sum_{a=1}^N \mathcal{O}_{a,i}^0 \mathcal{O}_{a,\tilde{i}}^0 (\zeta_k^a)^2. \quad (4.9)$$

The above formula is nothing but an eigendecomposition 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  $a = 1, \dots, N$ . A solution of the source separation problem is thus given by the diagonalization of one  $k$ -Shraiman matrix (e.g., taking  $k = 3$  or  $k = 4$ ).

**4.3 A Simple Solution Using Fourth-Order Cumulants.** We now consider the solution based on fourth-order cumulants (Cardoso, 1989), which is directly related to the results obtained in section 3.2.1. Let us consider the following cumulants of the input data projected onto the principal components,  $\mathbf{h}^0$ :

$$\mathbf{C}_4{}_{i,\tilde{i}} \equiv \left\langle h_i^0 \sum_{\tilde{i}'=1}^N h_{\tilde{i}'}^0 h_{\tilde{i}'}^0 h_{\tilde{i}}^0 \right\rangle_c. \quad (4.10)$$

In term of the orthogonal matrix  $\mathcal{O}^0$  and of the cumulants  $\zeta_4^a$ , it reads

$$(\mathbf{C}_4)_{i,\tilde{i}} = \sum_{a=1}^N \mathcal{O}_{a,i}^0 \zeta_4^a \sum_{\tilde{i}'=1}^N (\mathcal{O}_{a,\tilde{i}'}^0)^2 \mathcal{O}_{a,\tilde{i}}^0. \quad (4.11)$$

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

$$\mathbf{C}_4 = \mathcal{O}^0 \mathbf{K}_4 \mathcal{O}^{0T}, \quad (4.12)$$

where  $\mathbf{K}_4$  is the diagonal matrix of the fourth-order cumulants,

$$(\mathbf{K}_4)_{a,b} = \delta_{a,b} \zeta_4^a. \quad (4.13)$$

This shows that  $\mathcal{O}^0$  can be found by solving for the eigendecomposition of the cumulant  $\mathbf{C}_4$ .

All of the algebraic solutions considered thus far in this section are based on the same two facts: (1) the diagonalization of a real positive symmetric matrix leaves an arbitrary orthogonal matrix, which can be used to diagonalize another symmetric matrix; (2) 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).

**4.4 The Joint Diagonalization Approach.** All of the algebraic solutions discussed so far 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. One may wonder whether it would be possible to work with an algebraic solution involving more equations than unknowns, in such a way that indetermination cannot occur. A positive answer is given by the joint diagonalization method of Cardoso and Souloumiac (1993) and Belouchrani (1993). Here we give a different (and slightly more general) presentation from the one in Cardoso and Souloumiac (1993). In particular, we will make use of the theorems of section 3.1. We will consider only correlations at equal time. The case of time correlations is discussed in Belouchrani (1993).

The basic idea is to joint diagonalize a family of matrices  $\Gamma^r$

$$\Gamma_{\alpha,\beta}^r = \langle h_\alpha^0 h_\beta^0 Q_r(\mathbf{h}^0) \rangle_c, \tag{4.14}$$

where  $\mathbf{h}^0$  is the principal component vector as defined in equation 2.13 and the  $Q_r$  are well-chosen scalar functions of it, the index  $r$  labeling the function (hence the matrix) in the family. One possible example is the family defined by taking for  $r$  all possible choices of  $k - 2$  indices with  $k \geq 3$ ,

$$r \equiv (\alpha_1, \dots, \alpha_{k-2}), 1 \leq \alpha_1 \leq \alpha_2, \dots, \leq \alpha_{k-2} \leq N \tag{4.15}$$

and

$$Q_{r=(\alpha_1, \dots, \alpha_{k-2})}(\mathbf{h}^0) = h_{\alpha_1}^0, \dots, h_{\alpha_{k-2}}^0. \tag{4.16}$$

The case considered in Cardoso and Souloumiac (1993) is  $k = 4$ . Using the expression of  $\mathbf{h}^0$  as a function of the normalized sources,

$$\mathbf{h}^0(t) = \mathcal{O}^{0T} \mathbf{K}^{0-1/2} \boldsymbol{\sigma}(t), \tag{4.17}$$

where  $\mathcal{O}^0$  is the orthogonal matrix that we want to compute (see section 2.2), one can write

$$\Gamma^r = \mathcal{O}^{0T} \boldsymbol{\Lambda}^r \mathcal{O}^0, \tag{4.18}$$

where  $\Lambda^r$  is a diagonal matrix with components

$$\Lambda_a^r = \zeta_k^a \mathcal{O}_{a,\alpha_1}^0, \dots, \mathcal{O}_{a,\alpha_{k-2}}^0. \quad (4.19)$$

As it is obvious in equation 4.18, the matrices  $\Gamma^r$  are jointly diagonalizable by the orthogonal matrix  $\mathcal{O}^0$ . However, if for at least a pair  $a, b$  one has  $\Lambda_a^r = \Lambda_b^r$  for every  $r$ ,  $\mathcal{O}^0$  is not the only solution (up to a sign permutation). Actually this never happens, as shown in Cardoso and Souloumiac (1993) for  $k = 4$ . Let us give a direct proof valid for any  $k$ .

We consider one particular orthogonal matrix  $\mathcal{O}$ , which jointly diagonalizes all the matrices of the family, and let  $\mathbf{h} = \mathcal{O}\mathbf{h}^0$ . By hypothesis, the matrix  $\mathcal{O}\Gamma^r\mathcal{O}^T$  is diagonal, that is

$$\langle h_i h_{i'} h_{\alpha_1}^0, \dots, h_{\alpha_{k-2}}^0 \rangle_c = 0 \text{ for } i \neq i', \quad (4.20)$$

and this for every choice of the  $k - 2$  indices. Multiplying the left-hand side by  $\mathcal{O}_{i_1,\alpha_1}, \dots, \mathcal{O}_{i_{k-2},\alpha_{k-2}}$  and summing over the Greek indices, one gets that for any choice of  $i_1, \dots, i_{k-2}$ ,

$$\langle h_i h_{i'} h_{i_1}, \dots, h_{i_{k-2}} \rangle_c = 0 \text{ for } i \neq i'. \quad (4.21)$$

We can now make use of Theorem 2: one can write equation 4.21 for the particular choice  $i_1 = i_2 = \dots = i_{k-2} = i'$ , which gives exactly the conditions (see equation 3.2) for  $k$  and  $m = 2$ , and we can thus apply Theorem 2 (equivalently, one can deduce from equation 4.21 that these cumulants are zero whenever any two indices are different, and then use equation 3.13, that is, Comon's 1994 result).

For the simplest case  $k = 3$ , these conditions are exactly those of Theorem 2: joint diagonalizing the  $N$  matrices  $\Gamma^\alpha = \langle \mathbf{h}^0 \mathbf{h}^{0T} h_\alpha^0 \rangle_c$  is strictly equivalent to imposing the conditions (see equation 3.2) for  $k = 3$  (and  $m = 2$ ) (note, however, that the number of conditions is larger than the minimum required according to Theorem 1). For  $k > 3$ , the number of conditions in equation 4.21 is larger than the number of conditions in equation 3.2. To conclude, one sees that at the price of having a number of conditions larger than the minimum required (in order to guarantee that no indetermination will occur), source separation can be done with an algebraic method, even when there are identical source cumulants.

**Remark.** In practical applications, cumulants are empirically computed, and thus the matrices under consideration are not jointly diagonalizable. For this reason, a criterion is considered in Cardoso and Souloumiac (1993) that, if maximized, provides the best possible approximation to joint diagonalization. In this article, we do not consider this aspect of the problem.

## 5 Cost Functions Derived from Information Theory

We switch now to the study of adaptive algorithms. To do so, one first has to define proper cost functions. Whereas in the next section we will consider cost functions based on cumulants, here we consider the particular costs derived from information theory. In both cases we will take advantage of the results obtained in section 3. 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 (Bienenstock, Cooper, & Munro, 1982).

**5.1 From Infomax to Redundancy Reduction.** Our starting point is the main result obtained in Nadal and Parga (1994), namely, that maximization of the mutual information between the input data and the output (neural code) leads to redundancy reduction, hence to source separation for both linear and nonlinear mixtures. To be more specific, we first give a short derivation of that fact (for more details see Nadal & Parga, 1994). We consider a network with  $N$  inputs and  $p$  outputs, and nonlinear transfer functions  $f_i, i = 1, \dots, p$ . Hence the output  $\mathbf{V}$  is given by a gain control after some (linear or nonlinear) processing:

$$V_i(t) = f_i(h_i(t)), \quad i = 1, \dots, p. \quad (5.1)$$

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

$$h_i(t) = \sum_{j=1}^N J_{i,j} S_j(t), \quad i = 1, \dots, p. \quad (5.2)$$

However, here the  $h_i(t)$  can be as well *any* deterministic (hence not necessarily linear) functions of the inputs  $\mathbf{S}(t)$ . In particular, we will make use of this fact in section 7: there,  $h_i$  will be the local field at the output layer of a one-hidden-layer network with nonlinear transfer functions. The mutual information  $\mathcal{I}$  between the input and the output is given by (see Blahut, 1988):

$$\mathcal{I} \equiv \int d^p \mathbf{V} d^N \mathbf{S} P(\mathbf{V}, \mathbf{S}) \log \frac{P(\mathbf{V}, \mathbf{S})}{Q(\mathbf{V}) P(\mathbf{S})}. \quad (5.3)$$

This quantity is well defined only if noise processing is taken into account (e.g., resolution noise). In the limit of vanishing additive noise, one gets that maximizing the mutual information is equivalent to maximizing the (differential) output entropy  $H(Q)$  of the output distribution  $Q = Q(\mathbf{V})$ ,

$$H(Q) = - \int d^p \mathbf{V} Q(\mathbf{V}) \log Q(\mathbf{V}). \quad (5.4)$$

On the right-hand side of equation 5.4, one can make the change of variable  $\mathbf{V} \rightarrow \mathbf{h}$ , using

$$\prod_{i=1}^p dV_i Q(\mathbf{V}) = \prod_{i=1}^p dh_i \Psi(\mathbf{h}) \quad (5.5)$$

and

$$dV_i = f'_i(h_i) dh_i, \quad i = 1, \dots, p. \quad (5.6)$$

This gives

$$H(Q) = - \int d\mathbf{h} \Psi(\mathbf{h}) \ln \frac{\Psi(\mathbf{h})}{\prod_{i=1}^p f'_i(h_i)}. \quad (5.7)$$

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

$$\Psi(\mathbf{h}) = \prod_{i=1}^p \Psi_i(h_i), \quad (5.8)$$

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

$$f'_i(h_i) = \Psi_i(h_i), \quad i = 1, \dots, p. \quad (5.9)$$

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

**5.2 The Specific Case of BSS.** Let us now return to the **BSS** problem for which the  $\mathbf{h}$  are taken as linear combinations of the inputs. By hypothesis, the  $N$ -dimensional input is a linear mixture of  $N$ -independent sources. In the following we consider only  $p = N$ .

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 Nadal and Parga (1994), 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 *first* to 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 that is used in order to build the network that maximizes information transfer. Still, if one considers that the transfer functions are chosen at each

instant of time according to equation 5.9, the mutual information becomes precisely equal to minus the redundancy cost function  $\mathcal{R}$ :

$$\mathcal{R} \equiv \int d\mathbf{h} \Psi(\mathbf{h}) \ln \frac{\Psi(\mathbf{h})}{\prod_{i=1}^p \Psi_i(h_i)}. \quad (5.10)$$

In the context of blind source separation the relevance of the redundancy cost function has been recognized by Comon (1994) (we also note a work by Burel [1992], where a different but related cost function is considered).

**Remark on the terminology.** The quantity (see equation 5.10), called here and in the literature related to sensory coding the *redundancy*, is called in the signal processing literature (in particular in Comon, 1994) the *mutual information*, short for the mutual information between the random variables  $h_i$  (the outputs). But this mutual information, that is the redundancy (see equation 5.10), should not be mistaken for the mutual information (see equation 5.3) we introduced above, which is defined in the usual way—that is, between the input and the output of a processing channel (Blahut, 1988). To avoid confusion, we will consistently use *redundancy* for equation 5.10, and *mutual information* for equation 5.3.

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, changing the variable  $\mathbf{h} \rightarrow \mathbf{S}$  in equation 5.7, as done in Bell and Sejnowski (1995). Since the input entropy  $H(P)$  is a constant (it does not depend on the couplings  $\mathbf{J}$ ), the quantity that has to be maximized is

$$\mathcal{E} = \ln |\mathbf{J}| + \sum_i \langle \log c_i(h_i) \rangle, \quad (5.11)$$

where  $|\mathbf{J}|$  is the absolute value of the determinant of the coupling matrix  $\mathbf{J}$  and  $\langle \cdot \rangle$  is the average over the output activity  $h_i$ . The function  $c_i$  can be given two interpretations: it is equal to either  $f_i'$ , if one considers the mutual information, or to  $\Psi_i$ , 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.

The cost (see equation 5.11) can be given another interpretation. In fact, it was first derived in a maximum likelihood approach (Gaeta & Lacoume, 1990; Pham, Garrat, & Jutten, 1992): it is easy to see that equation 5.11, with  $c_i = \Psi_i$ , is equal to the (average of) the log likelihood of the observed data (the inputs  $\mathbf{S}$ ), given that they have been generated as a linear combination of independent sources with the  $\Psi_i$  as marginal distributions.

In the two following subsections we consider practical approaches.

**5.3 Working with a Given 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 these distributions are not known—and cannot be estimated without first performing source separation—the choice of a proper parameterized 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 practical way to search for the best  $f_i^*$  is to restrict the search to an a priori chosen family of transfer functions. In Pham et al. (1992), a practical algorithm is proposed, based on a particular choice combined with an expansion of the cost close to a solution. Another, and very simple, strategy has been tried in Bell and Sejnowski (1995), where very promising results have been obtained on some specific applications. Their numerical simulations suggest that one can take transfer functions with a simple behavior (that is, for example, with one peak in the derivative when the data show only one peak in their distribution), and to optimize just the gain and the threshold in each transfer function, which means fitting the location of the peak and its height.

**5.4 Cumulant Expansion of the Marginal Distributions.** When working with the redundancy, one would like to avoid having to estimate the marginal distributions from histograms, since this would take a lot of time. One may parameterize 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, considered in Gaeta and Lacoume (1990) and Comon (1994), is to replace each marginal by a simple probability distribution with the same first cumulants as the ones of the actual distribution. Recently this approach has been used in Amari et al. (1996). We consider this expansion here with a slightly different point of view in order to relate this approach to the results of section 3.

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  $\sigma_\alpha$ —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 3, which is that conditions on a limited set of cumulants are sufficient in order to obtain factorization, strongly suggests replacing the unknown distribution with a simple distribution having the same first cumulants up to some given order.

Let us consider the systematic close-to-gaussian cumulant expansion of

$\Psi_i(h_i)$  (Abramowitz & Stegun, 1972). At first nontrivial order, it is given by

$$\Psi_i(h_i) \approx \Psi_i^1(h_i) \equiv \Psi^0(h_i) \left[ 1 + \lambda_i^{(3)} \frac{h_i(h_i^2 - 3)}{6} \right], \quad (5.12)$$

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

$$\Psi^0(h_i) \equiv \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{h_i^2}{2}\right), \quad (5.13)$$

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

$$\lambda_i^{(3)} \equiv \langle h_i^3 \rangle_c. \quad (5.14)$$

In expression 5.12, we have taken into account that, as explained in section 2, one can always take

$$\langle h_i \rangle = 0. \quad (5.15)$$

and

$$\langle h_i^2 \rangle_c = 1. \quad (5.16)$$

In the cost function (see equation 5.11), that is,

$$\mathcal{E} = \ln |\mathbf{J}| + \sum_i \int dh_i \Psi_i(h_i) \log \Psi_i(h_i), \quad (5.17)$$

we replace  $\Psi_i(h_i)$  by  $\Psi_i^1(h_i)$ , and expand the logarithm

$$\ln \left[ 1 + \lambda_i^{(3)} \frac{h_i(h_i^2 - 3)}{6} \right].$$

Then the quantity to be maximized is, up to a constant,

$$\mathcal{E} = \ln |\mathbf{J}| + \frac{1}{6} \sum_{i=1}^N [\lambda_i^{(3)}]^2. \quad (5.18)$$

Since optimization has to be done under the constraints in equation 5.16, we add Lagrange multipliers:

$$\mathcal{E}(\rho) = \mathcal{E} - \sum_{i=1}^N \frac{1}{2} \rho_i (\langle h_i^2 \rangle_c - 1). \quad (5.19)$$

Taking into account equation 5.15, one then obtains the updating equation for a given synaptic efficacy  $J_{ij}$ :

$$\begin{aligned} \Delta J_{ij} &\propto -\frac{d\mathcal{E}(\rho)}{dJ_{ij}} \\ \frac{d\mathcal{E}(\rho)}{dJ_{ij}} &= -J_{ij}^{-1} - \langle h_i^3 \rangle_c \langle (h_i^2 - 1) S_j \rangle + \rho_i \langle h_i S_j \rangle. \end{aligned} \quad (5.20)$$

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

$$\delta_{i\bar{i}} = \rho_i \langle h_i h_{\bar{i}} \rangle_c - \langle h_i^3 \rangle_c \langle h_{\bar{i}}^2 h_{\bar{i}} \rangle_c \quad (5.21)$$

together with  $\langle h_i^2 \rangle_c = 1$  for every  $i$ . The parameters  $\rho_i$  are obtained by writing the fixed-point equation (5.21) at  $i = \bar{i}$ , that is,

$$\rho_i = 1 + \langle h_i^3 \rangle_c^2. \quad (5.22)$$

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

It follows from the result (see equation 3.1) of section 3 that the exact, desired solutions are particular solutions of the fixed-point equation (5.21), 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 equation 5.21 is satisfied with nondiagonal matrices  $\langle h_i h_{\bar{i}} \rangle_c$  and  $\langle h_i^2 h_{\bar{i}} \rangle_c$ .

**Remark.** One may wonder what happens if one first performs whitening, computing the  $\mathbf{h}^0$ , and then uses the mutual information between  $\mathbf{h}^0$  and  $\mathbf{h}$ . This is what is studied in Comon (1994), where at lowest order, the cost function (see equation 5.18) is found to be the sum of the square of the third cumulants. This can be readily seen from equation 5.18, where  $\mathbf{J}$  is now the orthogonal matrix that takes  $\mathbf{h}^0$  to  $\mathbf{h}$ , and thus  $\ln |\mathbf{J}|$  is a constant.

**5.5 Link with the BCM Theory of Synaptic Plasticity.** Let us now consider a possible stochastic implementation of the gradient descent (see equation 5.20). 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 instantaneous activities only. Still, by removing one of the averages in equation 5.20 one gets the following updating rule:

$$\Delta J_{ij} = \epsilon \{ J_{ij}^{-1} - \rho_i h_i S_j + \langle h_i^3 \rangle_c h_i^2 S_j \}, \quad (5.23)$$

where  $\epsilon$  is a parameter controlling the rate of variation of the synaptic efficacies. The parameters  $\rho_i$  can be taken at each time according to the fixed-point

equation (5.22). It is quite interesting to compare the updating equation (5.23) with the BCM theory of synaptic plasticity (Bienenstock et al., 1982). 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 nonlinear variant of the Hebbian covariance rule. One of its possible implementations reads, in our notation:

$$\Delta J_{ij} = \epsilon \gamma_i \{-h_i S_j + \Theta_i h_i^2 S_j\}, \quad (5.24)$$

where  $\gamma_i$  and  $\Theta_i$  are parameters possibly depending on the current statistics of the cell activities. The particular choices  $\Theta_i = \langle h_i^2 \rangle$ ,  $\gamma_i = 1$  or  $\Theta_i^{-1}$  have been studied with some detail (Intrator & Cooper, 1992; Law & Cooper, 1994). The two main features of the BCM rule are: (1) there is a synaptic increase or decrease depending on the postsynaptic activity relative to some threshold  $\Theta_i$ , which itself varies according to the cell mean activity level; (2) at low activity, there is no synaptic modification. Since  $\rho_i$  is positive, the rule we derived above is quite similar to equation 5.24, with a threshold  $\Theta_i$  equal to

$$\frac{\langle h_i^3 \rangle_c}{\rho_i}.$$

The main difference is in the constant (that is, activity-independent) term  $J_{ij}^{T-1}$ . This term plays a crucial role: it couples the  $N$  neural cells. Note in fact that 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 (Scofield & Cooper, 1985). Note also that in our formulas we have always assumed zero mean activity ( $\langle S_j \rangle = 0$ , hence also  $\langle h_i \rangle = 0$ ). If this were not the case, the corresponding averages have to be subtracted ( $S_j \rightarrow S_j - \langle S_j \rangle$  for every  $j$ ,  $h_i \rightarrow h_i - \langle h_i \rangle$  for every  $i$ ).

Finally we note that if the third cumulants are zero, one has to make the expansion up to the 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 equation 5.23, that is with the same qualitative behavior as the BCM rule.

## 6 Adaptive Algorithms from Cost Functions Based on Cumulants \_\_\_\_\_

**6.1 A Gradient Descent Based on Theorem 1.** Among the algorithms using correlations at equal time, only algebraic solutions (discussed in section 4) and the recently proposed deflation algorithm (Delfosse & Loubaton, 1995), 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, like the Herault-Jutten algorithm, and gradient methods based on some cost

function (most often a combination of cumulants), may have unwanted fixed points (see Comon, 1994; Delfosse & Loubaton, 1995).

Whatever the algorithm one is working with, the conditions derived in section 3 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 taking the sum of the square of every cumulant that 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. Of course, many other families of cumulants could be used for the same purpose. The possible interest of the one we are dealing with is that it involves a small number of terms. However, this does not imply a priori any particular advantage as far as efficiency is concerned.

For illustrative purposes, we consider with more detail a gradient descent for a particular choice of cost based on Theorem 1 in section 3. Specifically, we ask for the diagonalization of the two-point correlation and the third-order cumulants  $\langle h_i h_j^2 \rangle_c$ . Here again we use the reduction to the search for an orthogonal transformation, as explained in section 2. We thus consider the optimization of the orthogonal matrix. The cost is then defined by

$$\mathcal{E} = \frac{1}{2} \sum_{i \neq j} \langle h_i h_j^2 \rangle_c^2 - \frac{1}{2} \text{Tr}[\rho(\mathcal{O}\mathcal{O}^T - \mathbf{1}_N)] \quad (6.1)$$

where  $\rho$  is a symmetric matrix of Lagrange multipliers, and  $h_i$  has to be written in term of  $\mathcal{O}$  (see equation 2.15):

$$h_i = \sum_{\alpha=1}^N \mathcal{O}_{i,\alpha} h_\alpha^0, \quad (6.2)$$

the  $h_i^0$  being the projections of the inputs onto the principal components, as given by equation 2.13.

The simplest gradient descent scheme is given by

$$\frac{d\mathcal{O}_{i,\alpha}}{dt} = -\varepsilon \frac{d\mathcal{E}}{d\mathcal{O}_{i,\alpha}}, \quad (6.3)$$

where  $\varepsilon$  is some small parameter. From equation 6.1, one derives the derivative of  $\mathcal{E}$  with respect to  $\mathcal{O}_{i,\alpha}$ :

$$\begin{aligned} \frac{d\mathcal{E}}{d\mathcal{O}_{i,\alpha}} = & \sum_{j(\neq i)} \left[ \langle h_i h_j^2 \rangle_c \langle h_\alpha^0 h_j^2 \rangle_c + 2 \langle h_j h_j^2 \rangle_c \langle h_j h_i h_\alpha^0 \rangle_c \right] \\ & - \sum_j \rho_{i,j} \mathcal{O}_{j,\alpha}. \end{aligned} \quad (6.4)$$

One can either adapt  $\rho$  according to

$$\frac{d\rho}{dt} = -\varepsilon \frac{d\mathcal{E}}{d\rho},$$

or choose  $\rho$  imposing at each time  $\mathcal{O}\mathcal{O}^T = \mathbf{1}_N$ , which we do here. The equation for  $\rho$  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:

$$\mathcal{O} d\mathcal{O}^T + d\mathcal{O} \mathcal{O}^T = 0. \quad (6.5)$$

Paying attention to the fact that  $\rho$  is symmetric, one gets

$$\begin{aligned} \rho_{i,i} = & \frac{1}{2} \sum_{k(\neq i)} \left[ \langle \mathbf{h}_i \mathbf{h}_k^2 \rangle_c \langle \mathbf{h}_i \mathbf{h}_k^2 \rangle_c + 2 \langle \mathbf{h}_k \mathbf{h}_i^2 \rangle_c \langle \mathbf{h}_k \mathbf{h}_i \mathbf{h}_i \rangle_c \right] \\ & + \frac{1}{2} \sum_{k(\neq i)} \left[ \langle \mathbf{h}_i \mathbf{h}_k^2 \rangle_c \langle \mathbf{h}_i \mathbf{h}_k^2 \rangle_c + 2 \langle \mathbf{h}_k \mathbf{h}_i^2 \rangle_c \langle \mathbf{h}_k \mathbf{h}_i \mathbf{h}_i \rangle_c \right]. \end{aligned} \quad (6.6)$$

Replacing in equation 6.4  $\rho$  by its expression (6.6), multiplying both sides of equation 6.4 by  $\mathcal{O}_{i,\alpha}$  for some  $i$  and summing over  $\alpha$ , and using equation 6.2, one gets the rather simple equations for the projections of the variations of  $\mathcal{O}$  onto the  $N$  vectors  $\mathcal{O}_i$ :

$$\sum_{\alpha} \mathcal{O}_{i,\alpha} \frac{d\mathcal{O}_{i,\alpha}}{dt} = -\varepsilon \sum_{\alpha} \mathcal{O}_{i,\alpha} \frac{d\mathcal{E}}{d\mathcal{O}_{i,\alpha}} \equiv -\varepsilon \eta_{i,i}, \quad (6.7)$$

where:

$$\begin{aligned} \eta_{i,i} = & \frac{3}{2} \left[ \langle \mathbf{h}_i^3 \rangle_c \langle \mathbf{h}_i \mathbf{h}_i^2 \rangle_c - \langle \mathbf{h}_i^3 \rangle_c \langle \mathbf{h}_i \mathbf{h}_i^2 \rangle_c \right] \\ & + \sum_k \langle \mathbf{h}_k \mathbf{h}_i \mathbf{h}_i \rangle_c \left[ \langle \mathbf{h}_k \mathbf{h}_i^2 \rangle_c - \langle \mathbf{h}_k \mathbf{h}_i^2 \rangle_c \right]. \end{aligned} \quad (6.8)$$

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  $[\mathcal{O}\Lambda_0^{-\frac{1}{2}}\mathcal{O}^0]_{i,j}$  and summing over  $i$ ).

Note that the Lagrange multiplier  $\rho$  ensures that, starting from an arbitrary orthogonal matrix  $\mathcal{O}(0)$  at time 0,  $\mathcal{O}(t)$  remains orthogonal. In practice, since this orthogonality is enforced only at first order in  $\varepsilon$ , an explicit normalization will have to be done from time to time. This is an efficient method used in statistical mechanics and field theory (Aldazabal, Gonzalez-Arroyo, & Parga, 1985).

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  $\langle \cdot \rangle$  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.

**6.2 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 (Jutten & Herault, 1991). One can in fact perform any given linear processing with either one or the other architecture; it is only the algorithmic implementation that 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 Jutten and Herault (1991). 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

$$\frac{du_i}{dt} = - \sum_j \mathcal{L}_{i,j} u_j + S_i. \quad (6.9)$$

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

$$\mathcal{L} = \mathbf{J}^{-1}. \quad (6.10)$$

In the particular case considered above, one gets the updating rule for  $\mathcal{L}_{j,i} = \mathbf{J}_{j,i}^{-1}$  by multiplying equation 6.7 by  $[\mathcal{O} \Lambda_0^{\frac{1}{2}} \mathcal{O}^0]_{i,j}$  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  $h_\alpha^0$  as input to the next layer), then the updating rule for the feedforward and the lateral connections is essentially the same. Indeed, the feedforward coupling matrix that allows going 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,  $\mathcal{O}^T$ .

## 7 Possible Extensions to Nonlinear Processing

---

Some work has already proposed redundancy reduction criteria for nonlinear data processing (Nadal & Parga, 1994; Haft, Schlang, & Deco, 1995; Parra, 1996), and for defining unsupervised algorithms in the context of automatic data clustering (Hinton, Dayan, Frey, & Neal, 1995). Here we just point out that all the criteria and cost functions discussed in this article may be applied to the *output* layer of a multilayer network for performing independent component analysis on nonlinear data. Indeed, if a multilayer network is able to perform ICA, this implies that in the layer preceding 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 nonlinear) transformation of the data leading to a set of independent features, we suggest two lines of research. The first 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) (Cybenko, 1989; Hornik, Stinchcombe, & White, 1991; Barron, 1993). Then there exists a network with one hidden layer and  $N$  outputs such that the  $i$ th output unit gives an approximation of the particular function that extracts the  $i$ th independent component from the data. Hence, we know that it should be enough to take a network with one hidden layer. A possible approach is to perform gradient descent onto a cost function defined for the output layer, which, if minimized, means that separation has been achieved (we know that the redundancy will do). If the algorithm does not give good results, then one may increase the number of hidden units.

Another approach is suggested by the study of the infomax-redundancy reduction criteria (Nadal & Parga, 1994). It is easy to see that the cost function (see equation 5.11) 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  $k$ th layer, and  $c_{ki}(h_{ki})$  the derivative of the transfer function (or the marginal distribution; see section 5.2) of the  $i$ th neuron in the  $k$ th layer, the mutual information between the input and the output of the multilayer network can be written as

$$\mathcal{E}_L = \sum_k \ln |\mathbf{J}_k| + \sum_k \sum_i (\log c_{ki}(h_{ki})). \quad (7.1)$$

Hence the cost  $\mathcal{E}_L$  is a sum of terms, each tending to impose factorization 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 factorization is not obtained,

one can add a second layer, and so on (note, however, that the couplings of a given layer have to be readapted each time a new layer is added).

## 8 Conclusion

---

In this article, 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 that, if fulfilled, guarantee that separation has been performed. These conditions are on a limited set of cross-cumulants and can be used 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.

We then 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 development of the early visual system (Bienenstock et al., 1982). We also showed how some cost functions could be conveniently used for nonlinear processing, that is, for, say, a multilayer network.

In all cases, we paid 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 analyzing 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 (Laheld & Cardoso, 1994).

## Appendix A: Proof of Theorem 1

---

Let us consider a matrix  $\mathbf{J}$  for which equation 3.1 is true. First we use the fact that  $\mathbf{J}$  diagonalize the two-point correlation. Hence, with the notation and results of section 2, we have to determine the family of orthogonal matrices  $\mathbf{X}$  such that, when

$$\mathbf{h} = \mathbf{X}\mathbf{K}^{0-1/2}\boldsymbol{\sigma} \quad (\text{A.1})$$

the  $k$ -order cumulants in equation 3.1 are zero. Using the above expression of  $\mathbf{h}$ , we have for any  $k$ -order cumulant in equation 3.1,

$$\langle h_i^{(k-1)} h_r \rangle_c = \sum_{a=1}^N (X_{i,a})^{(k-1)} X_{r,a} \zeta_k^a, \quad (\text{A.2})$$

where the  $\zeta_k^a$  are the normalized  $k$ -cumulants

$$\zeta_k^a = \frac{\langle \sigma_a^k \rangle_c}{\langle \sigma_a^2 \rangle_c^{k/2}}, \quad a = 1, \dots, N. \tag{A.3}$$

**A.1 The Case of  $N$  Nonzero Source Cumulants.** We first consider the case when for every  $a$ ,  $\zeta_k^a$  is not zero. Since we want the  $k$ -order cumulants to be zero whenever  $i \neq i'$ , we can write

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

for some yet indeterminate constants  $\Delta_i$ . Using the fact that  $\mathbf{X}$  is orthogonal, we multiply both sides of this equation by  $X_{i',a} = X_{a,i'}^T$  for some  $a$  and sum over  $i'$ . This gives

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

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

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

where  $\varepsilon_{i,a}$  is 1 or 0. For  $k$  odd, one then has

$$X_{i,a} = \varepsilon_{i,a} \left[ \frac{\Delta_i}{\zeta_k^a} \right]^{\frac{1}{k-2}}. \tag{A.7}$$

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  $i$   $\Delta_i$  is nonzero. Second, for every pair  $i \neq i'$ ,  $\sum_a X_{i,a} X_{i',a} = 0$ . Then, from equation A.7, we have

$$\sum_a \varepsilon_{i,a} \varepsilon_{i',a} (\zeta_k^a)^{\frac{2}{k-2}} = 0. \tag{A.8}$$

The left-hand side is a sum of positive terms; hence each 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  $\pm 1$ :  $\mathbf{X}$  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 nonzero for every source.

**Remark.** For  $k$  even, there is a sign indetermination when going from  $X_{i,a}^{(k-2)}$  to  $X_{i,a}$ . Hence one cannot write that  $\sum_a X_{i,a} X_{i,a}$  is a sum of positive numbers. In fact, taking  $k = 4$  and  $N$  even, one can easily build an example where the conditions in equation 3.1 are fulfilled with at least one solution  $\mathbf{X}$  that is not a sign permutation. For instance, let  $N = 4$  and  $\zeta_4^a = z$  for every  $a$ . Then the equations 3.1 are fulfilled for  $\mathbf{X}$  defined by:

$$\mathbf{X} = \frac{1}{2} \begin{pmatrix} 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \quad (\text{A.9})$$

**A.2 The Case of  $L < N$  Nonzero Source Cumulants.** 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 ( $a = 1, \dots, L$ ). We have to reconsider the preceding argument. It started with equation A.4 in which the right-hand side can be considered as a sum over  $a = 1, \dots, L$ . It follows that a particular family of solutions is the set of block-diagonal matrices  $\mathbf{X}$ , with an  $L \times L$  block being a sign permutation matrix, followed by an  $(N-L) \times (N-L)$  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 the following possibilities for each  $i$ :

1.  $\Delta_i = 0$  and  $X_{i,a} = 0$ ,  $a = 1, \dots, L$ .
2.  $\Delta_i \neq 0$ ,  $X_{i,a} = 0$ ,  $a > L$ , and for  $a = 1, \dots, L$  equation A.6 is valid with at least one  $\varepsilon_{i,a}$  nonzero.

By applying an appropriate permutation, we can assume that it is the first  $\ell$  indices,  $i = 1, \dots, \ell$ , for which  $\Delta_i \neq 0$ . Hence  $\mathbf{X}$  has a nonzero upper-left  $\ell \times L$  block,  $\mathbf{X}^1$ , which satisfies all the equations derived previously (in particular,  $\mathbf{X}^1 \mathbf{X}^{1T} = \mathbf{1}_N$ ), but with  $i = 1, \dots, \ell$  and  $a = 1, \dots, L$ ; and a nonzero 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 nonzero cumulants that  $\mathbf{X}^1$  has a single nonzero 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  $1 \times 1$  matrix, whose unique element is thus  $\pm 1$ . Hence, in that case, it is in fact all the  $N$  sources that are separated.

**Appendix B: Proof of Theorem 2**

In the first part of the proof, we proceed exactly as in appendix A. We reduce the problem to the search of an orthogonal matrix  $\mathbf{X}$  such that, when

$$\mathbf{h} = \mathbf{X}\mathbf{K}^{-1/2}\boldsymbol{\sigma} \quad (\text{B.1})$$

the  $k$ -order cumulants in equation 3.2 are zero. Using the above expression of  $\mathbf{h}$ , we have for any  $k$ -order cumulant in equation 3.2,

$$\langle h_i^{(k-m)} h_{i'}^{m-1} h_{i''} \rangle_c = \sum_{a=1}^N (X_{i,a})^{(k-m)} X_{i',a}^{m-1} X_{i'',a} \zeta_k^a \quad (\text{B.2})$$

where the  $\zeta_k^a$  are the normalized  $k$ -cumulants as in equation A.3. Now we write that these quantities are zero whenever at least two of the indices  $i, i', i''$  are different:

$$\Delta_i \delta_{i,i'} \delta_{i,i''} = \sum_{a=1}^N (X_{i,a})^{(k-m)} X_{i',a}^{m-1} X_{i'',a} \zeta_k^a \quad (\text{B.3})$$

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

$$\Delta_i X_{i',a} \delta_{i,i'} = (X_{i,a})^{(k-m)} (X_{i',a})^{(m-1)} \zeta_k^a. \quad (\text{B.4})$$

Now, either  $X_{i',a} = 0$ , or  $X_{i',a} \neq 0$  and then

$$\Delta_i \delta_{i,i'} = (X_{i,a})^{(k-m)} (X_{i',a})^{(m-2)} \zeta_k^a. \quad (\text{B.5})$$

At this point the proof differs from the one of Theorem 1 and is in fact simpler. For  $i \neq i'$  the right-hand side of equation B.5 is 0. Because  $X_{i',a} \neq 0$  and  $k$  is strictly greater than  $m$ , we obtain  $X_{i,a} = 0$ . Hence, for every  $a$  such that  $\zeta_k^a \neq 0$ , there is at most one  $i$  for which  $X_{i,a} \neq 0$ . Since  $\mathbf{X}$  is orthogonal, this implies that its restriction to the subspace of nonzero  $\zeta_k^a$  is a sign permutation.

**Acknowledgments**

This work was partly supported by the French-Spanish program Picasso, the E.U. grant CHRX-CT92-0063, the Universidad Aut3noma de Madrid, the Universit3 Paris VI, and Ecole Normale Sup3rieure. NP and JPN thank, respectively, the Laboratoire de Physique Statistique (ENS) and the Departamento de F3sica Te3rica (UAM) for their hospitality. We thank B. Shraiman for communicating his results prior to publication and P. Del Giudice and A. Campa for useful discussions. We thank an anonymous referee for useful comments, which led us to elaborate on the joint diagonalization method.

## References

---

- Abramowitz, M., & Stegun, I. A. (1972). *Handbook of mathematical functions*. New York: Dover.
- Aldazabal, G., Gonzalez-Arroyo, A., & Parga, N. (1985). The stochastic quantization of  $U(N)$  and  $SU(N)$  lattice gauge theory and Langevin equations for the Wilson loops. *Journal of Physics*, *A18*, 2975.
- Amari, S. I., Cichocki, A., & Yang, H. H. (1996). A new learning algorithm for blind signal separation. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems 8*. Cambridge, MA: MIT Press.
- Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *NETWORK*, *3*, 213–251.
- Attneave, F. (1954). Informational aspects of visual perception. *Psychological Review*, *61*, 183–193.
- Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. Rosenblith (ed.), *Sensory communication* (p. 217). Cambridge, MA: MIT Press.
- Barlow, H. B., Kaushal, T. P., & Mitchison, G. J. (1989). Finding minimum entropy codes. *Neural Comp.*, *1*, 412–423.
- Bar-Ness, Y. (1982, November). Bootstrapping adaptive interference cancelers: Some practical limitations. In *The Globecom Conf.* (pp. 1251–1255). Paper F3.7, Miami, Nov. 1982.
- Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. I. T.*, *39*.
- Bell, A., & Sejnowski, T. (1995). An information-maximisation approach to blind separation and blind deconvolution. *Neural Comp.*, *7*, 1129–1159.
- Belouchrani, A., & M. (1993). Séparation aveugle au second ordre de sources corrélées. In *GRETSI'93*, Juan-Les-Pins, 1993.
- Bienenstock, E., Cooper, L., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal Neurosciences*, *2*, 32–48.
- Blahut, R. E. (1988). *Principles and practice of information theory*. Reading, MA: Addison-Wesley.
- Burel, G. (1992). Blind separation of sources: A nonlinear neural algorithm. *Neural Networks*, *5*, 937–947.
- Cardoso, J.-F. (1989) Source separation using higher-order moments. In *Proc. Internat. Conf. Acoust. Speech Signal Process.-89* (pp. 2109–2112). Glasgow.
- Cardoso, J.-F., & Souloumiac, A. (1993). Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, *140*(6), 362–370.
- Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, *36*, 287–314.
- Cybenko, G. (1989). Approximations by superpositions of a sigmoidal function. *Math. Contr. Signals Syst.*, *2*, 303–314.
- Delfosse, N., & Loubaton, Ph. (1995). Adaptive blind separation of independent sources: A deflation approach. *Signal Processing*, *45*, 59–83.

- Del Giudice, P., Campa, A., Parga, N., & Nadal, J.-P. (1995). Maximization of mutual information in a linear noisy network: A detailed study. *NETWORK*, 6, 449–468.
- Déville, Y., & Andry, L. (1995). Application of blind source separation techniques to multi-tag contactless identification systems. *Proceedings of NOLTA'95* (pp. 73-78). Las Vegas.
- Féty, L. (1988). *Méthodes de traitement d'antenne adaptée aux radio-communications*. Unpublished doctoral dissertation, ENST, Paris.
- Gaeta, M., & Lacoume, J. L. (1990). Source separation without apriori knowledge: The maximum likelihood approach. In L. Tores, E. MasGrau, & M. A. Lagunas (eds.), *Signal Processing V, proceedings of EUSIPCO 90* (pp. 621-624).
- Haft, M., Schlang, M., & Deco, G. (1995). Information theory and local learning rules in a self-organizing network of ising spins. *Phys. Rev. E*, 52, 2860–2871.
- Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The wake-sleep algorithm for unsupervised neural networks. *Science*, 268, 1158–1160.
- Hopfield, J. J. (1991). Olfactory computation and object perception. *Proc. Natl. Acad. Sci. USA*, 88, 6462–6466.
- Hornik, K., Stinchcombe, M., & White, H. (1991). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2, 359–366.
- Intrator, N., & Cooper, L. (1992). Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5, 3–17.
- Jutten, C., & Herault, J. (1991). Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24, 1–10.
- Laheld, B., & Cardoso, J.-F. (1994). Adaptive source separation without prewhitening. In *Proc. EUSIPCO* (pp. 183–186). Edinburgh.
- Law and Cooper, L. (1994). Formation of receptive fields in realistic visual environments according to the BCM theory. *Proc. Natl. Acad. Sci. USA*, 91, 7797–7801.
- Li, Z., & Atick, J. J. (1994). Efficient stereo coding in the multiscale representation. *Network: Computation in Neural Systems*, 5, 1–18.
- Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, 21, 105–117.
- Molgedey, L., & Schuster, H. G. (1994). Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72, 3634–3637.
- Nadal, J.-P., & Parga, N. (1994). Nonlinear neurons in the low-noise limit: A factorial code maximizes information transfer. *NETWORK*, 5, 565–581.
- Parra, L. C. (1996). Symplectic nonlinear component analysis. *Preprint*.
- Pham, D.-T., Garrat, Ph. & Jutten, Ch. (1992). Separation of a mixture of independent sources through a maximum likelihood approach. In *Proc. EUSIPCO* (pp. 771–774).
- Redlich, A. N. (1993). Redundancy reduction as a strategy for unsupervised learning. *Neural Comp.*, 5, 289–304.
- Rospars, J.-P., & Fort, J.-C. (1994). Coding of odor quality: Roles of convergence and inhibition. *NETWORK*, 5, 121–145.
- Scofield, C. L., & Cooper, L. (1985). Development and properties of neural networks. *Contemporary Physics*, 26, 125–145.

- Shraiman, B. (1993). *Technical memorandum* (AT&T Bell Labs TM-11111-930811-36).
- Tong, L., Soo, V., Liu, R., & Huang, Y. (1990). Amuse: A new blind identification algorithm. In *Proc. ISCAS*. New Orleans.
- van Hateren, J. H. (1992). Theoretical predictions of spatiotemporal receptive fields of fly lmc8, and experimental validation. *J. Comp. Physiol. A*, *171*, 157–170.

---

Received June 26, 1996; accepted February 19, 1997.