I.C.A.: conditions on cumulants and information theoretic approach

Jean-Pierre Nadal
Laboratoire de Physique Statistique de l'E.N.S.,* E.N.S.,
24, rue Lhomond, F-75231 Paris Cedex 05, France
and
Néstor Parga
Departamento de Física Teórica, U.A.M.,
Cantoblanco, 28049 Madrid, Spain

Abstract. In this contribution we present new conditions on cross cumulants which guarantee that blind source separation has been performed, and we relate these conditions to the maximization of mutual information as a criterion for ICA.

Submitted to ESANN97

1. Introduction

Independent Component Analysis (ICA), and in particular Blind Source Separation (BSS), can be obtained from the maximization of mutual information, as first shown in [Nadal and Parga, 1994][1]. This result was obtained for a deterministic processing system, with an arbitrary input-output relationship. Technically, a small additive noise was considered in order to define the mutual information between the inputs and the outputs variable. Then the zero noise limit was taken in order to extract the relevant "contrast" cost function for the deterministic case. The relevance for BSS was stressed out: in the particular case where the inputs are linear combinations of independent random variables ("sources"), one can use a feedforward network (with no hidden layer), and nonlinear transfer functions; then the outputs of the system will give the independent components if both the weigths and the transfer functions are adapted in such a way that mutual information is maximized.

The practical interest of this information theoretic based cost function was then demonstrated by [Bell and Sejnowski, 1995 and 1996][2, 3] in several BSS applications. Since then, it has also been realized [4, 5] that, for this BSS case, the cost function in the form written in [2] is in fact identical to the one derived several years before from a maximum likelihood approach [6].

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

In the present contribution, after a short reminder of [1], we report on recent results based on such information theoretic approach, and on new conditions on cross cumulants which guarantee that BSS has been performed.

2. Expansion of the mutual information

2.1. Link between infomax, loglikelihood, and ICA

Our starting point is the main result obtained in [1], 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 non linear mixtures. We consider a network with N inputs and p outputs, and *nonlinear transfer functions* $f_i, i = 1, ..., p$. Hence the output $\mathbf{V}$ is given by a gain control after some (linear or non linear) processing:

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

where the h_i's are arbitrary deterministic functions of the inputs $\mathbf{S}$, $h_i(t) = h_i[\mathbf{S}](t)$. For a noiseless channel, one gets that maximizing the mutual information is equivalent to maximizing the (differential) output entropy $H(Q)$ of the output distribution $Q = Q(\mathbf{V})$,

$$H(Q) = -\int d^pV\ Q(\mathbf{V})\ \log Q(\mathbf{V}), \tag{2}$$

Making the change of variable $\mathbf{V} \rightarrow \mathbf{h}$, one gets

$$H(Q) = -\int d\mathbf{h}\Psi(\mathbf{h})\ \ln \frac{\Psi(\mathbf{h})}{\prod_{i=1}^{p} f_i'(h_i)} \tag{3}$$

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

$$\Psi(\mathbf{h}) = \prod_{i=1}^{p}\ \Psi_i(h_i), \tag{4}$$

and at the same time for each output neuron the transfer function f_i has its derivative equal to the corresponding marginal probability distribution: $f_i'(h_i) = \Psi_i(h_i), \quad i = 1, ..., p$. As a result, infomax implies redundancy reduction: the optimal neural representation is a factorial code. In particular, this result applies for a linear mixture of independent components, in which case h_i has to be linear in the inputs:

$$h_i(t) = \sum_{j=1}^{N} J_{i,j} S_j(t),\ i = 1, ..., p \tag{5}$$

Recently, we showed that this result extends to stochastic outputs [?].

It is convenient to rewrite the output entropy, making in (3) the change of variable $\mathbf{h} \to \mathbf{S}$. Since the input entropy is a constant the quantity which has to be maximized is

$$\mathcal{E} = < \ln \mathcal{J} > \; + \sum_i < \log \Psi_i(h_i) > \tag{6}$$

where $< . >$ is the average over the output activity h_i, and $\mathcal{J}$ is the Jacobian of the transformation $\mathbf{S} \to \mathbf{h}$. For a linear rule as in (5), this is the change of variable done by Bell and Sejnowski [2]. In this case the Jacobian $\mathcal{J}$ is just $\mid \mathbf{J} \mid$, the absolute value of the determinant of the coupling matrix $\mathbf{J}$, and one has

$$\mathcal{E} = \ln \mid \mathbf{J} \mid \; + \sum_i < \log \Psi_i(h_i) > \tag{7}$$

In fact, this cost (7) was first derived in a maximum likelihood approach [6]: it is easy to see that (7) is equal to the (average of) the loglikelihood of the observed data (the inputs $\mathbf{S}$), given that they have been generated as a linear combination of independent sources with the Ψ_i as marginal distributions.

The cost (6) can be conveniently used for nonlinear ICA. If one uses a multilayer feedforward network, from the chain rule for derivatives the term $< \ln \mathcal{J} >$ takes the simple form of a sum of terms, one for each layer [4].

¿From now on we restrict to the case of the architecture needed for BSS, that is with linear h_i's as in (5).

2.2. Close to Gaussian expansion

To get some insight on BSS based on infomax, we consider the maximisation of the output entropy making a close-to-Gaussian cumulant expansion of $\Psi_i(h_i)$ [**?**, 4]. At first non trivial order, it is given by

$$\Psi_i(h_i) \approx \Psi_i^1(h_i) \equiv \; \Psi^0(h_i)\, [1 + \lambda_i^{(3)} \; \frac{h_i(h_i^2 - 3)}{6}], \tag{8}$$

where $\Psi^0(h_i)$ is the normal distribution $\Psi^0(h_i) \equiv \frac{1}{\sqrt{2\pi}} \exp(-\frac{h_i^2}{2})$, and $\lambda_i^{(3)}$ is the third (true) cumulant of h_i:

$$\lambda_i^{(3)} \equiv < h_i^3 >_c . \tag{9}$$

In the above expression (8) we have taken into account that one can always take

$$< h_i > = 0 \quad \text{and} \quad < h_i^2 >_c = 1. \tag{10}$$

In the cost function (7) we replace $\Psi_i(h_i)$ by $\Psi_i^1(h_i)$, and expand at first order in λ. One then gets that the quantity to be maximized is, up to a constant,

$$\mathcal{E} = \ln \mid \mathbf{J} \mid \; + \frac{1}{6} \sum_{i=1}^{N} [\lambda_i^{(3)}]^2 - \sum_{i=1}^{N} \frac{1}{2} \rho_i (< h_i^2 >_c - 1) \tag{11}$$

where ρ_i are Lagrange multiplyers introduced in order to enforce (10). One then derives the updating equation for a given synaptic efficacy J_{ij} as $\Delta J_{ij} \propto -\frac{d\mathcal{E}(\rho)}{dJ_{ij}}$. Considering the fixed point equations, that is $\Delta J_{ij} = 0$, one obtains[4] that any $\mathbf{J}$ solution of these equations satisifies

$$\delta_{ii'} \;=\; \rho_i \; < h_i h_{i'} >_c \; - \; < h_i^3 >_c < h_i^2 h_{i'} >_c \tag{12}$$

together with $< h_i^2 >_c = 1$ for every i. The parameters ρ_i are obtained by writing the fix point equation (12) at $i = i'$, that is $\rho_i \;=\; 1 + \; < h_i^3 >_c^2$. Note that, in particular, $\rho_i > 0$ for all i.

One can then ask whether any solution of (12) is a proper solution of the BSS problem, or whether one can have additional solutions of the fixed point equations which do not extract the independent components. Since an approximation has been made in order to derive these equations, the presence of such spurious solutions would not be a surprise. However, one may ask what happens if every term in (12) is set to zero, that is if

$$< h_i h_{i'} >_c = 0 \;\text{ and }\; < h_i^2 h_{i'} >_c = 0 \;\text{ for every } i \neq i'. \tag{13}$$

In the next section, we will see that these conditions (13) does imply that source separation has been obtained, since they appear as a particular case of a more general theorem.

3. BSS from new conditions on cross-cumulants

We thus assume that the data are a linear superposition of independent sources:

$$S_j(t) = \sum_{a=1}^{N} M_{j,a} \sigma_a(t), \; j = 1, ..., N \tag{14}$$

(in vector form $\mathbf{S} = \mathbf{M}\boldsymbol{\sigma}$) where the σ_a are N independent random variables, of unknown probability distributions, and $\mathbf{M}$ is an unknown, constant, $N \times N$ matrix, called the *mixture matrix*. By hypothesis, all the source cumulants are diagonal, in particular the two point correlation at equal time $\mathbf{K}^0$, $K^0_{a,b} \equiv < \sigma_a(t)\sigma_b(t) >_c = \delta_{a,b} K^0_a$ where $\delta_{a,b}$ is the Kronecker symbol.

In this section we claim that for $\mathbf{J}$ to be a solution of the BSS problem, it is sufficient (and of course necessary) that $\mathbf{J}$ performs whitening and sets altogether to zero a given set of cross-cumulants of some given order k, the number of which being only of order N^2. 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 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',

$$\left\{ \begin{array}{l} < h_i h_{i'} >_c = \delta_{i,i'} \\ < h_i^{(k-1)} h_{i'} >_c = 0 \textit{ for } i \neq i'. \end{array} \right. \tag{15}$$

where $\mathbf{h}$ *is the output vector as defined in (5).*

- *(ii) If only $1 \leq L \leq N-2$ k-order cumulants are nonzero, then any solution* $\mathbf{J}$ *of (15) is the product of a sign-permutation by a matrix which separates the L sources having non zero k- cumulants, and such that the restriction of* $\mathbf{J}\,\mathbf{M}\mathbf{K}^{0\,\frac{1}{2}}$ *to the space of the $N-L$ other sources is still an arbitrary $(N-L) \times (N-L)$ orthogonal matrix.*

The detailed proof will be published elsewhere[4]. Remark: for k even, one can easily find an example showing that the conditions (15) are not sufficient.

An interesting application of this theorem concerns the algorithm of Herault and Jutten [?]. In its simplest version, this algorithm aims at setting to zero the two point correlation and the cross-cumulants $< h_i^2 h_{i'} >_c$ 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 anologous 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 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'',

  $$\left\{ \begin{array}{l} < h_i h_{i'} >_c = \delta_{i,i'} \\ < h_i^{(k-m)} h_{i'}^{m-1} h_{i''} >_c = 0 \textit{ for at least two non identical indices.} \end{array} \right. \tag{16}$$

  where $\mathbf{h}$ *is the output vector as defined in (5).*

- *(ii) If only $1 \leq L \leq N-2$ k-order cumulants are nonzero, then any solution* $\mathbf{J}$ *of (16) is the product of a sign-permutation by a matrix which separates the L sources having non zero k- cumulants, and such that the restriction of* $\mathbf{J}\,\mathbf{M}\mathbf{K}^{0\,\frac{1}{2}}$ *to the space of the $N-L$ other sources is still an arbitrary $(N-L) \times (N-L)$ orthogonal matrix.*

Again the proof is given in [4]. One can show that the case $m = 2$ is somehow related to the joint diagonalization approach of [?]. One should note that Theorem 2 is not exhausted by Theorem 1, as it may seem since the conditions (15) are a subset of the conditions (16). But in Theorem 2 no condition on the parity of k (the order of the cumulants) is required.

4. Conclusion

We have presented recent results obtained for ICA, and in particular BSS. We mentionned that the mutual information, shown to be as a tool for performing ICA [1], can be used for both linear and nonlinear as well as for deterministic or stochastic networks. In the case of BSS (one layer feedforward network), we gave new conditons on cross cumulants[4]. These conditions show that it is sufficient to set to zero a limited number of cross-cumulants of a given order. In addition we showed that some of these conditions appear implicitely in the mutual information criterion.

References

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

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

[3] A. Bell and T. Sejnowski. The 'independent components' of natural scenes are edge filters. *to appear in Vision Research*, 1996.

[4] J.-P. Nadal and N. Parga. Redundancy reduction and independent component analysis: Algebraic and adaptive approaches. *submitted to Neural Computation*, 1996.

[5] J-F Cardoso. Infomax and maximum likelihood for blind separation. *to appear in IEEE Signal Processing Letters*, 1997.

[6] D.-T. Pham, Ph. Garrat, and Ch. Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In *Proc. EUSIPCO*, pages 771–774, 1992.

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

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

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

[10] Y. Dan, J. J. Atick, and R. C. Reid. Efficient coding of natural scenes in the lgn: experimental test of a computational theory. *The Journal of Neuroscience*, pages 3351–3362, 1996.

[11] Stein. The information capacity of nerve cells using a frequency code. *Biophys. Journal*, 1967.

[12] N. Brunel and J.-P. Nadal. Optimal tuning curves in a simple model of spiking neurons. *in preparation*, 1996.

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

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

[15] S. B. Laughlin. A simple coding procedure enhances a neuron's information capacity. *Z. Naturf.*, C 36:910–2, 1981.

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