Neural Codes and Independent Component Analysis: Information Theoretic Approach and Conditions on Cumulants

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

Abstract

In this contribution we review recent results obtained on blind source separation (BSS) and independent component analysis (ICA). In particular we show that maximisation of mutual information can lead to ICA, and we present new conditions on cross cumulants which guarantee that blind source separation has been performed.

TAINN'97, Ankara, may 1997

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 [1]. This result was obtained for a deterministic processing system, with an arbitrary input-output relationship. 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 feed-forward 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 in [2, 3] in several BSS applications. Since then, it has also been realized [5, 7] 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 [8].

Making a close to Gaussian approximation, one finds that the maximisation of mutual information can be reached by setting to zero a limited set of crosscumulants. From this remark one can then establish several conditions on cross-cumulants which, if fulfilled, guarantee that source separation has been obtained[5, 6].

In this contribution, we first give a short reminder of [1, 4], and then we report on these new conditions on cross cumulants for performing BSS[5, 6].

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

2 Link between infomax, loglikelihood, and ICA

We first present 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* $f_i, i = 1, ..., p$. More explicitely, 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 network, one gets that maximizing the mutual information is equivalent to maximizing the (differential) output entropy $H(Q)$ of the output distribution $Q = Q(\mathbf{V})$. Writting $\mathbf{V}$ in term of $\mathbf{h}$, one finds [5] that this entropy can be written as

$$H(Q) = -\int d\mathbf{h}\Psi(\mathbf{h}) \ \ln \frac{\Psi(\mathbf{h})}{\prod_{i=1}^{p} f_i'(h_i)} \tag{2}$$

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

$$\Psi(\mathbf{h}) = \prod_{i=1}^{p} \ \Psi_i(h_i), \tag{3}$$

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

$$f_i'(h_i) = \Psi_i(h_i), \quad i = 1, ..., p. \tag{4}$$

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 [4].

It is convenient to rewrite the output entropy, making in (2) 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 [8]: it is easy to see [5, 7] 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 (see [7] for a more detailed comparison of infomax and maximum likelihood approaches).

The cost (6) can be conveniently used for nonlinear ICA. If one uses a multilayer feed-forward 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 [5].

3 BSS from new conditions on cross-cumulants

From now on we restrict to the case of the architecture needed for BSS, that is with linear h_i's as in (5). We thus assume that the data are a linear superposition of independent sources:

$$S_j(t) = \sum_{a=1}^{N} M_{j,a}\sigma_a(t), \;\; j = 1, ..., N \tag{8}$$

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), if and only if one has:*

  for every i, i',

$$\begin{cases} < h_i h_{i'} >_c = \delta_{i,i'} \\ < h_i^{(k-1)} h_{i'} >_c = 0 \text{ for } i \neq i'. \end{cases} \tag{9}$$

  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 (9) 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 appear elsewhere[5]. Remark: for k even, one can easily find an example showing that the conditions (9) are not sufficient.

An interesting application of this theorem concerns the algorithm of Herault and Jutten [10]. 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. For example, one can take the cumulants $< h_i h_{i'} >_c$ and $<$

$h_i^{(k-m)} h_{i'}^{m-1} h_{i''} >_c$, with m at least equal to 2 and k strictly greater than m (and *no* condition on the parity of k, the order of the cumulants) [5, 6]. One can show that the case $m = 2$ is somehow related to the joint diagonalization approach of [11].

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[5]. These conditions show that it is sufficient to set to zero a limited number of cross-cumulants of a given order.

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, N. Brunel and N. Parga. Nonlinear feedforward networks with stochastic ouputs: infomax implies redundancy reduction. *Submitted to NETWORK.*

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

[6] J.-P. Nadal and N. Parga. I.C.A.: conditions on cumulants and information theoretic approach, *accepted to ESANN'97.*

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

[8] 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.

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

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

[11] J.-F. Cardoso and A. Souloumiac Blind beamforming for non Gaussian signals *IEE Proceedings-F*, 140(6):362–370, 1993.