NEURAL NETWORK IMPLEMENTATIONS OF INDEPENDENT COMPONENT ANALYSIS

Radu Mutihac, Marc M. Van Hulle

K. U. Leuven, Labo voor Neuro- en Psychofysiologie, Campus Ghastuisberg,
Herestraat 49, B-3000 Leuven, Belgium

ABSTRACT

The performance of six neuromorphic adaptive structurally different algorithms was analyzed in blind separation of independent artificially generated signals using the stationary linear independent component analysis (ICA) model. The estimated independent components were ranked and compared among different ICA approaches. All algorithms were run with different contrast functions, which were optimally selected on the basis of maximizing the sum of individual negentropies of the network outputs. Both subgaussian and supergaussian one-dimensional time series were employed throughout the numerical simulations.

INTRODUCTION

In many areas like data analysis, signal processing, and neural networks, a common task is to find an adequate representation of multivariate data for subsequent processing and interpretation. Linear transforms are often invoked due to their computational and conceptual simplicity. ICA has emerged as an extension of a linear transform called *Principal Component Analysis* (PCA), which has been developed in context with *Blind Source Separation* (BSS) in *Digital Signal Processing* (DSP) and array processing [1]. In its full generality, ICA amounts to blind model identification with minimal suppositions.

ICA MODEL

Our stationary linear ICA model considered hereafter (Fig. 1) assumes $\mathbf{x}(t)$, $\mathbf{n}(t) \in \Re^N$, and $\mathbf{s}(t) \in \Re^M$ three random vectors with zero mean and finite covariance, with the components of $\mathbf{s}(t)$ being statistically independent and at most one gaussian, whereas $\mathbf{A}$ is a rectangular constant full column rank $N \times M$ matrix with at least as many rows as columns ($N \geq M$):

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) = \sum_{i=1}^{M} s_i(t)\mathbf{a}_i + \mathbf{n}(t) \tag{1}$$

where the columns $\mathbf{a}_i$, $i = 1,2,...,M$ of the mixing matrix **A** are the basis vectors of ICA. The sample index *t* is assumed to take discrete values $t = 1,2,...,T$. Mixing is supposed to be instantaneous, so there is no time delay between the (latent) source variable $s_i(t)$ mixing into an observable variable $x_j(t)$. Within this framework, the ICA problem can be formulated as follows [2]: given *T* realizations of $\mathbf{x}(t)$, estimate both the matrix **A** and the corresponding realizations of $\mathbf{s}(t)$. In BSS the task is to find the waveforms $\{s_i(t)\}$ of the sources knowing only the mixtures $\{x_j(t)\}$. The noise-free ICA model corresponds to the absence of noise term $\mathbf{n}(t)$.

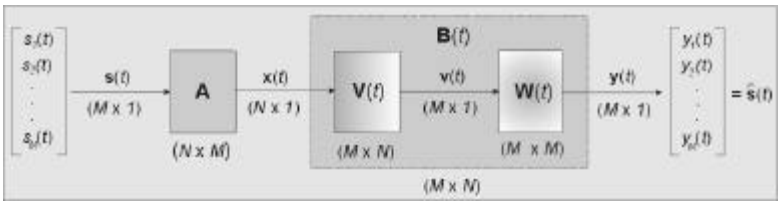

Figure 1: Mixing (**A**) and separating (**B**) source signals (**s**).

If the size of $\mathbf{x}(t)$ is greater than the size of $\mathbf{s}(t)$, that is $M < N$, the problem is over-determined and the extra data can be used for reducing noise. This is accomplished by projecting the input data $\mathbf{x}(t)$ into its *M*-dimensional signal subspace using for example PCA whitening. Contrarily, if the ICA problem is under-determined ($M > N$), then we expect the most energetic independent components still to be separated and the rest to come out as linear combinations in the remained estimates.

Adaptive source separation consists in updating a $M \times N$ separating matrix $\mathbf{B}(t)$, without resorting to any information about the spatial mixing matrix **A**, so that the vector

$$\mathbf{y}(t) = \mathbf{B}(t)\mathbf{x}(t) \tag{2}$$

becomes an estimate $\mathbf{y}(t) = \hat{\mathbf{s}}(t)$ of the original independent source signals $\mathbf{s}(t)$. In neural implementations, $\mathbf{y}(t)$ is the output vector of the network, and the full separating matrix $\mathbf{B}(t)$ is the total weight matrix between the input and the output layers. The estimate $\hat{s}_i(t)$ of the *i*-th source signal may appear in any component $y_j(t)$ of $\mathbf{y}(t)$. The ICA model can be resolved up to the product of a permutation and a diagonal matrix, because without prior information on the amplitude of the source signals nor on the matrix **A,** the scale of each source signal is unobservable. The permutation indeterminacy stems from the immateriality of labeling the source signals.

Since ICA deals with higher-order statistics it is justified to normalize in some sense the first- and second-order moments. The effect is that the separating matrix is divided in two parts dealing with dependencies in the first two moments, e.g.

the *whitening* matrix $\mathbf{V}(t)$, and the dependencies in higher-order statistics, e.g. the *orthogonal separating matrix* $\mathbf{W}(t)$ in the whitened space (Fig. 1). If we assume zero-mean observed data $\mathbf{x}(t)$, then by whitening we get a vector $\mathbf{v}(t) = \mathbf{V}(t)\mathbf{x}(t)$ with decorrelated components. The subsequent linear transform $\mathbf{W}(t)$ seeks the solution by an adequate rotation in the space of component densities and yields $\mathbf{y}(t) = \mathbf{W}(t)\mathbf{v}(t)$, which is the relationship between the whitening and the output layer of the network (Fig. 2). The total *separation matrix* between the input and the output layer becomes $\mathbf{B}(t) = \mathbf{W}(t)\mathbf{V}(t)$.

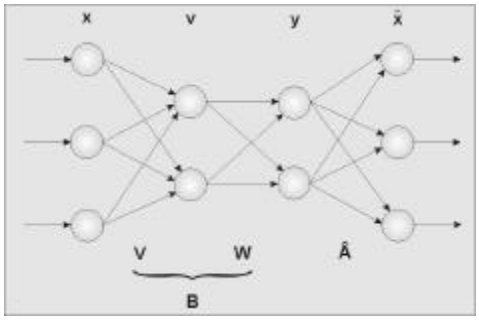

Figure 2: The architecture of a feedforward neural network performing BSS and providing the basis vectors of ICA as columns of the estimated mixing matrix $\hat{A}$.

In the standard stationary case, the whitening and the orthogonal separating matrices converge to some constant values during learning. The same model can nevertheless be used in nonstationary cases by keeping these matrices time-varying. Standard PCA is often used for whitening because information can be optimally compressed in the mean-square error sense and filter possible noise out.

Note that if the source signals are temporally correlated, then the spatial blind deconvolution (separation) may be based on *2*nd-order statistics only [3], and separation of gaussian sources is possible.

STATISTICAL INDEPENDENCE AND ESTIMATION PRINCIPLES FOR ICA

If a multidimensional random variable $\mathbf{x} \in \Re^N$ has the probability density function $f_{\mathbf{x}}(\mathbf{x})$, then the independence of the N scalar random variables x_i, $i = 1,2,...,N$, that is, the components of $\mathbf{x}$, having the probability density functions $f_{x_i}(x_i)$, respectively, is defined by the factorization of the joint density:

$$f_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^{N} f_{x_i}(x_i) \tag{3}$$

A meaningful treatment of the concept of independence relies on information theory, which means deriving the criterion for statistical independence from the statistical properties of data. Entropy is such a criterion based on the amount of information contained in some occurrences of a random variable. In the case of a

multidimensional continuous random variable $\mathbf{x}$ with density $f_{\mathbf{x}}(\mathbf{x})$, the *differential entropy* is defined as:

$$H(\mathbf{x}) = -\int f_{\mathbf{x}}(\mathbf{u}) log\ f_{\mathbf{x}}(\mathbf{u})\, d\mathbf{u} \tag{4}$$

Differential entropy is invariant to orthogonal transforms and it is upper bound, but is no longer invariant to invertible transforms as *entropy* is. Therefore, two other concepts are employed as contrast functions that are endowed with the invariance property, namely *negentropy* and *mutual information*.

For a multidimensional continuous random variable $\mathbf{x}$ with the density $f_{\mathbf{x}}(\mathbf{x})$, to which is associated a gaussian variable $\mathbf{x}_G$ with the same covariance matrix like $\mathbf{x}$, the *negentropy* is defined in terms of differential entropy:

$$J(\mathbf{x}) = H(\mathbf{x}_G) - H(\mathbf{x}) = \int f_{\mathbf{x}}(\mathbf{u}) log \frac{f_{\mathbf{x}}(\mathbf{u})}{f_{\mathbf{x}_G}(\mathbf{u})} d\mathbf{u} = K(\mathbf{x} \mid \mathbf{x}_G) \tag{5}$$

which can be interpreted as the distance from gaussianity expressed in the form of Kullback-Leibler (KL) divergence. Though not really a distance since it is not symmetric, the KL divergence behaves as a statistical measure of "distance" between two distributions. It is always nonnegative and takes the value *0* iff the distributions are identical. Hence negentropy is always nonnegative, reaches its minimum for a gaussian random variable, and it is invariant to linear invertible transforms.

The *mutual information* (MI) is also related with (differential) entropy. For the general case of N (scalar) random variables $x_i, i = 1,2,...,N$ the mutual information is given by:

$$I(x_1, x_2, ..., x_N) = \sum_i H(x_i) - H(\mathbf{x}) = \int f_{\mathbf{x}}(\mathbf{x}) log \frac{f_{\mathbf{x}}(\mathbf{x})}{\prod_i f_{\mathbf{x}_i}(x_i)} d\mathbf{x} = K\left(x \mid \prod_i x_i \right) \tag{6}$$

It turns out that MI is symmetric, zero iff the factorization of the joint density $f_{\mathbf{x}}(\mathbf{x})$ holds (e.g. the components are independent), and it is strictly positive otherwise. Comparing the form of negentropy in (5) and MI in (6), it comes out that if a gaussian multivariate is a reasonable approximation to the product of the marginal densities, then negentropy is a means to estimate the MI and, implicitly, a measure of independence.

If we consider the basic linear ICA model $\mathbf{y} = \mathbf{B}\mathbf{x}$, then MI between the estimated independent components $y_i, i = 1,2,...,N$ becomes:

$$I(y_1, y_2, ..., y_N) = \sum_{i=1}^{N} H(y_i) - H(\mathbf{y}) = \sum_{i=1}^{N} H(y_i) - H(\mathbf{x}) - log|det(\mathbf{B})| \tag{7}$$

If the scalar random variables $y_i, i = 1,2,...,N$ are constrained to be uncorrelated, then $det\,\mathbf{B}$ is constant.

$$I(y_1, y_2, ..., y_N) = \sum_{i=1}^{N} H(y_i) + constant \tag{8}$$

where the constant that does not depend on **B**. Moreover, since the estimations y_i are assumed of unit variance, entropy and negentropy differ only by a constant and the sign. Therefore, MI and negentropy differ by a constant only:

$$I(y_1, y_2, ..., y_N) = -\sum_{i=1}^{N} J(y_i) + constant \tag{9}$$

which explicitly shows that minimizing pairwise the MI of the random variables $y_1, y_2, ..., y_N$ equates to maximizing the sum of their individual negentropies $H(y_i) = -E\{log\ f_{y_i}(y_i)\}$. But because a gaussian density has maximal (differential) entropy, this also means *minimizing the gaussianities* of the random variables y_i, $i = 1,2,...,N$.

It is nevertheless quite difficult to compute both the KL divergence and negentropy. The approximations for negentropy introduced by Hyvärinen [4] for a scalar random variable *y* with zero mean, unit variance, and *p* functions G_i are:

$$J(y) \approx \sum_{i=1}^{p} k_i [E\{G_i(y)\} - E\{G_i(y_G)\}]^2 \tag{10}$$

where G_i are practically any nonquadratic functions, k_i are some positive constants, and y_G is a gaussian variable with zero mean and unit variance as *y*. When using only two nonlinear functions, $G_1 = y\,exp(-y^2/2)$, which measures the asymmetry, and an even one, $G_2 = |y|$, which measures the sparsity/bimodality of an *1D* nongaussian distribution, the approximate entropy becomes simpler [4]:

$$H(y) \approx H(y_G) - [k_1 (E\{G_1(y)\})^2 + k_2 (E\{G_2(y)\} - E\{G_2(y_G)\})^2] \tag{11}$$

Two practical implementations of (11) used in our experimental evaluations were:

$$\begin{aligned} H_a(y) &= H(y_G) - \left[k_1 \left[E\{y\, exp(-y^2/2)\} \right]^2 + k_2^a \left[E\{|y|\} - \sqrt{2/p} \right]^2 \right] \\ H_b(y) &= H(y_G) - \left[k_1 \left[E\{y\, exp(-y^2/2)\} \right]^2 + k_2^b \left[E\{exp(-y^2/2)\} - \sqrt{1/2} \right]^2 \right] \end{aligned} \tag{12}$$

We based our algorithm ranking on the strict monotonicity of negentropy computed according to (12).

ALGORITHMS FOR ICA

Apart from the estimation principle of ICA expressed in the form of an objective function subject to optimization, an algorithm is needed for implementing the necessary computations. Since nonquadratic functions are generally involved by the estimation methods, numerical algorithms are needed, which are quite computationally demanding. The current algorithms for ICA can loosely be classified in two categories. One category contains *adaptive algorithms* generally

based on stochastic gradient methods and implemented in neural networks [5], [6], [7]. Adaptive algorithms may also be based either on optimization of cumulant-based contrast functions [8], or on "estimating equations" involving nonlinear distortions of the output $y(t)$ [9]. The neural adaptive algorithms exhibit slow convergence and their convergence depends crucially on the correct choice of the learning rate parameters. The second category relies on *batch computation* optimizing some relevant criterion functions [1], [10]. Generally, they imply complex matrix or tensorial operations. Neuromorphic block technique algorithms based on *2*nd- and *4*th-order cumulants [11], as well as (quasi)-likelihood approaches were also proposed [12].

Algorithm assessment

Having known the original source components, the accuracy of separating power of the independent components of an ICA algorithm can be measured by means of various indexes. We will assume hereafter $M = N$. One index used as a global figure of merit for the separation performance may be defined as *signal-to-interference ratio* such as:

$$SIR = -\frac{1}{N}\sum_{i=1}^{N} 10\log_{10}\frac{max(Q_i)^2}{Q_i^T Q_i - max(Q_i)^2} \quad \text{(dB)} \tag{13}$$

where $\mathbf{Q} = \mathbf{BA}$ is the overall transforming matrix of the source components, Q_i is the *i*-th column of **Q,** $max(Q_i)$ is the maximum element of Q_i, and N is the number of source signals. The higher *SIR* is, the better the separation performance of the algorithm. A second index, *CTE*, which was used to measure the accuracy of retrieving the independent components, is the distance between the overall transforming matrix **Q** and an ideal permutation matrix, which is interpreted as the *cross-talking error* [13]:

$$CTE = \sum_{i=1}^{N}\left(\sum_{j=1}^{N}\frac{|Q_{ij}|}{max|Q_i|} - 1\right) + \sum_{j=1}^{N}\left(\sum_{i=1}^{N}\frac{|Q_{ij}|}{max|Q_j|} - 1\right) \tag{14}$$

Above, Q_{ij} is the *ij*-th element of **Q**, $max|Q_i|$ is the maximum absolute valued element of the row *i* in **Q**, and $max|Q_j|$ is the maximum absolute valued element of the column *j* in **Q**. A permutation matrix is defined so that on each of its rows and columns, only one of the elements equals to unity while all the other elements are zero. It means that *CTE* attains its minimum value zero for an exact permutation matrix.

Ranking the estimates

Ranking the estimated independent components was another criterion used for assessing the reliability of ICA algorithms. Friedman [14] proposed a robust *structural* measure to arrange the ICA basis vectors. The idea is to first sphere the

data and then to map them into the interval $[0,1]$ with the gaussian cumulative density function $\Phi(v)$. For T realizations of a (scalar) random variable $y_k(t)$ the proposed scheme leads to the index for the *k*-th estimated independent component:

$$E_1(y_k) = \sum_{t=1}^{T} \left[y_{s_t} - \left(\frac{2t}{T} - 1 \right) \right]^2 \tag{15}$$

where $\{s_t\}$ are the indexes of the ordered $\{y_k(t)\}$ in such a way that $y_k(i) \le y_k(j)$ iff $s_i \le s_j$. The higher $E_1(y_k)$ is, the more structural information contains the *k*-th estimated independent component.

It is meaningful sorting the components by the extent of their contribution to the original data. The *contribution* of the estimated component $y_k(t)$ can be estimated by the root mean square (RMS) of the data set reconstructed solely from this component $\hat{\mathbf{x}} = \hat{\mathbf{A}}\mathbf{y}$ in which **y** has only one nonzero row corresponding to the appropriate component, or as the RMS error introduced per data point when the data **x** are reconstructed without this component:

$$E_2(y_k) = \frac{1}{T \cdot N} \left[\sum_{j=1}^{N} \sum_{t=1}^{T} \left(C_{jt}^k \right)^2 \right]^{1/2} \tag{16}$$

where C_{jt}^k is the element of an $N \times T$ matrix computed from the outer product of the *k*-th independent component and the *k*-th column of $\hat{\mathbf{A}}$, that is $C_{jt}^k = B_{jk}^{-1} Y_{kt}$. The higher $E_2(y_k)$ is, the higher the contribution of the component $y_k(t)$ to the observed data.

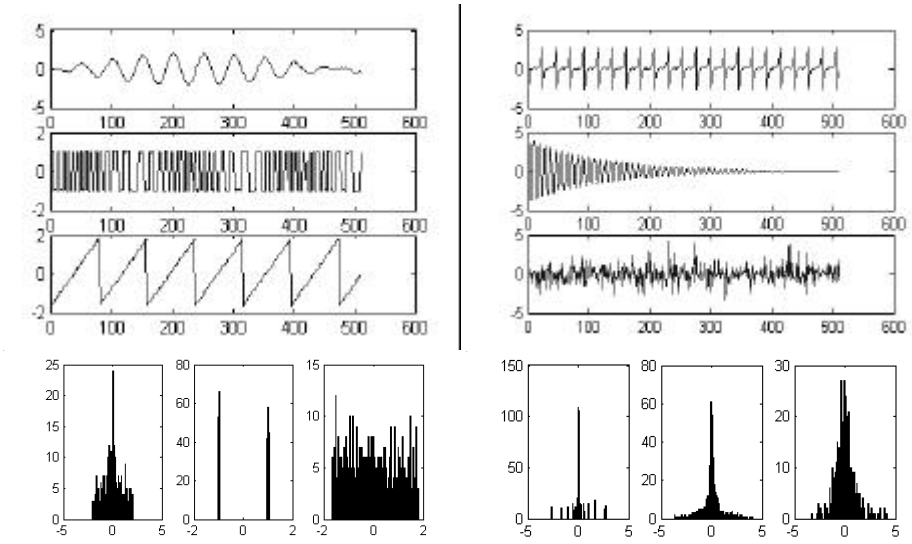

Figure 3: Artificially generated signals (up) and their histograms (bottom). Both subgaussian (left) and supergaussian (right) signal distributions were considered.

RESULTS AND DISSCUSSION

In our simulations, we used 6 different artificially generated time series of 512 samples each, both subgaussian and supergaussian (Fig. 3).

Table 1. The analytical form and the *3*-rd and *4*-th order cumulants of the sources**.**

Source signal	*Skewness*	*Kurtosis*
Modulated sinusoid: $S(1) = 2 * \sin(t/149) * \cos(t/8)$	*0.024637*	*–0.551312*
Square waves: $S(2) = sign(\sin(12 * t + 9 * \cos(2/29)))$	*0.015638*	*–1.996568*
Saw-tooth: $S(3) = (rem(t,79) - 17)/23$	*0.101021*	*–1.191073*
Impulsive curve: $S(4) = ((rem(t,23) - 11)/9)^5$	*–0.011980*	*2.353211*
Exponential decay: $S(5) = 5 * \exp(-t/121) * \cos(37 * t)$	*0.055131*	*3.410776*
Spiky noise: $S(6) = ((rand(1,T) < .5) * 2 - 1) * \log(rand(1,T))$	*0.464295*	*2.228476*

Table 2. The origin of code sources for the neurally implemented ICA algorithms.

Algorithms	*Type*	*Source*
FastICA	original	http://www.cis.hut.fi/projects/ica/fastica
BS	modified	http://www.sccn.ucsd.edu/~scott/ica.html and [15]
ACY	personal	as described in [7]
EASI	modified	http://sig.enst.fr/~cardoso/guidesepsou.html
Pearson-ICA	modified	http://wooster.hut.fi/statsp/papers/Pearson_ICA.zip
EGLD-ICA	modified	http://wooster.hut.fi/statsp/papers/EGLD_ICA.zip

Table 3. Nonlinearities and the sum of negentropies of the algorithms under study.

Algorithms	*Nonlinearities and score functions*	*Sum of negentropies*	
		J_a	J_b
FastICA	$g(y) = y \exp(-y^2/2)$	1.67 ± 0.028	1.25 ± 0.012
BS-Infomax	$g(y) = y \pm \tanh(y)$	1.0 ± 0.23	0.8 ± 0.35
ACY	$g(y) = y - \tanh(y)$	0.8 ± 0.50	0.6 ± 0.19
EASI	$g(y) = -\tanh(y)$	0.5 ± 0.48	0.4 ± 0.35
Pearson-ICA	$j(y) = -\frac{y - a}{b_0 + b_1 y + b_2 y}$	0.96 ± 0.075	0.8 ± 0.66
EGLD-ICA	$F^{-1}(p) = l_1 + \frac{p^{l_3} - (1-p)^{l_4}}{l_2}$	1.02 ± 0.095	0.89 ± 0.23

All codes were in MATLAB 6.0 and simulations were run on a PC machine with Pentium 4 processor and CPU at 1.5 MHz. The higher-order statistics of the source signals are presented in Table 1. The sources for the MATLAB codes are indicated in Table 2, whereas the optimal contrast functions and the sum of the individual negentropies for each algorithm are presented in Table 3. The separation performance of the algorithms under test is resumed in Table 4, where the indexes of the retrieved estimates correspond to the input data sequence.

Table 4. Indexes of performance and the retrieval sequence of the source signals.

Algorithms	*SIR* [dB]	*CTE*	E_1	E_2
FastICA	17.1 ± 3.51	0.65 ± 0.12	$y_6, y_3, y_5, y_2, y_1, y_4$	$y_3, y_5, y_2, y_1, y_4, y_6$
BS	12.2 ± 4.72	1.12 ± 0.33	$y_6, y_3, y_5, y_2, y_1, y_4$	$y_3, y_5, y_2, y_1, y_4, y_6$
ACY	13.5 ± 3.86	0.90 ± 0.41	$y_6, y_3, y_5, y_2, y_1, y_4$	$y_3, y_5, y_2, y_1, y_4, y_6$
EASI	7.02 ± 5.30	2.30 ± 1.39	$y_6, y_3, y_5, y_2, y_4, y_1$	$y_3, y_5, y_2, y_1, y_4, y_6$
Pearson	7.93 ± 2.88	2.14 ± 1.74	$y_6, y_3, y_5, y_2, y_4, y_1$	$y_3, y_5, y_2, y_1, y_4, y_6$
EGLD	8.21 ± 3.75	2.10 ± 1.56	$y_6, y_3, y_5, y_2, y_1, y_4$	$y_3, y_5, y_2, y_1, y_4, y_6$

CONCLUSIONS

According to our simulations with six artificially generated time series the best ICA algorithm in terms of convergence, computational requirements, and parameters to be tuned is the FP FastICA with *symmetric* orthogonalization and *exponential* nonlinearity. Its stabilized version converges always to a definite subspace of meaningful components even if the statistical independence is weak. Though the BS and ACY algorithms are theoretically optimal in terms of mutual information, their computational cost is higher whereas the results are similar with the FP. Moreover, like all neural unsupervised algorithms, both BS and ACY algorithms are heavily dependent on the learning rates and their convergence is quite slow. The Pearson and EGLD algorithms employing the ML principle separate a relative wide class of nongaussian source signals of large interest, even skewed distributions with zero kurtosis. However, Pearson system's ability is to model distributions that are close to normal distribution constrains its applications since it has no particular advantages for modeling distributions far from normality. As both estimators for parameters and score function are simple rational functions both Pearson-ICA and EGLD algorithms are computationally fast. However, the error margins are sensibly larger than in the case of FP, BS and ACY algorithms.

ACKNOWLEDGMENTS

R.M. is supported by a postdoc grant from the European Community, FP5 (QLG3-CT-2000-30161). M.M.V.H. is supported by research grants received from the Fund for Scientific Research (G.0185.96N), the National Lottery (Belgium)

(9.0185.96), the Flemish Regional Ministry of Education (Belgium) (GOA 95/99-06; 2000/11), the Flemish Ministry for Science and Technology (VIS/98/012), and the European Community, FP5 (QLG3-CT-2000-30161 and IST-2001-32114).

REFERENCES

[1] P. Comon, "Independent component analysis, A new concept?" **Signal Proces**., vol. 36, pp. 287-314, 1994.

[2] Cardoso, J.-F., "Blind signal separation: Statistical principles," **Proc**. **IEEE**, vol. 9, no. 10, pp. 2009-2025, 1998.

[3] K. Abed Meraim, A. Belouchrani, J.-F. Cardoso, E. Moulines, "Asymptotic performance of second order blind source separation," in **Proc. ICASSP**, IV, 1994, pp. 277-280.

[4] A. Hyvärinen, "New approximations of differential entropy for ICA and projection pursuit," in **Advances in Neural Information Processing Systems** (NIPS'97), MIT Press, vol. 10, pp. 273-279, 1998.

[5] A.J. Bell and T.J. Sejnowski, "An information maximization-approach to blind separation and blind deconvolution," **Neural Comput**., vol. 7, pp. 1129-1159, 1995.

[6] C. Juten and J. Hérault, "Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture," **Signal Proces**., vol. 24, no. 1, pp. 1-10, 1991.

[7] S. Amari, A. Cichocki and H. Yang, "A new learning algorithm for blind source separation," in **Advances in Neural Information Processing 8** (**Proc. NIPS'95**), Cambridge: MIT Press, 1996.

[8] O. Moreau and O. Macchi, "New self-adaptive algorithms for source separation based on contrast functions," in **Proc. IEEE Signal Processing Workshop on Higher Order Statistics**, Lake Tahoe, NV, 1993, pp. 215-219.

[9] J.-F. Cardoso, A. Belouchrani, and B. Lahed, "A new composite criterion for adaptive and iterative blind source separation," in **Proc. ICASSP**, vol. 4, pp. 273-276, 1994.

[10] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," **Neural Comput**., vol. 9, pp. 1483-1492, 1997.

[11] J.-F. Cardoso, and A. Souloumiac, "Blind beamforming for nongaussian signals," **IEE Proc.-F**, vol. 140, no. 6, pp. 362-370.

[12] D.-T. Pham, P. Garrat, and C. Jutten, Separation of a mixture of independent sources through a maximum likelihood approach," in **Proc. EUSIPCO**, pp. 771-774, 1992.

[13] H. Yang and S. Amari, "Adaptive online learning algorithms for blind separation: maximum entropy and minimum mutual information," **Neural Comput**. vol. 9, pp. 1457-1482, 1997.

[14] J.H. Friedman, "Exploratory projection pursuit," **J. of the American Statistical Association**, vol. 82, no. 397, pp. 249-266, 1987.

[15] T.-W. Lee, M. Girolami, A.J. Bell, and T.J. Sejnowski, "A unifying information-theoretic framework for Independent Component Analysis, **Computers and Mathematics with Applications**, vol. 39, pp. 1-21, 2000.