# Information-theoretic learning for FAN network applied to eterokurtic component analysis

S. Fiori

**Abstract:** The paper presents a novel approach for performing independent component analysis of mixed plati-kurtic and lepto-kurtic source signals, which is referred to as the 'eterokurtic' blind source separation problem. The approach employs a neural network formed by adaptive activation function neurons, which provide the statistics required for learning by the extended INFOMAX theory. Through computer simulations conducted on both synthetic and real-world data, the proposed approach is assessed and its effectiveness is illustrated.

## 1 Introduction

Over recent years, blind source separation by independent component analysis (ICA) has received attention owing to its potential applications, for instance in speech recognition systems, telecommunications, pollution monitoring, fault detection and non-invasive evaluation, medical imaging, financial data market analysis, and other signal/data processing fields [1–7]. The aim of ICA is to recover unobservable independent source signals from sensor observations, which are unknown mixtures of them, by reducing high-order statistical correlation.

The classical example used to explain the blind separation problem is the 'cocktail party' scenario: let us imagine that $m$ people stand in a room and speak together; if the room is equipped with a number of microphones, each sensor receives a different superposition of the speech signals uttered from each person in the room, so that the set of received signals may be described by a linear mixture model that takes into account the geometry of the emitters and sensors and the pressure waves propagation phenomena. The aim of blind source separation algorithm is to recover, from sensor observations only, the single independent uttered signals.

The pioneering work on blind source separation was carried out by Hérault and Jutten [8, 9], who introduced an adaptive algorithm in a feedback neuromimetic multiple filter. Their approach was further developed in [10]. Comon [11] developed the concept of independent component analysis and proposed a class of cost functions termed discriminant contrasts. The contrast-based approach has been further studied [12, 13] and simplified contrasts for lepto-kurtic and plati-kurtic source signals (i.e. for signals with positive or negative kurtosis) introduced. On the basis of simplified contrast functionals (i.e. on kurtosis optimisation), Cardoso and Laheld [14] proposed an adaptive algorithm relying on relative gradient, which has been proven to enjoy very desirable properties such as equivariancy and fast convergence; also, Delfosse and Loubaton [15] proposed an algorithm based on a deflation procedure, an idea that has recently been developed, resulting in a cascade neural network [16].

In parallel to blind source separation studies, unsupervised learning rules for artificial neural networks, based on information theory, were proposed by Linsker [17] and Plumbley [18]. The aim was to maximise the mutual information between the inputs and the outputs of a neural network, so that each neuron should match features being as statistically independent as possible from other neurons in the network. Nadal *et al.* [19] showed that in the low-noise case, the maximisation of the mutual information between the input and the output of a network implies that the output joint probability density function (pdf) factorises as a product of marginal pdfs. Roth and Baram [20] and Bell and Sejnowski [21], independently derived stochastic gradient learning rules for mutual information maximisation with application to forecasting and time-series analysis, blind separation of sources and blind deconvolution, respectively. Their approach is commonly referred to as 'INFOMAX'. Girolami and Fyfe [22] employed neural exploratory projection pursuit algorithms for achieving separation. Generalised Hebbian learning algorithms for ICA have been developed by Fiori [23, 24], Karhunen *et al.* [25], Hyvärinen and Oja [26]. It is interesting to mention a new theory, recently developed by Sagi *et al.* [27], based on the 'cortronic' neural network: it provides signal pre-processing in advance of pattern recognition and is based on a biologically feasible neural model. Non-conventional neural optimisation techniques have recently been applied to blind separation. As three examples, the present author employed mechanical-type learning algorithms to blind separation by the ICA (recently extended to complex-valued sources) [28–30]; Prieto and Puntonet developed a purely geometrical approach to apply in the presence of unimodal-symmetrical source distributions [31], while Yoshioka and Omatu applied a genetic algorithm to minimise ICA cost functions [32].

Other algorithms have been proposed from different perspectives by several authors, and simulations have been performed to illustrate the usefulness and effectiveness of the separating algorithms, as for instance in EEG data processing (showing that the algorithms can extract EEG



*IEE Proceedings* online no. 20020652

*DOI:* 10.1049/ip-vis:20020652

Paper first received 19th October 2001 and in revised form 8th April 2002

The author is with the Neural Networks and Signal Processing Group, IED, University of Perugia, Via Pentima bassa, 21-05100, Terni, Italy

activation and isolate artefacts), and for exploring independent features in natural images and sounds. Moreover, it is known that independent component analysis relies on some assumptions which limit its fields of applications. Researchers have recently tried to overcome this problem, especially by trying to extend the existing algorithms to convolutional, underdetermined, and nonlinear mixtures.

In this paper we deal with instantaneous linear mixtures and focus on the stream of INFOMAX learning algorithms. The paper is devoted to the separation of mixed independent signals from their linear mixtures when the observations are mixed plati-kurtic *and* lepto-kurtic signals, referred to as a hybrid or *eterokurtic sources* problem. We propose the use of networks formed by unsupervised adaptive activation function neurons (FAN), which provide a natural way of estimating the high-order statistical features required to achieve separation. Through numerical and analytical studies the effectiveness of the presented approach is also illustrated and discussed.

*Notation:* Throughout the paper lower-case bold letters denote column-vectors, while upper-case bold letters denote matrices; symbol $E_{\boldsymbol{u}}[f(\boldsymbol{u})]$ denotes mathematical expectation with respect to the statistics of the multivariate random variable (or stationary random process) $\boldsymbol{u}$; $f'(u)$ stands for the first derivative of $f(u)$.

## 2 Problem statement

The aim of the present Section is to formally present the blind separation problem and to recall the INFOMAX approach to its solution. Indeed, we recall an improved version of INFOMAX, termed flexible, introduced by Xu *et al.* in [33, 34]. As we introduce another kind of flexibility, Xu's algorithm is retained as a comparison term in the following.

### 2.1 The flexible INFOMAX approach

In the blind source separation problem, a mixture of independent source signals is supposed to be observed:

$$\boldsymbol{m}(t) = \boldsymbol{M}^T\boldsymbol{s}(t) \tag{1}$$

where $\boldsymbol{M}$ is a constant real-valued full-rank $m \times m$ mixing matrix and $\boldsymbol{s}(t)$ is the vector-stream containing the $m$ source signals to be separated. The only hypotheses made on the unknown sources are:

(i) every $s_j(t)$ is an independent identically distributed stationary random process;
(ii) the $s_j(t)$ are statistically independent at any time;
(iii) the $s_j$ are symmetrically distributed (i.e. denoting with $p_{s_j}(u)$ the probability density function of $s_j$, it holds that $p_{s_j}(u) = p_{s_j}(-u)$);
(iv) at most one among the source signals is allowed to be Gaussian.

To separate out the linearly mixed independent sources, a neural network is used, with $m$ inputs and $m$ outputs, described by the relationships:

$$\boldsymbol{x}(t) = \boldsymbol{W}^T(t)\boldsymbol{m}(t) \tag{2}$$

$$\boldsymbol{y}(t) = \boldsymbol{\Psi}(\boldsymbol{x}(t)) = [\Psi_1(x_1(t))\ \Psi_2(x_2(t))\ \cdots\ \Psi_m(x_m(t))]^T \tag{3}$$

where $\boldsymbol{m}$ is the network input vector, $\boldsymbol{x} = [x_1\ x_2\ \cdots\ x_m]^T$ denotes the net vector, $\boldsymbol{W}$ is the weight matrix, and $\boldsymbol{\Psi}$ represents the activation operator, which may contain adaptive parameters, at time $t$. As the mixing model is linear, a linear separating structure is effective, thus the first-layer output $\boldsymbol{x}(t)$ in (2) is taken as an estimate of the true source stream $\boldsymbol{s}(t)$.

The basic principle that the independent component analysis technique is based on is that after application of the mixing model (1), the observed signals $m_j(t)$ are no longer statistically independent; thus, in order to achieve separation, the weight matrix may be learnt so that the network's outputs (2) become as independent as possible, that is, they meet the complete factorisation property, which for the network's linear part outputs gives:

$$p_{\boldsymbol{x}}(\boldsymbol{x}) = p_{x_1}(x_1)p_{x_2}(x_2)\cdots p_{x_m}(x_m) \tag{4}$$

where $p_{\boldsymbol{x}}(\boldsymbol{x})$ denotes the joint probability density function of the activations, and

$$p_{x_j}(x_j) \stackrel{\text{def}}{=} \int_{\mathcal{R}^{m-1}} p_{\boldsymbol{x}}(\boldsymbol{x})dx_1\cdots dx_{j-1}dx_{j+1}\cdots dx_m$$

A way to achieve separation is thus to define a measure of the mismatch between the two sides of (4), and a learning algorithm to iteratively adjust the weight matrix to minimise this disagreement. Several possible mismatch measures have been reported in the scientific literature during recent years. Among others, here we recall a very interesting and fruitful one which relies on mutual information (MI). Using our notation, the MI of network activations is defined as

$$\mathcal{I}_{\boldsymbol{x}} \stackrel{\text{def}}{=} \int_{\mathcal{R}^m} p_{\boldsymbol{x}}(\boldsymbol{x}) \log \frac{p_{\boldsymbol{x}}(\boldsymbol{x})}{p_{x_1}(x_1)p_{x_2}(x_2)\cdots p_{x_m}(x_m)} d\boldsymbol{x} \tag{5}$$

With some algebra we obtain:

$$\mathcal{I}_{\boldsymbol{x}} = -\mathcal{H}_{\boldsymbol{m}} - |\det(\boldsymbol{W})|^{-1} - \sum_{j=1}^{m} E_{\boldsymbol{m}}[\log p_{x_j}(x_j)] \tag{6}$$

where $\mathcal{H}_{\boldsymbol{m}}$ denotes the Shannon differential entropy of the multivariate random process $\boldsymbol{m}$, which does not depend upon matrix $\boldsymbol{W}$.

The marginal entropies $-E_{\boldsymbol{m}}[\log p_{x_j}(x_j)]$ cannot be explicitly computed: as $x_j(t) = \boldsymbol{w}_j^T\boldsymbol{M}^T\boldsymbol{s}(t)$, the pdf $p_{x_j}(x_j)$ is a complicated function of the $p_{s_j}(s_j)$ that, by definition, are unknown; thus, their suitable approximations are needed and can be computed on the basis of the available quantities. Here we propose to employ the following approximations relying on time-varying (parametric) functions:

$$p_{x_j}(x_j) \sim \psi_j(x_j) = \psi_j(x_j; \boldsymbol{q}_j) \tag{7}$$

where each $\boldsymbol{q}_j$ represents a vector of suitable size containing the adjustable parameters for the $j$th approximating function. By denoting with $\boldsymbol{Q}$ the matrix whose columns are the aforementioned vectors $\boldsymbol{q}_j$, an approximation of criterion (5) can be defined as

$$\hat{\mathcal{I}}_{\boldsymbol{x}}(\boldsymbol{W}, \boldsymbol{Q}) = -\mathcal{H}_{\boldsymbol{m}} - |\det(\boldsymbol{W})|^{-1} - \sum_{j=1}^{m} E_{\boldsymbol{m}}[\log \psi_{x_j}(\boldsymbol{w}_j^T\boldsymbol{m}; \boldsymbol{q}_j)] \tag{8}$$

It is worth noting that both $\boldsymbol{W}$ and $\boldsymbol{Q}$ are network parameters that are iteratively adjusted through a suitable learning algorithm. This makes the functions $p_{x_j}(x_j)$ time-varying. This is an important feature of the proposed architecture that makes it able to be reactive in a non-stationary environment and makes it useful in non-stationary signal processing.

The nonlinear functions mentioned are related by $\psi_j(x_j; \boldsymbol{q}_j) = \Psi_j'(x_j; \boldsymbol{q}_j)$; any $\Psi_j$ should approximate the cumulative distribution function of the $j$th source signal, whose shape closely recalls the classical saturating 'sigmoidal' function extensively used in the neural network literature.

To iteratively adjust the weight matrices to minimise the MI, we use the following learning rules:

$$\Delta \boldsymbol{W} = -\eta_W \frac{\partial \hat{\mathcal{I}}}{\partial \boldsymbol{W}} (\boldsymbol{W}^T \boldsymbol{W}) \tag{9}$$

$$\Delta \boldsymbol{Q} = -\eta_Q \frac{\partial \hat{\mathcal{I}}}{\partial \boldsymbol{Q}} \tag{10}$$

The first equation represents a natural-gradient descent flow [35]; the constant $\eta_W$ represents the learning step-size for $\boldsymbol{W}$, while constant $\eta_Q$ denotes the learning step-sizes for the entries of $\boldsymbol{Q}$.

If we define, for easier notation, the new functions:

$$d_j(x_j; \boldsymbol{a}_j) \stackrel{\text{def}}{=} \frac{\psi_j'(x_j; \boldsymbol{q}_j)}{\psi_j(x_j; \boldsymbol{q}_j)} = \frac{\Psi_j''(x_j; \boldsymbol{q}_j)}{\Psi_j'(x_j; \boldsymbol{q}_j)} \tag{11}$$

and replace in the above formulas the expectation by sample mean, we obtain the following stochastic-gradient learning rules:

$$\Delta \boldsymbol{W} = \eta_W [\boldsymbol{I}_m + \boldsymbol{d}(\boldsymbol{x}) \boldsymbol{x}^T] \boldsymbol{W} \tag{12}$$

$$\Delta \boldsymbol{q}_j = \eta_Q \frac{1}{\psi_j(x_j)} \frac{\partial \psi_j(x_j)}{\partial \boldsymbol{q}_j} \tag{13}$$

It is worth noting that criterion (6) may be derived in different ways and is closely related to maximum entropy and minimal Kullback–Leibler divergence criteria [35].

### 2.2 Existing approaches to eterokurtic component analysis and neural density estimation

Some neural blind separation techniques are at present available, which allow one to perform eterokurtic component analysis [15, 16, 26, 36, 37]. They rely on parallel/cascade/hierarchical linear neural structures adapted through kurtosis optimisation. As the sign of the kurtosis determines whether it has to be minimised or maximised, it is necessary to recursively estimate the sign by means of a specific algorithm. Such a necessity is problematic, as it is difficult to guarantee that the joint learning processes converge for parallel running algorithms; also, to avoid dangerous oscillations in non-parallel implementations, a cooling scheme seems often to be necessary (i.e. learning step-sizes drop to zero during the learning phase), making the algorithms ineffective when non-stationary signal processing tasks are dealt with. Furthermore, an important drawback arising from kurtosis optimisation is that algorithms relying on it are ineffective when kurtosis is zero or does not exist [34] (note that a signal need not be Gaussian to have zero kurtosis).

A starting point for getting rid of these drawbacks was the INFOMAX approach by Bell and Sejnowski. However, in its earliest version [21], the approximating function (7) was chosen so that $\Psi(x_j) = \tanh(x_j)$, that is a fixed nonlinearity. Both theoretical studies and extensive computer simulations have shown that this approach was not reliable in that the performances of the algorithm strongly depends on the statistics of the sources. Bell and Sejnowski suggested themselves in [21] the use of a family of flexible (non-adaptive) nonlinear functions, to overcome this problem.

The key point is that the algorithm would require reliable estimates of the probability density functions of the sources that cannot be represented by functions fixed *a priori*.

Since 1996 some researchers in the ICA field started claiming that the use of adaptive nonlinear functions could help solve the eterokurtic source separation problem. Pearlmutter and Parra [38] proposed the use of linear combinations of parametric basis functions; Taleb and Jutten [39], and Roth and Baram [40] employed a MLP to adaptively estimate the 'score function' or the probability density function of the sources, respectively; Gustaffson [41] proposed the use of an adjustable linear combination of quasi-Dirac's-delta-functions, while Xu *et al.* [33, 34] presented a 'mixture of kernel' based approximation technique; recently, Welling and Webber [42] proposed a maximum-likelihood approach to density estimation applied to blind separation based on the well-known expectation/maximisation (EM) algorithm, while Karvanen *et al.* [43] proposed employing the Pearson system to estimate the source distributions. These methods have in common the feature that such flexible functions may be 'learnt' so that they approximate the required statistical functions, helping the separation algorithm to perform better.

In this paper we propose a neural technique for density estimation that meets the basic requirements of good approximation ability and low architectural and computational complexity. It is based on adaptive activation function neural units, which are able to provide good adaptive estimation of density functions from raw available data.

## 3 Adaptive activation function neuron (FAN)

Given a random process $x(t) \in \Gamma_x$, stationary and ergodic, endowed with a continuous pdf $p_x(x)$, an approximation of $p_x(x)$ is required by means of a parametric function $\psi(x; \boldsymbol{q})$, where $\boldsymbol{q} \in \mathcal{Q}$ contains the searched parameters. A possible approach relies on using the nonlinear transformation $y = \Psi(x; \boldsymbol{q}) \in \Gamma_y$, which has to be chosen so that

$$\frac{d\Psi(x; \boldsymbol{q})}{dx} = \psi(x; \boldsymbol{q}) \geq 0 \quad \lim_{x \to -\infty} \Psi(x; \boldsymbol{q}) = 0 \quad \lim_{x \to +\infty} \Psi(x; \boldsymbol{q}) = 1$$

The nonlinear adaptive function may be implemented by a neuron with adaptive activation function, which in general provides good flexibility while retaining fairly low computational complexity in comparison with other more complex structures. A detailed discussion on these topics appeared in [44–46].

### 3.1 Pseudo-polynomial adaptive activation function units

Here we propose the following nonlinear structure:

$$x = \boldsymbol{w}^T \boldsymbol{m} \qquad \Psi(x; \boldsymbol{q}) \stackrel{\text{def}}{=} \text{sgm}\left[\sum_{k=1}^{r} \phi(v_k) x^{2k-1} + v_0\right] \tag{14}$$

where $\boldsymbol{q} = [v_0 \ \boldsymbol{v}^T]^T$, $\boldsymbol{v} \stackrel{\text{def}}{=} [v_1 \cdots v_r]^T$, $\text{sgm}(\cdot)$ is a sigmoidal function, $\phi(\cdot)$ is a non-negative regular function, and $r$ is the expansion degree, which determines the neuron's approximation ability.

Let us define the following vectors:

$$\boldsymbol{f}(x) \stackrel{\text{def}}{=} \begin{bmatrix} x \\ x^3 \\ \vdots \\ x^{2r-1} \end{bmatrix} \qquad \boldsymbol{f}'(x) \stackrel{\text{def}}{=} \begin{bmatrix} 1 \\ 3x^2 \\ \vdots \\ (2r-1)x^{2r-2} \end{bmatrix}$$

$$\boldsymbol{f}''(x) \stackrel{\text{def}}{=} \begin{bmatrix} 0 \\ 6x \\ \vdots \\ (2r-1)(2r-2)x^{2r-3} \end{bmatrix}$$

Then, the required quantities in (12) and (13) are as follows [Note 1]:

$$\Psi(x;\boldsymbol{q}) = \mathrm{sgm}[\phi^T(\boldsymbol{v})\boldsymbol{f}(x) + v_0]$$
$$\psi(x;\boldsymbol{q}) = \mathrm{sgm}'[\phi^T(\boldsymbol{v})\boldsymbol{f}(x) + v_0][\phi^T(\boldsymbol{v})\boldsymbol{f}'(x)]$$
$$\psi'(x;\boldsymbol{q}) = \mathrm{sgm}''[\phi^T(\boldsymbol{v})\boldsymbol{f}(x) + v_0][\phi^T(\boldsymbol{v})\boldsymbol{f}'(x)]^2 + \mathrm{sgm}'[\phi^T(\boldsymbol{v})\boldsymbol{f}(x) + v_0][\phi^T(\boldsymbol{v})\boldsymbol{f}''(x)]$$
$$\frac{\partial\psi}{\partial v_0} = \mathrm{sgm}''[\phi^T(\boldsymbol{v})\boldsymbol{f}(x) + v_0][\phi^T(\boldsymbol{v})\boldsymbol{f}'(x)]$$
$$\frac{\partial\psi}{\partial \boldsymbol{v}} = \mathrm{sgm}''[\phi^T(\boldsymbol{v})\boldsymbol{f}(x) + v_0][\phi^T(\boldsymbol{v})\boldsymbol{f}'(x)]\frac{\partial\phi(\boldsymbol{v})}{\partial\boldsymbol{v}}\boldsymbol{f}(x) + \mathrm{sgm}'[\phi^T(\boldsymbol{v})\boldsymbol{f}(x) + v_0]\frac{\partial\phi(\boldsymbol{v})}{\partial\boldsymbol{v}}\boldsymbol{f}'(x)$$

Hence, the quantities required for adaptive activation function neuron learning, in a separating network, are:

$$\frac{1}{\psi}\frac{\partial\psi}{\partial v_0} = \frac{\mathrm{sgm}''[\phi^T(\boldsymbol{v})\boldsymbol{f}(x) + v_0]}{\mathrm{sgm}'[\phi^T(\boldsymbol{v})\boldsymbol{f}(x) + v_0]} \tag{15}$$

$$\frac{1}{\psi}\frac{\partial\psi}{\partial\boldsymbol{v}} = \frac{\mathrm{sgm}''[\phi^T(\boldsymbol{v})\boldsymbol{f}(x) + v_0]}{\mathrm{sgm}'[\phi^T(\boldsymbol{v})\boldsymbol{f}(x) + v_0]}\nabla_{\boldsymbol{v}}\phi(\boldsymbol{v})\boldsymbol{f}(x) + \frac{\nabla_{\boldsymbol{v}}\phi^T(\boldsymbol{v})\boldsymbol{f}'(x)}{\phi(\boldsymbol{v})\boldsymbol{f}'(x)} \tag{16}$$

The function (11), which has been proven useful for a separating network, in this case reads:

$$d(x) = \frac{\mathrm{sgm}''[\phi^T(\boldsymbol{v})\boldsymbol{f}(x) + v_0]}{\mathrm{sgm}'[\phi^T(\boldsymbol{v})\boldsymbol{f}(x) + v_0]}\phi^T(\boldsymbol{v})\boldsymbol{f}(x) + \frac{\phi^T(\boldsymbol{v})\boldsymbol{f}''(x)}{\phi^T(\boldsymbol{v})\boldsymbol{f}'(x)} \tag{17}$$

An exemplary partial schematic of the adaptive activation function neuron, having the mentioned mathematical structure, is shown in Fig. 1 for $r=3$.

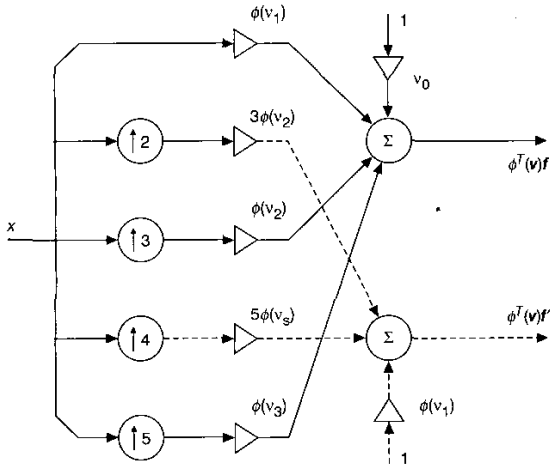


**Fig. 1** *Exemplary partial schematic of the FAN neuron*
$r=3$; blocks marked $\uparrow n$ compute $x^n$

Note 1: It should be noted that $\phi(\boldsymbol{v}) = [\phi(v_1)\ \phi(v_2)\cdots\phi(v_r)]^T$, thus $\nabla_{\boldsymbol{v}}\phi = \partial\phi/\partial\boldsymbol{v}$ is a diagonal matrix whose $i$th in-diagonal entry equals $\phi'(v_i)$.

### 3.2 The mixture-of-kernel (MOK) approach

Following Xu *et al.* [33, 34], the adaptive function $\Psi(x)$ can be assumed of the following form:

$$\Psi(x;\boldsymbol{q}) = \sum_{j=1}^{n}\alpha_j\beta_j(x),\quad \beta_j(x) \stackrel{\mathrm{def}}{=} \beta[b_j(x - c_j)] \tag{18}$$

where $\beta(\cdot)$ is called the 'basis kernel', $b_j \in \mathcal{R}$ are called 'slopes', $c_j \in \mathcal{R}$ are termed 'centres', and $\alpha_j \in \mathcal{R}$ are termed 'weights'; the integer number $n$ determines the dimension of the basis and the approximation flexibility degree. Here the parameters vector $\boldsymbol{q}$ contains all triples $(\alpha_j, b_j, c_j)$.

Any coefficient $\alpha_j$ plays the role of the probability that the term $\beta_j(x)$ gives a contribution to the representation, hence we require that they satisfy the restrictions $\alpha_j > 0$ and $\sum_{j=1}^{n}\alpha_j = 1$ for the representation to be consistent. Following [33, 34] we assume $\beta(u)$ to be the standard sigmoidal function. We than have:

$$\psi(x;\boldsymbol{q}) = \sum_{j=1}^{n}\alpha_j b_j[\beta_j^2(x) - \beta_j(x)]$$

To force parameters $\alpha_j$ to fulfil the conditions above, the softmax transformation

$$\alpha_j = \frac{\exp(\varrho_j)}{\sum_k \exp(\varrho_k)}$$

is employed, where parameters $\varrho_j \in \mathcal{R}$ are new unconstrained variables used instead of the true weights $\alpha_j$.

The learning rules for the parameters are recalled in the following for the sake of completeness:

$$\Delta\varrho_j = k_\varrho\frac{1}{\psi}\frac{\partial\psi}{\partial\varrho_j} = \frac{k_\varrho}{\psi}\sum_{i=1}^{n} b_i(\beta_i - \beta_i^2)\alpha_j(\delta_{ij} - \alpha_i) \tag{19}$$

$$\Delta c_j = k_c\frac{1}{\psi}\frac{\partial\psi}{\partial c_j} = -\frac{k_c}{\psi}\alpha_j b_j^2(1 - 2\beta_j)(\beta_j - \beta_j^2) \tag{20}$$

$$\Delta b_j = k_b\frac{1}{\psi}\frac{\partial\psi}{\partial b_j} = \frac{k_b\alpha_j}{\psi}[1 + b_j(x_j - c_j)(1 - 2\beta_j)] \times (\beta_j - \beta_j^2) \tag{21}$$

where $k_\varrho$, $k_c$ and $k_b$ are positive learning step-sizes. In this case the nonlinear function $d(x)$ is given by:

$$d(x) = \frac{1}{\psi(x)}\sum_{j=1}^{n}\alpha_j b_j^2[1 - 2\beta_j(x)][\beta_j(x) - \beta_j^2(x)] \tag{22}$$

The recalled learning theory will, in the following, be referred to as MOK.

## 4 FAN structures and their complexity

The aim of this Section is to discuss some possible choices of the functions $\phi(\cdot)$ and $\mathrm{sgm}(\cdot)$, from a computational complexity perspective, on the basis of the proposals already discussed by the author in preliminary reports [47–49].

A possible choice of the warping function for the coefficients of the polynomials is $\phi(u) = e^u$; in this case we have $\nabla_{\boldsymbol{v}}\phi = \mathrm{diag}(\phi(\boldsymbol{v}))$. Clearly this function meets exactly the requirement of being strictly positive, so that function $\Psi(x)$ is strictly monotonic. However, it introduces complexity in the neural architecture. A simpler choice could be $\phi(u) = 0.5u^2$, which is a function that vanishes at $u=0$. In this case $\nabla_{\boldsymbol{v}}\phi = \mathrm{diag}(\boldsymbol{v})$.

**Table 1: Complexity comparison of MOK and FAN**

| Algorithm | Multiplications | Divisions | Nonlinear Funcs | Parameters to adapt |
|---|---|---|---|---|
| MOK | $14mn+3$ | $nm+1$ | $2mn$ | $3mn$ |
| TANH/EXP | $7rm$ | $m$ | $(r+2)m$ | $(r+1)m$ |
| ERF/EXP | $7rm$ | $m$ | $(r+1)m$ | $(r+1)m$ |
| TANH/QUAD | $(9r-1)m$ | $m$ | $m$ | $(r+1)m$ |
| ERF/QUAD | $(9r-1)m$ | $m$ | 0 | $(r+1)m$ |

TANH: sgm = tanh; ERF: sgm = erf; EXP: $\phi(u)=e^u$; QUAD: $\phi(u)=0.5u^2$

For function $\mathrm{sgm}(u)$, we tried to employ the standard sigmoidal function $\mathrm{sgm}(u)=0.5+0.5\tanh(u)$. A drawback related to this function is that

$$\frac{\mathrm{sgm}''(u)}{\mathrm{sgm}'(u)}=2-2\mathrm{sgm}(u)$$

thus its computation is necessary. A more interesting sigmoidal function is $\mathrm{sgm}(u)=0.5+0.5\mathrm{erf}(u)$, which leads to:

$$\frac{\mathrm{sgm}''(u)}{\mathrm{sgm}'(u)}=-2u$$

Notice that with this choice, computation of the error function $\mathrm{erf}(\cdot)$ is not necessary.

In Table 1, the number of operations (multiplication, divisions and any generic nonlinearity computation) involved (and strictly needed) in the learning equations is evaluated. The structures discussed are listed in descending complexity order. Regarding the number of nonlinear functions to be computed, its weight in the comparison depends on which way they are implemented. In the usual truncated Taylor series implementation, any nonlinearity requires several multiplication operations, while if the nonlinear functions are represented as look-up tables [50], their evaluation roughly costs as much as a linear interpolation. In any case, it is quite apparent that the ERF/QUAD adaptive activation neuron structure is the simplest one.

## 5 Experimental results

In this part of the paper, we show computer simulation results that confirm the effectiveness of the proposed approach, and present a numerical comparison of MOK and FAN-based separation methods.

As convergence figure of merit, an interference residual measure has been defined as the sum of the $m^2-m$ smallest squared entries of the product $\boldsymbol{P}\stackrel{\mathrm{def}}{=}\boldsymbol{W}^T\boldsymbol{M}^T$, as [26]. In fact, since $\boldsymbol{P}$ represents the overall source-to-output transfer matrix, perfect separation would imply that only one entry per row should differ from zero [11]; in a real-world context, however, some residual interference should be tolerated. Also, the noise-to-signal ratio (NSR) at the output of each neuron measures the power of the residual interference (the 'noise') with respect to the power of the separated source signal pertaining to that neuron. Formally, it is defined as

$$\mathrm{NSR}_j\stackrel{\mathrm{def}}{=}10\log_{10}\left(\frac{\sum_{i\neq k}P_{ji}^2}{P_{jk}^2}\right)\qquad k\stackrel{\mathrm{def}}{=}\arg\max_{\{i\}}P_{ji}^2$$

### 5.1 Blind source separation tests on FAN

Three experiments have been performed with the four possible FAN structures to evaluate their numerical performances. Details of the experiments are reported in Table 2. The mixing matrices for the $2\times2$ and $3\times3$ cases were, respectively:

$$\boldsymbol{M}_2=\begin{bmatrix}1 & 0.6\\ 0.7 & 1\end{bmatrix}\qquad \boldsymbol{M}_3=\begin{bmatrix}1 & 0.6 & 0.2\\ 0.8 & 1 & 0.3\\ 0.4 & 0.9 & 1\end{bmatrix}$$

as in [11, 34]. The results achieved using the four FAN structures discussed are reported in Table 3, along with the value of the learning step-size $\eta_Q$.

The ERF/QUAD architecture seems to give the best performance. In order to assess the proposed separation algorithm based on the ERF/QUAD FAN structure, it has been tested with an input stream which is a mixture of

**Table 2: Details of experiments on FAN structures**

| Experiment | Value of $r$ | No. of sources | Source signals |
|---|---|---|---|
| 1 | 1 | 2 | Two uniform noises |
| 2 | 1 | 2 | Two speech signals |
| 3 | 2 | 3 | A uniform noise, a speech signal, a sinusoid |

**Table 3: NSRs achieved by the four FAN structures after 50 000 iterations**

| Experiment | ERF/QUAD | ERF/EXP | TANH/QUAD | TANH/EXP |
|---|---|---|---|---|
| 1 | $\mathrm{NSR}_1=-40.55$<br>$\mathrm{NSR}_2=-44.66$<br>($\eta_Q=0.001$) | $\mathrm{NSR}_1=-39.57$<br>$\mathrm{NSR}_2=-42.33$<br>($\eta_Q=0.001$) | $\mathrm{NSR}_1=-39.96$<br>$\mathrm{NSR}_2=-47.59$<br>($\eta_Q=0.001$) | $\mathrm{NSR}_1=-35.68$<br>$\mathrm{NSR}_2=-43.21$<br>($\eta_Q=0.001$) |
| 2 | $\mathrm{NSR}_1=-41.62$<br>$\mathrm{NSR}_2=-30.06$<br>($\eta_Q=0.0005$) | $\mathrm{NSR}_1=-36.30$<br>$\mathrm{NSR}_2=-51.76$<br>($\eta_Q=0.0005$) | $\mathrm{NSR}_1=-49.67$<br>$\mathrm{NSR}_2=-32.95$<br>($\eta_Q=0.0005$) | $\mathrm{NSR}_1=-48.53$<br>$\mathrm{NSR}_2=-34.16$<br>($\eta_Q=0.0005$) |
| 3 | $\mathrm{NSR}_1=-54.96$<br>$\mathrm{NSR}_2=-29.58$<br>$\mathrm{NSR}_3=-68.11$<br>($\eta_Q=0.001$) | $\mathrm{NSR}_1=-63.86$<br>$\mathrm{NSR}_2=-40.64$<br>$\mathrm{NSR}_3=-28.61$<br>($\eta_Q=0.001$) | $\mathrm{NSR}_1=-51.07$<br>$\mathrm{NSR}_2=-24.24$<br>$\mathrm{NSR}_3=-63.80$<br>($\eta_Q=0.03$) | $\mathrm{NSR}_1=-55.34$<br>$\mathrm{NSR}_2=-43.05$<br>$\mathrm{NSR}_3=-35.23$<br>($\eta_Q=0.001$) |

All NSR values are given in dB

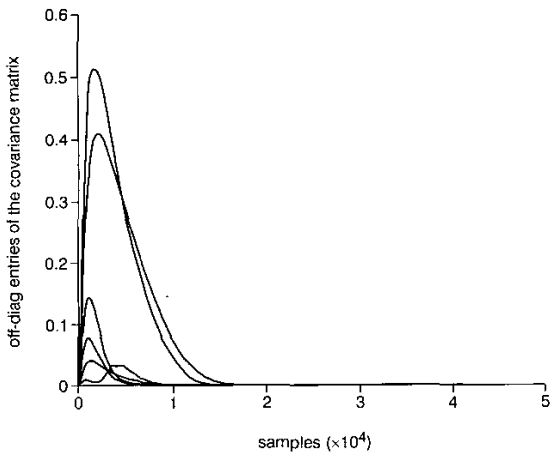


**Fig. 2** *Three sources problem: off-diagonal entries of network's output estimated covariance matrix (FAN-based algorithm)*

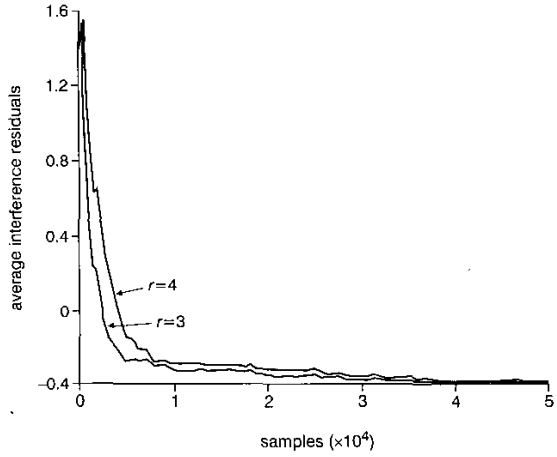


**Fig. 4** *Three sources problem: interference residuals of FAN algorithm obtained with $r=3$ and $r=4$*

three signals: $s_1(t)=\text{sign}[\cos(500t+9\cos(50t))]$, $s_2(t)$ is a uniformly distributed white noise in the range $[-1, +1]$, and $s_3(t)$ is an 8 kHz sampled speech signal. The signal $s_1(t)$ has been chosen because in [41] Gustaffson reported that the original Bell–Sejnowski algorithm may be ineffective in the presence of it. The learning step-size was $\eta_W=10^{-4}$, the neural network had three inputs, three outputs and thus three adaptive neurons, with $r=2$.

As the separation technique used does not require prewhitening, it is worth verifying that it exhibits a self-whitening ability; in Fig. 2 the off-diagonal entries of the estimated network's output covariance matrix is shown versus time. As they tend to zero, the FAN-based algorithm proves to provide data whitening with no additional computational effort. Fig. 3 shows the entries of the separation product $\boldsymbol{P}(t)$. In this case, the separation product has nine entries: six of them converge to about zero; note that in this technique the power of the network's outputs is uncontrolled, thus the three remaining non-zero elements of $\boldsymbol{P}(t)$ may converge to arbitrary values. It is interesting to note that the algorithm is able to operate in the presence of non-stationary signals, such as the sampled speech. This is clearly evidenced by the top trace which, after convergence, is slowly varying: it corresponds to the extracted speech source signal, which possesses non-stationary statistics that the learning algorithm is able to follow, i.e. the network parameters keep adjusting after convergence to a separating connection pattern to lock the signal's variable pdf; the same does not happen for the other two source signals that are wide-sense stationary.

An important structural question concerns the choice of the expansion degree $r$. It is quite apparent that an underestimated expansion degree renders the network unable to approximate the true pdfs in a suitable way, thus an overestimated degree is preferred, instead. To test for the effect of using a larger polynomial degree than required, we performed two simulations with the three source signals of the previous experiment 3, using $r=3$ and $r=4$. The results are shown in Fig. 4. They show that the separation algorithm is able to converge to the expected solution with no degradation of performance.

It would also be interesting to investigate the behaviour of the proposed separating neural structure on a larger number of real-world signals. We ran the FAN-based algorithm on five source-streams: a pure tone, a pulse train, a female speech signal and a male speech signal sampled at 11 kHz, and a piano sound sampled at 11 kHz. The interference residual during learning is shown in Fig. 5.

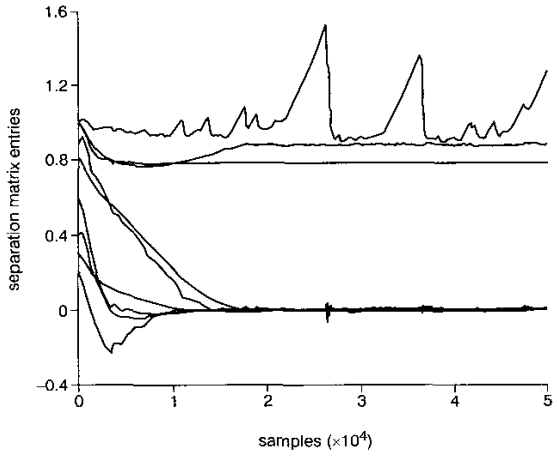


**Fig. 3** *Three sources problem: entries of the separation product $\boldsymbol{P}(t)$ (FAN-based algorithm)*

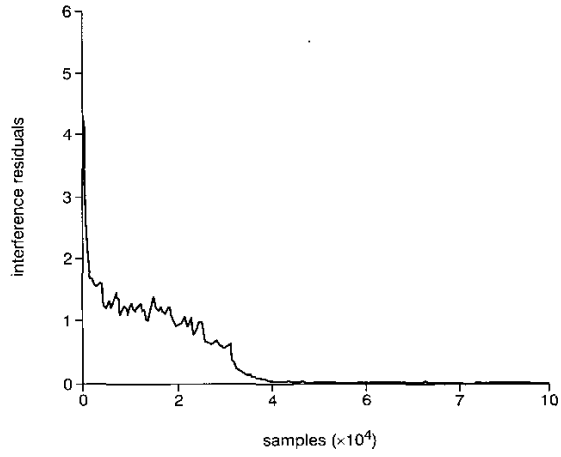


**Fig. 5** *Five sources problem: interference residual for FAN-based algorithm*

**Table 4: Results of MOK on the three experiments after 50 000 iterations**

| Experiment | 1 | 2 | 3 |
|---|---|---|---|
| SNR (dB) | $NSR_1 = -48.09$<br>$NSR_2 = -40.88$ | $NSR_1 = -35.56$<br>$NSR_2 = -44.55$ | $NSR_1 = -56.80$<br>$NSR_2 = -33.93$<br>$NSR_3 = -61.67$ |
| Kernels ($n$) | 2 | 7 | 5 |

In this case the mixing matrix was generated randomly and we used 100 000 mixed signal samples because more difficult convergence was expected. Even in this case, the algorithm was capable of separating out the five mixed sources from the mixtures.

### 5.2 Comparison of MOK and FAN-based separation algorithms

As a first comparison of the behaviours of MOK and FAN-based separation algorithms, in the following the results obtained by running the MOK algorithm in the three cases as described in Table 2 are presented. Table 4 summarises the NSR resulting from the experiments with the indicated number of kernels $n$. These results show that similar performances may be achieved, but the difference lies in the cost implied by the use of $n = 5$, 7 kernels for MOK against the low-order ($r = 1$, 2) polynomials of FAN.

So far we have considered the static performances only, i.e. NSR after 50 000 iterations. It could be interesting to analyse the average dynamical behaviour of the algorithms. The interference residuals shown in simulations have been averaged over 10 trials; over each trial, the mixing matrix was randomly generated by uniformly picking its entries in the interval [0, 1].

In experiment 1, for each neuron we had $n = 7$, $r = 1$ and $\eta_Q = 0.001$. Fig. 6 shows the interference residuals of MOK and FAN algorithms.

In experiment 2, the testing conditions were similar to those of experiment 1, except for the learning rates that were set to $\eta_Q = 0.0005$. The curves, Fig. 7, have the same meaning as in the preceding experiment.

In experiment 3, for the MOK algorithm, each neuron had $n = 5$, while for the FAN-based algorithm, each neuron had $r = 2$ and $\eta_Q = 0.001$. The results are shown in Fig. 8.

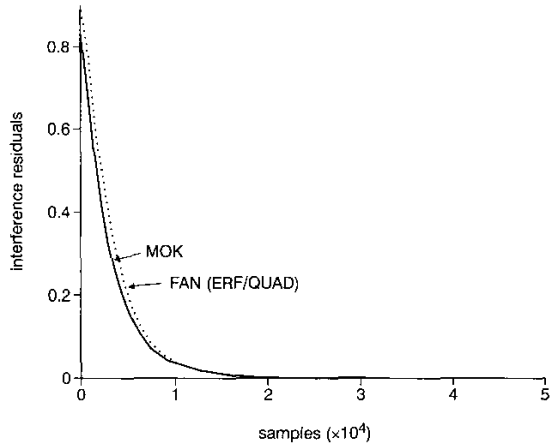


**Fig. 6** *Comparative experiment 1*

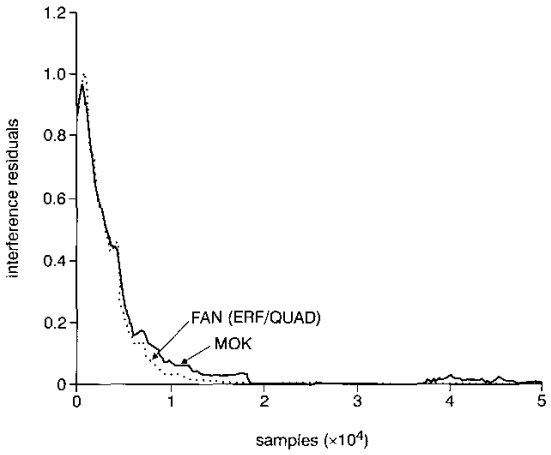


**Fig. 7** *Comparative experiment 2*

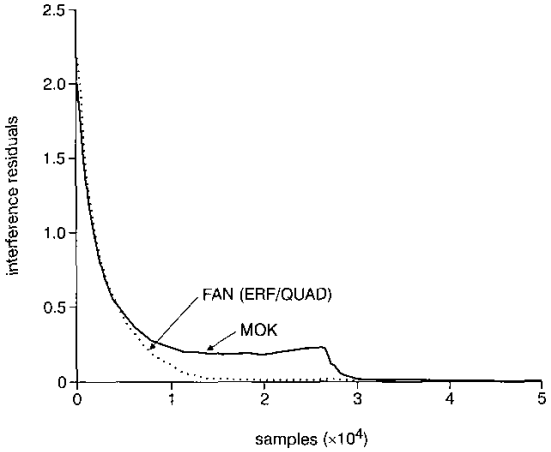


**Fig. 8** *Comparative experiment 3*

Along with the non-aggregate average results, it would be interesting to give the mean separation performance and its standard deviation: for the mean we found about −45 dB with standard deviation of about 11 dB. In spite of the high variability of the results, they are usually below −30 dB, which is a good value in, for example, speech processing.

## 6 Conclusions

In this paper we have focused on the problem of blind source separation by the independent component analysis technique, when eterokurtic source signals are dealt with. A solution was proposed based on an unsupervised adaptive activation function network which adapts through an information-theoretic based learning theory. A new eterokurtic ICA algorithm was developed which is as accurate as Xu's (as shown in Figs. 6 and 7) but is computationally simpler and has fewer tunable parameters (as shown in Table 1). In fact, both numerical simulation results, performed on synthetic and real-world data, and structural comparisons, showed the proposed approach to be effective and interesting from a computational complexity point of view. Applications of the presented theory are at present being extended to blind system deconvolution and channel equalisation, as well as to complex-valued data processing.

## 7 Acknowledgments

The author wishes to thank Dr P. Baldassarri who kindly prepared the experimental setup for Section 5, the anonymous reviewers and the Honorary Editor Dr S. McLaughlin who provided in-depth comments and very useful suggestions for the final version of the paper.

## 8 References

1 CLEMENTE, R.M., and ACHA, J.I.: 'Blind separation of sources using a new polynomial equation', *Electron. Lett.*, 1997, **33**, (3), pp. 176–177
2 FIORI, S., and BURRASCANO, P.: 'Electromagnetic environmental pollution monitoring: source localization by the independent component analysis'. Proceedings of Third International Conference on Independent component analysis and signal separation, 2001, pp. 575–580
3 FIORI, S., and BURRASCANO, P.: 'ECT-data fusion by the independent component analysis for non-destructive evaluation of metallic slabs'. Proceedings of Third International Conference on Independent component analysis and signal separation, 2001, pp. 323–327
4 GIANNAKOPOULOS, X., KARHUNEN, J., and OJA, E.: 'An experimental comparison of neural algorithms for independent component analysis and blind separation', *Int. J. Neural Syst.*, 1999, **9**, (2), pp. 99–114
5 HYVÄRINEN, A., KARHUNEN, J., and OJA, E.: 'Independent component analysis' (John Wiley & Sons, 2001)
6 LEE, T.-W.: 'Independent component analysis — theory and practice' (Kluwer Academic Publisher, 1998)
7 LIU, R.-W.: 'Blind signal processing: An introduction'. Proceedings of International Symposium on Circuits and systems (IEEE-ISCAS), 1996, vol. 2, pp. 81–84
8 JUTTEN, C., and HÉRAULT, J.: 'Independent component analysis versus principal component analysis'. Proceedings of European Symposium on Signal processing (EUSIPCO), 1988, vol. 2, pp. 643–646
9 JUTTEN, C., and HÉRAULT, J.: 'Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture', *Signal Process.*, 1991, **24**, pp. 1–10
10 CICHOCKI, A., and UNBEHAUEN, R.: 'Robust neural networks with on-line learning for blind identification and blind separation of sources', *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, 1996, **43**, pp. 894–906
11 COMON, P.: 'Independent component analysis, A new concept?', *Signal Process.*, 1994, **36**, pp. 287–314
12 COMON, P., and MOREAU, E.: 'Improved contrast dedicated to blind separation in communications'. Proceedings of International Conference on Acoustics, speech and signal processing, 1997, pp. 3453–3456
13 MOREAU, E., and MACCHI, O.: 'High-order contrasts for self-adaptive source separation', *Int. J. Adapt. Control Signal Process.*, 1996, **10**, pp. 19–46
14 CARDOSO, J.-F., and LAHELD, B.: 'Equivariant adaptive source separation', *IEEE Trans. Signal Process.*, 1996, **44**, (12), pp. 3017–3030
15 DELFOSSE, N., and LOUBATON, P.: 'Adaptive blind separation of independent sources: A deflation approach', *Signal Process.*, 1995, **45**, pp. 59–83
16 THAWONMAS, R., CICHOCKI, A., and AMARI, S.-I.: 'A cascade neural network for blind signal extraction without spurious equilibria', *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, 1998, **E81-A**, (9), pp. 1833–1846
17 LINSKER, R.: 'Local synaptic rules suffice to maximize mutual information in a linear network', *Neural Comput.*, 1992, **4**, pp. 691–702
18 PLUMBLEY, M.D.: 'Efficient information transfer and anti-hebbian neural networks', *Neural Netw.*, 1993, **6**, pp. 823–833
19 NADAL, J.P., BRUNEL, N., and PARGA, N.: 'Nonlinear feedforward networks with stochastic inputs: Infomax implies redundancy reduction, Netw.', *Comput. Neural Syst.* 1998, **9**
20 BARAM, Y., and ROTH, Z. 'Density shaping by neural networks with application to classification, estimation and forecasting'. Tech. Rep. CIS-94-20, Center for Intelligent Systems, Technion, Israel Institute for Technology, Haifa, 1994
21 BELL, A.J., and SEJNOWSKI, T.J.: 'An information maximisation approach to blind separation and blind deconvolution', *Neural Comput.*, 1996, **7**, (6), pp. 1129–1159
22 GIROLAMI, M., and FYFE, C.: 'Extraction of independent signal sources using a deflationary projection pursuit network with lateral inhibition', *IEE Proc., Vis. Image Signal Process.*, 1997, **14**, (5), pp. 299–306
23 FIORI, S.: 'Blind separation of circularly distributed source signals by the neural extended APEX algorithm', *Neurocomputing*, 2000, **34**, (1–4), pp. 239–252
24 FIORI, S.: 'On blind separation of complex-valued sources by extended hebbian learning', *IEEE Signal Process. Lett.*, 2001, **8**, (8), pp. 217–220
25 KARHUNEN, J., OJA, E., WANG, L., VIGÁRIO, R., and JOUTSENSALO, J.: 'A class of neural networks for independent component analysis', *IEEE Trans. Neural Netw.*, 1997, **8**, (3), pp. 486–504
26 HYVÄRINEN, A., and OJA, E.: 'Independent component analysis by general nonlinear hebbian-like rules', *Signal Process.*, 1998, **64**, (3), pp. 301–313
27 SAGI, B., NEMAT-NASSER, S.C., KERR, R., HAYEK, R., DOWNING, C., and HECHT-NIELSEN, R.: 'A biologically motivated solution to the cocktail party problem', *Neural Comput.*, 2001, **13**, (7), pp. 1575–1602
28 FIORI, S.: 'A Theory for learning by weight flow on Stiefel–Grassman manifold', *Neural Comput.*, 2001, **13**, (7), pp. 1625–1647
29 FIORI, S.: 'A theory for learning based on rigid bodies dynamics', *IEEE Trans. Neural Netw.*, 2002, **13**, (3), pp. 521–531
30 FIORI, S.: 'Complex-weighted one-unit 'rigid-bodies' learning rule for independent component analysis', *Neural Process. Lett.*, 2002, **15**, (3), pp. 275–282
31 PRIETO, A., and PUNTONET, B.: 'A neural learning algorithm for blind separation of sources based on geometric properties', *Signal Process.*, 1998, **64**, (3), pp. 315–331
32 YOSHIOKA, M., and OMATU, S.: 'Signal separation method using genetic algorithm'. Proceedings of International Joint Conference on Neural networks (IJCNN'98), 1998, pp. 909–912
33 XU, L., CHEUNG, C.C., RUAN, J., and AMARI, S.-I.: 'Nonlinearity and separation capability: Further justifications for the ICA algorithm with a learned mixture of parametric densities'. Proceedings of European Symposium on Artificial neural networks, 1997, pp. 291–296
34 XU, L., CHEUNG, C.C., and AMARI, S.-I.: 'Learned parametric mixture based ICA algorithm', *Neurocomputing*, 1998, **22**, (1–3), pp. 69–80
35 YANG, H.H., and AMARI, S.-I.: 'Adaptive online learning algorithms for blind separation: Maximum entropy and minimum mutual information', *Neural Comput.*, 1997, **9**, pp. 1457–1482
36 EVERSON, R.M., and ROBERTS, S.J.: 'ICA: A flexible non-linearity and decorrelating manifold approach', *Neural Comput.*, 1999, **11**, (8), pp. 1957–1983
37 FRUSLEBEN, B., HAGEN, C., and BORSCBACH, M.: 'A neural network for the blind separation of non-Gaussian sources'. Proceedings of International Joint Conference on Neural networks (IJCNN'98), 1998, pp. 837–842
38 PEARLMUTTER, B.A., and PARRA, L.C.: 'Maximum likelihood blind source separation: A context-sensitive generalization of ICA'. Proceedings of Neural Information Processing System (NIPS*9), Denver, CO, USA, 1996, pp. 613–619
39 TALEB, A., and JUTTEN, C.: 'Entropy optimization—Application to source separation'. Proceedings of International Conference on Artificial neural networks, Lausanne, Switzerland, 1997, pp. 529–534
40 ROTH, Z., and BARAM, Y.: 'Multidimensional density shaping by sigmoids', *IEEE Trans. Neural Netw.*, 1996, **7**, (5), pp. 1291–1298
41 GUSTAFSON, M.: 'Gaussian mixture and kernel based approach to blind source separation using neural networks'. Proceedings of International Conference on Artificial neural networks, 1998, Springer-Verlag, vol. 2, pp. 869–874
42 WELLING, M., and WEBBER, M.: 'A constrained E.M. algorithm for independent component analysis', *Neural Comput.*, 2001, **13**, (3)
43 KARVANEN, J. ERIKSSON, J. KOIVUNEN, V.: 'Pearson system based method for blind separation'. Proceedings of Second International Workshop on Independent component analysis and blind signal separation (ICA), 2000, Helsinki, Finland, pp. 585–590
44 CHEN, C.T., and CHANG, W.D.: 'A feedforward neural network with function shape autotuning', *Neural Netw.*, 1996, **9**, (4), pp. 627–641
45 FIORI, S.: 'Probability density function learning by unsupervised neurons', *Int. J. Neural Syst.*, 2001, **11**, (5), pp. 399–417
46 VECCI, L., PIAZZA, F., and UNCINI, A.: 'Learning and approximation capabilities of adaptive spline activation function neural networks', *Neural Netw.*, 1998, **11**, (2), pp. 259–270
47 FIORI, S.: 'Blind source separation by new M-WARP algorithm', *Electron. Lett.*, 1999, **35**, (4)
48 FIORI, S., BALDASSARRI, P., and PIAZZA, F.: 'An efficient architecture for independent component analysis'. Proceedings of International Symposium on Circuits and systems, 1999, vol. V, pp. 335–338
49 FIORI, S., and BUCCIARELLI, P.: 'Probability density estimation using adaptive activation function neurons', *Neural Process. Lett.*, 2001, **13**, (1), pp. 31–42
50 PIAZZA, F., UNCINI, A., and ZENOBI, M.: 'Neural networks with digital LUT activation function'. Proc. of International Joint Conference on Neural Networks (IJCNN'93), Nagoya, Japan, Oct. 1993, pp. 1401–1404