Neural Independent Component Analysis by 'Maximum-Mismatch' Learning Principle

Simone Fiori

Faculty of Engineering - University of Perugia
Loc. Pentima bassa, 21, I-05100 Terni (Italy)

E-mail: sfr@unipg.it

Keywords:
Generalized Hebbian learning; Sudjianto-Hassoun theory;
Independent component analysis; Telecommunication signals.

Pages: 42, **Figures**: 18, **References**: 39.

Contribution to appear on: Neural Networks

Submitted February 2002. Revised September 2002.
Accepted January 2003.

Neural Independent Component Analysis by 'Maximum-Mismatch' Learning Principle

Simone Fiori

Abstract

The aim of the present paper is to apply Sudjanto-Hassoun theory of Hebbian learning to neural independent component analysis. The basic learning theory is first recalled and expanded in order to make it suitable for a network of non-linear complex-weighted neurons; then its interpretation and application is shown in the context of blind separation of complex-valued sources. Numerical results are given in order to assess the effectiveness of the proposed learning theory and the related separation algorithm on telecommunication signals; a comparison with other existing techniques finally helps assessing the performances and computational requirements of the proposed algorithm.

1 Introduction

Since the pioneering work of D.O. Hebb, several generalizations have been proposed to simple correlation rule, leading to interesting and elegant adapting theories.

The research efforts about generalized Hebbian rules have given remarkable results especially in the field of unsupervised learning, by providing local and structurally-simple self-organization algorithms. Perhaps, the most know theory in this field concerns the Oja-Amari principal-component learning rule [2, 32], which appears as a stabilized Hebbian paradigm, and allows extracting the principal eigenvector of the covariance matrix of a neuron's multivariate input signal.

Formally, the generalization with respect to the original correlation rule consists in considering different non-linear activation functions and non-quadratic learning criteria, which are designed to drive a neuron to learn different tasks. An interesting example is robust (non-linear) principal component analysis: Some notes have been given for instance in [13, 33].

In particular, within the paper [39], Sudjianto and Hassoun experimentally investigate the behavior of non-classical Hebbian learning over two mixtures of a signal and a disturbance, and report the ability of the neuron to extract the

useful signal from the linear combinations provided that the noise is statistically characterized. In turn, the Sudjianto-Hassoun learning principle is a particular case of learning in non-classical constrained Hebbian networks studied by Oja, Ogawa and Wangviwattana in [35], where the networks are used for feature extraction from input signals in an unsupervised or self-organizing mode, assuming that the features are some non-linear functions of the inputs themselves.

It is interesting to note that in [39] an innovative learning paradigm was introduced for unsupervised signal filtering which, instead of trying to match a neuron to the desired signal, tends to match it to the disturbance and force the neuron to filter out the statistics of what is different from the disturbance.

We believe it is interesting to investigate this learning paradigm, termed Sudjianto-Hassoun 'maximum-mismatch' (SHMM) principle, in order to better understand its capabilities and properties, in relation to blind signal separation by the independent component analysis.

The aim of blind source separation is to recover independent constituents from a complex multiple signal. The activity about this problem finds its roots in known biological facts, related to the manner that the nerves transmit mixed information to the brain, and to the way the brain interprets such mixed signals [29]. However, this research stream has found important engineering applications and has benefited from cross-fertilization among neural networks and information-theoretic signal processing fields.

Examples of well-established applications of blind source separation by the independent component analysis find in speech recognition, telecommunications, fault detection, medical imaging, financial data market analysis, remote sensing and other (for a recent review see e.g. [11, 13, 16, 20, 21, 26]).

Blind source separation by extensions of classical Hebbian learning theories has recently received particular attention: It has been proven by many papers that adding non-linearity to linear Hebbian learning neural networks makes them able to improve the independence of their outputs so as to allow blind separation of real-valued independent sources.

Also, some of the independent component analysis approaches may work well on complex-valued signals. This holds for example for Cardoso's JADE and EASI [6, 7], the 'fixed-point' algorithm that can be applied to complex signals [5], and the ψ–APEX algorithms [12] that may be extended to the complex-valued case [10].

In this work we formally derive a new learning algorithm as a non-linear

complex-valued extension of generalized Hebbian rule [38] for a linear feed-forward network and discuss the choice of the non-linearity under the theoretical framework proposed by Sudjianto and Hassoun [39]. Then we show – analytically and numerically – how a particular non-linearity, arising from the Rayleigh distribution, allows the neural network to perform blind separation of complex-valued circular source signals.

In order to numerically assess the behavior of the proposed method, computer simulation results are presented and discussed, along with a numerical comparison among closely related algorithms drawn from the scientific literature.

Organization of the paper: Section 2 is devoted to the explanation of the basic SHMM principle and to its formalization, aimed at the subsequent detailed study of general applicability and suitability to drive a network to perform independent component analysis (ICA); such section also contains a discussion on closely-related theories drawn from the scientific literature; the general concepts of ICA for complex-valued sources is also surveyed in this section, along with the basic observations about ICA by generalized Hebbian algorithms. Section 3 is devoted to the detailed study of SHMM principle in two meaningful cases, in which we are allowed to write the complicated learning criterion in closed form and henceforth to analytically study its essential properties. In section 4, the application of SHMM theory to blind separation of complex-valued sources by the independent component analysis is issued; as from sections 2-3 to section 4 the fundamental extension of SHMM principle from a real-valued single-unit system to a complex-weighted multiple-unit separation system is carried out, a necessary general result ensuring the suitability of the novel theory to blind separation is given. Section 5 is devoted to the presentation of numerical results aimed at the assessment of the effectiveness of the proposed learning theory and of the related separation algorithm for telecommunication signals; a comparison with other existing techniques also helps assessing the performances and computational requirements of the proposed algorithm. Section 6 concludes the paper.

2 Principles and Basics

In the present section we formally recall, expand and discuss the Sudjianto-Hassoun learning principle in order to prepare the subsequent application to

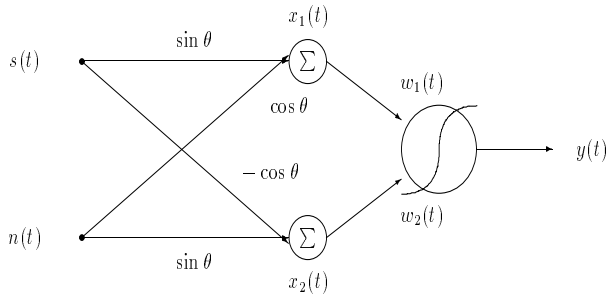

Figure 1: Schematic of signal generation network and neuron model.

independent component analysis. We also describe the blind source separation problem, with particular emphasis to the separation of complex-valued sources. The two research streams conjunct in section 4.

2.1 Sudjianto-Hassoun 'maximum-mismatch' learning principle

Let us consider two stationary zero-mean random signals $s(t) \in \mathcal{R}$ and $n(t) \in \mathcal{R}$, where $n(t)$ denotes a disturbance (or noise) with known probability density function $p_n(n)$. Let us further suppose that both signals are singularly unobservable, while only linear combinations of them are available, namely $x_1(t) \in \mathcal{R}$ and $x_2(t) \in \mathcal{R}$. The aim is to separate out the useful signal $s(t)$ from the combinations by means of a neural unit endowed with two inputs, two learnable weights denoted by $w_1 \in \mathcal{R}$ and $w_2 \in \mathcal{R}$, and one output denoted by $y(t) \in \mathcal{R}$, as graphically shown in Figure 1. The neuron transfer function writes $y = \sigma(z)$, where $z(t) \in \mathcal{R}$ denotes the net input to the neuron and computes as $z = w_1 x_1 + w_2 x_2$, while $\sigma(\cdot)$ stands for the activation function of the neural unit.

Following Sudjianto and Hassoun, we make the neuron learn through a criterion $\mathbb{E}_{s,n}[y^2]$ under the necessary consistency constraint $w_1^2 + w_2^2 = 1$, where the symbol $\mathbb{E}_x[f(x)]$ denotes the expectation value of a function of the random variable x with respect to its statistics and the constraint prevents the weights from vanishing to zero or growing indefinitely as may happen in the original Hebbian-learning setting.

It is worth recalling that the original Hebb's learning paradigm arises as a correlation rule between the stimuli and the answers of neuronal pairs. Formally, such rule was found to lead to unstable synapses, therefore Oja proposed in [32]

a stabilized Hebbian learning rule obtained as a first-order approximation of a normalized Hebbian rule – endowed with a limited-resource criterion – in the (small) learning step-size parameter. Years before, in [2] Amari had proposed an optimization principle for unsupervised learning that led to Oja's rule, while later on Karhunen and Oja revisited the same learning principle in order to define useful generalization of Oja's rule (see e.g. [30] and references therein).

The constrained expression of the learning criterion is:

$$C(w_1, w_2) = \mathbb{E}_{s,n}[\sigma^2(w_1 x_1 + w_2 x_2)] + \gamma(w_1^2 + w_2^2 - 1) \ , \tag{1}$$

where γ is the Lagrange multiplier introduced in order to enforce the consistency constraint.

Again following Sudjianto and Hassoun, the non-linear activation function is chosen in a particular way, that is:

$$\sigma(z) = Q_n(z) \stackrel{\text{def}}{=} 2 \int_{-\infty}^{z} p_n(\nu) d\nu - 1 \ . \tag{2}$$

It is easily recognizable that, apart from the scaling and shifting factors, the $Q_n(\cdot)$ has been chosen as the cumulative distribution function associated to the random noise $n(t)$. It is also worth noting that, irrespective of the values taken on by the signals and the weights, the neuron output value in this case always range in $[-1, +1]$. This choice for the neuron's activation function represents indeed the key-point in the SHMM theory, as demonstrated in the continuation of the discussion, and it also represents the key-point enabling us to properly select the non-linear activation function for the separating neurons in the generalized-Hebbian ICA net under construction.

The circumstance that $\sigma(z)$ is bounded by fixed limits plays a central role in SHMM theory, due to the following observation: Let us suppose the neuron weights are hand-crafted to the combination which makes $z(t) = n(t)$; in this case, the random noise enters the non-linearity $Q_n(\cdot)$ which, in turn, generates a new random signal that can be easily characterized from a statistical point of view. In fact, it is known that the cumulative distribution function associated to a signal warps the signal itself producing a *uniformly distributed noise* [37], in this case ranging in $[-1, +1]$; under fair conditions, we are ensured that, for a given signal, this function is the only one enjoying such noticeable property. In other words, the non-linearity behaves as a probabilistic filter that is able to detect a signal whose probability density function is matched to its shape by producing a uniform noise.

Let us now qualitatively consider what is the behavior of the neuron when the learnable weights adapt by maximizing the criterion function (1). The neuron seeks for a combination of the weights that maximizes its response variance; as the response is limited between -1 and $+1$ it is probable that the values of neuron's output tend to accumulate around such bounds, that is equivalent to say that the distribution $p_y(y)$ of the response values becomes *U-shaped* [39]. If this is the result of neuron's learning, we can guess that $z(t)$ will likely be different from $n(t)$, because we have already demonstrated that $z(t) = n(t)$ corresponds to a flat-shaped output distribution. From these informal considerations, Sudjianto and Hassoun infer that the maximization of criterion (1) might lead the neuron to prefer the unmatched signal $s(t)$, that is the useful one. Because of its peculiarity, we refer to this principle as Sudjianto-Hassoun maximum-mismatch.

As mentioned, these informal considerations will lead to working algorithms which allow separating a number of independent signals from their linear mixtures under the hypothesis that one of them is a disturbance of known statistics, as can be environmental noise in speech processing or channel noise in telecommunications. The aim of the first part of this paper is to formally investigate the consistency of the SHMM principle.

The SHMM learning equations for the mentioned neural structure are derived from a gradient-based optimization of the criterion (1), namely:

$$\Delta w_1 = +\frac{\mu}{2}\left(\frac{\partial C}{\partial w_1}\right)^{\text{opt}} \ , \ \Delta w_2 = +\frac{\mu}{2}\left(\frac{\partial C}{\partial w_2}\right)^{\text{opt}} \ , \tag{3}$$

where $\mu > 0$ denotes a learning step-size. Provided that the parameters space is endowed with the Euclidean metric, the optimal gradient computes by the gradient of the criterion C:

$$\frac{\partial C}{\partial w_1} = 2\mathbb{E}_{s,n}[\sigma(z)\sigma'(z)x_1] + 2\gamma w_1 \ , \ \frac{\partial C}{\partial w_2} = 2\mathbb{E}_{s,n}[\sigma(z)\sigma'(z)x_2] + 2\gamma w_2 \ ,$$

where the optimal Lagrange multiplier finds by the steady-state condition[1]

[1] The employed condition comes from the well-known elimination rule for the Lagrange multipliers: We can imagine a particle whose position is described by the coordinate-pair (w_1, w_2), whose equations of motion read $\dot{w}_1 = \frac{1}{2}\frac{\partial C}{\partial w_1}$, $\dot{w}_2 = \frac{1}{2}\frac{\partial C}{\partial w_2}$; if the coordinates must satisfy the constraint $w_1^2 + w_2^2 = 1$ at any time, by deriving both members of this equation we get the corresponding constraint for the velocities, namely $w_1\dot{w}_1 + w_2\dot{w}_2 = 0$; by replacing the equations of motion into the velocity-constraints, the mentioned steady-state condition is readily obtained.

$w_1 \frac{\partial C}{\partial w_1} + w_2 \frac{\partial C}{\partial w_2} = 0$, which gives:

$$\gamma^{\mathrm{opt}} = -\mathbb{E}_{s,n}[\sigma(z)\sigma'(z)z] \ .$$

Consequently, the optimal gradient components write:

$$\left(\frac{\partial C}{\partial w_1}\right)^{\mathrm{opt}} = 2\mathbb{E}_{s,n}[\sigma(z)\sigma'(z)(x_1 - zw_1)] \ , \tag{4}$$

$$\left(\frac{\partial C}{\partial w_2}\right)^{\mathrm{opt}} = 2\mathbb{E}_{s,n}[\sigma(z)\sigma'(z)(x_2 - zw_2)] \ . \tag{5}$$

It is worth noting that the learning theory derivation is essentially similar to that of Oja-Amari, except for the non-linear neural transfer function shape; in fact, by letting $\sigma(z) = z$ the equations particularize into the well-known principal-component rule [2, 32].

In order to carry out a formal analysis of the behavior of SHMM principle applied to a single neuron's learning, let us consider the following technical hypotheses on the described signals and separation problem:

- The signal s and noise n possess symmetric statistical distributions, namely $p_n(n) = p_n(-n)$ and $p_s(s) = p_s(-s)$;

- The composition of the signals is orthonormal, i.e. the relationship between s, n, x_1 and x_2 writes:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} \sin\theta & \cos\theta \\ -\cos\theta & \sin\theta \end{bmatrix} \begin{bmatrix} s(t) \\ n(t) \end{bmatrix} \ , \tag{6}$$

  where the angle θ is the only parameter describing the combination. Note that any full-rank non-orthonormal linear combination may be reduced to orthonormal combination by pre-whitening the observed signals [10, 14];

- Because of the consistency constraint imposed on the neuron's weights, we may assume for the weight-set the natural parameterization $[w_1 \ w_2] = [\cos\psi \ \ \sin\psi]$, where the angle ψ is the only neuron's free parameter.

On the basis of these settings, the neuron's activation writes:

$$z = s\cos\varepsilon + n\sin\varepsilon \ , \tag{7}$$

where the angle $\varepsilon \stackrel{\mathrm{def}}{=} \theta - \psi$ denotes neuron's angular deviation.

By denoting as $p_z(z;\varepsilon)$ the probability density function of neuron's activation, the learning criterion (1) recasts into:

$$L(\varepsilon) \stackrel{\text{def}}{=} \int_{-\infty}^{+\infty} \sigma^2(z) p_z(z;\varepsilon) dz \ . \tag{8}$$

The density function $p_z(z;\varepsilon)$ may be computed exactly by using the theorems about probability transform upon scaling and summation of random variables, and has the expression:

$$p_z(z;\varepsilon) = \frac{1}{|\cos\varepsilon \sin\varepsilon|} \int_{-\infty}^{+\infty} p_s\left(\frac{\zeta}{\cos\varepsilon}\right) p_n\left(\frac{z-\zeta}{\sin\varepsilon}\right) d\zeta \ . \tag{9}$$

It is useful to note that the function $p_z(z;\varepsilon)$ has some symmetry in ε. In fact, it is easily seen that if $\varepsilon \in [0,\pi]$ then $p_z(z;-\varepsilon) = p_z(z;\varepsilon)$: This accounts for the well-known sign-blindness of independent component analysis. Moreover, let $\varepsilon = \frac{\pi}{2} \pm \alpha$, with $\alpha \in [0,\pi/2]$: straightforward calculi show that:

$$p_z(z;\frac{\pi}{2} \pm \alpha) = \frac{1}{|\cos\alpha \sin\alpha|} \int_{-\infty}^{+\infty} p_s\left(\frac{\zeta}{\sin\alpha}\right) p_n\left(\frac{z-\zeta}{\cos\alpha}\right) d\zeta \ .$$

As a consequence, we may restrict our analysis of the criterion function $L(\varepsilon)$ to the interval $[0,\pi/2]$ without loss of generality.

With this convention, the criterion function writes:

$$\begin{aligned} L(\varepsilon) &= \int_{-\infty}^{+\infty} Q_n^2(\xi) \int_{-\infty}^{+\infty} \frac{1}{\cos\varepsilon} p_s\left(\frac{\zeta}{\cos\varepsilon}\right) \frac{1}{\sin\varepsilon} p_n\left(\frac{\xi-\zeta}{\sin\varepsilon}\right) d\zeta d\xi \\ &= \iint_{\mathcal{R}^2} Q_n^2(u\sin\varepsilon + v\cos\varepsilon) p_s(u) p_n(v) du dv \ . \end{aligned} \tag{10}$$

It is important to note that the last expression in (10) may be obtained from expressions (8)-(9) by variable changes which make sense only for $\varepsilon \neq 0$ and $\varepsilon \neq \pi/2$; for these two limit-cases the following results hold:

$$\lim_{\varepsilon\to 0} \frac{1}{\sin\varepsilon} p_n\left(\frac{x}{\sin\varepsilon}\right) = \lim_{\varepsilon\to \pi/2} \frac{1}{\cos\varepsilon} p_s\left(\frac{x}{\cos\varepsilon}\right) = \delta(x) \ ,$$

thanks to the properties of the probability density functions, where the symbol $\delta(x)$ denotes the Dirac's delta. This allows concluding that the equivalence between the two rows of equation (10) holds in the whole interval $[0,\pi/2]$.

From expression (7) it is readily recognized that in order for the neuron's activation to coincide with the signal $s(t)$, the contribution of the noise should peter out, which ultimately means $\varepsilon = 0$. As a conclusion, the SHMM principle

is effective when and only when the maximum of the criterion $L(\varepsilon)$ finds in the origin.

The criterion $L(\varepsilon)$ has in general a complex structure that prevents us from carrying out any general investigation. Two meaningful cases-study, aimed at clarifying the behavior of SHMM principle and to extract some general considerations on its applicability, are investigated with details in section 3.

The only useful general consideration is that for $\varepsilon = \pi/2$ it holds:

$$L\left(\frac{\pi}{2}\right) = \iint_{\mathcal{R}^2} Q_n^2(\xi) p_s(\zeta) p_n(\xi) d\zeta d\xi = \int_0^1 Q_n^2 dQ_n = \frac{1}{3} \ . \tag{11}$$

Unfortunately, the value of $L(0)$ cannot be computed in closed form unless p_n and p_s are specified, but it is clear that a necessary condition in order for $L(\varepsilon)$ to possess a only maximum in $\varepsilon = 0$ is $L(0) > L(\pi/2)$, namely $3L(0) > 1$.

2.2 Blind separation of complex-valued sources

In the present context, blind source separation concerns the problem of recovering source signals mixed by a linear operator when the mixing is unknown and very little information is available on the sources. Formally, the model of observed signals writes:

$$\mathbf{x}(t) = \mathbf{M}^H \mathbf{s}(t) \ , \ t \in \mathcal{T} \ , \tag{12}$$

where $\mathbf{s}(t) \in \mathcal{K}^q$ contains the source signals, $\mathbf{M} \in \mathcal{K}^{q \times m}$ is the mixing operator and $\mathbf{x}(t) \in \mathcal{K}^m$ is the only observable (sensor) signal. The field $\mathcal{K}$ may be either $\mathcal{R}$ or $\mathcal{C}$ depending on the real-valued or complex-valued nature of the problem, and superscript "$\cdot^H$" denotes Hermitian transposition (ordinary transposition is instead denoted by superscript "$\cdot^T$"). In order for the separation problem to be consistent, the following hypotheses should be met [9]:

- Number $m \geq q$: The number of observations should exceed (or equate) the number of sources; in some special cases this hypothesis may be relaxed (e.g. in presence of binary sources [3]; also see the over-complete representation theory [31]);
- The matrix $\mathbf{M}$ should be full-column rank: In this case all the observations are not redundant and carry on valuable information on the sources;
- Each source signal $s_k(t)$ is ergodic;

- The s_k's are statistically independent at any time, meaning that $s_k(t)$ is statistically independent of $s_j(t)$ for any $k \neq j$ and any $t \in \mathcal{T}$;
- At most one, among the source signals, may be Gaussian.

Provided that these hypotheses are met, a way to estimate the source signals from the observed mixtures through an adaptive linear neural structure can be envisaged. The linear inverse model writes:

$$\mathbf{z}(t) = \mathbf{B}^H \mathbf{x}(t) \ , \ t \in \mathcal{T} \ , \tag{13}$$

where $\mathbf{z}(t) \in \mathcal{K}^q$ is the network response at time t and matrix $\mathbf{B}(t) \in \mathcal{K}^{m \times q}$ denotes the neural connection pattern at time t. Also, the time-set $\mathcal{T}$ may coincide with a subset of $\mathcal{R}$ or $\mathcal{Z}$. As we are interested in computer-based signal processing, in the present paper we always consider $\mathcal{T} \subset \mathcal{Z}$.

Solving a blind separation problem means finding the matrix $\mathbf{B}_\star$ such that $\mathbf{B}_\star^H \mathbf{M}^H = \mathbf{QD}$, where $\mathbf{Q}$ is an arbitrary permutation matrix and $\mathbf{D} \in \mathcal{K}^{q \times q}$ is diagonal invertible; it is in fact known [9] that, in general, the sources can be recovered but for scale and phase-shift distortions.

From the above discussion it is readily seen that, in general, $\mathbf{B}_\star$ has not any special form, but for the case that $\mathbf{M}$ is an orthogonal matrix, i.e. $\mathbf{M}^H \mathbf{M} = \mathbf{I}_q$, that implies $\mathbf{B}_\star$ is orthonormal, too; in this case the problem to solve is simpler.

It has been theoretically proven [9] that any arbitrary separation problem can be reduced to an orthogonal separation problem by *pre-whitening* the observations $\mathbf{x}(t)$, that means computing a pair of matrices $(\mathbf{E}, \mathbf{A})$ such that the new signal $\mathbf{v}(t) = \mathbf{A}^{-\frac{1}{2}} \mathbf{E}^H \mathbf{x}(t)$, $t \in \mathcal{Z}$, is spatially white, i.e. $\mathbb{E}[\mathbf{v}(t)\mathbf{v}^H(t)] = \mathbf{I}_q$. The matrix $\mathbf{V} \stackrel{\text{def}}{=} \mathbf{E}\mathbf{A}^{-\frac{1}{2}}$ defines the pre-whitening network[2] and the relationship $\mathbf{B}^H = \mathbf{V}^H \mathbf{W}^H$ defines the global separating network, where connection-matrix $\mathbf{W} \in \mathcal{K}^{q \times q}$ denotes the orthonormal separating network to be learnt. This approach has two important advantages:

- If the whitening operation is performed by means of optimal compression algorithms (e.g. through principal component analysis by the decomposition $\mathbb{E}[\mathbf{x}\mathbf{x}^H] = \mathbf{E}\mathbf{A}\mathbf{E}^H$), it gives an indirect way to estimate the true number of sources q (by thresholding the eigenvalues in $\mathbf{A}$), and therefore to reduce the dimension of the problem to the minimum. This approach is commonly used to filter out the observation noise (thus to increment the

[2]Note that $\mathbf{A}$ is diagonal and real (as well as positive-definite).

signal-to-noise ratio); it is worth noting that on-line principal component algorithms also allow *tracking* the actual number of sources in a non-stationary context [36]; it is also worth recalling that ad-hoc whitening techniques have been developed that do not exhibit any optimal compression ability but are simpler to implement and converge rapidly [6].

- After whitening, the equivalent mixing operator is orthonormal (and possibly square) thus its inversion is easier; also, the white multivariate signal $\mathbf{v}(t)$ has well-conditioned and bounded covariance matrix, that from a numerical point of view is a profitable advantage; for these reasons, also those blind separation algorithms that do not strictly require pre-whitening are often run on whitened data [27].

In the following we shall develop a learning theory under the hypothesis that the separation matrix is orthonormal; in the experiments, the general case will be tackled by properly pre-whitening the observations.

2.3 Existing approaches to blind separation by extended Hebbian learning

Since the pioneering work of Jutten and Hérault [28], several contributions have appeared on the scientific literature about blind source separation; in this paper we focus on the research stream related to extended Hebbian learning. Some contributions on the use of non-classical principal component techniques to blind source separation have been given in [5, 10, 25, 30] where various extensions of Hebbian and anti-Hebbian basic principles have been presented and adapted to signal separation and representation; in [8, 10, 28] new network architectures and learning procedures have been developed on the basis of Hebbian learning principle; in [34, 39] some non-heuristic non-linear functions have been suggested on the basis of maximum-likelihood principle and on a probabilistic filtering criterion, and in [27] the performances of various algorithms have been illustrated through a wide comparison.

The main weakness points of the mentioned approaches are that:

- The non-linear functions are generally chosen on the basis of existing contributions from robust statistics and on heuristic observations/findings;
- Their choice generally rely on the fact that using non-linear functions add to the system high-order statistical features but give little insight into how

these features are related to the separation problem;

- Because of the strong non-linearity of the learning equations, it is difficult to give any analytical proof of convergence nor detailed studies about the features of the employed learning criteria. Some formal results and notes on this topic are available in [4].

The main contribution of the SHMM learning theory to independent component analysis will be shown to allow designing the proper structure of non-linear part of the neurons in the separating network provided that some information is known on the statistical structure of signals that the source signals differ from. In other terms, this principle is useful when the statistical structure of the involved signals is not known in advance, but the structure of signals different from the ones to be separated out, such as noises, are available.

2.4 Relationship with exploratory projection pursuit

Exploratory projection pursuit (EPP) is a statistical data-analysis method aimed at identifying structure in high dimensional data. It relies on the projection of the data onto a low dimensional subspace in which the search for structure is carried out. However, not all projections reveal the data structure equally well, therefore in EPP an index is defined that measures how "interesting" a given projection is, and then the data are represented in terms of projections that maximize the mentioned index.

What constitutes "interesting" structure is usually defined according to the observation of Diaconis and Freedman (see e.g. [22] and references therein) that many projections of high-dimensional data onto arbitrary lines give almost Gaussian distributions. This suggests that if "interesting" features are looked for being identified in the data, we should look for those directions onto which the data-projections are as non-Gaussian as possible.

Two simple measures of deviation from a Gaussian distribution are based on the high-order moments of the distribution, such as skewness and kurtosis: Skewness is based on the normalized third-order moment of the distribution and measures the deviation of the distribution from symmetry, while kurtosis is based on the normalized fourth-order moment of the distribution and measures the heaviness of the tails of a distribution. As an exemplary application, a bimodal distribution often has negative kurtosis, therefore negative kurtosis can signal that a particular distribution shows evidence of clustering.

In [22], Fyfe proposed a single-neuron implementation of an EPP algorithm. The proposed EPP method is essentially a non-linear modification of Oja's principal-component algorithm. The input data $\mathbf{x} \in \mathcal{R}^m$ is fed forward via the weights, $\mathbf{w} \in \mathcal{R}^m$, to the output neuron where a simple summation is performed; the output neuron's activation is then fed back via the same weights to the inputs as inhibition; then a (non-linear) function of the weights is calculated and used in the updating of the weights by the Hebbian learning rule. Formally, the Fyfe's neuronal exploratory projection pursuit learning algorithm writes:

$$z = \mathbf{w}^T \mathbf{x} \ , \ \mathbf{e} = \mathbf{x} - \mathbf{w}z \ , \ r = f(z) \ ,$$
$$\Delta \mathbf{w} = \mu \, \mathbb{E}_{\mathbf{x}}[r\mathbf{e}] = \mathbb{E}_{\mathbf{x}}[f(z)(\mathbf{x} - z\mathbf{w})] \ .$$

where $z(t)$ denotes the activation of the output neuron and $r(t)$ is the value of the function $f(\cdot)$ on the output neuron. The vector $\mathbf{e}$ may be referred to as 'residual', and contains the remaining part of input-signal structure after deflation.

From the viewpoint of learning as optimization procedure, it was shown that the above algorithm maximizes the following index:

$$P(\mathbf{w}) \stackrel{\text{def}}{=} \mathbb{E}_{\mathbf{x}} \left[\int f(z) dz \right] \ ,$$

under the constraint of normality of the weight-vector $\mathbf{w}$.

The last observation explains the connection between Fyfe's exploratory-projection-pursuit neuron and the learning theory (3)+(4)+(5): By identifying $P(\mathbf{w})$ with $\mathbb{E}[\sigma^2(z)]$ the two theories collapse in a unified learning paradigm.

It is worth mentioning that in successive works the EPP algorithm for a one-unit network has been extended by Fyfe and Girolami to a complete neural network and the blind-signal-separation abilities of the obtained structure has been investigated, as well as the capabilities of the extended EPP algorithm which embodies a network-output decorrelation rule which ensures the different neurons to become sensitive to the single source signals.

As a further note, it might be interesting to observe the close relationship among the discussed learning theories and blind deconvolution adapting laws for adaptive filters, recently discussed in the works [15, 19]. In these papers a detailed survey of 'Bussgang' filtering has been presented, along with a study on generalized Bayesian estimators for blind deconvolution and equalization applications; in particular, in [19] it has been shown that 'Bussgang' filtering

can be reformulated in terms of non-linear minor-component-analysis learning, which is closely related to extended anti-Hebbian learning.

2.5 Relationship with maximum-likelihood Hebbian rules

It is worthwhile reviewing a Hebbian-like learning paradigm, termed *maximum-likelihood Hebbian learning*, (MLH) recently introduced by Fyfe and Corchado in [23], and to briefly investigate the relationships among this new learning paradigm and the SHMM one. Remarkably, what actually turns out is that MHL is a direct counterpart of SHMM, in a sense.

The starting point in the development of MLH theory is the observation that many Hebbian-type (i.e. non-linear) learning paradigms may be derived by the minimization of a convex scalar function of the residual $\mathbf{e}(t)$ introduced in the previous section. If the convex warping function has a parabolic shape, the standard stabilized Hebbian learning rule (i.e. Oja-Amari's) is obtained. In turn, this is equivalent to minimizing the negative log-likelihood of the residual, provided that the $\mathbf{e}(t)$ has a Gaussian distribution. Namely, if $p_{\mathbf{e}}(\mathbf{e}) \propto \exp(-\mathbf{e}^T\mathbf{e})$ then the negative log-likelihood $-\mathbb{E}[\log p_{\mathbf{e}}(\mathbf{e})]$ equals $\mathbb{E}[\mathbf{e}^T\mathbf{e}]$ apart from an unessential additive constant.

This observation readily suggests a generalization of stabilized Hebbian learning rule: If the probability density function of the residuals is known in advance, this knowledge can be advantageously employed in order to determine the optimal cost function which, in turn, gives an optimal learning rule. This suggests a family of learning rules which are derived from the family of exponential distributions.

Let the residual after deflation possess probability density function $p_{\mathbf{e}}(\mathbf{e}) \propto \exp(-\sum_k |e_k|^p)$, with p being the generalized-Gaussian exponential. In this case, the negative log-likelihood writes $\mathbb{E}[\sum_k |e_k|^p]$ and its gradient-based minimization leads [23] to the generalized Hebbian rule:

$$\Delta w_k = \mu\, \mathbb{E}_{\mathbf{x}}[\mathrm{sign}(e_k)\,|e_k|^{p-1} z] \ .$$

In the context of independent component analysis extraction, it is expected that for leptokurtotic residuals (which are more kurtotic than a Gaussian), values of $p < 2$ would be appropriate, while for platikurtotic residuals (less kurtotic than a Gaussian), values of $p > 2$ would be more appropriate. The experiments in [23] and in the more extended paper [24] tend to support the intuition that accuracy and speed of convergence are improved when the choice of p is accurate.

It is interesting to note that, in this case, the requirement is to have a good match of the neuron characteristic non-linearity, $\sum_k |e_k|^p$, with the statistical distribution of the component that is looked for, therefore, in a sense, the MLH principle might be referred to as 'maximum-match' paradigm. In a network of neurons, however, statistical matching of a neuron to a particular component reads statistical mismatch to the other components, thus, again, the two principles may be viewed as collapsing into each other.

3 SHMM Learning Theory: Two Cases-Study

The aim of the present part is to illustrate the Sudjianto-Hassoun 'maximum-mismatch' learning principle for two cases-study and to carry out a numerical analysis of the static properties of the related learning algorithms.

3.1 The case of sinusoidal excitation corrupted by Gaussian noise

The case of sinusoidal signal corrupted by Gaussian noise has been considered from a numerical point of view by Sudjianto and Hassoun in [39].

Our analysis of the properties of criterion $L(\varepsilon)$ is based on the observation that generally the function $Q_n^2(x)$ is *V-shaped* around $x = 0$ and saturates to 1 for relatively large values of the argument.

For a sinusoidal signal $s(t)$ and a Gaussian noise $n(t)$ we have:

$$p_s(s) = \frac{1}{\pi\sqrt{1-s^2}} \ , s \in]-1,+1[, \text{ and } p_s(s) = 0 \text{ otherwise} \ , \qquad (14)$$

$$p_n(n) = \frac{1}{\sqrt{2\pi}\rho} \exp\left(-\frac{n^2}{2\rho^2}\right) \ , \qquad (15)$$

where ρ^2 denotes the variance of the Gaussian noise. Consequently, the non-linear function $Q_n(x)$ takes on the expression:

$$Q_n(x) = \text{erf}\left(\frac{x}{\sqrt{2}\rho}\right) \ . \qquad (16)$$

In order to facilitate the computation of the double-integral (10), it is interesting to note that the quantity $Q_n^2(x)$ can be easily approximated. In fact we have:

$$Q_n^2(x) = \frac{4}{2\pi\rho^2} \int_0^x \int_0^x e^{-\frac{\xi^2+\eta^2}{2\rho^2}} d\xi d\eta \ .$$

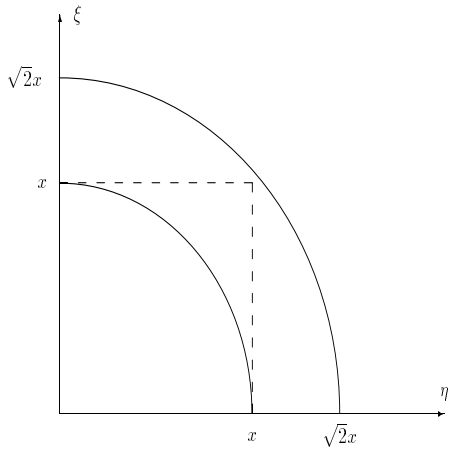

Figure 2: Plane-slice used for computing an approximation of integral Q_n^2.

By the variable change $\xi = r\cos\varphi$, $\eta = r\sin\varphi$, we obtain an easier polar-coordinate integral. The domain of integration in Cartesian coordinates is the square denoted by the dashed-lines in Figure 2; by replacing the actual domain with the sector of angular amplitude $\pi/2$ and radius x or $\sqrt{2}x$, denoted by solid-lines in the Figure, two approximated integral values are obtained, namely:

$$\frac{2}{\pi\rho^2}\int_0^{\pi/2}\int_0^{x} e^{-\frac{r^2}{2\rho^2}} r dr d\varphi \leq Q_n^2(x) \leq \frac{2}{\pi\rho^2}\int_0^{\pi/2}\int_0^{\sqrt{2}x} e^{-\frac{r^2}{2\rho^2}} r dr d\varphi \ .$$

By computing the double-integrals in polar coordinates we obtain:

$$1 - e^{-x^2/2\rho^2} \leq Q_n^2(x) \leq 1 - e^{-x^2/\rho^2} \ ,$$

therefore, we found useful to consider the following approximation:

$$Q_n^2(x) \approx 1 - e^{-\kappa^2 x^2} \ , \ x \in \mathcal{R} \ , \tag{17}$$

where the constant κ^2 should be adjusted in order for the approximation to be as tight as possible and should be chosen in the range $0.5 \leq (\kappa\rho)^2 \leq 1$.

With this approximation, the approximated learning criterion (10) writes:

$$\hat{L}(\varepsilon) \stackrel{\text{def}}{=} 1 - \frac{1}{\pi\sqrt{2\pi}\rho}\int_{-1}^{+1}\int_{-\infty}^{+\infty} e^{-\kappa^2(s\cos\varepsilon + n\sin\varepsilon)^2}\frac{1}{\sqrt{1-s^2}} e^{-n^2/2\rho^2} dn ds \ . \tag{18}$$

Some mathematical work allows expressing this double integral in closed form; in fact, it may be proven that:

$$\hat{L}(\varepsilon) = 1 - \frac{1}{\sqrt{2\kappa^2\rho^2\sin^2\varepsilon + 1}} \exp\left(\frac{-\kappa^2\cos^2\varepsilon}{2\kappa^2\rho^2\sin^2\varepsilon + 1}\right) I_0\left(\frac{-\kappa^2\cos^2\varepsilon}{2\kappa^2\rho^2\sin^2\varepsilon + 1}\right) \ ,$$

where $I_\nu(u)$ denotes the modified Bessel function of the first kind [1].

Now, the optimal value of the constant κ may be defined on the basis of the knowledge that $L(\pi/2) = 1/3$; in fact, by imposing that:

$$\hat{L}(\pi/2) = 1 - \frac{1}{\sqrt{2\kappa^2\rho^2 + 1}} = \frac{1}{3} \ ,$$

it is readily found that $\kappa^2\rho^2 = 5/8$; note that $1/2 < 5/8 < 1$, as required. In conclusion, the approximation found for the criterion $L(\varepsilon)$ reads:

$$\hat{L}(\varepsilon) = 1 - \frac{2}{\sqrt{5\sin^2\varepsilon + 4}} \exp\left(\frac{-5\cos^2\varepsilon}{4\rho^2(5\sin^2\varepsilon + 4)}\right) I_0\left(\frac{-5\cos^2\varepsilon}{4\rho^2(5\sin^2\varepsilon + 4)}\right) \ . \quad (19)$$

In order to evaluate the error owing to the approximation used, we first inspect the deviation between functions $Q_n^2(x)$ and $1 - e^{-\kappa^2 x^2}$ for the chosen value of κ: In Figure 3 the true and approximating functions are depicted, along with their absolute difference, for three values of the parameter ρ. The approximation looks quite good in all the three cases considered and the maximal absolute deviation for these cases is about 7×10^{-3}.

It might now be interesting to observe the shape of the criterion function $\hat{L}(\varepsilon)$. For the same three values of the noise variance ρ^2, these curves are depicted in Figure 4. The obtained results reveal that for small values of ρ the criterion function is indeed monotonically decreasing from $L(0)$ to $L(\pi/2)$, while for increasing values of ρ the criterion function first tends to become flat, and then to become monotonically *increasing* for ε ranging from 0 to $\pi/2$. This fact means that there exists a critical value ρ_c such that as long as $\rho < \rho_c$ the Sudjianto-Hassoun principle is valid, while for $\rho \geq \rho_c$ it ceases to be applicable.

An estimation of the critical standard deviation ρ_c of the noise may be obtained in the following way: The value of the approximated criterion in $\varepsilon = 0$ is given by:

$$\hat{L}(0) = 1 - \exp\left(\frac{-5}{16\rho^2}\right) I_0\left(\frac{-5}{16\rho^2}\right) \ .$$

In order for the criterion function to be nearly-flat, the equality $L(0) = L(\pi/2)$ should hold, thus we claim that a non-monotonic-increasing condition writes:

$$\exp(-\bar{x}) I_0(-\bar{x}) = \frac{2}{3} \text{ with } \bar{x} = \frac{5}{16\rho_c^2} \ .$$

This equation can be solved numerically and gives $\rho_c \approx 0.827$; this result explains the inversion of behavior observed in the Figure 4.

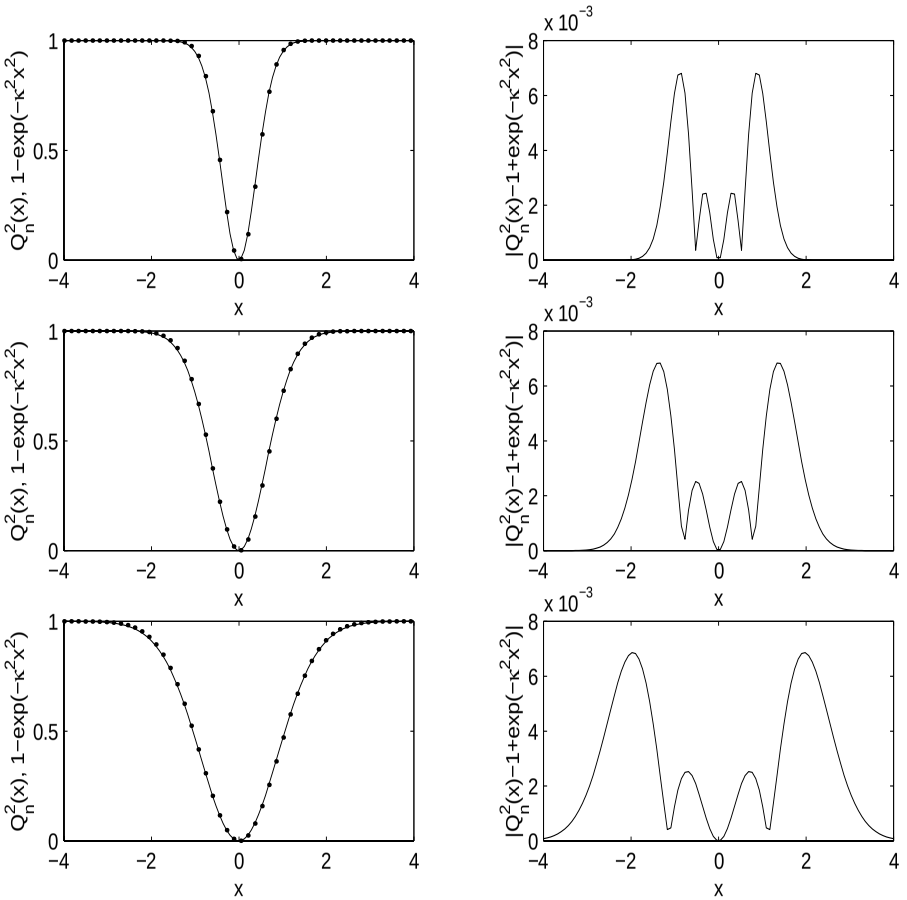

Figure 3: Numerical example about approximation (17). The left-hand columns show the function $Q_n^2(x)$ (solid-line) and its approximation (dotted-line) superimposed, while the right-hand columns show their absolute difference. Top-row: $\rho = 0.45$; Middle-row: $\rho = 0.70$; Bottom-row: $\rho = 1.00$.

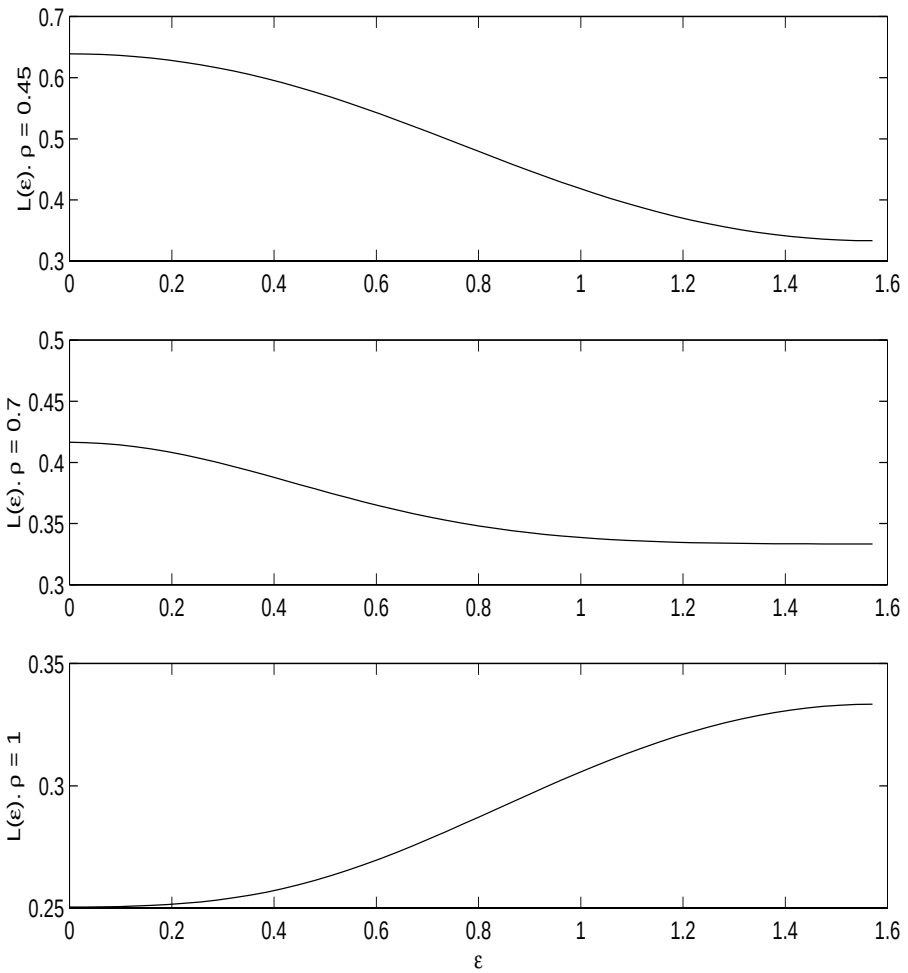

Figure 4: Shapes of criterion function $\hat{L}(\varepsilon)$ for the case of sinusoidal signal corrupted by Gaussian noise.

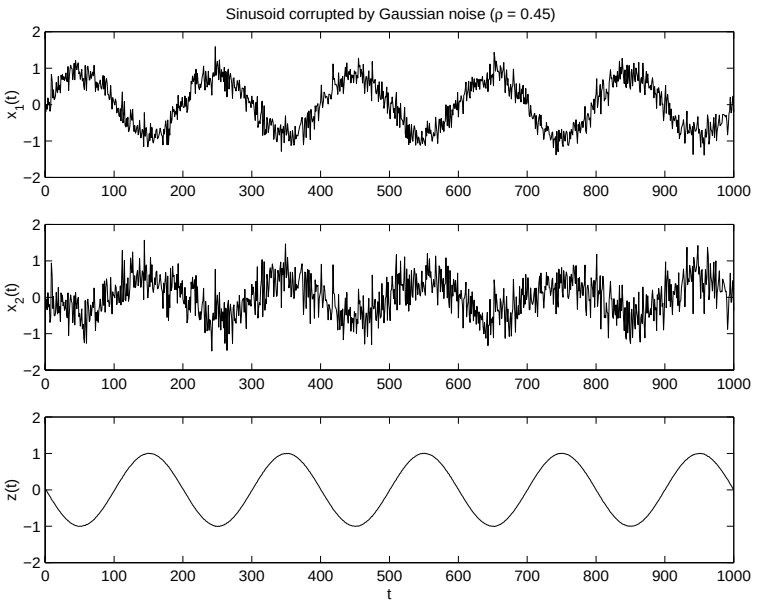

Figure 5: SHMM behavior example: Sinusoidal signal corrupted by Gaussian noise. Neuron's input signals and activation response after learning.

A computer-based simulation of the extended Hebbian learning terms (4)-(5) obtains by considering a sinusoidal signal as $s(t)$ and a Gaussian noise for $n(t)$, that is the same case considered in [39]. In this case, the stochastic learning algorithm has the structure:

$$\Delta \mathbf{w} = \mu \ \mathrm{erf}\left(\frac{z}{\sqrt{2}\rho}\right) \exp\left(-\frac{z^2}{2\rho^2}\right) (\mathbf{x} - z\mathbf{w}) \ , \tag{20}$$

with μ being a positive learning step-size, $\mathbf{x} \stackrel{\mathrm{def}}{=} [x_1 \ x_2]^T$, and $\mathbf{w} \stackrel{\mathrm{def}}{=} [w_1 \ w_2]^T$.

Figure 5 displays the neuron inputs $x_1(t)$ and $x_2(t)$ and the neuron's activation signal value $z(t)$ after learning, while Figure 6 shows the values of the neuron's weights during the learning phase. These experiments refer to the values $\rho = 0.45$ and $\mu = 0.025$. Also, Figure 7 shows the histograms that approximate the probability density functions of the neuron's activation values during learning, obtained by 'freezing' the neuron state and passing the whole training set to the neural unit. These numerical results, completely similar to those reported in the original paper [39], confirm the soundness of the presented 'maximum-mismatch' learning theory.

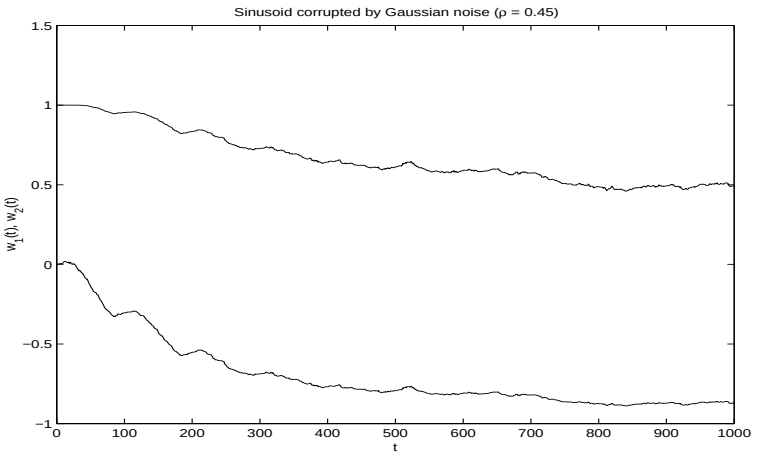

Figure 6: SHMM behavior example: Sinusoidal signal corrupted by Gaussian noise. Neuron's weights values during learning.

3.2 The case of Laplacean signal corrupted by uniform noise

A case-study that allows investigating the separation of a uniformly-distributed disturbance from a sparse Laplacean signal (i.e. a prototype for e.g. speech signal) is considered in what follows; it has the interesting property of giving rise to tractable mathematics, which enables us to write the quantities of interest in closed form.

Formally, we consider the following distributions for the random signal $s(t)$ and the random noise $n(t)$:

$$p_s(s) = \frac{\lambda}{2}\exp(-\lambda|s|) \ , \ p_n(n) = \frac{H(n+1) - H(n-1)}{2} \ , \tag{21}$$

where $H(x)$ denotes the Heaviside (unit-step) function, and λ is a positive constant characterizing the Laplacean distribution $p_s(s)$.

In order to evaluate the behavior of the criterion $L(\varepsilon)$ in this case, it is first necessary to compute the squared function $Q_n^2(x)$ that assumes the expression:

$$Q_n^2(x) = \begin{cases} 1 & |x| \geq 1 \ , \\ x^2 & |x| < 1 \ . \end{cases} \tag{22}$$

Thanks to the simple form of neurons' activation function, some long but straightforward mathematical work allows computing the integral (10), that

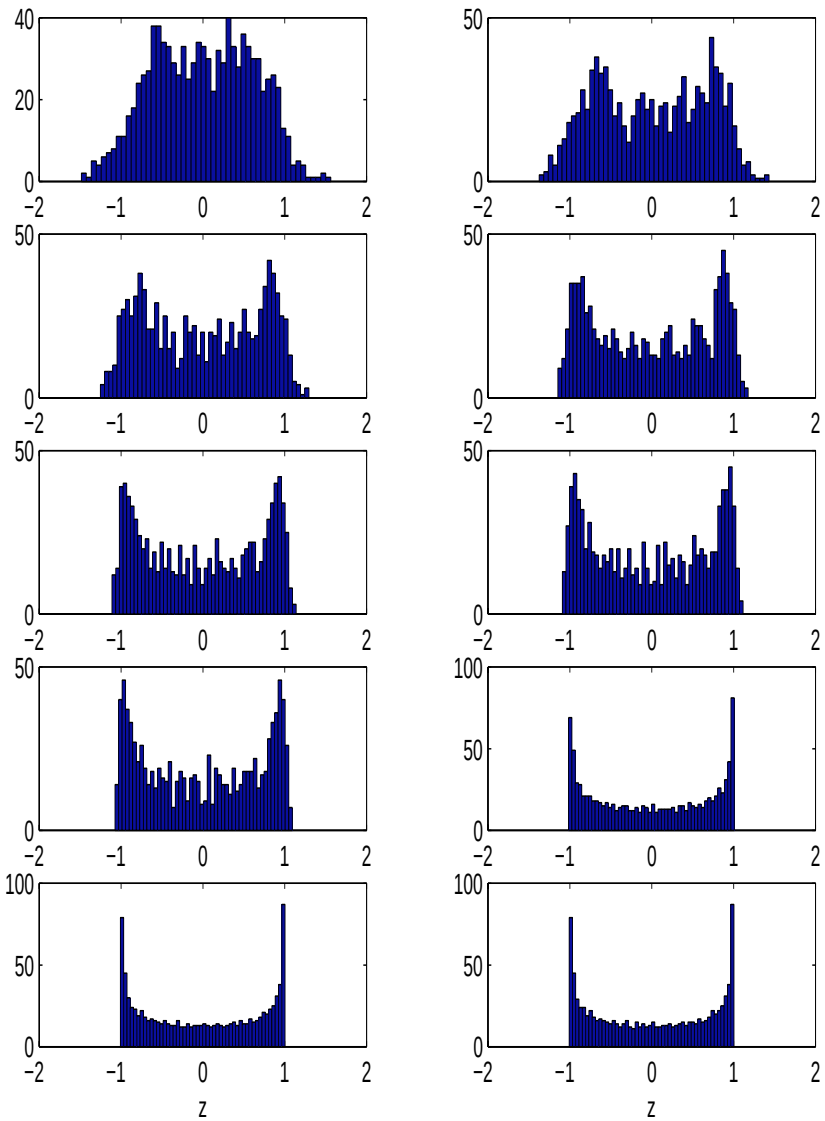

Figure 7: SHMM behavior example: Sinusoidal signal corrupted by Gaussian noise. Distribution of neuron's activation values, estimated every 100 learning iterations (to be read in lexicographic way).

writes:

$$L(\varepsilon) = \frac{2\cos^2\varepsilon}{\lambda^2} + \frac{\sin^2\varepsilon}{3} - \frac{\sinh(\lambda\tan\varepsilon)}{\sin\varepsilon}\frac{2\cos^2\varepsilon}{\lambda^2}\left(1+\frac{\cos\varepsilon}{\lambda}\right)\exp\left(-\frac{\lambda}{\cos\varepsilon}\right) \ . \tag{23}$$

The criterion $L(\varepsilon)$ is continuous in the edges $\varepsilon = 0$ and $\varepsilon = \pi/2$, and the limits in these points have values:

$$\lim_{\varepsilon\to 0} L(\varepsilon) = \frac{2}{\lambda^2} - \left(\frac{2}{\lambda}+\frac{2}{\lambda^2}\right)e^{-\lambda} \ , \ \lim_{\varepsilon\to\pi/2} L(\varepsilon) = \frac{1}{3} \ .$$

In this case too a monotonic decreasing/increasing behavior may be envisaged for the criterion function, broken by a critical value λ_c of Laplacean parameter. Given the above limits, the critical λ may be found by solving the non-linear equation:

$$\frac{2}{\lambda_c^2} - \left(\frac{2}{\lambda_c}+\frac{2}{\lambda_c^2}\right)e^{-\lambda_c} = \frac{1}{3} \ .$$

Numerically, we found $\lambda_c \approx 1.790$.

With the aim to numerically assess the present case, we considered as useful excitation $s(t)$ to the neuron a 0.5s segment of speech signal, corrupted by uniformly distributed noise $n(t)$, observed twice, in order to form signals $x_1(t)$ and $x_2(t)$.

In order to make the above theoretical results profitable for the analytical investigation of the neuron's behavior in this case, we first estimate the (sparse) probability density function of the speech signal and fit it to the Laplacean one: A very good fit is found for $\lambda = 1.7$, as can be seen in the Figure 8. The Figure also shows the shape of the learning criterion function $L(\varepsilon)$ for the chosen value of parameter λ. This function has the maximum value in $\varepsilon = 0$, thus the learning algorithm should be able to extract the speech signal from the noisy mixtures. It is worth noting that λ is very close to λ_c, in fact the learning criterion is quite flat; this makes the separation task rather difficult to the neuron.

Thanks to the simple shape of the squashing function $Q_n(\cdot)$, the learning algorithm has the very simple structure:

$$\Delta\mathbf{w} = \mu\ z H(1-|z|)(\mathbf{x} - z\mathbf{w}) \ , \tag{24}$$

with μ being again a positive learning step-size. It would be interesting to note that the above learning algorithm closely resembles the Oja-Amari principal component rule, except for the limitations occurring when the absolute value of neuron's activation z exceeds the bound 1. In the simulations we used $\mu = 0.005$.

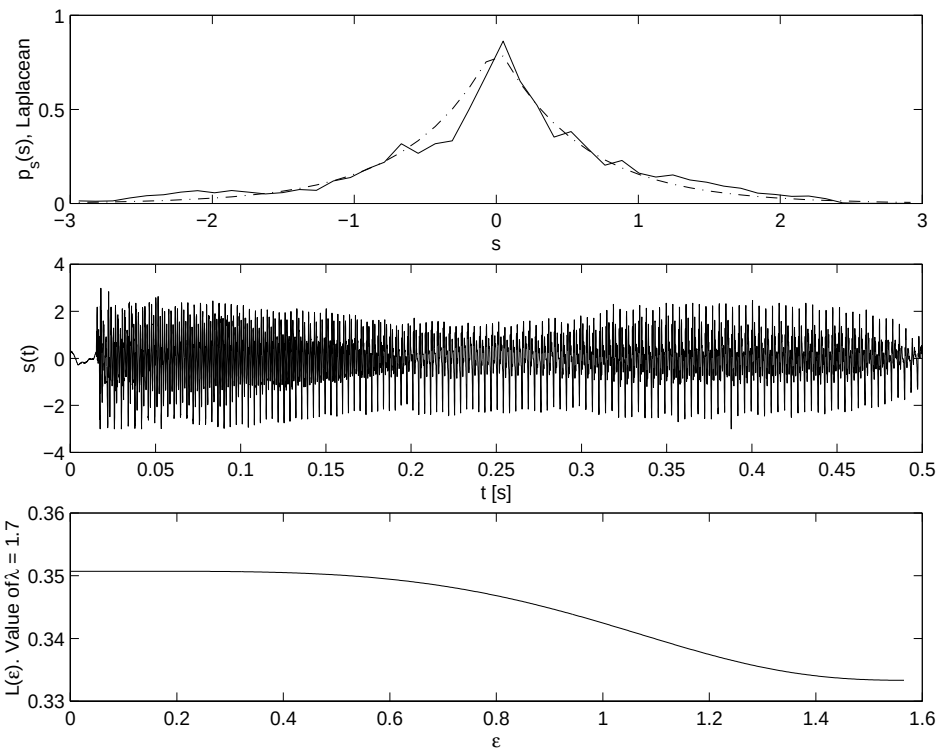

Figure 8: Top: True probability density function of a 0.5s segment of speech signal (solid line) and fitted Laplacean (dot-dashed line). Middle: Speech segment. Bottom: Shape of the criterion $L(\varepsilon)$ in the present case.

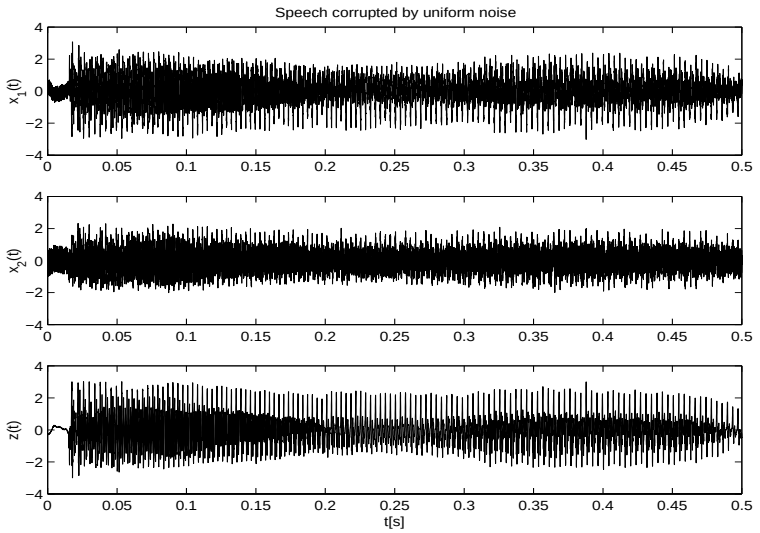

Figure 9: SHMM behavior example: Speech signal corrupted by uniform noise. Neuron's input signals and activation response after learning. Note that the learning criterion is inherently sign-blind, thus the response signal may differ from the sign of $s(t)$, as it just happens in this case.

The neuron's input waveforms and output signal after learning are depicted in the Figure 9, from which it readily emerges that the neuron's linear part output, i.e. signal $z(t)$, closely resembles the speech segment, thus the noise has been almost completely removed from the neuron's response. Auditory experiments further confirm this conclusion. Figure 10 illustrates the values of the neuron's weights during the learning phase, showing the good and plain convergence to the correct connection pattern. It pays to consider that the speech signal is in general non-stationary, therefore the neuron is somewhat forced to keep adjusting the weights values during the whole learning phase to follow the fluctuations of signals' statistics (this may be seen even at the beginning of learning, when a short silence-segment precedes the actual uttered signal).

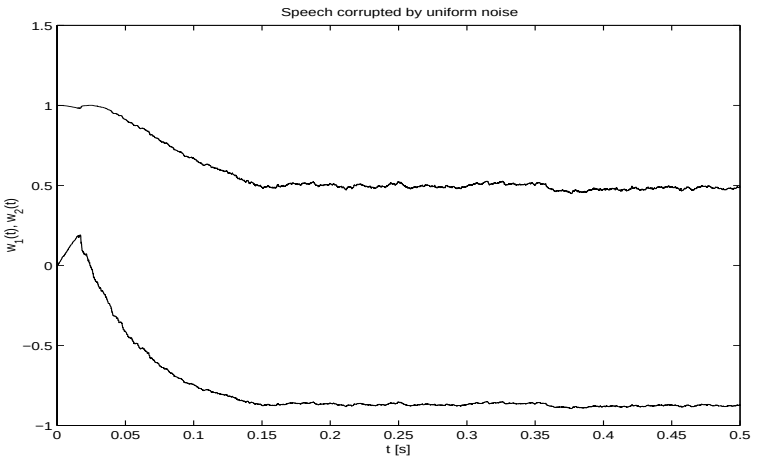

Figure 10: SHMM behavior example: Speech signal corrupted by uniform noise. Neuron's weights values during learning.

4 Application of SHMM Principle to Separation of Complex-Valued Circular Sources

Within the present second part of the paper, we consider the application of SHMM principle to blind separation of an arbitrary number of complex-valued circular sources, particularly useful in the telecommunication context.

A complex-valued signal $s(t) = \rho(t)e^{i\theta(t)}$ (with $i \stackrel{\text{def}}{=} \sqrt{-1}$) is termed 'circularly-distributed' when its probability density function writes $p_s(s) = p_\rho(\rho)p_\theta(\theta)$; for instance in the PSK (Phase Shift Keying) or QAM (Quadrature Amplitude Modulation) case, the constellation of symbols are such that:

$$p_\rho(\rho) = \sum_r \text{Prob}(\rho = r)\delta(\rho - r) \ , \ p_\theta(\theta) = \sum_\tau \text{Prob}(\theta = \tau)\delta(\theta - \tau) \ , \qquad (25)$$

because the values of s lie on the complex plane over circumferences with different radii; the probabilities $\text{Prob}(\rho = r)$ and $\text{Prob}(\theta = \tau)$ represent, respectively, the a-priori discrete distributions of the moduli and phases of the symbols in the constellation.

In the present section we extend the real-valued one-unit version of SHMM principle to a complex-weighted network of separating neurons; the learning equations are formulated in the general case, but under the hypothesis that the noise affecting the mixture is a complex Gaussian disturbance, we are able to

formally prove that the extended SHMM theory may allow the neural network to separate out m contributions from m mixtures.

4.1 Extended Hebbian learning algorithm (EHA)

We consider a complex-weighted single-layer neural network, formed by m units, which performs extended ICA of complex-valued signals. The network is described by the input vector $\mathbf{x} \in \mathcal{C}^m$, a set of weight-vectors $\mathbf{w}_k \in \mathcal{C}^m$ and outputs $z_k \stackrel{\text{def}}{=} \mathbf{w}_k^H \mathbf{x}$. In section 2.2 we considered the general case of the extraction of q sources out of m observations. Hereafter we consider only square mixtures, namely (invoking the definitions of section 2.2) the case that $q = m$.

An extended Hebbian learning algorithm for the mentioned complex-weighted network may be devised as follows. Let the following learning criterion be defined:

$$C(\mathbf{w}_k) \stackrel{\text{def}}{=} U(\mathbf{w}_k) + \Lambda(\mathbf{w}_k) \ . \tag{26}$$

The criterion $C(\cdot)$ contains a nonlinear function of the k^{th} neuron's output and is defined as:

$$U(\mathbf{w}_k) \stackrel{\text{def}}{=} \mathbb{E}_{\mathbf{x}}[g(|z_k|)] \ , \tag{27}$$

with $g : \mathcal{R}_0^+ \to \mathcal{R}_0^+$ being continuously differentiable almost everywhere, non-decreasing with a unique minimum in 0. It should be optimized under the constrain of orthonormality of the weight-vectors. The orthogonality conditions can be written conveniently by defining the function $\Lambda(\cdot)$ as [14]:

$$\Lambda(\mathbf{w}_k) \stackrel{\text{def}}{=} \gamma_{k\,k}(\mathbf{w}_k^H \mathbf{w}_k - 1) + \sum_{j=1}^{k-1} \Re\{\gamma_{kj}^\star \mathbf{w}_k^H \mathbf{w}_j\} \ , \tag{28}$$

where a set of complex Lagrange multipliers $\{\gamma_{kj}\}$ has been introduced, and "$\cdot^\star$" denotes complex conjugation.

To search for optimal weights maximizing the criterion (26), a gradient steepest ascent learning algorithm may be employed. The optimal multipliers as functions of the $\mathbf{w}_k$'s can be found by the standard elimination procedure. In conclusion, by defining:

$$\mathbf{P}_k \stackrel{\text{def}}{=} \mathbf{I}_m - \sum_{j=1}^{k} \mathbf{w}_j \mathbf{w}_j^H \ , \ G(x) \stackrel{\text{def}}{=} \frac{dg(x)}{dx}\frac{1}{x} \text{ with } x > 0 \ , \tag{29}$$

the new complex non-classic counterpart of GHA learning rule (here referred to as *EHA*) writes:

$$\Delta \mathbf{w}_k = \mu \ \mathbf{P}_k \mathbb{E}[G(|z_k|) z_k^\star \mathbf{x}] \ , \ k = 1, 2, \ldots, m \ , \tag{30}$$

with μ being a positive learning step-size. The factor $\mathbb{E}[G(z_k)z_k^\star\mathbf{x}]$ may be interpreted as a complex non-classical Hebbian term, while projector $\mathbf{P}_k$ plays the role of a deflating factor. About function $g(\cdot)$, it can be chosen on the basis of the specific task for which the network is used. It deserves to note that assuming $g(x) = \frac{1}{2}x^2$ yields $G(x) = 1\ \forall x > 0$, thus in this case and in presence of real-valued data, the algorithm (30) coincides to well-known GHA rule by Sanger [38].

Let us consider now the extension of the SHMM theory for a single-unit neural system to the complex-valued case. We may define the cost function:

$$C_c(\mathbf{w}) \stackrel{\text{def}}{=} \mathbb{E}[\sigma^2(|z|)] + \gamma_c(\mathbf{w}^H\mathbf{w} - 1) \ , \tag{31}$$

for a complex-weighted neuron with output $z = \mathbf{w}^H\mathbf{x}$, where γ_c denotes the Lagrangean multiplier for the usual energy-conservation consistency constraint. Its gradient-ascent maximization yields the learning rule:

$$\Delta\mathbf{w} = \mu\ (\mathbf{I}_m - \mathbf{w}\mathbf{w}^H)\mathbb{E}\left[\sigma(|z|)\sigma'(|z|)\frac{z^\star}{|z|}\mathbf{x}\right] \ , \tag{32}$$

that closely recalls equation (30) for $k = 1$. In analogy to section 2, here it is assumed:

$$\sigma(|z|) \stackrel{\text{def}}{=} \int_0^{|z|} p(\zeta)d\zeta \ ,$$

where $p(\cdot)$ represents the discriminant probability density function that should be filtered out (discarded) from the data.

We possess now the instruments for developing a SHMM theory for a complete neural network formed by complex-weighted neurons. It obtains by merging the general complex-valued extended Hebbian learning theory with the one-unit SHMM principle extended to the complex-valued case. By comparing, in fact, definitions (31) and (26) for $k = 1$, we arrive at the conclusion that $g(x) = \sigma^2(x)$, from which the relationship between $p(\cdot)$ and $g'(\cdot)$:

$$g'(x) = 2p(x)\int_0^x p(\xi)d\xi \tag{33}$$

readily stems.

4.2 Application to blind separation of telecommunication-related sources

The explained principle can be employed for separating out independent complex-valued circular signals from their linear mixtures. In the following two sub-

sections we present the non-linearity arising from the assumption of complex Gaussian noise and a study of the convergence properties of the EHA learning algorithm in this case.

4.2.1 Non-linearity pertaining to a complex Gaussian noise

Let us suppose input $\mathbf{x}$ contains a complex linear mixture of statistically independent signals [6], and that one of these signals is a Gaussian noise of the form $n = n_I + in_Q$, where both in-phase n_I and in-quadrature n_Q components are zero-mean Gaussian random variables of variance ρ^2. Then it is known that the modulus $|n|$ follows the Rayleigh distribution:

$$p_{\mathrm{R}}(|n|) = \frac{|n|}{\rho^2} \exp\left(-\frac{|n|^2}{2\rho^2}\right) \ . \tag{34}$$

Then by formula (33) we find:

$$g'_{\mathrm{R}}(x) = \frac{2x}{\rho^2}\left[\exp\left(-\frac{x^2}{2\rho^2}\right) - \exp\left(-\frac{x^2}{\rho^2}\right)\right] H(x) \ , \tag{35}$$

where $H(x)$ denotes again the Heaviside function; for example, the Figure 11 depicts the non-linearity $g'_{\mathrm{R}}(x)$ for three different values of noise power.

By assuming in (30) the function $G(|z_k|)$ as the quantity $\frac{g'_{\mathrm{R}}(|z_k|)}{|z_k|}$, it would then be possible to separate out independent complex-valued signals mixed by a unitary operator.

4.2.2 Convergence properties of the EHA rule under complex Gaussian noise assumption

In section 3 we presented a formal analysis of the learning criterion that ensured the convergence of the algorithm to the proper (separating) solution, provided that the established conditions on noise power are met. Now we need to extend these results to the present multi-unit, complex-valued case. Under the hypothesis of complex Gaussian noise, it is possible to give formal results about the local properties of the learning criterion (26) at the equilibrium.

Our results are based on a recent study presented by Bingham and Hyvärinen in [5]. In addition to the basic assumptions on the sources and on the signals model summarized in section 2.2, the restrictions on the separation problem in order for Bingham-Hyvärinen result to hold true are:

- The independent source signals in $\mathbf{s}(t)$ have zero mean, unit variance and uncorrelated real and imaginary parts of equal variance (i.e. $\mathbb{E}[\mathbf{s}\mathbf{s}^T] = \mathbf{0}_m$);

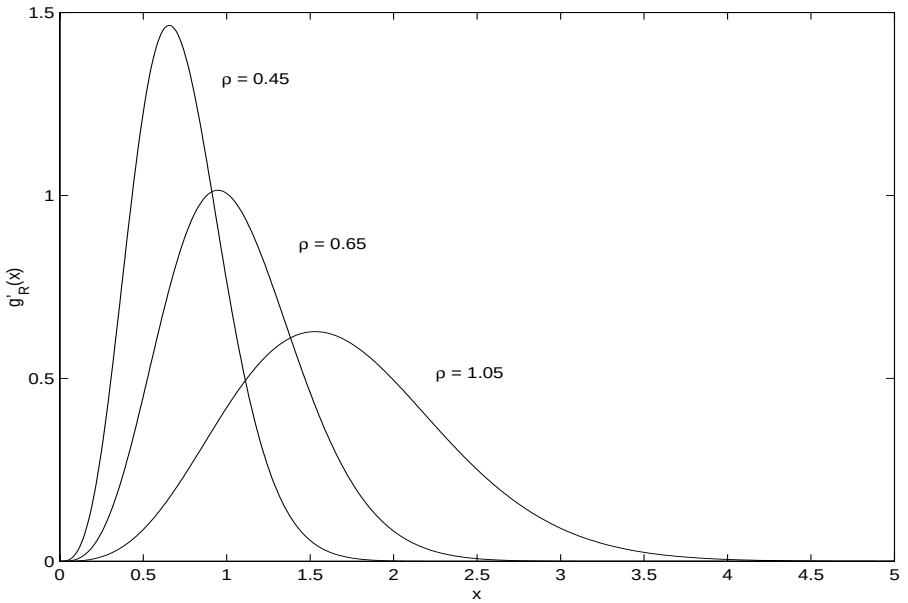

Figure 11: Function $g'_{\mathrm{R}}(\cdot)$ for three values of noise power ρ^2.

- The observed vector stream $\mathbf{x}(t)$ is spatially white, i.e. $\mathbb{E}[\mathbf{x}\mathbf{x}^H] = \mathbf{I}_m$;
- The separating-network learning phase is driven by the optimization of a criterion function of network's linear-part output signals, namely $\mathbb{E}[F(|z_k|^2)]$, whose kernel $F(\cdot)$ is an even, smooth function.

Under these hypotheses, the local maxima (minima) of $\mathbb{E}[F(|z_k|^2)]$ under constraint $\|\mathbf{w}_k\|^2 = 1$ correspond to those independent components that satisfy:

$$\mathbb{E}[B(|s_k|^2)] = \int_0^\infty B(\xi^2) p_{|s_k|}(\xi) d\xi < 0 \ (> 0) \ , \tag{36}$$

where function $B(\cdot)$ depends on the shape of $F(\cdot)$ and is defined as:

$$B(x^2) \stackrel{\mathrm{def}}{=} (1 - x^2) F'(x^2) + x^2 F''(x^2) \ , \tag{37}$$

and the s_k's denote again the source signals.

Technically speaking, this important result may be obtained by computing the Jacobian and Hessian of the criterion $\mathbb{E}[F(|z_k|^2)]$ on the expected solutions and by showing that, if the above conditions are met, then the Jacobian vanishes and the Hessian matrix is negative (positive) definite [5].

This result may be advantageously employed to study the behavior of the proposed EHA algorithm (29)+(30)+(33), because the EHA rule arises as an optimization algorithm based on a driving criterion function and constraints of orthonormality. In fact, such important result adapts to our case provided that $F(x^2)$ is expressed as a function of $g(x)$ and by observing that, by definition of EHA algorithm, it always seeks for the (constrained) maximum of criterion $\mathbb{E}[g(|z_k|)]$, thus we need to show in which case $\mathbb{E}[B(|s_k|^2)] < 0$.

The link-equation between the two criterion-functions is $F(x^2) = g(x)$, and by taking into account the relationship (33) between the criterion function $g(\cdot)$ and the discriminant probability density function $p(\cdot)$, we ultimately obtain:

$$B(x^2) = \frac{p^2(x)}{2} + \frac{g'(x)p'(x)}{4p(x)} - \left(\frac{1}{2x} - x\right)\frac{g'(x)}{2} \ , \tag{38}$$

The found expression for $B(x^2)$ may now be particularized for the Rayleigh case, namely $p(x) = p_{\mathrm{R}}(x)$ and $g'(x) = g'_{\mathrm{R}}(x)$ as given by equations (34) and (35). With these relationships, the quantity $B(\cdot)$ particularizes into:

$$B_{\mathrm{R}}(x^2;\rho) = \left(\frac{x^2}{\rho^4} + \frac{1}{2\rho^2} - \frac{x^2}{\rho^2}\right)\exp\left(-\frac{x^2}{\rho^2}\right) + \left(\frac{x^2}{\rho^2} - \frac{1}{2\rho^2} - \frac{x^2}{2\rho^4}\right)\exp\left(-\frac{x^2}{2\rho^2}\right) \ . \tag{39}$$

Clearly condition (36) depends on the statistics of $|s_k|$, thus the fulfillment of the separation conditions require the knowledge of the probability density function of independent signals' moduli.

It is worth noting that by plugging the found expression of $B(\cdot)$ in the condition (36) we obtain an inequality in the only free parameter of the neurons' activation functions, namely in ρ. This suggests that there exist intervals of values of ρ for which the algorithm may converge to a separating solution. Let us inspect the behavior of EHA in concrete cases related to the separation of telecommunication signals.

As mentioned, in the present paper we consider the cases of PSK/QAM4 and QAM16 modulations as sources $|s_k|$, whose statistical descriptions may be given in the following terms:

- **PSK/QAM4**: In this case we have $p_{|s|}^{\mathrm{PSK}}(x) = \delta(x-1)$, because the symbols are concentrated on the unit-circle in the complex plane.
- **QAM16**: The required probability density function writes: $p_{|s|}^{\mathrm{QAM16}}(x) = \frac{1}{4}\delta(x - \frac{\sqrt{2}}{3}) + \frac{1}{2}\delta(x-1) + \frac{1}{4}\delta(x-\sqrt{2})$, because the symbols lie on three circumferences with different probabilities.

Now the expectation of function (39) for the two cases may be computed in closed-form and read, respectively:

$$\begin{aligned} \mathbb{E}_{\mathrm{PSK}}[B_{\mathrm{R}}(x^2;\rho)] &= B_{\mathrm{R}}(1;\rho) \ , \\ 4\mathbb{E}_{\mathrm{QAM16}}[B_{\mathrm{R}}(x^2;\rho)] &= 2B_{\mathrm{R}}(1;\rho) + B_{\mathrm{R}}(2;\rho) + B_{\mathrm{R}}\left(\frac{2}{9};\rho\right) \ . \end{aligned}$$

The Figure 12 shows the shape of the above functions of ρ for PSK/QAM4 and QAM16 modulations: It clearly emerges that there exists a range for ρ such that the functions take on negative values (for instance, it is not difficult to prove that $B_{\mathrm{R}}(1;0.5) \approx -6e^{-2}$); this confirms that the EHA algorithm can converge to the expected solutions.

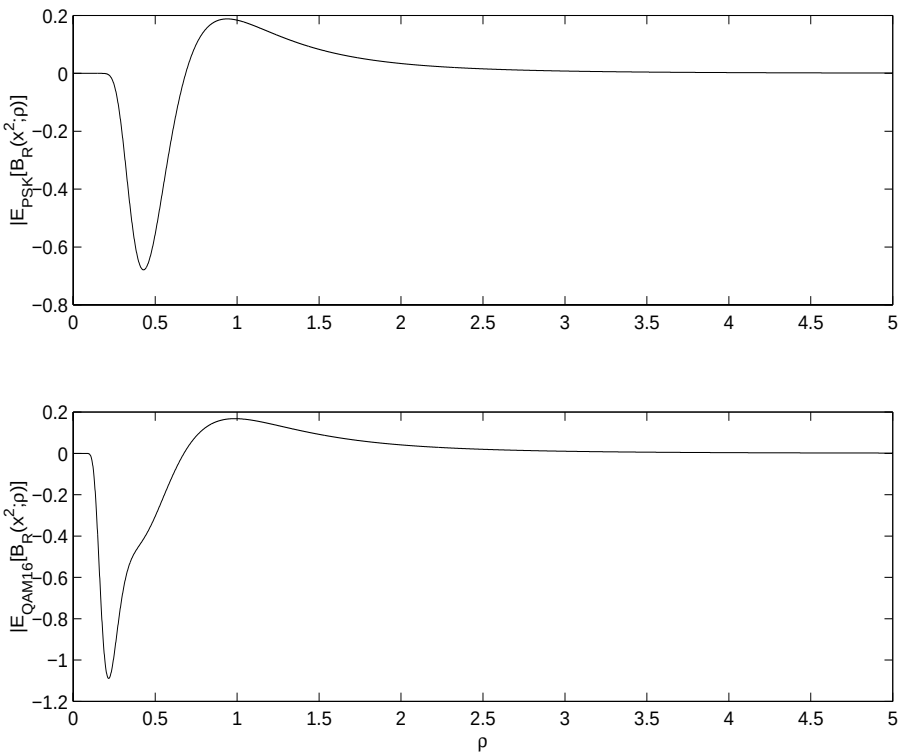

Figure 12: $\mathbb{E}_x[B_{\mathrm{R}}(x^2;\rho)]$ for PSK/QAM4 and QAM16 modulations, as function of the standard deviation ρ of the discriminant complex Gaussian noise.

5 Experimental Results

The assessment of the proposed blind source separation method was conducted through experiments aimed at evaluating the performance of the EHA and at comparing its features with the ones pertaining to similar algorithms drawn from the scientific literature.

In order to objectively evaluate the separation capability of the considered algorithm, we compute the *separation matrix* $\mathbf{S}$, defined by $\mathbf{S} \stackrel{\text{def}}{=} \mathbf{W}^H \mathbf{V}^H \mathbf{M}^H$, which should resemble a quasi-identity after learning, where again $\mathbf{W}$ is the unitary separation matrix, $\mathbf{V}$ denotes the whitening matrix, and $\mathbf{M}$ is the mixing matrix. Then, as separation measure, we take the standard figure-of-merit SIR (signal-to-interference ratio), defined as:

$$SIR \stackrel{\text{def}}{=} \sum_{k=1}^{m} \sum_{j=1}^{m} \frac{|S_{kj}|^2}{\max_\ell \{|S_{\ell j}|^2\}} - m \ .$$

Perfect separation would imply SIR=0, but in the present context values of the order of -25 or -30dB guarantee perfect recognition of the symbols in the constellation, which ultimately is the aim of source recovering from the mixtures of transmitted signals in telecommunications.

5.1 Single-trial experiments on EHA

In the first experiment, we considered input $\mathbf{x} \in \mathcal{C}^4$ formed by a mixture of four independent signals in $\mathbf{s} \in \mathcal{C}^4$. Signals s_1 to s_3 are QAM16 modulations; signal s_4 is a complex-valued Gaussian disturbance. The first row of Figure 13 depicts the independent signals while the second row shows the obtained four mixtures computed according to observable signals model (12).

Blind complex-valued-source separation results are shown in Figure 14: The first row depicts the result of pre-whitening performed by means of standard eigenvalue normalization (as explained in subsection 2.2); the second row shows the last 100 output samples of the network trained by (30) on the pre-whitened data with $\mu = 0.002$ and confirms the symbols of the QAM16 constellation are now clearly recognizable. The Figure 15 shows the absolute values of source-to-output separation matrix and confirms only one entry per column significantly differs from zero, thus the network has been able to recover the independent signals from the mixtures.

Another interesting example concerns blind separation from a high-dimensional

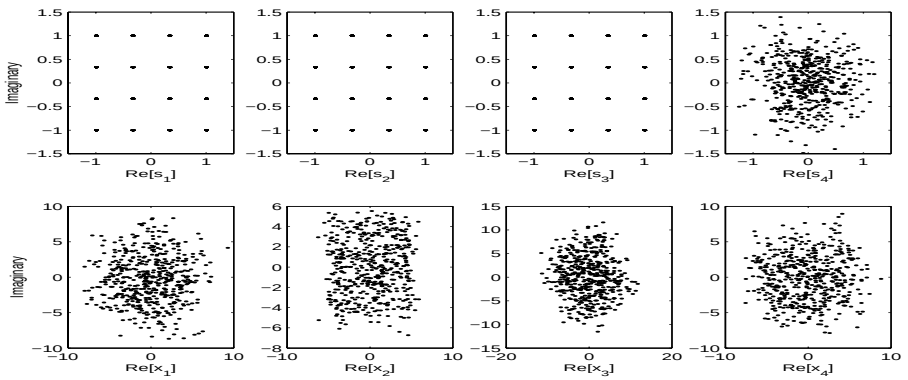

Figure 13: Four independent complex-valued signals and four complex-valued mixtures of them.

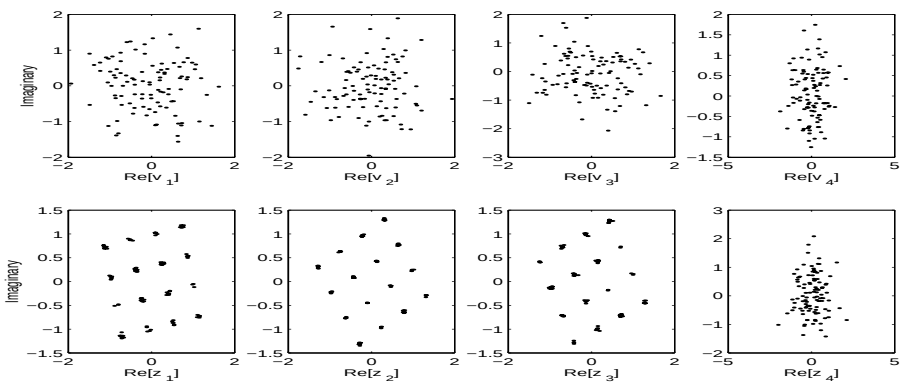

Figure 14: Whitened mixtures and network's outputs for (four-sources problem).

mixture. Let us consider the separation of 4 QAM4 signals, 3 PSK8 signals and 3 QAM16 signals. This example enable us to show an interesting issue that is quite clear from the theory but which is worth remarking from a numerical point of view: We chose as discriminant distribution the probability density function of a complex Gaussian noise, which makes the algorithm able to separate out the different independent contributions that differ from such disturbance; however, this *does not necessarily mean the Gaussian noise needs to be present in the mixture.*

Figure 17 shows 100 network's output samples recovered, while Figure 16 depicts the separation product matrix $\mathbf{S}$; the algorithm was run on 50,000 source samples with $\mu = 0.001$.

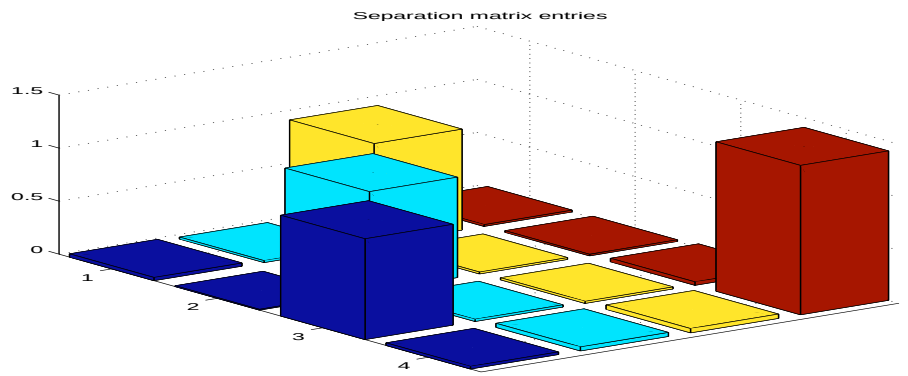

Figure 15: Source-to-output separation matrix **S** (four-sources problem). (Absolute values are represented.)

In this example, all the signals in the mixture are useful contributions, and the algorithm has been able to successfully separate them out.

5.2 A comparison of six algorithms

We consider for comparison the EHA algorithm, two among the ψ–APEX algorithms from [10] (here the $|y|$–APEX algorithm, referred to as EAPEX1, and the 0–APEX algorithm, referred to as EAPEX0, were employed), and the MEC algorithm, introduced in [17] as a novel learning engine for real-valued signal processing, and recently extended to complex-valued networks in [18]. We also consider the JADE algorithm [7] and the complex-valued version of 'FastICA' package [5], hereafter referred to as CFP algorithm; it is worth noting that the first four algorithms are iterative (i.e. they process the data on a sample-by-sample basis), while the last two algorithms are of batch type.

Figure 18 summarizes the results of simulation on a six-source separation problem (2 QAM4, 2 QAM16, 1 PSK, plus a complex-valued Gaussian noise as a disturbance).

The top graph shows the signal-to-interference ratios at each iteration (except for JADE and CFP algorithms for which the final SIR values only are reported). The parameters (such as learning step-size) in the algorithms have been set in order for the learning speed and performance to be nearly the same for all the techniques. The different algorithms show very good performances, being able to attain quite low values of signal-to-interference ratios, ranging

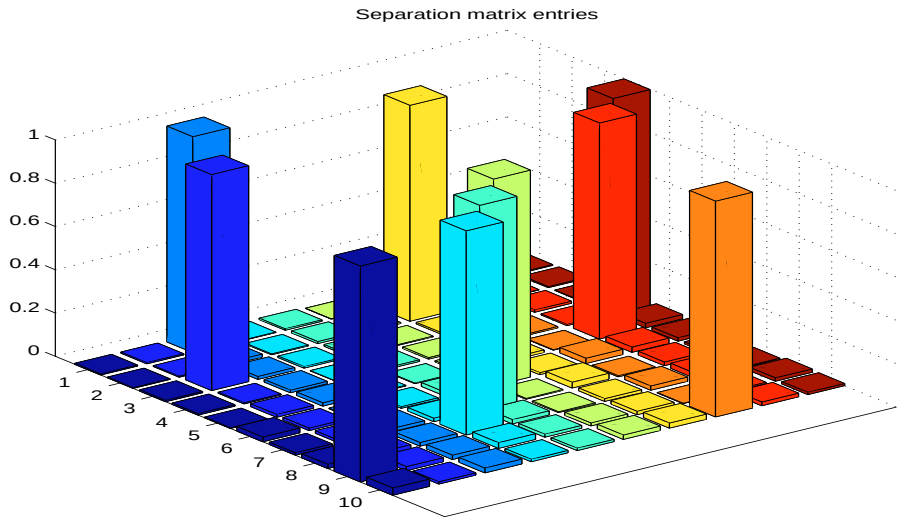

Figure 16: Source-to-output separation matrix **S** (ten-sources problem). (Absolute values are represented.)

from −30dB to −40dB.

The middle graph in the Figure shows a comparison of computational complexity of the six considered algorithms, expressed in terms of the number of floating point operations (flops) demanded per sample on a common platform (500 MHz clock, 64MB RAM), while the bottom graph shows a comparison of computation (CPU) time required for the codes to run on the same platform, expressed in milliseconds per sample.

Both the flops and CPU-time may be assumed as measures of the computational complexity of the algorithms, as they take into account different complexity aspects, such as computational effort and memory-storage requirements.

In our analysis, EHA algorithm exhibits the lowest computational requirements in comparison to the other algorithms, while guaranteeing comparable separation performances.

6 Conclusion

The aim of the present paper was to recall and formalize the Sudjanto-Hassoun theory of non-classical Hebbian learning, to extend this theory to complex-weighted neural networks, and to show how it allows blind source separation by

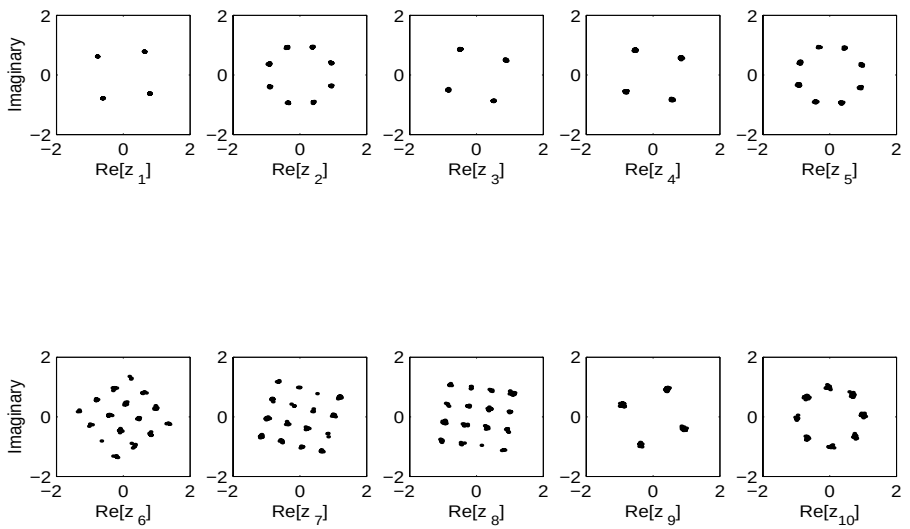

Figure 17: Neuron output for ten source signals.

the independent component analysis. Formal results about the SHMM learning principle are given in order to better understand its properties and its main features, and formal results about the capability of a complex-weighted neural network, equipped with a SHMM learning algorithm, to perform independent component analysis are given too. The main contribution of the SHMM learning theory to ICA is to allow designing the proper structure of non-linear part of neurons *provided that some information is known on the statistical structure of signals that the source signals differ from*; this principle is very useful in blind signal processing, where the statistical structure of the involved signals is not known in advance, but it is easily known the structure of signals different from the useful ones, such as noises.

Numerical results have also been presented to assess the effectiveness of the proposed learning theory. They refer to telecommunication signals and have been presented under the form of single trials and a comparison of five existing techniques with the one proposed in the present paper. The found numerical results show that the proposed technique allows obtaining comparable separation performance at a lower global computational cost.

Acknowledgments

I am indebted with the anonymous Reviewers for some detailed and careful comments that allowed me to consistently improve the presentation and organization of the original manuscript. I am also grateful to Prof. Colin Fyfe (University of Paisley, Scotland) for sharing an interesting preprint on recent advancements in the theory of exploratory projection pursuit and maximum-likelihood Hebbian learning, which helped me in the enrichment of the content of the old version of the paper, along with some computer codes on the implementation of these methods.

References

[1] M. Abramowitz and C.A. Stegun (Ed.s), *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, (9^{th} printing), pp. 374 – 377, 1972. New York: Dover

[2] S.-I. Amari, *Neural theory of association and concept-formation*, Biological Cybernetics, Vol. 26, pp. 175 – 185, 1977

[3] S.-i. Amari, *Natural gradient learning for over- and under-complete bases in ICA*, Neural Computation, Vol. 11, pp. 1875 – 1883, 1999

[4] S.-i. Amari, T.-p. Chen, and A. Cichocki, *Stability Analysis of Learning Algorithms for Blind Source Separation*, Neural Networks, Vol. 10, No. 8, pp. 1345 – 1351, 1997

[5] E. Bingham and A. Hyvärinen, *A fast fixed-point algorithm for independent component analysis of complex valued signals*, International Journal of Neural Systems, Vol. 10, No. 1, pp. 1 – 8, Feb. 2000

[6] J.F. Cardoso and B. Laheld, *Equivariant adaptive source separation*, IEEE Trans. on Signal Processing, Vol. 44, No. 12, pp. 3017 – 3030, Dec. 1996

[7] J.-F. Cardoso, *High-order contrasts for independent component analysis*, Neural Computation, Vol. 11, No. 1, pp. 157 – 192, 1999

[8] A. Cichocki, R. Unbehauen, and E. Rummert, *Robust Learning Algorithm for Blind Separation of Signals*, Electronics Letters, Vol. 30, No. 17, pp. 1386 – 1387, Aug. 1994

[9] P. Comon, *Independent Component Analysis, a new concept ?*, Signal Processing, Vol. 36, pp. 287 – 314, 1994

[10] S. Fiori, *Blind Separation of Circularly Distributed Source Signals by the Neural Extended APEX Algorithm*, Neurocomputing, Vol. 34, No. 1-4, pp. 239 – 252, Aug. 2000

[11] S. Fiori, *Blind Signal Processing by the Adaptive Activation Function Neurons*, Neural Networks, Vol. 13, No. 6, pp. 597 – 611, Aug. 2000

[12] S. Fiori and F. Piazza, *A General Class of ψ–APEX PCA Neural Algorithms*, IEEE Trans. on Circuits and Systems – Part I, Vol. 47, No. 9, pp. 1394 – 1398, Sept. 2000

[13] S. Fiori, *A Theory for Learning by Weight Flow on Stiefel-Grassman Manifold*, Neural Computation, Vol. 13, No. 7, pp. 1625 – 1647, July 2001

[14] S. Fiori, *On Blind Separation of Complex-Valued Sources by Extended Hebbian Learning*, IEEE Signal Processing Letters, Vol. 8, No. 8, pp. 217 – 220, Aug. 2001

[15] S. Fiori, *A Contribution to (Neuromorphic) Blind Deconvolution by Flexible Approximated Bayesian Estimation*, Signal Processing, Vol. 81, No. 10, pp. 2131 – 2153, Sept. 2001

[16] S. Fiori, *Hybrid Independent Component Analysis by Adaptive LUT Activation Function Neurons*, Neural Networks, Vol. 15, No. 1, pp. 71 – 80, Feb. 2002

[17] S. Fiori, *A Theory for Learning Based on Rigid Bodies Dynamics*, IEEE Trans. on Neural Networks, Vol. 13, No. 3, pp. 521 – 531, May 2002

[18] S. Fiori, *Complex-Weighted One-Unit 'Rigid-Bodies' Learning Rule for Independent Component Analysis*, Neural Processing Letters, Vol. 15, No. 3, pp. 275 – 282, June 2002

[19] S. Fiori, *Blind Deconvolution by Simple Adaptive Activation Function Neuron*, Neurocomputing, Vol. 48, pp. 763 – 778, Oct. 2002

[20] S. Fiori, *Overview of Independent Component Analysis Technique with an Application to Synthetic Aperture Radar (SAR) Imagery Processing*, Neural Networks (Special Issue on 'Neural Networks for Analysis of Complex Scientific Data: Astronomy, Geology and Geophysics'). Accepted for publication.

[21] S. Fiori, L. Albini, A. Faba, E. Cardelli and P. Burrascano, *Numerical Modeling for the Localization and the Assessment of Electromagnetic Field Sources*, IEEE Trans. on Magnetics. Accepted for publication.

[22] C. Fyfe, *A comparative study of two neural methods of exploratory projection pursuit*, Neural networks, Vol. 10, No. 2, pp. 257 – 262, 1997

[23] C. Fyfe and E. Corchado, *Maximum-likelihood Hebbian learning*, Proc. of the Tenth European Symposium on Artificial Neural Networks (ESANN'02), pp. 143 – 148, Bruges (Belgium), Apr. 2002

[24] C. Fyfe and D. MacDonald, *Epsilon-insensitive Hebbian learning*, Neurocomputing, pp. 35 – 57, Vol. 47, No. 1-4, Aug. 2002

[25] A. Hyvärinen and E. Oja, *Independent component analysis by general non-linear Hebbian-like rules*, Signal Processing, Vol. 64, No. 3, pp. 301 – 313, 1998

[26] A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001

[27] X. Giannakopoulos, J. Karhunen, and E. Oja, *An Experimental Comparison of Neural Algorithms for Independent Component Analysis and Blind Separation*, International Journal of Neural Systems, Vol. 9, No. 2, pp. 99 – 114, Apr. 1999

[28] C. Jutten and J. Hérault, *Independent Component Analysis (INCA) Versus Principal Component Analysis*, Proc. EUSIPCO, Vol. 2, pp. 643 – 646, 1988

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

[30] J. Karhunen and J. Joutsensalo, *Representation and separation of signals using nonlinear PCA type learning*, Neural Networks, Vol. 7, No. 1, pp. 113 – 127, 1994

[31] M.S. Lewicki and T. Sejnowski, *Learning non-linear overcomplete representations for efficient coding*, Advances in Neural Information Processing Systems (NIPS*10), pp. 815 – 821, 1998

[32] E. Oja, *A simplified neuron model as a principal component analyzer*, Journal of Mathematics and Biology, Vol. 15, pp. 267 – 273, 1982

[33] E. Oja, *Beyond PCA: Statistical expansions by non-linear neural networks*, Proc. International Conference on Artificial Neural Networks (ICANN), Vol. 2, pp. 1049 – 1054, 1994

[34] E. Oja, *Nonlinear PCA Criterion and Maximum Likelihood in Independent Component Analysis*, Proc. ICA'98, pp. 143 – 148, 1998

[35] E. Oja, H. Ogawa, and J. Wangviwattana, *Learning in non-linear constrained Hebbian networks*, Proc. of ICANN'91, pp. 385 – 390, 1991

[36] A. Paraschiv-Ionescu, C. Jutten, and G. Bouvier, *Neural Network Based Processing for Smart Sensor Arrays*, Proc. of International Conference on Artificial Neural Networks (ICANN), pp. 565 – 570, 1997

[37] V.K. Rohatgi, *Statistical Inference*, J. Wiley & Sons, New York, 1984

[38] T.D. Sanger, *Optimal unsupervised learning in a single-layer linear feedforward neural network*, Neural Networks, Vol. 2, pp. 459 – 473, 1989

[39] A. Sudjianto and M.H. Hassoun, *Non-linear Hebbian rule: A statistical interpretation*, Proc. of IEEE-ICNN, Vol. 2, pp. 1247 – 1252, 1994

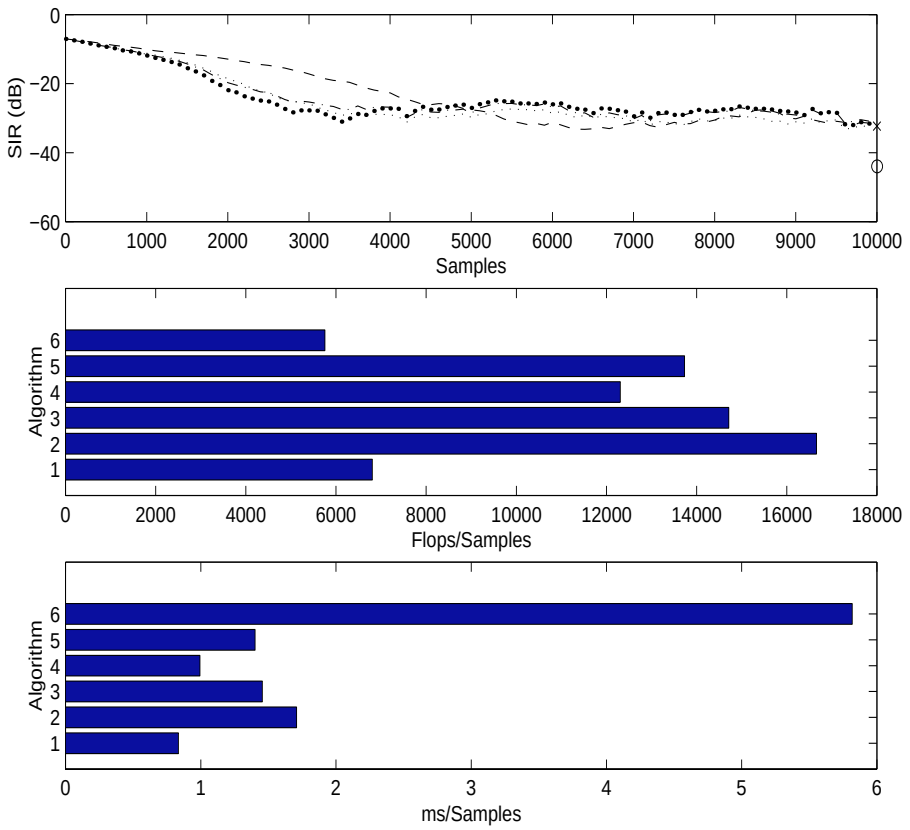

Figure 18: EHA (Dotted-line, algorithm 1), EAPEX1 (Dot-dashed line, algorithm 2), EAPEX0 (Pointed-line, algorithm 3), MEC (Dashed-line, algorithm 4), JADE (Symbol 'o', algorithm 5) and CFP (Symbol 'x', algorithm 6) employed to separate out 6 mixed sources. Top: Compared interference residuals (in dB scale). Middle: Flops per sample demanded by the six algorithms. Bottom: CPU time required by the MATLAB codes to run, expressed in milliseconds (ms) per sample.