A Linear Feedforward Neural Network with Lateral Feedback Connections for Blind Source Separation

Seungjin CHOI
Lab for Artificial Brain Systems
Frontier Research Program, RIKEN
2-1 Hirosawa, Wako-shi
Saitama 351-01, Japan
schoi@lion.riken.go.jp

Andrzej CICHOCKI
Lab for Artificial Brain Systems
Frontier Research Program, RIKEN
2-1 Hirosawa, Wako-shi
Saitama 351-01, Japan
cia@hare.riken.go.jp

Abstract

We presents a new necessary and sufficient condition for the blind separation of sources having non-zero kurtosis, from their linear mixtures. It is shown here that a new blind separation criterion based on both odd ($f(y) = y^3$) and even ($f(y) = y^2$) functions, presents desirable solutions, provided that all source signals have negative kurtosis (sub-Gaussian) or have positive kurtosis (super-Gaussian). Based on this new separation criterion, a linear feedforward network with lateral feedback connections is constructed. Both theoretical and computer simulation results are presented.

1. Introduction

Blind source separation is a fundamental problem encountered in many applications such as array signal processing, sonar, digital communications, some biomedical applications. The goal of blind source separation is to recover the independent source signals from their linear mixtures without the knowledge of mixing matrix. Specially when the propagation media is slowly changing (the mixing matrix is slowly changing), an adaptive system for blind source separation is necessary.

An adaptive blind source separation was first introduced by Jutten and Herault [11]. Later it was further developed by others [12, 7, 4, 1, 2, 8, 5]. Most of existing methods mentioned above are based on nonlinear odd functions (for example, $f(y) = y^3$ or $f(y) = tanh(y)$) because they assume that the probability distributions of all source signals are symmetric. Recently Choi *et al* [6] have shown that a new learning algorithm based on even nonlinearity (e.g. a quadratic function, $f(y) = y^2$) enables the separation of source signals having non-zero skewness (the 3rd-order cumulant) without spurious equilibria. We extend this result to the separation of source signals having the same sign of non-zero kurtosis (the 4th-order cumulant). A new necessary and sufficient condition for blind source separation is presented. It is shown here that the source signals can be separated by a linear transformation, if and only if all the 2nd- and 4th-order *cross*-cumulants of the output of the network are zero, provided that the source signals are statistically independent and each of them has the same sign of non-zero kurtosis. Based on this criterion, a linear feedforward network with lateral feedback connections is constructed with associated adaptive learning algorithms. The proposed learning algorithms are based on both cubic and quadratic functions to force all 2nd- and 4th-order cross-cumulants of the output of the network to vanish to zero.

2 A New Blind Source Separation Criterion

Consider the case where the observation vector $\mathbf{x}(t) \in \mathbb{R}^n$ and the source vector $\mathbf{s}(t) \in \mathbb{R}^n$ are related by

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is the mixing matrix. The problem of blind sources separation is to recover the source signals $\mathbf{s}(t)$ from the observation vector $\mathbf{x}(t)$ without the knowledge of the mixing matrix $\mathbf{A}$. Let $\mathbf{y}(t)$ be the output of the network, i.e., $\mathbf{y}(t) = \mathbf{W}(t)\mathbf{x}(t) = \mathbf{W}(t)\mathbf{A}\mathbf{s}(t)$. It is desired to update $\mathbf{W}(t)$ such that the global system $\mathbf{G}(t) = \mathbf{W}(t)\mathbf{A}$ converges to a matrix,

$$\mathbf{G} = \mathbf{P}\mathbf{\Lambda} \tag{2}$$

as $t \to \infty$, for some permutation matrix $\mathbf{P}$ and nonsingular diagonal matrix $\mathbf{\Lambda}$.

Throughout this paper, the following assumptions are made:

A1: $\mathbf{A} \in \mathbb{R}^{n\times n}$ is nonsingular.
A2: At each time t, the components of $\mathbf{s}(t)$ are statistically independent.
A3: Each component of $\mathbf{s}(t)$ is a zero mean ergodic stationary process with a non-zero variance.
A4: Each component of $\mathbf{s}(t)$ has a non-zero kurtosis, i.e.,

$$\begin{aligned} cum_4\{s_i(t)\} &= E\{s_i^4(t)\} - 3E^2\{s_i^2(t)\}, \\ &\neq 0, \quad \text{for } i = 1, \cdots, n, \end{aligned} \tag{3}$$

where E denotes the expectation operator.

Without loss of generality, we can assume that the variance of each source signal is unity, i.e., $\mathbf{R}_{ss} = E\{\mathbf{s}(t)\mathbf{s}^T(t)\} = \mathbf{I}$, where $\mathbf{I}$ denotes identity matrix. Let the vector sequence $\mathbf{y}(t)$ be a linear mixture of the source vector sequence $\mathbf{s}(t)$,

$$\mathbf{y}(t) = \mathbf{G}\mathbf{s}(t). \tag{4}$$

Let us define two different 4th-order cumulant matrices of $\mathbf{y}(t)$ as follows:

$$\begin{aligned} [\mathbf{C}_{13y}]_{ij} &= cum_{13}\{y_i(t), y_j(t)\} \\ &= cum\{y_i(t), y_j(t), y_j(t), y_j(t)\}, \end{aligned} \tag{5}$$

$$\begin{aligned} [\mathbf{C}_{22y}]_{ij} &= cum_{22}\{y_i(t), y_j(t)\} \\ &= cum\{y_i(t), y_i(t), y_j(t), y_j(t)\}, \end{aligned} \tag{6}$$

where $[\cdot]_{ij}$ denotes the (i,j)th component of the matrix. The covariance matrix $\mathbf{R}_y = E\{\mathbf{y}(t)\mathbf{y}^T(t)\}$ is decomposed as

$$\mathbf{R}_y = \mathbf{G}\mathbf{G}^T. \tag{7}$$

For the following decomposition, we introduce Hadamard product (element-wise product). Let $\circ$ denotes the Hadamard product. For example, the (i,j)th element of $\mathbf{G} \circ \mathbf{G}$ is g_{ij}^2.

Property 1 *The 4th-order cumulant matrices $\mathbf{C}_{13y}$ and $\mathbf{C}_{22y}$ have the following decompositions:*

$$\mathbf{C}_{13y} = \mathbf{G}\mathbf{K}_s[\mathbf{G} \circ \mathbf{G} \circ \mathbf{G}]^T, \tag{8}$$

$$\mathbf{C}_{22y} = [\mathbf{G} \circ \mathbf{G}]\mathbf{K}_s[\mathbf{G} \circ \mathbf{G}]^T, \tag{9}$$

where $\mathbf{K}_s$ is the nonsingular diagonal matrix whose ith diagonal element is $cum_4\{s_i(t)\}$.

Proof: It can be proved by using multi-linearity of cumulant. Consider the (i,j)th element of $\mathbf{C}_{13y}$. Time index t is omitted throughout the proof. ($y_i = y_i(t),\ s_i = s_i(t)$)

$$\begin{aligned} & cum_{13}\{y_i, y_j\} \\ &= cum\{\sum_{l_1} g_{il_1}s_{l_1}, \sum_{l_2} g_{jl_2}s_{l_2}, \sum_{l_3} g_{jl_3}s_{l_3}, \sum_{l_4} g_{jl_4}s_{l_4}\} \\ &= \sum_{l_1} g_{il_1} \sum_{l_2} g_{jl_2} \sum_{l_3} g_{jl_3} \sum_{l_4} g_{jl_4} cum\{s_{l_1}, s_{l_2}, s_{l_3}, s_{l_4}\} \\ &= \sum_k g_{ik}g_{jk}^3 cum_4\{s_k\} \end{aligned}$$

Similarly, the decomposition (9) can be proved.

Theorem 1 *Let the source signals $\mathbf{s}(t)$ satisfy the assumptions given in A1 through A4. Suppose that the kurtosis of all source signals $\mathbf{s}(t)$ are either positive or negative. Then $\mathbf{G}$ has decomposition (2), i.e.,*

$$\mathbf{G} = \mathbf{P}\mathbf{\Lambda}, \tag{10}$$

if and only if the following conditions are satisfied:

$$\mathbf{G}\mathbf{G}^T = \mathbf{\Lambda_1}, \tag{11}$$

$$\mathbf{G}\mathbf{K}_s[\mathbf{G} \circ \mathbf{G} \circ \mathbf{G}]^T = \mathbf{\Lambda_2}, \tag{12}$$

$$[\mathbf{G} \circ \mathbf{G}]\mathbf{K}_s[\mathbf{G} \circ \mathbf{G}]^T = \mathbf{\Lambda_3}, \tag{13}$$

where $\mathbf{\Lambda_1}$, $\mathbf{\Lambda_2}$, and $\mathbf{\Lambda_3}$ are nonsingular diagonal matrices.

Proof: For the sake of simplicity, let us assume that $\mathbf{\Lambda_1} = \mathbf{I}$. Then, $\mathbf{G}$ is an orthogonal matrix. Let $\mathbf{\Lambda_2} = diag\{\alpha_1, \cdots, \alpha_n\}$ and $\mathbf{K}_s = diag\{\kappa_1, \cdots, \kappa_n\}$. From (12), the (i,j)th element of the matrix $\mathbf{G}$, g_{ij} should satisfy $g_{ij} = 0$ or $g_{ij}^2 = \frac{\alpha_i}{\kappa_j}$. Using this fact, the (i,j)th element $(i \neq j)$ of LHS in (13) is given by

$$\alpha_i\alpha_j \sum_{l\in\mathcal{K}} \frac{1}{\kappa_l}, \tag{14}$$

where $\mathcal{K}$ is a set for the collection of all possible cases where $\mathbf{G} \neq \mathbf{P}\ \mathring{\mathbf{I}}$ for some permutation matrix $\mathbf{P}$ and the diagonal matrix $\mathring{\mathbf{I}}$ whose diagonal elements are either +1 or -1. Note that $\alpha_i \neq 0$ for $\forall i$ and $\sum_{l\in\mathcal{K}} \frac{1}{\kappa_l} \neq 0$. Thus, $\mathcal{K}$ is empty set. Therefore, $\mathbf{G} = \mathbf{P}\ \mathring{\mathbf{I}}$ □

Note that most of the existing algorithms satisfy the first two conditions (11), (12) so that they could converge to spurious equilibria. For instance, it can be easily shown that the following $\mathbf{G}$ satisfies (11) and (12), but it is not a solution to blind source separation:

$$\mathbf{G} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}. \tag{15}$$

3. Implementation

We present two different learning algorithms whose equilibria satisfy the conditions given in (11), (12), and (13). The network as shown in Figure 1, consists of a feedforward network $\mathbf{W}(t)$ and a feedback network $\mathbf{U}(t)$. The feedback network $\mathbf{U}(t)$ can be a lower triangular matrix with zero diagonal elements (this feedback connections are referred to be lateral feedback connections) or can be a full matrix (a fully recurrent network).

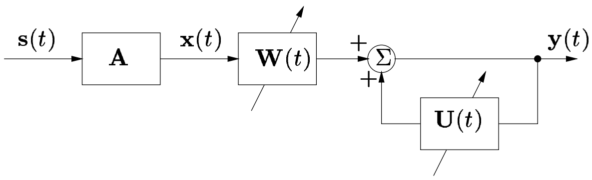

Figure 1. Our approach to blind source separation

3.1. The network with lateral feedback connections

We consider the case where $\mathbf{U}(t)$ is a lower triangular matrix with zero diagonal elements as shown in Figure 2. The output of the network $y_i(t),\ i = 1, \cdots, n$ is described by

$$y_i(t) = \sum_{j=1}^{n} w_{ij}(t)x_j(t) + \sum_{j<i} u_{ij}(t)y_j(t), \quad (16)$$

where $w_{ij}(t)$ is the synaptic weight between the ith output $y_i(t)$ and the jth input $x_j(t)$ and $u_{ij}(t)$ is the lateral connection between $y_i(t)$ and $y_j(t)$. Or in matrix form,

$$\mathbf{y}(t) = \mathbf{W}(t)\mathbf{x}(t) + \mathbf{U}\mathbf{y}(t), \quad (17)$$

where $\mathbf{W}(t) = [w_{ij}(t)]_{n\times n}$ is a full matrix and $\mathbf{U} = [u_{ij}(t)]_{n\times n}$ is a lower triangular matrix with $u_{ij}(t) = 0$ for $i \leq j$.

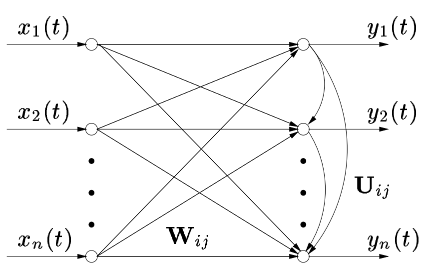

Figure 2. The structure of the feedforward network with lateral feedback connections

For such a neural network, we have developed the following adaptive learning algorithm:

$$\begin{aligned}\frac{d\mathbf{W}(t)}{dt} &= \eta_w(t)\{\alpha\mathbf{I} - \alpha\mathbf{y}(t)\mathbf{y}^T(t) \\ &\quad - sgn(\kappa_s)\Lambda(t) \\ &\quad + sgn(\kappa_s)[\mathbf{y}(t)\circ\mathbf{y}(t)\circ\mathbf{y}(t)]\mathbf{y}^T(t)\}\mathbf{W}(t), \quad (18)\\ \frac{du_{ij}(t)}{dt} &= \eta_u(t)\{1 - y_i^2(t)y_j^2(t)\}, \quad \text{for } i > j, \quad (19)\end{aligned}$$

where $\mathbf{\Lambda}(t)$ is a diagonal matrix whose ith diagonal element is $y_i^4(t)$. For sub-Gaussian source signals, $sgn(\kappa_s) = -1$, and for super-Gaussian source signals, $sgn(\kappa_s) = +1$. When the convergence of learning algorithm (18) and (19) is achieved, we have

$$\begin{aligned} E\{y_i(t)y_j(t)\} &= 0, \quad (20)\\ E\{y_i(t)y_j^3(t)\} &= 0, \quad (21)\\ E\{y_i^2(t)y_j^2(t)\} &= 1, \quad (22)\end{aligned}$$

for $i \neq j$, and

$$E\{y_i^2(t)\} = 1, \quad \text{for } i = 1, \cdots, n. \quad (23)$$

Note that

$$\begin{aligned} cum_{22}\{y_i(t), y_j(t)\} &= E\{y_i^2(t)y_j^2(t)\} - E\{y_i(t)y_j(t)\} \\ &\quad -2E\{y_i^2(t)\}E\{y_j^2(t)\}. \quad (24)\end{aligned}$$

It can be easily seen that when these conditions (20), (21), (22), and (23) are satisfied, all 2nd- and 4th-order cross-cumulants of $\mathbf{y}(t)$ become zero. By Theorem 1, these equilibria are desirable solutions.

3.2. The network with full feedback connections

We consider the case where the feedback connection matrix $\mathbf{U}(t)$ is a full matrix (a fully recurrent network including the self-inhibitions connections from each output node back to itself). The output of the network $\mathbf{y}(t)$ is still given by

$$\mathbf{y}(t) = \mathbf{W}(t)\mathbf{x}(t) + \mathbf{U}(t)\mathbf{y}(t). \quad (25)$$

The feedforward connections $\mathbf{W}(t)$ is trained to force all 4th-order cross-cumulants of the output $\mathbf{y}(t)$ to vanish to zero. The decorrelation learning algorithm with the feedback connections $\mathbf{U}(t)$ was motivated from the natural gradient-based algorithm [1]. We have developed the following learning algorithm

$$\begin{aligned}\frac{d\mathbf{W}(t)}{dt} &= \eta_w(t)\{\mathbf{\Gamma}(t) - [\mathbf{y}(t)\circ\mathbf{y}(t)][\mathbf{y}(t)\circ\mathbf{y}(t)]^T \\ &\quad - sgn(\kappa_s)\mathbf{y}(t)[\mathbf{y}(t)\circ\mathbf{y}(t)\circ\mathbf{y}(t)]^T \\ &\quad + sgn(\kappa_s)[\mathbf{y}(t)\circ\mathbf{y}(t)\circ\mathbf{y}(t)]\mathbf{y}^T(t)\}\mathbf{W}(t), \quad (26)\\ \frac{d\mathbf{U}(t)}{dt} &= \eta_u(t)\{\mathbf{I} - \mathbf{U}(t)\}\{\mathbf{I} - \mathbf{y}(t)\mathbf{y}^T(t)\}, \quad (27)\end{aligned}$$

where the ith diagonal elements of $\mathbf{\Gamma}(t)$ is $\mathbf{y}_i^4(t)$ and all off-diagonal elements of $\mathbf{\Gamma}(t)$ are unity. It can be shown that stable equilibria of (26) and (27) satisfy the conditions (11), (12), and (13).

4. Discussion

We will discuss the extension of Theorem 1 to the case where sub-Gaussian (negative kurtosis) and super-Gaussian (positive kurtosis) source signals are mixed. Even in this case, Theorem 1 still holds if any possible sum of the inverse of kurtosis of more than two source signals, is not zero, i.e., $\sum_{l \in \mathcal{F}} \frac{1}{\kappa_l} \neq 0$, where $\mathcal{F}$ is a set of all combination of indices greater than or equal to two, out of $(1, 2, \cdots, n)$. This condition guarantees that all diagonal elements of $\mathbf{\Lambda_3}$ are non-zero. However, in situation where $\mathbf{x}(t)$ contains the mixtures of both sub-Gaussian and super-Gaussian source signals, there is a possibility that the sum of the inverse of the kurtosis might be zero. Moreover, the learning algorithms become unstable. Thus additional techniques are required [3].

The learning algorithm (18) and (19) associated with the first proposed network is working well for sub-Gaussian sources and super-Gaussian sources (not for mixtures of sub- and super-Gaussian). For the stability of learning algorithm, we introduce the function $sgn(\kappa_s)$. At this moment, we do not have theoretical result for stability analysis.

The neural network architecture shown in Figure 1 and 2 can be used for extraction of source signals one by one. Maximization or minimization of the normalized kurtosis with deflation approach is found in [9, 10]. We are developing the learning algorithm for sequential extraction of source in the hierarchical network as shown in 2.

5. Computer Simulations

The computer simulations are conducted to evaluate the performance of the proposed algorithm (18), (19) and (26), (27). The global system $\mathbf{G} = (\mathbf{I} + \mathbf{U})^{-1}\mathbf{W}\mathbf{A}$ is evaluated to check the efficiency of the algorithm. The global system $\mathbf{G}$ approach the generalized permutation when the learning algorithm converges to desirable solutions.

5.1. Computer Simulation 1

Three i.i.d. binary sources drawn from uniform distribution (negative kurtosis) are used. The mixing matrix $\mathbf{A}$ is chosen randomly as

$$\mathbf{A} = \begin{bmatrix} 0.1549 & 0.1405 & 0.3916 \\ 0.5258 & 0.2041 & 0.9370 \\ 0.2047 & 0.5108 & 0.4310 \end{bmatrix} \qquad (28)$$

All elements of $\mathbf{G}(t) = (\mathbf{I} + \mathbf{U}(t))^{-1}\mathbf{W}(t)\mathbf{A}$ are plotted as shown in Figure 3 and 4.

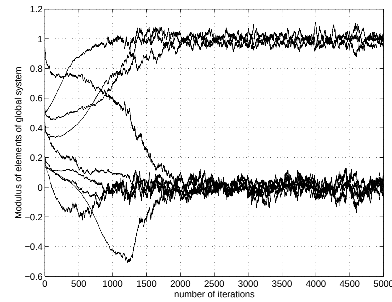

Figure 3. The modulus of each element of the global system $\mathrm{G}(t)$ when the first proposed network with the learning algorithm (18), (19). The learning rates were set $\eta_w(t) = .003$, $\eta_u(t) = .0007$.

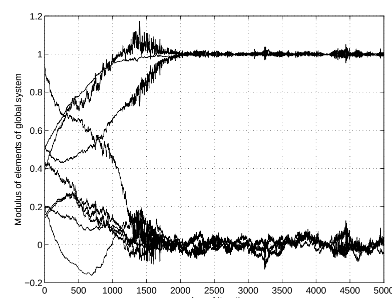

Figure 4. The modulus of each element of the global system $\mathrm{G}(t)$ when the second proposed network with the learning algorithm (26), (27). The learning rates were set $\eta_w(t) = .003$, $\eta_u(t) = .009$.

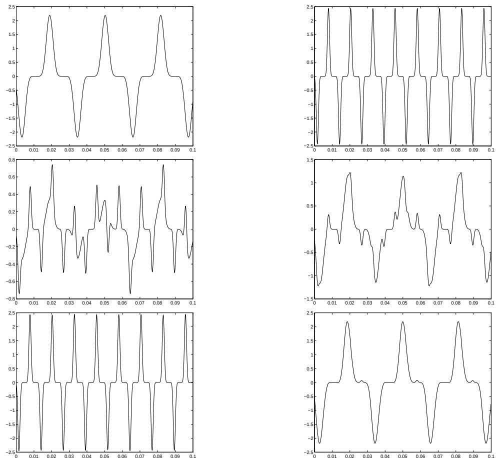

Figure 5. Sampling rate was 10kHz. Each plot shows the signal for the duration of .1 second. The original source signals are shown in the first row. Two sensor outputs are shown in the second row. The recovered signals are shown in the third row by using the learning algorithm (18) and (19). The similar result can be obtained by using (26) and (27).

5.2. Computer Simulation 2

Two independent source signals having positive kurtosis were used to generate the observation data $\mathbf{x}(t)$ by using randomly generated mixing matrix $\mathbf{A}$. Two different source signals are given by

$$\begin{aligned} s_1(t) &= sin^7(\omega_1 t), \\ s_2(t) &= sin^9(\omega_2 t). \end{aligned} \tag{29}$$

Figure 5 shows original source signals $\mathbf{s}(t)$, the sensor outputs $\mathbf{x}(t)$, and the recovered signals $\mathbf{s}(t)$.

6. Conclusion

In this paper, we presented a new necessary and sufficient condition for blind source separation. It was shown that if all 2nd- and 4th-order cross-cumulants of the output $\mathbf{y}(t)$ become zero, then the source signals can be recovered. It was proved by algebraic properties of 2nd- and 4th-order cumulants for the case of n sources. Based on this criterion, we constructed a linear feedforward network followed by a feedback network. Two different learning algorithms were presented and confirmed by the computer simulations.

References

[1] S. Amari, A. Cichocki, and H. H. Yang, “A new learning algorithm for blind signal separation,” *Advances in Neural Information Processing Systems*, Vol. 8, 1995.

[2] S. Amari, A. Cichocki, and H. H. Yang, “Recurrent neural networks for blind separation of sources,” in *Proc. NOLTA*, Vol. 1, pp. 37-42, 1995.

[3] S. Amari, T.-P. Chen, and A. Cichocki, “Stability analysis of learning algorithms for blind source separation,” *Neural Networks*, 1997. (in print)

[4] A. Bell and T. Sejnowski, “An information maximisation approach to blind separation and blind deconvolution,” *Neural Computation* 7, pp. 1129-1159, 1995.

[5] J. -F. Cardoso and B. H. Laheld, “Equivariant adaptive source separation,” *IEEE Trans. on Signal Processing*, pp. 3017-3030, December 1996.

[6] S. Choi and R. Liu, “A learning algorithm for the blind separation of non-zero skewness source signals with no spurious equilibria,” in *Proc. ESANN* April 1997. (in print)

[7] A. Cichocki, R. Unbehauen, L. Moszczynski, and E. Rummert, “A new on-line adaptive learning algorithm for blind separation of source signals,” in *Int. Symp. on Artificial Neural Networks*, pp. 406-411 1994.

[8] A. Cichocki, S. Amari, M. Adachi, and W. Kasprzak, “Self-adaptive neural networks for blind separation of sources' ,” in *Proc. ISCAS*, Vol.2, pp. 157-160, 1996.

[9] A. Cichocki, S. Amari and R. Thawonmas, “Blind signal extraction using self adaptive non-linear Hebbian learning rule,” in *Proc. NOLTA*, pp. 377-380, October 1996.

[10] A. Cichocki, R. Thawonmas and S. Amari “Sequential blind signal extraction in order specified by stochastic properties”, *Electronics Letters*, Vol.33, No.1, pp.64-65, 1997.

[11] C. Jutten and J. Herault, “Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture,” *Signal Processing*, 24, pp. 1-10, 1991.

[12] E. Moreau and O. Macchi, “Two novel architectures for the self adaptive separation of signals,” in *IEEE Conf. Comm.*, pp. 1154-1159, 1993.

[13] E. Oja, J. Karhunen, and A.Hyvarinen, “From Neural PCA to Neural ICA,” *Neural Information Processing Systems Workshop on Blind Signal Processing and its Applications*, December 1996.