# SELF-ADAPTIVE INDEPENDENT COMPONENT ANALYSIS for SUB-GAUSSIAN and SUPER-GAUSSIAN MIXTURES with an UNKNOWN NUMBER of SOURCES and ADDITIVE NOISE

Andrzej CICHOCKI [1], Ireneusz SABALA [2], Seungjin CHOI [3]
Bruno ORSIER [1] and Ryszard SZUPILUK [2]

[1] *Brain Science Institute Riken, Laboratory for Open Information Systems, JAPAN, cia@brain.riken.go.jp,*
[2] *Warsaw University of Technology, POLAND, isa@nov.iem.pw.edu.pl*
[3] *Chungbuk National University, KOREA, schoi@engine.chungbuk.ac.kr*

***Abstract***— In this paper we derive and analyze un-supervised adaptive on - line algorithms for instantaneous blind separation of sources (BSS) in the case when sensors signals are noisy and they are mixture of unknown number of independent source signals with unknown statistics. Nonlinear activation functions are rigorously derived assuming that source have generalized Gaussian, Cauchy or Rayleigh distributions. Extensive computer simulations confirmed that the proposed family of learning algorithms are able to separate sources from mixture of sub and super-Gaussian sources.

## I. INTRODUCTION

The problem of independent component analysis (ICA) and/or blind separation or extraction of source signals from their mixtures has become increasingly important due to still some opened theoretical problems and many potential applications, e.g. in speech recognition and enhancements, telecommunication and biomedical signal analysis and processing (EEG, MEG, ECG). While several recently developed algorithms have shown promise to solve practical tasks, they may fail to separate on-line (non-stationary) signal mixtures containing both sub- and super-Gaussian distributed source signals, especially when number of sources is unknown and change dynamically over the time. The problem of on-line estimation of sources in the case when the number of sources is unknown is relevant in many practical applications like analysis of EEG signals and "cocktail party problem" where the number of source signals change usually over the time. In this paper we propose solution to this problem under assumption that the number of sources is less or equal to the number of sensors.

Mathematically the problem is formulated as follows: the mixing model is described by matrix equation [1]- [17]

$$\mathbf{x}(k) = \mathbf{H}(k)\mathbf{s}(k) + \boldsymbol{\nu}(k), \tag{1}$$

where $\mathbf{s}(k)$ is an n-dimensional vector of unknown stochastically independent of source signals, $\mathbf{x}(k)$ is am m - dimensional observable sensor vector, (with $m \geq n$), $\mathbf{H}(k)$ is a full rank $m \times n$ mixing matrix and $\boldsymbol{\nu}(k)$ is $m$ dimensional uncorrelated with sources Gaussian noise vector. It is assumed that only the sensor vector is available and is necessary to design a feed-forward neural network and associated adaptive learning algorithm which enables estimation of sources and/or identification of mixing matrix $\mathbf{H}$ with good tracking abilities. The problem is often referred as noisy ICA (independent component analysis): the ICA of a noisy random vector $\mathbf{x} = [x_1 \cdots x_m]^T$ is obtained by finding an $n \times m$, full rank, linear transformation (un-mixing) matrix $\mathbf{W}$ such that the output signal vector $\mathbf{y} = [y_1 \cdots y_n]^T$, defined as $\mathbf{y} = \mathbf{W}\mathbf{x}$ contains component that are as independent as possible, as measured by an information - theoretic cost function such as minimum Kullback - Leibler divergence. In other words, it is required to adapt the synaptic weights of $w_{ij}$ of $m \times m$ matrix $\mathbf{W}$ of a linear system $\mathbf{y}(t) = \mathbf{W}(t)\mathbf{x}(t)$ (often referred to as a single-layer feed-forward neural network) to combine the observation $x_i(t)$ to form estimates of source signals $\hat{s}_j(t) = y_j(t) = \sum_{i=1}^{m} w_{ji}x_i(t)$. The optimal weights correspond to the statistical independence of output signals $y_j(t)$.

## II. DERIVATION OF BASIC ADAPTIVE ALGORITHMS

In order, to solve the above formulated ICA problems a key task is to formulate appropriate loss (cost) function which should be the function of the parameters of the specified neural network model. Minimization of such loss function should lead to satisfy desired conditions (stochastic independence and/or temporal

and spatial mutual de-correlation) of the output extracted signals. Recently several researchers (see e.g. Amari [1, 2, 3], Inouye [15], Matsuoka [17], Cardoso [7], Bell and Sejnowski [6], Girolami and Fyfe [14]) proposed useful criteria and loss functions (called also contrast, cost or energy functions) for BSS. Differential entropy maximization (DEM), independent component analysis (ICA) and maximization of likelihood (ML) lead to the same type of expected loss function which is measure of mutual stochastic independence of output signals [6, 12, 3, 7]. The natural gradient search method developed by Amari [1, 2, 3] has emerged as a particularly-useful technique for solving iterative optimization problems. The suitable expected loss or risk function could be defined as Kullback-Leibler divergence [1, 3]

$$E\{\phi(\mathbf{y},\mathbf{W})\} = -\int p_y(\mathbf{y}) \log \frac{p_y(\mathbf{y})}{p_M(\mathbf{y})} d\mathbf{y}, \qquad (2)$$

where $p_y(\mathbf{y})$ is the joint probability distribution of output signals and $p_M(\mathbf{y}) = \prod p_i(y_i)$ is the marginal distribution of $p_y(\mathbf{y},\mathbf{W})$.

Such risk function, which is in fact, equal to mutual information among outputs components of $\mathbf{y}$ leads to a simple loss (cost) function

$$\phi(\mathbf{y},\mathbf{W}) = -\log\det(\mathbf{W}^T\mathbf{W}) - \sum_{i=1}^{m} \log p_i(y_i), \qquad (3)$$

where $p_i(y_i)$ are probability density functions (p.d.f.) of output signals, $\det(\mathbf{W})$ means the determinant of matrix $\mathbf{W}$ and $(\cdot)^T$ is a transpose operator. Taking into account that gradient of the loss function can expressed as

$$\nabla_{\mathbf{W}}\phi(\mathbf{W}) = \frac{\partial\phi(\mathbf{W})}{\partial\mathbf{W}} = \mathbf{f}(\mathbf{y})\mathbf{x}^T - \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1} \qquad (4)$$

and applying natural gradient approach developed by Amari we can derive basic learning rule [1, 3, 4, 8, 10]

$$\begin{aligned}\Delta\mathbf{W}(k) &= \mathbf{W}(k+1) - \mathbf{W}(k) = -\eta\frac{\partial\phi(\mathbf{W})}{\partial\mathbf{W}}\mathbf{W}^T\mathbf{W} \\ &= \eta(k)[\mathbf{\Lambda}(k) - \mathbf{f}(\mathbf{y}(k))\mathbf{y}^T(k)]\mathbf{W}(k), \qquad (5)\end{aligned}$$

where $\mathbf{\Lambda}(k) = diag\{\lambda_1(k)\cdots\lambda_n(k)\}$ is any positive-definite diagonal scaling matrix, e.g. $\mathbf{\Lambda} = \mathbf{I}$ or $\mathbf{\Lambda}(k) = \sigma^2\mathbf{I} + diag\{\mathbf{f}(\mathbf{y}(k))\mathbf{y}^T(k)\}$ (where $\sigma^2$ represents variance of additive Gaussian noise) and $\mathbf{f}(\mathbf{y}) = [f_1(y_1),\cdots,f_n(y_n)]^T$ with nonlinearities $f_i(y_i) = -d\log(p_i(y_i))/dy_i = -\dot{p}_i(y_i)/p_i(y_i)$.

Alternatively, we can use the following (preconditioning) filtering gradient [5, 8, 10] :

$$\begin{aligned}\Delta\mathbf{W}(k) &= -\eta(k)\mathbf{W}\left[\frac{\partial\phi(\mathbf{W})}{\partial\mathbf{W}}\right]^T\mathbf{W} \\ &= \eta(k)[\mathbf{\Lambda}(k) - \mathbf{y}(k)\mathbf{f}(\mathbf{y}^T(k))]\mathbf{W}(k). \qquad (6)\end{aligned}$$

The above two learning rule could be combined together to build up more general and flexible (universal) learning rule (see Cichocki at al. [8, 10]:

$$\Delta\mathbf{W}(k) = \eta(k)[\mathbf{\Lambda}(k) - \mathbf{f}[\mathbf{y}(k)]\mathbf{g}(\mathbf{y}^T(k))]\mathbf{W}(k), \qquad (7)$$

where $\mathbf{f}(\mathbf{y})$ and $\mathbf{g}(\mathbf{y})$ are suitably designed nonlinear functions. In the special case when the number of sources is known we can assume that un-mixing matrix $\mathbf{W}$ is $n\times m$, however, in the general case when number of sources is unknown we assume that $\mathbf{W}$ is $m\times m$ quadratic matrix.

## III. DESIGN OF ACTIVATION FUNCTIONS

The performance of the learning algorithms strongly depends on shape of activation functions. Optimal selection of nonlinearities depend on p.d.f. of source signals. It can be proved that for specific nonlinearity $f(y_i) = \alpha_i y_i + tanh(\gamma_i y_i)$ the learning rule (5) is able to successfully separate signals if all of them are super-Gaussian signals while the learning rule (6) could separate them if all of them are sub-Gaussian signals. Analogously for $f(y_i) = \alpha_i y_i + y_i^3$ algorithm (5) separates sub-Gaussian signals while the algorithm (6) super-Gaussian signals. However, if the measured signals $x_i(k)$ contains mixtures of both sub-Gaussian and super-Gaussian sources then these algorithms may fail to separate these signals reliably (cf. [13, 14]).

In this paper we propose a new rigorous strategy for design flexible, near optimal activation functions that enable source signals from arbitrary non-Gaussian distributions to be extracted from the measurements of the mixed signals.

Let us assume that source signals have generalized Gaussian distributions of the form: $p_i(y_i) = \exp(\lambda_{0i} - 1)\exp\left(-\lambda_{1i}\frac{|y_i|^{q_i}}{\sigma_i^2}\right)$, where Lagrange multipliers and variance are expressed as follows: $\exp(\lambda_{0i} - 1) = \frac{q_i}{2(q_i\sigma_i^2)^{1/q_i}\Gamma(1/q_i)}$, $\lambda_{1i} = 1/q_i$, $\sigma_i^2 = \langle|y_i|^{q_i}\rangle$.

The locally optimal flexible normalized nonlinear activation functions can be expressed in such case as

$$f_i(y_i) = -\frac{d\log(p_i(y_i))}{dy_i} = |y_i|^{q_i-1}\mathrm{sign}(y_i) \quad q_i \geq 1. \qquad (8)$$

The parameter $q_i$ can change from 1 (Laplace distribution, through $q_i = 2$ - standard Gaussian distribution) to $q_i$ going to infinity (for uniform distribution). In the general case, when we do not have a priory knowledge about distribution of sources we can start from standard Gaussian density ($q_i = 2$ - linear activation functions) and adaptively change these parameters depending on estimated distance of density of actual output signals $y_i(k)$ from Gaussianity.

Analogous nonlinearity can be derived, for example, for generalized Cauchy distribution as $f_i(y_i) = [(vq_i +$

$1)/(v|A(p)|^{q_i} + |y_i|^{q_i})]|y_i|^{q_i-1}\text{sign}(y_i)$ and for generalized Rayleigh distribution as $f_i(y_i) = |y_i|^{q_i-2}y_i$ for complex - valued signals and coefficients.
Alternatively, we can consider a novel 'robust' generalized Gaussian distributions

$$p_i(y_i) = e^{\lambda_{0i}-1} \exp\left(-\lambda_{1i} \frac{|\log(\cosh(\beta_i y_i))/\beta_i|^{q_i}}{\sigma_i^2}\right) \quad (9)$$

where $\sigma_i^2 = \langle \log(\cosh(\beta_i y_i))/\beta_i \rangle$, $\beta_1 \geq 2$.
Such model of p.d.f. (for $q_i = 1$) is especially useful for noisy natural speech signals. For small variance of the signal $y$ we have $\log(\cosh(\beta y))/\beta \approx y^2$ so we can approximate the distribution (9) by the standard Gaussian distribution $p_y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{y^2}{2\sigma^2}\right)$ where $\sigma^2 = \langle y^2 \rangle$ (silent period). For large variance of the signal $\log(\cosh(\beta y))/\beta \approx |y|$ so we can approximate the distribution (9) by Laplace distribution $p_y(y) = \frac{1}{2\sigma^2} \exp\left(-\frac{|y|}{\sigma^2}\right)$ where $\sigma^2 = \langle |y| \rangle$.
Using this model and Amari's natural gradient approach for noisy data we derived a new algorithm of the form (7) with entries of diagonal matrix $\mathbf{\Lambda}(k)$ as

$$\lambda_i(k) = \sigma_i^2 + \frac{1}{\beta_i} \log(\cosh(\beta_i y_i(k))) \quad (10)$$

and flexible (adaptive) nonlinear activation functions

$$f_i(y_i) = \begin{cases} \tanh(\beta_i y_i) & \text{for } \kappa_4(y_i) > \delta \\ y_i & \text{for } \kappa_4(y_i) < \delta \end{cases} \quad (11)$$

$$g_i(y_i) = \begin{cases} y_i & \text{for } \kappa_4(y_i) > -\delta \\ \tanh(\beta_i y_i) & \text{for } \kappa_4(y_i) < -\delta, \end{cases} \quad (12)$$

where $\kappa_4(y_i) = E\{y_i^4\}/E^2\{y_i^2\} - 3$ is normalized value of kurtosis and $\delta \geq 0$ is a threshold. The value of kurtosis can be evaluated on-line as

$$E\{y_i^q(k+1)\} = (1-\eta)E\{y^q(k)\} + \eta|y_i|^q \ (q = 2, 4)$$

The above learning algorithm (7), (10)-(13) monitors and estimates the statistics of each output signal and depending on sign or value of its normalized kurtosis automatically select (or switch) suitable nonlinear activation functions, such that successful (stable) separation of all non-Gaussian source signals is possible. In this approach activation function are adaptive time-varying nonlinearities. It can be shown by mathematical analysis and computer simulation experiments that it is sufficient to use robust (in respect to outliers and spiky noise) nonlinearities of the form: $f_i(y_i) = \tanh(\beta_i y_i)$ or $g_i(y_i) = \tanh(\beta_i y_i)$ and it is not necessary to use nonlinearities of the form $f_i(y_i) = \text{sign}(y_i)|y_i|^{q_i-1}$ which are rather very sensitive to outliers for $q_i \geq 3$.

## IV. COMPUTER SIMULATIONS

Extensive computer simulation experiments confirm validity and high performance of the proposed algorithms In the case where $m \geq n$ and mixing sensors are noiseless the $n$ of output signals estimate all n unknown sources, while the additional $(m-n)$ outputs decay quickly to zero.

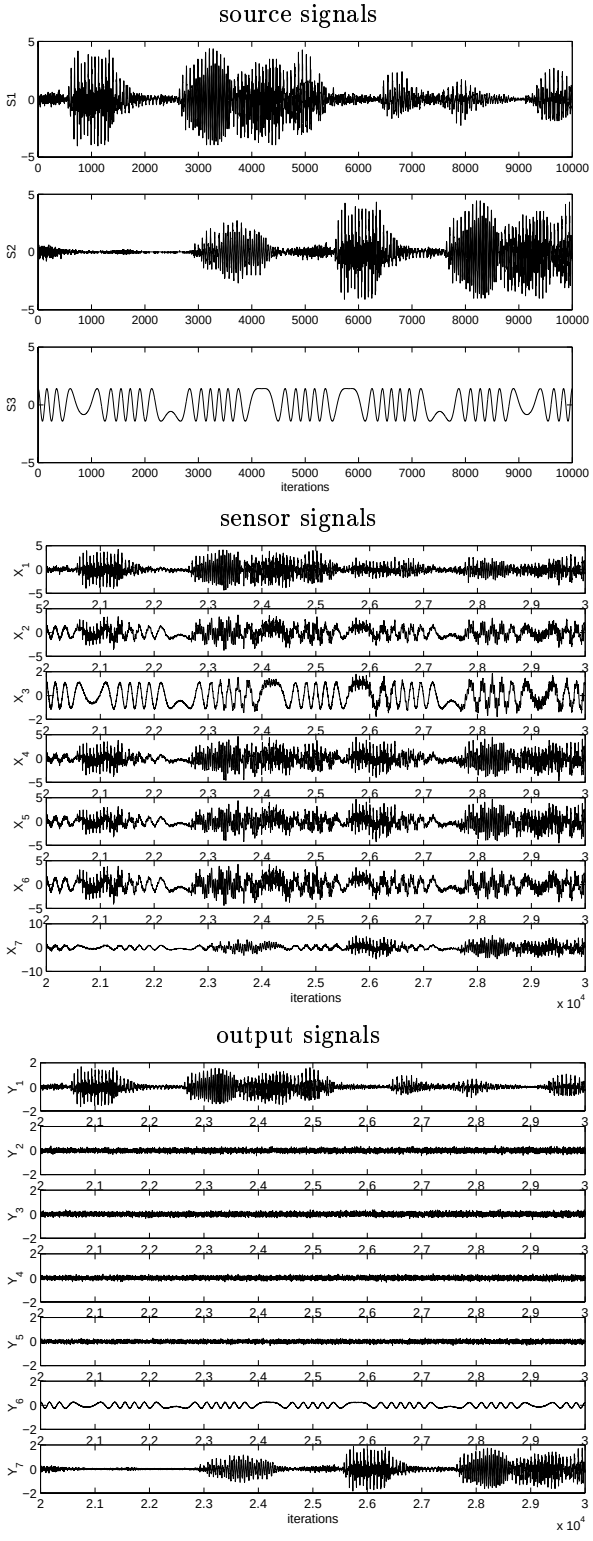


Fig.1 Exemplary computer simulation results

In more realistic scenario of noisy sensor signals the additional $(m-n)$ outputs collects only additive noises

(under condition that the noise is not too large, usually less than 3%) while the other outputs estimates the desired source signals with reduced noise. This feature is very desirable property of the proposed learning algorithms. Fig.1 illustrates typical simulation results. Three unknown acoustical signals (two natural speech signals with positive kurtosis and a single tone signal with negative kurtosis) where mixed using randomly selected $7 \times 3$ full rank mixing matrix $\mathbf{H}$. To sensors signals were added 2% i.i.d. Gaussian noises. The mixing matrix, the number of sources as well as their statistics were assumed to be completely unknown. The learning algorithm (7), (10) - (13) with self- adaptive learning rate $\eta(k)$ [9] and parameters $\sigma^2 = 0.025$, $\delta = 0.1$ and fixed $\beta_i = 10\ \forall i$ was able successfully estimate the number of active sources and their waveforms and also to 'shift' noise signals to free channels.

## V. CONCLUSIONS

In this paper we have proposed on - line adaptive learning algorithms for independent component analysis in the general case when sensor signals are noisy and the number of sources as well as their statistics are completely unknown. Locally optimal nonlinear functions are derived for various generalized distributions models. Extensive computer simulations have confirmed validity and excellent performance of the developed learning algorithms.

### References


[1] S. Amari, A. Cichocki, and H.H. Yang, "A new learning algorithm for blind signal separation", *Advances in Neural Information Processing Systems 1995* (Boston, MA: MIT Press, 1996), pp. 752-763.

[2] S. Amari "Natural gradient works efficiently in learning", *Neural Computation*, 1997, (in print).

[3] S. Amari, T.-P. Chen, and A. Cichocki, "Stability analysis of adaptive blind source separation", Neural Networks, 1997 (in print).

[4] S. Amari, T.-P. Chen, and A. Cichocki, "Nonholonomic orthogonal learning algorithms for blind separation" ICONIP-97 (in print).

[5] J. J. Attick and A. N. Redlich, "Convergent algorithm for sensory receptive fields development", *Neural Computation*, Vol. 5, 1993, pp.45-60.

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

[7] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation", *IEEE Trans. Signal Processing,* Vol. SP-44, no. 12, December 1996, pp. 3017-3030.

[8] A. Cichocki, R. Unbehauen, L. Moszczyński and E. Rummert, "A new on-line adaptive learning algorithm for blind separation of source signals", in *Proc. of ISANN-94*, Taiwan, 1994, 406–411.

[9] A. Cichocki, S. Amari, M, Adachi, W. Kasprzak, "Self-adaptive neural networks for blind separation of sources", *1996 IEEE International Symposium on Circuit and Systems*, IEEE, Piscataway, Vol.2, 1996, pp.157-160.

[10] A.Cichocki and R. Unbehauen, "Robust neural networks with on-line learning for blind identification and blind separation of sources", IEEE Trans. on Circuits and Systems - I Vol.43, Nov. 1996 (submitted in June 1994), pp.894-906.

[11] A. Cichocki, "Blind separation and extraction of source signals – recent results and open problems", Proc of 41th Annual Conference of the Institute of Systems, Control and Information Engineers, ISCIE-97 May 21-23, Osaka-Japan, May 1997, pp. 43- 48.

[12] P. Comon, "Independent component analysis: a new concept?", Signal Processing, vol.36, 1994, pp.287-314.

[13] S.C. Douglas, A. Cichocki and S. Amari, "Multichannel blind separation and deconvolution of sources with arbitrary distributions", in Proc. IEEE Workshop on Neural Networks for Signal Processing, Sept. 97 IEEE New York, pp. 436-444.

[14] M. Girolami and C. Fyfe, "Extraction of independent signals sources using a deflationary exploratory projection pursuit network with lateral inhibition", IEE Proceedings on Vision, Image and Signal processing, 1997 (in print).

[15] Y. Inouye and T. Sato, "On line algorithms for blind deconvolution of multichannel linear time-invariant systems". In Proc. IEEE Signal Processing Workshop on HOS, pp. 204-208, 1997.

[16] Z. Maluche and O. Macchi, "Adaptive separation of unknown number of sources", Proc. of IEEE Workshop on Higher Order Statistics, IEEE Computer Society, Los Almitos 1997, pp. 295-299.

[17] K. Matsuoka, M. Ohya and M. Kawamoto, "A neural net for blind separation of non-stationary signals", Neural Networks, 1995, Vol.8,No.3, pp.411-419.