AN OLD WORKFRAME FOR A NEW PROBLEM: A CLASSICAL MODEL FOR UNSUPERVISED ADAPTIVE FILTERING

M. Elsabrouty [*]

ABSTRACT

A classical adaptive filtering model of the problem of instantaneous blind signal separation, or what is formally known as unsupervised adaptive filtering is presented. This classical form helps understanding the well-known superior behaviour of the natural gradient solution to the blind separation problem. A new RLS-based algorithm is developed using this classical model. The algorithm provides improved on-line separation speed under the same steady state error compared to the natural gradient algorithm without requiring pre-whitening.

KEY WORDS

Blind Signal Separation, Natural Gradient, Recursive Least Square, Unsupervised Adaptive Filtering

[*] Faculty of Information and Engineering Technology, The German University in Cairo,
e-mail: maha.elsabrouty@guc.edu.eg

I. INTRODUCTION

Blind Source Separation (BSS) is intended to separate or recover a set of unobservable sources from another set of mixtures, without having access to the mixing matrix coefficients. One main category of cost functions for blind separation of instantaneous mixtures is information theory based methods [1]. This category is preferred due to its robustness against outliers, meaning that a single or a few highly erroneous observations should have little impact on the overall estimate of the sources. Information theory approaches aim at separating the mixtures into their basic components using estimates of the sources' probability densities. One of the main cost functions in the information theory category is the Maximum-Likelihood (ML) cost function.

In [2], the idea of applying differential geometry to the LMS-update of the maximum-likelihood cost function for blind separation led to the well-known natural gradient update which exhibits equivariant properties, meaning that the gradient is better directed to the minima of the cost function on the curved optimization manifold, i.e., Riemannian manifold.

In [3], the idea of iterative inversion of the mixing matrix using the maximum-likelihood cost function for blind separation was introduced. One of the main contributions in [3] is the ability to develop the natural gradient algorithm without the use of Riemannian geometry.

This paper presents another view of the blind source separation based on maximum-likelihood. We introduce a classical filtering analogy to the separation process by splitting the stochastic gradient into equivalent "input" and "error" components. Using this "input/error" representation enables the interpretation of the maximum-likelihood update as a classical adaptive filtering LMS-update.

A new RLS-algorithm stemming from this analogy is then introduced that can be reduced to the natural gradient update. The new algorithm, in its original non-reduced form, and the natural gradient are compared when applied to different audio mixtures. Simulation results show that the new proposed RLS-based algorithm achieves better convergence speed while producing the same steady state error as the natural gradient algorithm.

The rest of this paper is organized as follows. In section 2, the principle of signal separation based on the maximum-likelihood is presented, along with a review of the Riemannian manifold approach to natural gradient. Section 3 presents the adaptive filtering interpretation of the maximum-likelihood update. The new proposed RLS algorithm for BSS is presented in Section 4 along with the reduction to the natural gradient algorithm. Section 5 is dedicated to the test setup and the presentation of the results. We finally conclude the paper in Section 6.

II. MAXIMUM-LIKELIHOOD PRINCIPLE

Let $s_i(t), i = 1 \cdots N$ be the scalar source inputs for a mixing matrix $\mathbf{A}$ at a time n, with the resulting mixtures $x_i(t), i = 1 \cdots N$. For simplicity, it is assumed that the mixing is linear and that the mixing matrix $\mathbf{A}$ is square, i.e. of size $N \times N$. The mixing model can be expressed as:

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

where $\mathbf{x}$ and $\mathbf{s}$ are the $N \times 1$ vectors of the mixtures and the sources, respectively. The purpose of blind signal separation algorithms is to estimate an $N \times N$ matrix $\mathbf{W}$ and estimate output signals $y_i(t), i = 1 \cdots N$:

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

such that $\mathbf{W} \times \mathbf{A} = \mathbf{P}$, where $\mathbf{P}$ is a permutation of a diagonal matrix. If this condition is met, then the outputs, $\mathbf{y}(t)$ or better a smoothed nonlinear version of the outputs $g(\mathbf{y}(t))$ are equal to the independent sources $s_i(n), i = 1 \cdots N$, except for order and scale ambiguity [10].

The Maximum-likelihood cost function attempts to perform the separation via maximizing the joint entropy $H(\mathbf{y}; \mathbf{W})$, where $\mathbf{y}$ is the $N \times 1$ vector of the outputs $y_i(n)$ of the separation process. The concept has been introduced in [4]. The cost function of the log-likelihood $L(\mathbf{W})$ can be expressed as:

$$L(\mathbf{W}) = \log|\det(\mathbf{W})| + E\left\{ \sum_{i=1}^{N} \log p(y_i) \right\} \tag{3}$$

where $E\{.\}$ refers to the expected value, y_i is the ith output and $p(y_i)$ is the probability density function of y_i. The gradient of $L(\mathbf{W})$ on the Euclidean space is:

$$\nabla L(\mathbf{W}) = -(\mathbf{W}^{-1})^T + E\left\{ \mathbf{g}(\mathbf{y})\mathbf{x}^T \right\} \tag{4}$$

where $g(y_i) = -\dfrac{p(y_i)'}{p(y_i)}$. The operation $(.)'$ denotes the derivative. A good choice for $g(y_i)$ is [5]:

$$g(y_i) = \begin{cases} 2 \tanh(y_i), & \text{if } y_i \text{ is supergauss ian} \\ y_i^3, & \text{if } y_i \text{ is subgaussia n} \end{cases} \tag{5}$$

The natural gradient update can be obtained by finding the steepest descent direction of the cost function $L(\mathbf{W})$ on the Lie-algebra space associated with $\mathbf{W}$ [4]. The modified differential along this tangent space is [5]:

$$d\mathbf{M} = d\mathbf{W}\,\mathbf{W}^{-1} \tag{6}$$

such that:

$$d\mathbf{y} = d\mathbf{W}\,\mathbf{x} = d\mathbf{W}\,\mathbf{W}^{-1}\,\mathbf{y} = d\mathbf{M}\,\mathbf{y} \tag{7}$$

Differentiation of the cost function $L(\mathbf{W})$ with respect to $d\mathbf{M}$ results in two terms. The first term is:

$$d\left|\log\left(\det(\mathbf{W})\right)\right| = -tr(d\mathbf{M}) \tag{8}$$

such that $\dfrac{d\left|\log\left(\det(\mathbf{W})\right)\right|}{d\mathbf{M}} = \mathbf{I}$. The second term is $\dfrac{\partial \sum\limits_{i=1}^{N} E\{\log|p(y_i)|\}}{\partial \mathbf{M}}$ which can be estimated as follows:

$$\frac{\partial \sum\limits_{i=1}^{N} E\{\log|p(y_i)|\}}{\partial \mathbf{M}} = -E\left\{ g(y_1)\frac{\partial y_1}{\partial \mathbf{M}} + \cdots + g(y_N)\frac{\partial y_N}{\partial \mathbf{M}} \right\} = -E\left\{ g(\mathbf{y})\mathbf{y}^T \right\} \tag{9}$$

The update along the Lie-algebra is given by:

$$\Delta\mathbf{M} = \mu\, E\{(I - g(\mathbf{y})\mathbf{y}^T)\} \tag{10}$$

This update is mapped to the Riemannian space of $\mathbf{W}$ via exponential mapping [6]:

$$\mathbf{W}(t) = \exp(-\Delta\mathbf{M})\,\mathbf{W}(t-1) \tag{11}$$

The update $\Delta\mathbf{W}$ can be approximated to be:

$$\Delta\mathbf{W} \cong \left[I - E\{\mathbf{g}(\mathbf{y})\mathbf{y}^T\} \right]\mathbf{W}(t-1) \tag{12}$$

A stochastic form of this algorithm can also be developed. This is achieved by replacing the expectation by the instantaneous value, and in each iteration of the algorithm a small step size $\mu \ll 1$ is used to add the update to the demixing matrix $\mathbf{W}$:

$$\mathbf{W}(t) = \mathbf{W}(t-1) + \mu\left[I - \mathbf{g}(\mathbf{y})\mathbf{y}^T \right]\mathbf{W}(t-1) \tag{13}$$

Using equation (3) as the starting point, we propose in the following section a classical adaptive filtering interpretation of the maximum-likelihood update. This classical view facilitates the analysis of natural gradient algorithm update and also helps giving rise to a new RLS-based update based on the multiplication of an input signal by the inverse of its correlation matrix. The new algorithm can provide faster convergence than the natural gradient in the regular instantaneous mixtures, square mixing matrix case.

III. CLASSICAL INTERPRETATION OF THE NATURAL GRADIENT UPDATE

We consider the gradient of the maximum-likelihood cost function:

$$\nabla L(\mathbf{W}) = -(\mathbf{W}^{-1})^T + E\{g(\mathbf{y})\mathbf{x}^T\} \tag{14}$$

This form can be re-written as:

$$\nabla L(\mathbf{W}) = -\mathbf{W}^{-T}\mathbf{R_{xx}}^{-1} E\{\mathbf{x}\mathbf{x}^T\} + E\{g(\mathbf{y})\mathbf{x}^T\} \tag{15}$$

with the autocorrelation $\mathbf{R_{xx}} = E\{\mathbf{x}\mathbf{x}^T\}$. An instantaneous version of the above update can then be re-written as:

$$\nabla L(\mathbf{W}) = \left(-\mathbf{W}^{-T}\mathbf{R_{xx}}^{-1}\mathbf{x} + g(\mathbf{y})\right)\mathbf{x}^T \tag{16}$$

As in [11], the above update can be interpreted as an error term $\left(-\mathbf{W}^{-T}\mathbf{R_{xx}}^{-1}\mathbf{x} + g(\mathbf{y})\right)$ multiplied by an input term $\mathbf{x}^T$.

$$\nabla L(\mathbf{W}) = \underbrace{\left(-\mathbf{W}^{-T}\mathbf{R_{xx}}^{-1}\mathbf{x} + g(\mathbf{y})\right)}_{error} \underbrace{\mathbf{x}^T}_{input} \tag{17}$$

Such that:

$$\mathbf{W}(t+1) = \mathbf{W}(t) + \mu \underbrace{\left(-\mathbf{W}^{-T}\mathbf{R_{xx}}^{-1}\mathbf{x} + g(\mathbf{y})\right)}_{error} \underbrace{\mathbf{x}^T}_{input} \tag{18}$$

which has a classical LMS-update form. A closer look at the error term $\left(-\mathbf{W}^{-T}\mathbf{R_{xx}}^{-1}\mathbf{x} + g(\mathbf{y})\right)$ reveals that $\mathbf{R_{xx}}^{-1}\mathbf{x}$ can be considered as a whitened version of the mixtures $\mathbf{X}$ given that the mixtures $\mathbf{X}$ have zero-mean and in this case the transpose of the inverse of the de-mixing matrix is equivalent to $\mathbf{W}$ under the orthogonality constraints [1]. In this case the error term can be written as:

$$error = \left(-\mathbf{W}\,\mathbf{R_{xx}}^{-1}\mathbf{x} + g(\mathbf{y})\right) \tag{19}$$

Thus, the error term can be analyzed as the difference between the estimated sources $g(\mathbf{y})$ and the output of the whitened input mixtures $\mathbf{X}$ applied on the demixing matrix $\mathbf{W}$.

The above interpretation of the maximum-likelihood update gives rise to a spectrum of variant algorithms based on classical adaptive filtering theory. In the following section an RLS-based update is presented, along with an interpretation of the natural gradient as a special case of this RLS-based algorithm.

IV. RLS-BASED BSS ALGORITHM

In this section, a new approach to blind signal separation is developed in analogy to the classical adaptive filtering RLS algorithm, which bases its work on a multiplication of the input signal by the inverse of its correlation matrix. The developed algorithm can be updated using the iterative inversion and the well-known matrix inversion lemma.

As a starting point, we consider the gradient of the maximum-likelihood cost function:

$$\mathbf{W}(t+1) = \mathbf{W}(t) + \mu \underbrace{\left(-\mathbf{W}^T \mathbf{R_{xx}}^{-1} \mathbf{x} + g(\mathbf{y}) \right)}_{error} \underbrace{\mathbf{x}^T}_{input} \tag{20}$$

To create the new algorithm based to the above LMS-based one, we modify the update $\nabla L(\mathbf{W}) = \left(-\mathbf{W}^{-T} \mathbf{R_{xx}}^{-1} \mathbf{x} + g(\mathbf{y}) \right) \mathbf{x}^T$ by post-multiplying it with the inverse of the correlation matrix $\mathbf{R_{xx}}$ of the mixtures input mixtures $\mathbf{X}$:

$$\nabla L(\mathbf{W}) = \left(-\mathbf{W}^{-T} \mathbf{R_{xx}}^{-1} \mathbf{x} + g(\mathbf{y}) \right) \mathbf{x}^T \mathbf{R_{xx}}^{-1} = \left(-\mathbf{W}^{-T} + g(\mathbf{y}) \mathbf{x}^T \right) \mathbf{R_{xx}}^{-1} \tag{21}$$

Assuming that the unobservable source signals are equivalent to the nonlinear outputs $g(\mathbf{y})$ [10] such that:

$$\mathbf{R}_{ss} = E\{\mathbf{s}\mathbf{s}^T\} = E\{g(\mathbf{y}) g(\mathbf{y})^T\} = \mathbf{R}_{g(\mathbf{y})g(\mathbf{y})} \tag{22}$$

the correlation matrix $\mathbf{R_{xx}}$ can be written as:

$$\mathbf{R_{xx}} = E\{\mathbf{x}\mathbf{x}^T\} = \mathbf{A} E\{\mathbf{s}\mathbf{s}^T\} \mathbf{A}^T \approx \mathbf{A} E\{g(\mathbf{y}) g(\mathbf{y})^T\} \mathbf{A}^T = \mathbf{A} \mathbf{R}_{g(\mathbf{y})g(\mathbf{y})} \mathbf{A}^T \tag{23}$$

Using equation (22), the inverse $\mathbf{R_{xx}}^{-1}$ of the correlation matrix $\mathbf{R_{xx}}$ can be estimated as:

$$\mathbf{R_{xx}}^{-1} = \mathbf{W}^T \mathbf{R}_{g(\mathbf{y})g(\mathbf{y})}^{-1} \mathbf{W} \tag{24}$$

Thus the RLS- update in (21) can be re-written as:

$$\mathbf{W}(t+1) = \mathbf{W}(t) + \mu \left(\mathbf{I} - g(\mathbf{y}) \mathbf{y}^T \right) \mathbf{R}_{g(\mathbf{y})g(\mathbf{y})}^{-1} \mathbf{W}(t) \tag{25}$$

The above RLS-form can reduce to the well-known natural gradient algorithm given in equation (13) as follows:

at the steady state point, i.e. when separation is reached, the output signals $\mathbf{y}$ are independent. As such:

$$\text{E}\{\, g(\mathbf{y})\, g(\mathbf{y})^{\mathrm{T}}\,\} = \begin{pmatrix} \alpha_1 & & \\ & \ddots & \\ & & \alpha_N \end{pmatrix} \tag{26}$$

where $\alpha_i = E\{g(y_i)g(y_i)\}$, $i=1\cdots N$ is a constant factor that only adds scaling factor to each separated source $y_i \to \alpha_i\, y_i$, $i=1\cdots N$. This scaling factor can be neglected as the separation process itself introduces scaling ambiguity in the separated sources. Taking this ambiguity into consideration, and applying this approximated optimal value of $\mathbf{R}_{g(\mathbf{y})g(\mathbf{y})}$ in equation (24), the resulting equation can be written as:

$$\mathbf{W}(t) = \mathbf{W}(t-1) + \mu \left(\mathbf{I} - g(\mathbf{y})\,\mathbf{y}^T \right) \mathbf{W}(t-1) \tag{27}$$

This is the same as the natural gradient equation developed in equation (13). Thus, the natural gradient can be considered as a special case of the developed RLS algorithm in (24).

However, the introduction of the inverse matrix $\mathbf{R}_{g(\mathbf{y})g(\mathbf{y})}^{-1}$ in the RLS algorithm (24) without approximating the matrix by its diagonal value at the optimal separation points helps improving the convergence speed of the algorithm. The estimation of the autocorrelation function $\mathbf{R}_{g(\mathbf{y})g(\mathbf{y})} = E\{g(\mathbf{y})g(\mathbf{y})^T\}$ is done by a running sum of the expected value updated from time instant (*t*-1) to time *t* using an exponential window with a forgetting factor λ as:

$$\mathbf{R}_{g(\mathbf{y})g(\mathbf{y})}(t) = \lambda\, \mathbf{R}_{g(\mathbf{y})g(\mathbf{y})}(t-1) + g(\mathbf{y}(t))g(\mathbf{y}(t))^T \tag{28}$$

Considering the additional computations required for this algorithm, it is easily seen that the main increase is in the matrix inversion process to estimate $\mathbf{R}_{g(\mathbf{y})g(\mathbf{y})}^{-1}$ which could be in the order of N^3 for *N* sources. However, using the matrix inversion lemma [9] to estimate $\mathbf{R}_{g(\mathbf{y})g(\mathbf{y})}^{-1}$ iteratively from one time instant to the next, this computational load can be reduced to the order N^2, which lies in the same range of complexity of the natural gradient algorithm.

V. SIMULATION RESULTS

In order to assess the improvement gained by the proposed RLS algorithm given by (25), it was compared to the natural gradient algorithm, given by (13). The step size of the natural gradient algorithm was set to 0.0005, while the forgetting factor λ of the RLS-algorithm was set to the fixed value of 0.999. The values of the step size for the natural gradient and the forgetting factor for the RLS-algorithm were chosen so that they converge to the same steady state error. The nonlinear function $g(\cdot)$ used was: $g(\mathbf{y}) = 2\tanh(\mathbf{y})$.

The tests were performed on audio data files from [7]. These files are sampled at 8 kHz and are of duration 6.5 sec. each, with the following two groups:

(a) Miscellaneous_4: two males, one female, and soft music.
(b) Male_4: 4 male speakers with similar characteristics.

We used the following performance index [8], [9]:

$$\eta = 20 \log \left(\sum_{i=1}^{N} \left(\sum_{j=1}^{N} \frac{\left|P_{ij}\right|}{\max_k \left|P_{ik}\right|} - 1 \right) + \sum_{j=1}^{N} \left(\sum_{i=1}^{N} \frac{\left|P_{ij}\right|}{\max_k \left|P_{kj}\right|} - 1 \right) \right)$$

(29)

where P_{ij} is one element of the matrix $\mathbf{P} = \mathbf{WA}$, which is supposed to be a scaled permutation of a diagonal matrix at the optimal point of separation. The above cost function thus penalizes the deviation of **P** from the permutation structure.

The test was performed for 100 different runs. The matrix $\mathbf{A}$ was randomly generated for each run. Figure 1 and Figure 2 illustrate the results of the simulations, showing the improved performance of the proposed algorithm for the same steady state error.
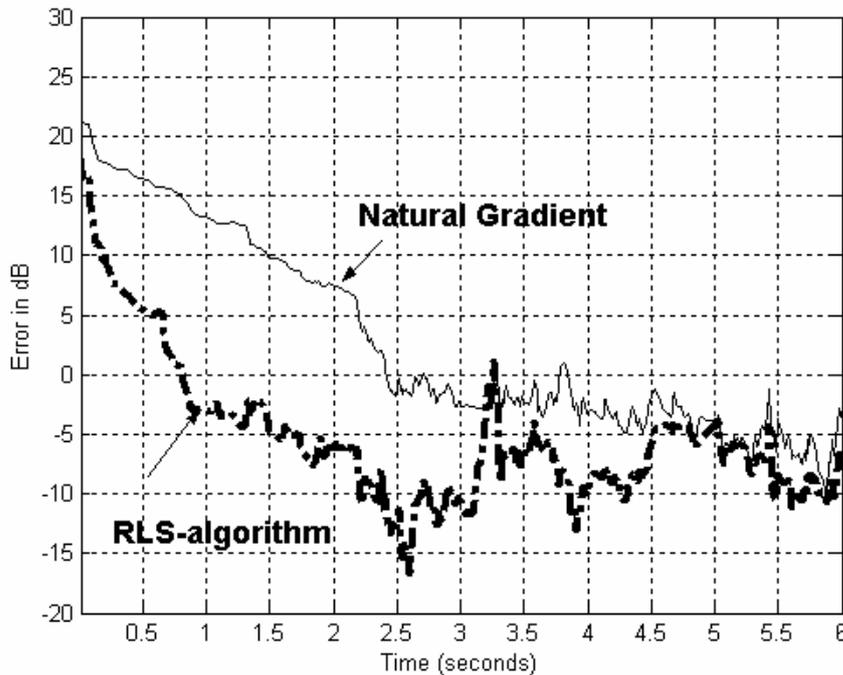

Fig1. RLS-algorithm vs. Natural Gradient for 100 different runs with the first group of sources (Miscelleneous_4)
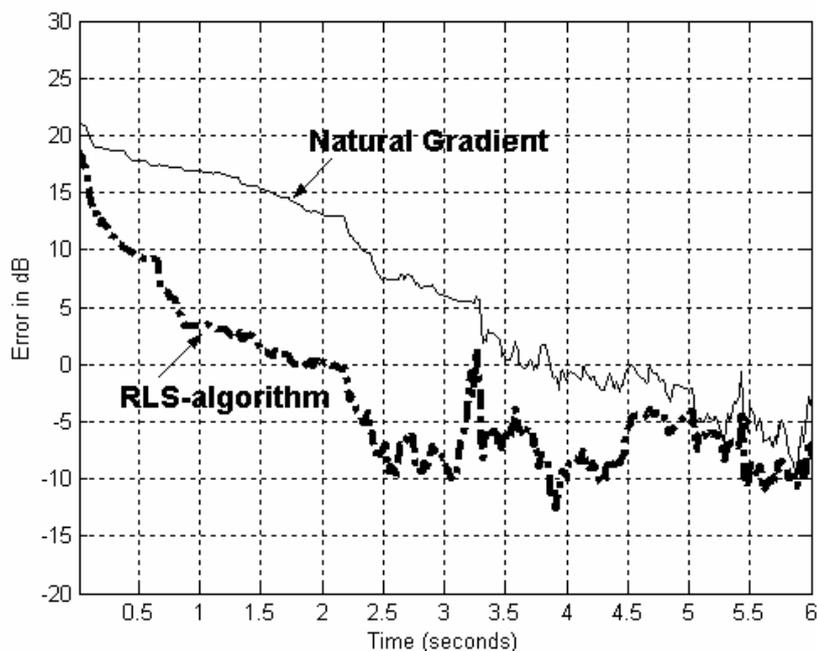
Fig2. RLS-algorithm vs. Natural Gradient for 100 different runs with the second group of sources (Male_4)

VI. CONCLUSION

This paper introduces a classical adaptive filtering view of the blind source separation process in its instantaneous form using the maximum-likelihood principle. This global classical view helps establishing a clear analogy between well known adaptive filtering algorithms and blind source separation algorithm. The regular Maximum-likelihood update was linked to the LMS update. This approach also gives an interpretation of the natural gradient algorithm and its improved performance without rendering to the Riemannian geometry mathematics. A new algorithm that resembles the classical RLS algorithm in its principle of work has also been developed. This algorithm modifies the maximum-likelihood update to perform a multiplication of the input signal by the inverse of its correlation matrix. The new RLS-based algorithm is capable of working on-line with good results, as confirmed by the assessment quality measure that was used. This improved performance can be achieved with a reasonable additional computational load.

REFERENCES

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

[2] S. Amari, A. Cichocki, and H.H. Yang, "A new learning algorithm for blind source separation". In *Advances in NeuralInformation Processing 8,* pp. 757-763, MIT Press, Cambridge, MA, 1996.

[3] S. Cruces-Alvarez, A. Cichocki, and L. Castedo-Ribas, "An iterative inversion approach to blind source separation", *IEEE Transactions on Neural Networks*, vol.11 , no.6 , pp. 1423 - 1437, Nov. 2000.

[4] A. Bell, and T. Sejowenski, "An information-maximization approach for blind separation and blind deconvolution". *Neural Comput.*, vol.7, pp. 1129-1159, 1995.

[5] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing*, John Wiley & Sons, NY, 2002.

[6] W. Boothby, *An Introduction to Differential Manifolds and Riemannian Geometry*, 2nd Ed., Academic Press 2002.

[7] Helsinki University of Technology, Laboratory of Computer and Information Science, Neural Network Research Center, "ICA Sources and Demonstration" *http://www.cis.hut.fi/projects/ica/cocktail/cocktailen.cgi*

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

[9] W. H. Press et al. "*Numerical Recipes in C++: The Art of Scientific Computing*". 2nd Ed., Cambridge University Press, 2002.

[10] M. Joho, *A Systematic Approach to Adaptive Algorithms for Multi-Channel System Identification, Inverse Modeling and Blind Identification*, PhD thesis, Swiss Federal Institute of Technology, Zurich, 2000.

[11] M. Bouchard, "New recursive-least-squares algorithms for active control of sound and vibration using neural networks", *IEEE Trans. on Neural Networks*, vol. 12, pp. 135-147, Jan. 2001.