ON SOME EXTENSIONS OF THE NATURAL GRADIENT ALGORITHM

Pando Georgiev [a], *Andrzej Cichocki* [b] *and Shun-ichi Amari* [c]

Brain Science Institute, RIKEN, Wako-shi, Saitama 351-01, Japan
[a] On leave from the Sofia University ”St. Kl. Ohridski”, Bulgaria
E-mail: georgiev@bsp.brain.riken.go.jp
[b] On leave from the Warsaw University of Technology, Poland
E-mail: cia@bsp.brain.riken.go.jp
[c] E-mail: amari@bsp.brain.riken.go.jp

ABSTRACT

Recently several novel gradient descent approaches like natural or relative gradient methods have been proposed to derive rigorously various powerful ICA algorithms. In this paper we propose some extensions of Amari’s Natural Gradient and Atick-Redlich formulas. They allow us to derive rigorously some already known algorithms, like for example, robust ICA algorithm and local algorithm for blind decorrelation. Furthermore, we hope they enable us to generate the family of new algorithms with improved convergence speed or performance for various applications. We present conditions for which the proposed general gradient descent dynamical systems are stable. We show that the nonholonomic orthogonal algorithm can not be derived from minimization of any cost function. We propose a stabilized nonholonomic algorithm, which preserves the norm of the demixing matrix.

1. INTRODUCTION

Gradient techniques are established and well known methods for adjusting a set of parameters to minimize or maximize a chosen cost function. However, simple standard gradient descent techniques can be rather slow and the system can stuck in local minima. Recently, in order to improve convergence speed for matrix algorithms, several novel gradient systems has been proposed and their dynamic properties have been investigated (see [1-11]).

In 1995, Amari (see [1-5]) introduced natural gradient approach which can be written in compact form as:

$$\frac{d\mathbf{W}}{dt} = -\mu \frac{\partial J}{\partial \mathbf{W}} \mathbf{W}^T \mathbf{W}, \tag{1}$$

where J is a suitable nonnegative cost function. Independently Cardoso [7] introduced the relative gradient which is equivalent to natural gradient formula.

In 1993, Atick and Redlich introduced the following gradient formula [6]:

$$\frac{d\mathbf{W}}{dt} = -\mu \mathbf{W} \left[\frac{\partial J}{\partial \mathbf{W}} \right]^T \mathbf{W} \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{n \times n}$ and $J(\mathbf{W}, \mathbf{y})$ is suitable chosen cost function.

The main objective of the paper is to investigate the basic dynamic properties of these gradient systems and to propose some extension and generalization which can be useful in deriving some ICA algorithms. We consider specific dynamical system and prove that some algorithms in ICA can be obtained by adjusting the parameters of this dynamical system. We give conditions under which this dynamical system possesses Lyapunov function, which proves rigorously the convergence of some algorithms.

2. CONVERGENCE PROOF VIA NONHOLONOMIC BASIS

In this section we shall demonstrate that the nonholonomic basis $d\mathbf{X}$ can be used for proving convergence of the natural gradient algorithm and the Atick-Redlich algorithm (under some conditions). We refer to [4] for the properties of nonholonomic basis $d\mathbf{X}$.

Using nonholonomic basis $d\mathbf{X} = d\mathbf{W}\mathbf{W}^{-1}$, (1) and (2) become respectively

$$\frac{d\mathbf{X}}{dt} = -\mu \frac{\partial J}{\partial \mathbf{W}} \mathbf{W}^T \tag{3}$$

and

$$\frac{d\mathbf{X}}{dt} = -\mu \mathbf{W} \left[\frac{\partial J}{\partial \mathbf{W}} \right]^T . \tag{4}$$

Putting $\mathbf{H} = \frac{\partial J}{\partial \mathbf{W}} \mathbf{W}^T$ we have $\mathbf{H} = \frac{\partial J}{\partial \mathbf{X}}$ (consult with [**?**] for a proof). By (3), we obtain

$$\frac{dJ(\mathbf{W}(t))}{dt} \quad = \quad -\mu \,\mathrm{trace}(\mathbf{H}^T \mathbf{H}) \tag{5}$$

$$= \quad -\mu \sum_{i,j=1}^{n} H_{i,j}^2 \leq 0,$$

as equality is achieved if and only if $\frac{\partial J}{\partial \mathbf{W}} = 0$ (assuming that $\mathbf{W}$ is nonsingular).

Analogously, by (4) we obtain

$$\begin{aligned} \frac{dJ(\mathbf{W}(t))}{dt} &= -\mu \operatorname{trace}(\mathbf{HH}) \qquad (6) \\ &= -\mu \sum_{i,j=1}^{n} H_{i,j} H_{j,i} \, . \end{aligned}$$

The equation (5) shows that J is a Lyapunov function of (1) (in wide sense).

The trace in (6) is not always positive. Let us decompose $\mathbf{H}$ as

$$\mathbf{H} = \mathbf{S} + \mathbf{A},$$

where $\mathbf{S}$ is symmetric and $\mathbf{A}$ is antisymmetric. Then we have

$$\begin{aligned} \operatorname{trace}(\mathbf{HH}) &= \sum_{i,j=1}^{n} H_{i,j} H_{j,i} \qquad (7) \\ &= \sum_{i,j=1}^{n} S_{i,j}^2 - \sum_{i,j=1}^{n} A_{i,j}^2 \\ &= \|\mathbf{S}\|^2 - \|\mathbf{A}\|^2. \end{aligned}$$

This gives a sufficient condition for convergence of Atick-Redlich algorithm, that is, $\|\mathbf{S}\| > \|\mathbf{A}\|$ for any $\mathbf{W}$ (the matrices $\mathbf{H}$ and $\mathbf{S}, \mathbf{A}$ depend on $\mathbf{W}$).

We note that, using this approach, it is possible to give another proof of Theorem 1 below.

3. EXTENSION OF THE NATURAL GRADIENT ALGORITHM AND ATICK-REDLICH FORMULA

In the following theorem we describe a general dynamical system defined by a matrix function.

Theorem 1 *Consider a dynamical system, described by the following differential equation*

$$\frac{d\mathbf{W}}{dt} = -\boldsymbol{\eta}\Big(\frac{\partial J(\mathbf{W})}{\partial \mathbf{W}}\Big) F(\mathbf{W})^T F(\mathbf{W}), \qquad (8)$$

where $\mathbf{W} \in \mathbb{R}^{n \times n}$ is a square matrix (depending on t), $J : \mathbb{R}^{n\times n} \to \mathbb{R}$ is differentiable function with matrix argument, bounded below, $F : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ is an arbitrary matrix-valued function with matrix arguments and nonsingular values, and $\boldsymbol{\eta} \in \mathbb{R}^{n\times n}$ is a symmetric positively definite matrix (possibly depending on t). Then J is a Lyapunov function (in wide sense) for the dynamical system (8).

Proof. Denote by $w_{ij}, f_{ij}, \eta_{ij}$ and b_{ij}, $i, j = 1, ..., n$, the elements of the matrices $\mathbf{W}, F(\mathbf{W}), \boldsymbol{\eta}$ and $\mathbf{B} = \Big(\frac{\partial J(\mathbf{W})}{\partial \mathbf{W}}\Big) F(\mathbf{W})^T$ respectively. We calculate:

$$\begin{aligned} \frac{dJ(\mathbf{W}(t))}{dt} &= \sum_{i,j=1}^{n} \frac{\partial J}{\partial w_{ij}} \frac{dw_{ij}}{dt} \\ &= -\sum_{i,j=1}^{n} \frac{\partial J}{w_{ij}} \sum_{l,k=1}^{n} \eta_{il} b_{lk} f_{kj} \\ &= -\sum_{i,k=1}^{n} b_{ik} \sum_{l=1}^{n} \eta_{il} b_{lk} \\ &= -\sum_{r=1}^{n} \mathbf{b}_r^T \boldsymbol{\eta} \mathbf{b}_r \\ &\leq 0, \end{aligned}$$

where $\mathbf{b}_r$ denotes the r-th vector-column of $\mathbf{B}$. It is easy to see that zero is achieved if and only if $\mathbf{b}_r = 0$ for every $r = 1, ..., n$, i.e. when $\frac{d\mathbf{W}}{dt} = 0$. ∎

Here we present similar extension of Atick-Redlich formula.

Theorem 2 *Consider a dynamical system, described by the following matrix differential equation*

$$\frac{d\mathbf{W}}{dt} = -F(\mathbf{W})\boldsymbol{\eta}\Big(\frac{\partial J(\mathbf{W})}{\partial \mathbf{W}}\Big)^T F(\mathbf{W}), \qquad (9)$$

where $J : \mathbb{R}^{n\times n} \to \mathbb{R}$ is a differentiable function with matrix argument $F : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ is an arbitrary matrix-valued function with matrix arguments and nonsingular values, and $\boldsymbol{\eta} \in \mathbb{R}^{n\times n}$ is a symmetric positively definite matrix (possibly depending on t). Assume that the function $L : \mathbb{R}^{n\times n} \to \mathbb{R}$ is a solution of the following system of differential equations:

$$(F(\mathbf{W}))^T \frac{\partial L(\mathbf{W})}{\partial \mathbf{W}} = \Big(\frac{\partial J(\mathbf{W})}{\partial \mathbf{W}}\Big)^T F(\mathbf{W}).$$

Then L is a Lyapunov function (in wide sense) for the dynamical system (9).

The proof is similar to that one of Theorem 1 and is omitted.

We note that when $F(\mathbf{W})\Big(\frac{\partial J}{\partial \mathbf{W}}\Big)^T$ is symmetric and $\boldsymbol{\eta}$ is scalar, then (9) is reduced to (8).

A sufficient condition of convergence of the trajectories of (9), when $\boldsymbol{\eta}$ is scalar, is similar as those for the original Atick-Redlich formula and is proved analogously: $\|\mathbf{S}\| > \|\mathbf{A}\|$ for a decomposition of the matrix $\mathbf{H} := \frac{\partial J}{\partial \mathbf{W}} \left(F(\mathbf{W})\right)^T$ as $\mathbf{H} = \mathbf{S} + \mathbf{A}$, where $\mathbf{S}$ is symmetric and $\mathbf{A}$ is antisymmetric (for any $\mathbf{W}$).

4. NONHOLONOMIC ALGORITHM IS NOT A GRADIENT ALGORITHM

In this section we state that the nonholonomic orthogonal algorithm [4] is not derived from minimization of any cost function. The main observation for proving this fact is that for a given diagonal matrix $\mathbf{D}$ (different from the identity matrix) there is no function $\varphi(\mathbf{W})$ such that

$$\frac{\partial\varphi(\mathbf{W})}{\partial\mathbf{W}} = \mathbf{D}\mathbf{W}^{-T}.$$

This fact follows from the criterium for existence of potential functions (see [12], Theorem 3.4).

5. NORM PRESERVING ALGORITHMS

In this section we consider the following nonholonomic algorithm, which stabilizes the norm of the matrix $\mathbf{W}$ to be fixed:

$$\frac{d\mathbf{W}}{dt} = \eta(t)\Big(\mathbf{F}(\mathbf{y}) - \mathrm{trace}\big(\mathbf{F}(\mathbf{y})\mathbf{W}\mathbf{W}^T\big)\mathbf{I}\Big)\mathbf{W}(t), \quad (10)$$

where F is activation function in sense of [2] and $F_{ii} = 0, i = 1, ..., n$. If the mixing matrix is orthogonal, then this algorithms is equivariant, since it can be written in the form:

$$\frac{d\mathbf{G}}{dt} = \eta(t)\Big(\mathbf{F}(\mathbf{y}) - \mathrm{trace}\big(\mathbf{F}(\mathbf{y})\mathbf{G}\mathbf{G}^T\big)\mathbf{I}\Big)\mathbf{G}(t),$$

where $\mathbf{G} = \mathbf{W}\mathbf{A}$.

It is easy to check that

$$\begin{aligned}&\frac{d\mathrm{trace}(\mathbf{W}^T\mathbf{W})}{dt} = \quad (11)\\ &= 2\eta(t)\mathrm{trace}(\mathbf{F}(\mathbf{y})\mathbf{W}\mathbf{W}^T)(1 - \mathrm{trace}(\mathbf{W}^T\mathbf{W})).\end{aligned}$$

So, if the initial matrix $\mathbf{W}(0)$ has unit norm, the norm of $\mathbf{W}(t)$ is preserved to be one. This property helps to recover the original signals, when their number is unknown (but less than the number of sensors). An example is shown in [4] for extraction of 4 signals from a linear mixture of 3 source signals; the usage of the standard natural algorithm leads to growing of the norm of the demixing matrix $\mathbf{W}$ to infinity; the usage of the nonholonomic algorithms reduces this explosion to a fluctuation. In general, stability of the norm of $\mathbf{W}$ in the standard nonholonomic algorithm is not proven, so the stabilizing form (10) can be used. Something more, the stability conditions of the standard nonholonomic algorithm [4] and (10) are the same (when $F_{i,j} = -\mathbf{f}_i(\mathbf{y}_i)\mathbf{y}_j^T, i \neq j, F_{i,i} = 0$).

6. ILLUSTRATIVE EXAMPLES

6.1. Robust Extended ICA Algorithm

Let us illustrate the application of Theorem 1 for deriving the class of robust algorithms for ICA [8], [9] with two nonlinear activation functions $\mathbf{f}(\mathbf{y})$ and $\mathbf{g}(\mathbf{y}) = \mathbf{D}(\mathbf{y})\mathbf{y}$:

$$\Delta\mathbf{W}(l) = \eta\left[\mathbf{I} - \left\langle\mathbf{f}(\mathbf{y})\mathbf{g}^T(\mathbf{y})\right\rangle\right]\mathbf{W}(l), \quad (12)$$

where $\mathbf{D}(\mathbf{y})$ is a diagonal matrix with positive entries, $\langle . \rangle$ is the expectation operator and $\mathbf{y}^l(k) = \mathbf{y}(k) = \mathbf{W}(l)\mathbf{x}(k)$.

It should be noted that the standard natural gradient leads to the following equivariant ICA algorithm [1]-[4]

$$\Delta\mathbf{W}(l) = \eta(l)\left[\mathbf{I} - \left\langle\mathbf{f}(\mathbf{y})\mathbf{y}^T\right\rangle\right]\mathbf{W}(l). \quad (13)$$

For this purpose consider a special form of Theorem 1 for $F(\mathbf{W}) = \mathbf{D}(\mathbf{y})^{1/2}\mathbf{W}$:

$$\Delta\mathbf{W} = -\eta\frac{\partial J(\mathbf{y},\mathbf{W})}{\partial\mathbf{W}}\mathbf{W}^T\mathbf{D}(\mathbf{y})\mathbf{W}, \quad (14)$$

where $J(\mathbf{y},\mathbf{W})$ is suitably selected cost function. Theorem 1 ensures stable gradient descent search of a local minimum of the cost function.

Let us consider, as example, the cost function

$$J(\mathbf{y},\mathbf{W}) = -\log|\det(\mathbf{W})| - \sum_{i=1}^{n}\log(p_i(y_i)). \quad (15)$$

The gradient is:

$$\frac{\partial J(\mathbf{y},\mathbf{W})}{\partial\mathbf{W}} = -\mathbf{W}^{-T} + \tilde{\mathbf{f}}(\mathbf{y})\mathbf{x}^T, \quad (16)$$

where $\tilde{\mathbf{f}}(\mathbf{y}) = [\tilde{f}_1(y_1), \cdots, f_n(y_n)]^T$ with $f_i(y_i) = -d\log(p(y_i)/dy_i$.

Applying formula (14) and taking average equation, we obtain the known robust learning rule

$$\Delta\mathbf{W}(l) = \tilde{\eta}\left[\langle\mathbf{D}\rangle - \left\langle\tilde{\mathbf{f}}(\mathbf{y})\mathbf{g}^T(\mathbf{y})\right\rangle\right]\mathbf{W}(l), \quad (17)$$

where $\tilde{\eta} = \eta\langle\mathbf{D}\rangle^{-1}$.

This can be written in the form (12) for $\mathbf{f}(\mathbf{y}) = \langle\mathbf{D}\rangle^{-1}\tilde{\mathbf{f}}(\mathbf{y})$.

Let us mention the special case, for symmetric pdf distributions of sources and odd activations functions $f_i(y_i)$ and

$$\mathbf{D} = \mathbf{diag}\{|y_1|^p, \cdots, |y_n|^p\}.$$

When $p = -1$ the algorithm simplifies to the median learning rule

$$\Delta\mathbf{W}(l) = \tilde{\eta}\left[\langle\mathbf{D}\rangle - \left\langle\tilde{\mathbf{f}}(y)[\mathrm{sign}(\mathbf{y})]^T\right\rangle\right]\mathbf{W}(l), \quad (18)$$

where $\mathrm{sign}(\mathbf{y}) = [\mathrm{sign}(y_1), \cdots, \mathrm{sign}(y_n)]^T$. Simulation results show that such median learning rule with sign activation function is more robust to additive noise.

6.2. Local Algorithm for Blind Decorrelation

If $p_i(y) = e^{y^2/2}$ in (15) and $\mathbf{D} = \mathbf{I}$ in (14), we obtain the Local Algorithm for Blind Decorrelation:

$$\frac{d\mathbf{W}}{dt} = -\mu \frac{\partial J}{\partial \mathbf{W}} \mathbf{W}^T \mathbf{W} = \mu \left(\mathbf{I} - \langle \mathbf{y}\mathbf{y}^T \rangle \right) \mathbf{W} \qquad (19)$$

The same result is obtained, if we use Atick-Redlich formula:

$$\frac{d\mathbf{W}}{dt} = -\mu \mathbf{W} \left[\frac{\partial J}{\partial \mathbf{W}} \right]^T \mathbf{W} = \mu \left(\mathbf{I} - \langle \mathbf{y}\mathbf{y}^T \rangle \right) \mathbf{W}, \quad (20)$$

since it reduces, in this case, to Amari's natural gradient formula, due to the fact that $\mathbf{W} \left[\frac{\partial J}{\partial \mathbf{W}} \right]^T$ is symmetric.

The corresponding discrete-time on-line algorithm can be written as:

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta \left[\mathbf{I} - \mathbf{y}(k)\mathbf{y}(k)^T \right] \mathbf{W}(k). \quad (21)$$

6.3. Derivation of simple local learning rule

The learning rule can be considerably simplified if we can assume that the decorrelation matrix $\mathbf{W}$ is symmetrical one. It is always possible to decorrelate vector $\mathbf{x}$ by using a symmetric matrix $\mathbf{W}$. To this end we can use a stable simple gradient formula

$$\frac{d\mathbf{W}}{dt} = -\mu \frac{\partial J}{\partial \mathbf{W}} \mathbf{W}^T = \mu \left[\mathbf{I} - \langle \mathbf{y}\mathbf{y}^T \rangle \right] \qquad (22)$$

which is obtained by (8) for J given in (15) with $p_i(y) = e^{y^2/2}$, putting $F(\mathbf{W}) = \mathbf{W}^{1/2}, \boldsymbol{\eta} = \eta_0 \mathbf{I}, \mathbf{W}(0)$ symmetric.

The above formula can be written in scalar form as

$$\frac{dw_{ij}}{dt} = \mu \left(\delta_{ij} - \langle y_i y_j \rangle \right). \qquad (23)$$

The discrete time on line local learning algorithm can be written as

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta(k)(\mathbf{I} - \mathbf{y}(k)\mathbf{y}^T(k)) \qquad (24)$$

or in scalar form as

$$w_{ij}(k+1) = w_{ij}(k) + \eta(k) \left[\delta_{ij} - y_i(k) y_j(k) \right]. \quad (25)$$

In addition to the merit that the algorithm (24) is much simpler to implement than (21), the local signal requirements of the algorithm in (24) make it ideal for hardware and VLSI implementations. However, the performances of (21) and (24) are not the same, and convergence speed of the local algorithm is usually much slower.

The update in (24) has an interesting property that it converges also for a suitable sequence of *negative* step sizes $\eta(k)$ [10]. To see this result, multiply both sides of (24) by (-1). By defining $\tilde{\mathbf{W}}(k) = -\mathbf{W}(k)$ and $\tilde{\mathbf{y}}(k) = -\mathbf{y}(k) = \tilde{\mathbf{W}}(k)\mathbf{x}(k)$ we obtain:

$$\tilde{\mathbf{W}}(k+1) = \tilde{\mathbf{W}}(k) - \eta(k)(\mathbf{I} - \tilde{\mathbf{y}}(k)\tilde{\mathbf{y}}^T(k)). \qquad (26)$$

This algorithm is equivalent to that in (24), and thus the coefficient matrix $\tilde{\mathbf{W}}(k)$ tends towards the solution obtained by $-\mathbf{W}(k)$ in the original algorithm. Summarizing, the local learning rule can be formulated in a more general form as

$$\mathbf{W}(k+1) = \mathbf{W}(k) \pm \eta(k)(\mathbf{I} - \langle \mathbf{y}(k)\mathbf{y}^T(k) \rangle), \quad (27)$$

where $\eta(k) > 0$.

7. FURTHER GENERALIZATION

We propose the following generalization of the Natural Gradient Algorithm

$$\mathbf{W}(l+1) = \exp\Big\{ -\eta \frac{\partial J}{\partial \mathbf{W}} \Big(\mathbf{y}(l), \mathbf{W}(l) \Big) \mathbf{W}(l)^T \Big\} \mathbf{W}(l).$$

In the standard situation, when J is given by (15), we obtain

$$\mathbf{W}(l+1) = \exp \Big\{ \eta \left[\mathbf{I} - \mathbf{f}\Big(\mathbf{y}(l)\Big)\mathbf{y}(l)^T \right] \Big\} \mathbf{W}(l). \quad (28)$$

If we develop the exponential term in infinite series

$$\exp(\mathbf{F}) = \sum_{k=0}^{\infty} \frac{1}{k!} \mathbf{F}^k,$$

where $\mathbf{F} = \eta[\mathbf{I} - \mathbf{f}(\mathbf{y})\mathbf{y}^T]$, and neglect nonlinear terms, we obtain the standard natural gradient algorithm.

An open question is a comparison of various matrix algorithms, obtained by replacing exponential function in (28) with other functions, for example $1 - \tanh$.

8. CONCLUSION

We present a general matrix dynamical systems and prove convergence of the continuous trajectories of a generalization of Amari's natural gradient algorithm for ICA and Atick-Redlich's algorithm (under some conditions). A nonholonomic basis is used to prove in another way the convergence of the above algorithms. The nonholonomic algorithm can not be derived from a minimization of a cost function. We propose norm stabilizing nonholonomic algorithm. As illustrative examples, the convergence of some decorrelation algorithms are proven. We emphasize, that the matrix $\boldsymbol{\eta}$ appearing in (8) is subject of further investigations in order to obtain fast convergence (like in Newton or quasi-Newton methods for scalar algorithms) but here $\boldsymbol{\eta} \in \mathbb{R}^{n \times n}$ (not $\boldsymbol{\eta} \in \mathbb{R}^{n^2 \times n^2}$ after vectorizing the matrix variables).

9. REFERENCES

[1] S. Amari, "Natural Gradient Works Efficiently in Learning", *Neural Computation* 10, pp. 251-276, 1998.

[2] S. Amari and J.-F. Cardoso, "Blind source separation - semiparametric statistical approach", *IEEE Trans. on Signal Processing*, 45(11), pp. 2692-2700, 1997.

[3] S. Amari, A. Cichocki, and H.H. Yang. "Unsupervised Adaptive Filtering", chapter *Blind Signal Separation and Extraction - Neural and Information Theoretic Approaches*, John Wiley, 1999.

[4] S. Amari, T.-P. Chen and A. Cichocki, "Nonholonomic Orthogonal Learning Algorithms for Blind Source Separation", *Neural Computation*, 12, pp. 1463-1484, 2000.

[5] S. Amari and S.C. Douglas. "Why Natural Gradient"' In *Proc. IEEE International Conference Acoustics, Speech, Signal Processing*, volume II, pp. 1213–1216, Seattle, WA, May 1998.

[6] J. J. Atick and A. N. Redlich. "Convergent Algorithm for Sensory Receptive Field Development" *Neural Computation*, 5(1):45–60, 1993.

[7] J. F. Cardoso and B. Laheld. "Equivariant Adaptive Source Separation" *IEEE Trans. on Signal Processing*, 44, pp. 3017-3030, 1996.

[8] A. Cichocki and R. Unbehauen. "Robust Neural Networks with On-Line Learning for Blind Identification and Blind Separation of Sources" *IEEE Trans Circuits and Systems I : Fundamentals Theory and Applications*, 43(11), pp. 894–906, Nov. 1996.

[9] S. Cruces, A. Cichocki, and L. Castedo. "An Iterative Inversion Approach to Blind Source Separation" *IEEE Trans. on Neural Networks*, Nov. 2000, pp.1423-1437, Nov. 2000,

[10] S.C. Douglas and A. Cichocki, "Neural Networks for Blind Decorrelation of Signals", *IEEE Trans. Signal Processing*, 45(11), pp. 2829–2842, Nov. 1997.

[11] A. Edelman, T. Arias and S. T. Smith, "The Geometry Of Algorithms With Orthogonality Constraints", *SIAM Journal on Matrix Analysis and Applications*, pp. 303-353. 1998.

[12] S. Lang, "Undergraduate Analysis", Second Edition , Springer, New York, 1997.

[13] A. Hyvarinen, J. Karhunen and E. Oja, "Independent Component Analysis", John Wiley & Sons, 2001.