Minimax Mutual Information Approach for ICA of Complex-Valued Linear Mixtures

Jian-Wu Xu, Deniz Erdogmus, Yadunandana N. Rao, and José Carlos Príncipe

CNEL, Electrical and Computer Engineering Department,
University of Florida, Gainesville, Florida 32611, USA
{jianwu,deniz,principe}@cnel.ufl.edu
http://www.cnel.ufl.edu

Abstract. Recently, the authors developed the Minimax Mutual Information algorithm for linear ICA of real-valued mixtures, which is based on a density estimate stemming from Jaynes' maximum entropy principle. Since the entropy estimates result in an approximate upper bound for the actual mutual information of the separated outputs, minimizing this upper bound results in a robust performance and good generalization. In this paper, we extend the mentioned algorithm to complex-valued mixtures. Simulations with artificial data demonstrate that the proposed algorithm outperforms FastICA.

1 Introduction

Independent Component Analysis (ICA), which may be viewed as an extension of Principle Component Analysis (PCA), is a method of finding a set of directions to minimize the statistical dependence of the projections of input random vector x on these directions. As a measure of independence between random variables, mutual information is considered as the natural criterion for ICA since minimizing mutual information would make the components of output as independent as possible. One commonly used definition of mutual information is Shannon's mutual information. Given n random variables $Y_1,\ldots\ldots,Y_n$ whose joint probability density function (pdf) is $f_{\mathbf{Y}}(\mathbf{y})$ and marginal probability density functions (pdfs) are defined as $f_1(y^1),\ldots,$ $f_n(y^n)$ respectively, then Shannon's mutual information [1] is defined as follows

$$I(\mathbf{Y})=\int_{-\infty}^{+\infty} f_{\mathbf{Y}}(\mathbf{y})\log\left(f_{\mathbf{Y}}(\mathbf{y})\middle/\prod_{o=1}^{n} f_o(y^o)\right)d\mathbf{y} \tag{1}$$

where the components y^i, i=1,…,n constitute the vector $\mathbf{y}$. Meanwhile, we can also write Shannon's mutual information as the sum of marginal and joint entropies [1] of these random variables as,

$$I(\mathbf{Y})=\sum_{o=1}^{n} H(Y_o)-H(\mathbf{Y}) \tag{2}$$

where Shannon's marginal and joint entropies [1] are given by

$$H(Y_o)=\int_{-\infty}^{+\infty} f_o(y^o)\log f_o(y^o)dy^o \qquad H(\mathbf{Y})=\int_{-\infty}^{+\infty} f_{\mathbf{Y}}(\mathbf{y})\log f_{\mathbf{Y}}(\mathbf{y})d\mathbf{y} \tag{3}$$

C.G. Puntonet and A. Prieto (Eds.): ICA 2004, LNCS 3195, pp. 311–318, 2004.

respectively. Three of most widely known algorithms for ICA, namely JADE [2], Infomax [3], and FastICA [4], use the diagonalization of cumulant matrices, maximization of output entropy, and fourth order cumulants separately, instead of using minimization of output mutual information. The difficulties associated with minimum mutual information are the lack of robust pdf estimators; most of them suffer from sensitivity to the underlying data samples.

A common method in developing information theoretic ICA algorithms is to use polynomial expansions to approximate the pdf of the signals, e.g. Gram-Charlier, Edgeworth, and Legendre polynomial expansions. In order to estimate the signal pdf, a truncated polynomial is taken, evaluated in the vicinity of a maximum entropy density [5]. Alternative techniques include Parzen windowing [6], and orthogonal basis functions [7]. Other researchers also use kernel estimates in ICA [8,9,10].

Recently, we used the minimum output mutual information method to develop an efficient and robust ICA algorithm, which is based on a density estimate stemming from Jaynes' maximum entropy principle, where estimated pdfs belong to the exponential family [11, 12]. This approach approximates the solution to a constrained entropy maximization problem and provides an approximate upper bound for the actual mutual information of the output signals, and hence the name Minimax Mutual Information. In addition, this method is related to ICA methods using higher order cumulants when a specific set of constraint functions are selected in the maximum entropy density estimation step.

In this paper, we extend this Minimax Mutual Information algorithm to complex-valued mixtures. The algorithm is compared with the complex-valued FastICA method. The simulations demonstrate that complex-valued Minimax ICA exhibits better performance.

2 The Problem Statement

Suppose that there are n mutual independent sources s, whose components are zero-mean complex-valued signals. We also assume the independence between real and imaginary parts of source signal. The source signal s is mixed by an unknown linear mixture of the form z = Hs to generate n observed random vector z, where the square matrix H is invertible. In this case, the original independent sources can be obtained from z by a two-stage process: spatial whitening to generate uncorrelated but not necessarily independent mixture x = Wz, and a coordinate system rotation in the n-dimensional mixture space to determine the independent components y = Rx [5,8,13]. The whitening matrix W is obtained from the eigendecomposition of the measurement of covariance matrix. Namely, $\mathbf{W} = \mathbf{\Lambda}^{-1/2}\mathbf{\Phi}^T$, where Λ denotes the diagonal eigenvalue matrix and Φ denotes the corresponding orthonormal eigenvector matrix of the mixture covariance matrix $\Sigma = E[zz^T]$ provided that the observations are zero mean. The coordinate rotation is determined by an orthonormal matrix R parameterized by Givens angles [15]. Specifically the procedure involves the minimization of the mutual information between the output signals [5]. Considering the fact that the joint entropy is invariant under rotations, the definition of mutual information in (2) reduces to the summation of marginal output entropies for this case. Namely,

$$J(\mathbf{\Theta}) = \sum_{o=1}^{n} H(Y_o) \tag{5}$$

where the vector $\mathbf{\Theta}$ is composed of Givens angles $\theta_{ij}, i = 1,....,n-1, j = i+1,....,n.$ The Givens parameterization of a rotation matrix involves the multiplication of in-plane rotation matrices. Each of the matrices $R^{ij}(\theta_{ij}^{1,2})$ for the complex-valued signal is constructed by starting with an $n \times n$ identity matrix and replacing the entries $(i,i)^{th}, (i,j)^{th}, (j,i)^{th}, (j,j)^{th}$ by $\cos\theta_{ij}^{1}\exp(j\theta_{ij}^{2})$, $\sin\theta_{ij}^{1}$, $-\sin\theta_{ij}^{1}$, and $\cos\theta_{ij}^{1}\exp(-j\theta_{ij}^{2})$, respectively, where θ^{1} is the angle for the real part and θ^{2} is for the imaginary part. The total rotation matrix is then defined as the product of these 2-dimensional rotations parameterized by $n(n$-1) Givens angles to be optimized:

$$\mathbf{R}(\mathbf{\Theta}) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} R^{ij}(\theta_{ij}^{1,2}) \tag{6}$$

The described whitening-rotation procedure through Givens angles parameterization of the rotation matrix is widely used in ICA algorithm, and many studies have been done on the efficient ways of dealing with the optimization of these parameters.

3 The Maximum Entropy Principle

Jaynes' maximum entropy principle states that one must maximize the entropy of the estimated distribution under certain constraints so that the estimated pdf fits the known data best without committing extensively to the unknown data because the entropy of a pdf is related with the uncertainty of the associated random variables.

Given the nonlinear moments $\alpha_k = E_X[f_k(X)]$, the maximum entropy pdf estimate for X is obtained by solving the following constrained optimization problem.

$$\max_{p_{\overline{X}}(.)} H = -\int_{C} p_{\overline{X}}(x)\log p_{\overline{X}}(x)dx \quad s.t.\ E_{\overline{X}}\left[f_k(\overline{X})\right] = \alpha_k \ \ k = 1,...,m \tag{7}$$

where $p_{\overline{X}} : C \rightarrow R$ is the pdf of a complex-valued variable, and f_k:C→R are the constraint functions defined *a priori*. Using calculus of variations and the Lagrange multipliers method, we can get the optimal pdf for the complex-valued signal [1]

$$p_{\overline{X}}(x) = C(\lambda)\exp\left(\sum_{l=1}^{m} \lambda_l f_l(x)\right) \tag{8}$$

where $\boldsymbol{\lambda} = [\lambda_1,....\lambda_m]^T$ is the Lagrange multiplier vector and $C(\lambda)$ denotes the normalization constant. It is not easy to solve the Lagrange multipliers simultaneously from the constraints in case of continuous random variables due to the infinite range of the definite integrals involved. We use the integration by parts method under the assumption that the actual distribution is close to the maximum entropy distribution. Consider the kth constraint equation,

$$\alpha_k = \int_{c} f_k(x)p(x)dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_k(x_r, x_i)p(x_r, x_i)dx_r dx_i \tag{9}$$

where $f_k(x_r, x_i)$ is the nonlinear moment of the real and imaginary parts of the signal, denoted by x_r, x_i. The integrand covers the whole real and imaginary ranges.

We first give the following definitions:

$$\begin{aligned} &F_k^{(0,1)}(x_r,x_i) = \int_{-\infty}^{+\infty} f_k(x_r,x_i)dx_i, \quad F_k^{(1,0)}(x_r,x_i) = \int_{-\infty}^{+\infty} f_k(x_r,x_i)dx_r \\ &f_l^{(0,1)}(x_r,x_i) = \frac{\partial}{\partial x_i} f_l(x_r,x_i), \quad f_l^{(1,0)}(x_r,x_i) = \frac{\partial}{\partial x_r} f_l(x_r,x_i) \end{aligned} \tag{10}$$

Integrating by parts over the real part the double integral in (9), we obtain

$$\alpha_k = \int_{-\infty}^{+\infty}\left[p(x_r,x_i)F_k^{(0,1)}(x_r,x_i)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{+\infty} F_k^{(0,1)}(x_r,x_i)\left(\sum_{l=1}^{m}\lambda_l f_l^{(0,1)}(x_r,x_i)\right)p(x_r,x_i)dx_r\right]dx_i \tag{11}$$

Meanwhile we can also do partial integration over the imaginary part such that

$$\alpha_k = \int_{-\infty}^{+\infty}\left[p(x_r,x_i)F_k^{(1,0)}(x_r,x_i)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{+\infty} F_k^{(1,0)}(x_r,x_i)\left(\sum_{l=1}^{m}\lambda_l f_l^{(1,0)}(x_r,x_i)\right)p(x_r,x_i)dx_i\right]dx_r \tag{12}$$

If the functions $f_l(x_r, x_i)$ are selected such that their integrals $F_l(x_r, x_i)$ do not diverge faster than the decay rate of the exponential pdf $p_{\overline{X}}(x)$, then the first terms on the right hand sides of (11) and (12) go to zero. For example, this condition would be satisfied if moments of the random variable were defined as the constraint functions since $F_l(x_r, x_i)$ will be a polynomial function and $p_{\overline{X}}(x)$ decays exponentially. Then adding (11) and (12) yields the expression for α_k

$$\begin{aligned} \alpha_k &= -\frac{1}{2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left[\begin{array}{l} F_k^{(0,1)}(x_r,x_i)\left(\sum_{l=1}^{m}\lambda_l f_l^{(0,1)}(x_r,x_i)\right)p(x_r,x_i) \\ +F_k^{(1,0)}(x_r,x_i)\left(\sum_{l=1}^{m}\lambda_l f_l^{(1,0)}(x_r,x_i)\right)p(x_r,x_i) \end{array}\right]dx_i dx_r \\ &= -\frac{1}{2}\sum_{l=1}^{m}\lambda_l E\left[F_k^{(1,0)}(x_r,x_i)f_l^{(0,1)}(x_r,x_i) + F_k^{(0,1)}(x_r,x_i)f_l^{(1,0)}(x_r,x_i)\right] \\ &\overset{\Delta}{=} -\frac{1}{2}\sum_{l=1}^{m}\lambda_l\beta_{kl} \end{aligned} \tag{13}$$

Note that the coefficients β_{kl} can be estimated using the sample mean. Finally, introducing the vector $\boldsymbol{\alpha} = [\alpha_1 .....\alpha_m]^T$ and the matrix $\boldsymbol{\beta} = [\beta_{kl}]$, the Lagrange multipliers are given by

$$\boldsymbol{\lambda} = -\frac{1}{2}\boldsymbol{\beta}^{-1}\boldsymbol{\alpha} \tag{14}$$

This method provides a simple way of finding the coefficients of the estimated pdf directly from the samples when $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are estimated using sample means.

4 Gradient Update Rule for the Givens Angles

Minimax ICA minimizes the cost function in (5) using the entropy estimate corresponding to the maximum entropy distribution described in the previous section. A gradient descent update rule for the Givens angles is employed to adapt the rotation matrix. The derivative of marginal entropy with respect to a Givens angle is

$$\frac{\partial H(Y_o)}{\partial \theta_{pq}^{1,2}} = -\sum_{k=1}^{m} \lambda_k^o \frac{\partial \alpha_k^o}{\partial \theta_{pq}^{1,2}} \tag{15}$$

where λ^o is the Lagrange multiplier parameter vector for the pdf of o^{th} output signal and α_k^o is the value of the k^{th} constraint for the pdf of the o^{th} output. Using (13) to get the solution for λ^o and the sample mean estimate

$$\alpha_k^o = \frac{1}{N}\sum_{l=1}^{N} f_k(y_{o,l}) \tag{16}$$

where $y_{o,l}$ is the l^{th} sample at the o^{th} output for the current angles, the derivative of α_k^o with respect to $\theta_{pq}^{1,2}$ is obtained as,

$$\begin{aligned}\frac{\partial \alpha_k^o}{\partial \theta_{pq}^{1,2}} &= \frac{1}{N}\sum_{l=1}^{N} f_k^{'}(y_{o,l}) \frac{\partial y_{o,l}}{\partial \theta_{pq}^{1,2}} = \frac{1}{N}\sum_{l=1}^{N} f_k^{'}(y_{o,l}) \left(\partial y_{o,l} / \partial \mathbf{R}_{o:}\right)^T \left(\partial \mathbf{R}_{o:} / \partial \theta_{pq}^{1,2}\right)^T \\ &= \frac{1}{N}\sum_{l=1}^{N} f_k^{'}(y_{o,l}) \mathbf{x}_l^T \left(\partial \mathbf{R} / \partial \theta_{pq}^{1,2}\right)_{o:}^T\end{aligned} \tag{17}$$

where the subscripts in $\mathbf{R}_{o:}$ and $\left(\partial \mathbf{R} / \partial \theta_{pq}^{1,2}\right)_{o:}$ denote the o^{th} row of the corresponding matrix. By the definition, the derivative of R with respect to an angle is

$$\begin{aligned}\partial \mathbf{R} / \partial \theta_{pq}^{1,2} = &\left(\prod_{i=1}^{p-1}\prod_{j=i+1}^{n} \mathbf{R}^{ij}(\theta_{ij})\right)\left(\prod_{j=o+1}^{q-1} \mathbf{R}^{pj}(\theta_{pj})\right)\left(\partial \mathbf{R}^{pq}(\theta_{pq}) / \partial \theta_{pq}^{1,2}\right) \\ &\left(\prod_{j=q+1}^{n} \mathbf{R}^{pj}(\theta_{pj})\right)\left(\prod_{i=p+1}^{n-1}\prod_{j=i+1}^{n} \mathbf{R}^{ij}(\theta_{ij})\right)\end{aligned} \tag{18}$$

Thus, the overall update rule for the Givens angles summing the contributions from each output is

$$\mathbf{\Theta}_{t+1} = \mathbf{\Theta}_t - \eta \sum_{o=1}^{n} \frac{\partial H(Y_o)}{\partial \mathbf{\Theta}} \tag{19}$$

where η is a small step size.

5 Discussion on the Algorithm

In the previous sections, we proposed an approximate numerical solution which replaces the expectation operator over the maximum entropy by a sample mean over the data distribution due to the difficulties associated with solving for the Lagrange multipliers analytically. In this section, we provide how to choose the constraint functions $f_k(.)$ in the formulation. Here we consider the moment constraints for both real and imaginary parts of the output $y_{o,l}$, namely

$$\alpha_k^o = E\left[{y_r^o}^{u_k} {y_i^o}^{v_k} \right] = \frac{1}{N} \sum_{l=1}^{N} {y_{r,l}^o}^{u_k} {y_{i,l}^o}^{v_k} \tag{20}$$

where y_r^o and y_i^o are the real and imaginary parts of o^{th} output, u_k, v_k are the moment order. Our brief investigation on the effect of other constraint functions suggests that the simple moment constraint yields significantly better solutions. One motivation to use moment constraint is the asymptotic properties of the exponential pdf estimates in (8).

Besides the desirable asymptotic convergence properties of the exponential family of density estimates, the moment constraint function gives simple gradient updates. Let $y_o = y_r + jy_i = (\mathbf{R}_r + j\mathbf{R}_i) \times (x_r + jx_i) = (\mathbf{R}_r x_r - \mathbf{R}_i x_i) + j(\mathbf{R}_r x_i + \mathbf{R}_i x_r)$. Here $\mathbf{R}_r$ and $\mathbf{R}_i$ are the real and imaginary parts of the rotation matrix R. Then, we can find the derivative of (17) with respect to the Givens angle $\theta_{pq}^{1,2}$ as

$$\frac{\partial \alpha_k^o}{\partial \theta_{pq}^{1,2}} = \frac{1}{N} \sum_{l=1}^{N} {y_{r,l}^o}^{(u_k-1)} {y_{i,l}^o}^{(v_k-1)} \left(u_k y_{i,l}^o \frac{\partial y_{r,l}^o}{\partial \theta_{pq}^{1,2}} + v_k y_{r,l}^o \frac{\partial y_{i,l}^o}{\partial \theta_{pq}^{1,2}} \right) \tag{22}$$

where the derivative of output with respect to angle is

$$\frac{\partial y_r^o}{\partial \theta_{pq}^{1,2}} = \frac{\partial \mathbf{R}_r^o}{\partial \theta_{pq}^{1,2}} x_r - \frac{\partial \mathbf{R}_i^o}{\partial \theta_{pq}^{1,2}} x_i \qquad \frac{\partial y_i^o}{\partial \theta_{pq}^{1,2}} = \frac{\partial \mathbf{R}_r^o}{\partial \theta_{pq}^{1,2}} x_i + \frac{\partial \mathbf{R}_i^o}{\partial \theta_{pq}^{1,2}} x_r \tag{23}$$

Furthermore, in the computation of (18), we can express $\partial \mathbf{R}^{pq}(\theta_{pq}) \big/ \partial \theta_{pq}^{1,2}$ as

$$\frac{\partial \mathbf{R}^{pq}(\theta_{pq})}{\partial \theta_{pq}^1} = \begin{bmatrix} -\sin\theta_{pq}^1 e^{j\theta_{pq}^2} & \cos\theta_{pq}^1 \\ -\cos\theta_{pq}^1 & -\sin\theta_{pq}^1 e^{-j\theta_{pq}^2} \end{bmatrix} \quad \frac{\partial \mathbf{R}^{pq}(\theta_{pq})}{\partial \theta_{pq}^2} = diag \begin{bmatrix} j\cos\theta_{pq}^1 e^{j\theta_{pq}^2} \\ -j\cos\theta_{pq}^1 e^{-j\theta_{pq}^2} \end{bmatrix} \tag{24}$$

Here $\partial \mathbf{R}_r^o / \partial \theta_{pq}^{1,2}$ and $\partial \mathbf{R}_i^o / \partial \theta_{pq}^{1,2}$ are the real and imaginary parts of $\left(\partial \mathbf{R} / \partial \theta_{pq}^{1,2} \right)^o$.

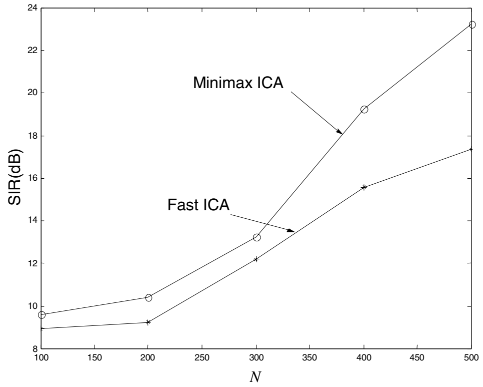

Fig. 1. Average SIR (dB) obtained by complex Minimax ICA and FastICA versus sample size.

6 Simulations

In this section, we present a simple comparison of the proposed complex Minimax ICA algorithm and the popular complex FastICA method [16]. In this controlled environment, the signal-to-interference ratio (SIR) is used as the performance measure:

$$SIR(dB) = \frac{1}{n}\sum_{o=1}^{n} 10\log_{10}\left(\max_k(\mathbf{O}_{ok}^{o}) \Big/ \left(\mathbf{O}_{o:}\mathbf{O}_{o:}^{T} - \max_k(\mathbf{O}_{ok}^{o}) \right)\right) \tag{25}$$

where O is the overall matrix after separation, i.e. O=RWH. This measure is the average ratio in decibels (dB) of the main signal power in the output channel to the total power of the interfering signals. Minimax ICA uses all complex moments up to order 4 as constraints, thus it considers kurtosis information as FastICA does.

For training set sample sizes (*N*) ranging from 100 to 500, a set of 100 Monte Carlo simulations are run for each sample size. In each run, *N* complex samples are generated artificially according to $s_j = r_j\left(\cos\phi_j + i\sin\phi_j\right)$, where r_1 is Gaussian, and r_2 and the phases ϕ_j are uniform. In this setup, the sources have independent real and imaginary parts with equal variance. The 2x2 mixing matrix is also complex-valued whose real and imaginary parts of entries are uniformly random in [-1,1].

Fig. 1 shows the SIR for both methods. While Minimax ICA is always better than FastICA, the difference in performance increasingly becomes significant as the sample size is increased. On the other hand, the computational requirement of Minimax ICA is much larger than that of FastICA, as one can assess from the previous sections.

7 Conclusions

In this paper, we extended the Minimax ICA algorithm to complex-valued signals. This algorithm is based on a density estimate stemming from Jaynes' maximum entropy principle. Thus, an approximate upper bound for the mutual information between the separated outputs is obtained from the samples and minimized through the optimization procedure. The density estimation stage utilizes integration by parts in a novel way to arrive at a set of linear equations that uniquely determine the Lagrange multipliers of the constrained maximum entropy density estimation problem.

Numerical simulations conducted using artificial mixtures suggest that the proposed complex Minimax ICA algorithm yields better separation performance compared to complex FastICA at the cost of additional computational burden.

Acknowledgments

This work is supported by NSF grant ECS-0300340.

References

1. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (1991)
2. Cardoso, J.F., Souloumiac, A.: Blind Beamforming for Non-Gaussian Signals. IEE Proc. F Radar and Signal Processing. 140 (1993) 362-370
3. Bell, A., Sejnowski, T.: An Information-Maximization Approach to Blind Separation and Blind Deconvolution. Neural Computation. 7 (1995) 1129-1159
4. Hyvarinen, A.: Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. IEEE Transactions on Neural Networks. 10 (1999) 626-636
5. Comon, P.: Independent Component Analysis, A New Concept? Signal Processing. 36 (1994) 284-314
6. Parzen, E.: On Estimation of a Probability Density Function and Mode. Annals of Mathematical Statistics. 33 (1962) 1065-176
7. Girolami, M.: Orthogonal Series Density Estimation and the Kernel Eigenvalue Problem. Neural Computation. 14 (2002) 1065-1076
8. Hild II, K.E., Erdogmus, D., Principe, J.C.: Blind Source Separation Using Renyi's Mutual Information. IEEE Signal Processing Letters. 8 (2001) 174-176
9. Xu, D., Principe, J.C., Fisher, J., Wu, H.C.: "A Novel Measure for Independent Component Analysis. Proc. ICASSP'98, Seattle, Washington. (1998) 1161-1164
10. Pham, D.T.: Blind Separation of Instantaneous Mixture of Sources via the Gaussian Mutual Information Criterion. Signal Processing. 81 (2991) 855-870
11. Erdogmus, D., Hild II, K.E., Rao, Y.N., Principe, J.C.: Independent Component Analysis Using Jaynes' Maximum Entropy Principle. Proc. ICA'03, Nara, Japan. (2003) 385-390
12. Erdogmus, D. Hild II, K.E., Rao, Y.N., Principe, J.C.: Minimax Mutual Information Approach for Independent Component Analysis. Neural Computation (2004) to appear
13. Cardoso, J.F.: High-Order Contrasts for Independent Component Analysis. Neural Computation. 11 (1999) 157-192
14. Hild, K.E., Erdogmus, D., Principe, J.C.: Blind Source Separation Using Renyi's Mutual Information. IEEE Signal Processing Letters. 8 (2001) 174-176
15. Golub, G., van Loan, C.: Matrix Computation. John Hopkins Univ. Press, Baltimore (1993)
16. Bingham, E., Hyvärinen, A.: A Fast Fixed-point Algorithm for Independent Component Analysis of Complex-Valued Signals. Int. J. of Neural Systems. 10 (2000) 1-8