ORDER STATISTICS BASED ESTIMATOR FOR RENYI'S ENTROPY

Anant Hegde[1], Tian Lan[2], Deniz Erdogmus[2]
[1] CNEL, Electrical & Computer Engineering Department, University of Florida, Gainesville, Florida, USA
[2] CSEE & BME Departments, OGI, Oregon Health & Science University, Portland, Oregon, USA

Abstract – Several types of entropy estimators exist in the information theory literature. Most of these estimators explicitly involve estimating the density of the available data samples before computing the entropy. However, the entropy-estimator using sample spacing avoids this intermediate step and computes the entropy directly using the order-statistics. In this paper, we extend our horizon beyond Shannon's definition of entropy and analyze the entropy estimation performance at higher orders of alpha, using Renyi's generalized entropy estimator. We show that the estimators for higher orders of alpha better approximate the true entropy for an exponential family of distributions. Practical application of this estimator is demonstrated by computing mutual information between functionally coupled systems. During the estimation process, the joint distributions are decomposed into sum of their marginals by using linear ICA.

Keywords – Order Statistics, Entropy, Mutual Information, Independent Component Analysis.

I. INTRODUCTION

Entropy estimators are needed in almost every discipline of scientific research such as engineering, biomedicine, biochemistry and physics, predominantly due to the need to measure information from a sequence of available data. Even though many entropy estimation techniques have been developed, they suffer from either convergence problems or with lack of large amount of data, a problem common to any estimator. Estimation also suffers if the data has long-range dependencies or trends. In addition, some of the estimators need to estimate the data distribution before estimating the entropy itself. Kernel estimation techniques using Renyi's entropy definitions, overcome the need to estimate pdf intermittently; however, they suffer from computational complexity. Recently, Miller *et al*. [1] proposed using a sample spacing entropy estimator derived by Vasicek [2] that alleviates the intermediate step of estimating the density of the samples and at the same time achieves computational simplicity. Specifically, this estimator uses order statistics in which the data sequence is sorted apriori to the entropy estimation. However in [1], Shannon's definition of entropy (order/alpha = 1) is used. One of the main problems with Shannon's entropy definition is that the estimate does not converge to the true entropy of the data sequence, even after the bias compensations [2]. We therefore propose the generalized Renyi's entropy definition that uses higher orders of alpha (>1). The paper is organized as follows: In section II we use the generalized Renyi's entropy definition and derive the entropy estimate on similar lines and assumptions as in [1]. Section III derives the entropy estimate for the case of a generalized exponential family of distributions. Section IV presents a simple synthetic simulation and also simulations with EEG data, demonstrating the utility of the proposed estimator. Finally, Section V. discusses potential avenues for further improvements.

II. ENTROPY ESTIMATOR

Suppose that N iid samples $\{x_1,x_2,\ldots,x_N\}$ of a one-dimensional random variable X is drawn from a distribution $p(x)$. The entropy of X can be defined in terms of the generalized definition of Renyi's entropy as:

$$\begin{aligned} H_\alpha &= \frac{1}{(1-\alpha)}\log\int p^\alpha(x)dx \\ &\approx \frac{1}{(1-\alpha)}\log\int \hat{p}^\alpha(x)dx \end{aligned} \tag{1}$$

We know from statistics that the samples of X written in ascending order of their sample values represent the order statistics of X, i.e., $x^{(1)}\le x^{(2)}\le\ldots\le x^{(N)}$. By defining m-spacing to be $x^{(i+m)}-x^{(i)}$ for $1\le i<i+m\le N$, the Renyi's entropy estimator can now be deduced as described below. Although the estimator that we will present next is based on the universal properties of order statistics and does not require an explicit density assumption, we find it more convenient and intuitive to motivate the estimator through the introduction of an implicit density model.

We begin by approximating the probability density of X, $p(x)$, by assigning equal weightings to each interval between the ordered data samples [1]. By assuming a uniform distribution across the intervals, we have

$$\hat{p}(x) = \frac{1}{(N+1)(x^{(i+1)}-x^{(i)})} \tag{2}$$

for $x^{(i)} \le x \le x^{(i+1)}$. This density approximation is based on the fact that the expected increment in the value of the empirical cdf at this interval is identical to (2) [1]. Using (1), the 1-spacing Renyi's entropy can now be written as

$$\begin{aligned}
H_{\alpha,1} &\overset{a}{\approx} \frac{1}{1-\alpha}\log\sum_{i=0}^{N}\int_{x_i}^{x_{i+1}}[(N+1)(x_{i+1}-x_i)]^{-\alpha}dx \\
&= \frac{1}{1-\alpha}\log\sum_{i=0}^{N}(N+1)^{-\alpha}(x_{i+1}-x_i)^{-\alpha}(x_{i+1}-x_i) \\
&= \frac{1}{1-\alpha}\log[(N+1)^{-\alpha}\sum_{i=0}^{N}(x_{i+1}-x_i)^{1-\alpha}] \\
&= \frac{1}{1-\alpha}\log[\frac{(N+1)^{1-\alpha}}{(N+1)}\sum_{i=0}^{N}(x_{i+1}-x_i)^{1-\alpha}] \\
&\overset{b}{\approx} \frac{1}{1-\alpha}\log[\frac{(N+1)^{1-\alpha}}{(N-1)}\sum_{i=1}^{N-1}(x_{i+1}-x_i)^{1-\alpha}] \\
&= \log(N+1)+\frac{1}{1-\alpha}\log[\frac{1}{(N-1)}\sum_{i=1}^{N-1}(x_{i+1}-x_i)^{1-\alpha}]
\end{aligned} \tag{3}$$

The first approximation (a) is due to the piecewise uniform density assumption of the distribution of x and the second assumption (b) arises from the fact that we do not have information on the data samples, $x^{(0)}$ and $x^{(N+1)}$. As pointed in [1], the 1-spacing estimate suffers from a high variance, due to the single-interval dependency of the order statistics. The bias-variance dilemma can be asymptotically eliminated by increasing the sample-spacings to m. As $m \to \infty$ and $(m/N) \to 0$, the estimator becomes consistent. Typically, $m=N^{1/2}$. The definition in (3) for the Renyi's 1-spacing entropy estimator can now be extended to m-spacings as

$$H_{\alpha,m} = \log\frac{(N+1)}{m} + \frac{1}{1-\alpha}\log[\frac{1}{(N-m)}\sum_{i=1}^{N-m}(x_{i+m}-x_i)^{1-\alpha}] \tag{4}$$

We know that the Shannon's m-spacing entropy estimator [1, 2] denotes the special case when α=1:

$$\hat{H}_{shannon} = \frac{1}{N-m}\sum_{i=1}^{N-m}\log\frac{(N+1)(x_{(i+m)}-x_{(i)})}{m} \tag{5}$$

Both estimators suffer from bias that varies with m. The bias-compensation term in the case of Shannon's m-spacing entropy estimator was evaluated in [7]. While we leave the calculation of this bias to an extended journal version of this paper, in many practical applications, such as ICA or other entropy-based optimal signal processing scenarios, this bias does not affect the solution, since it is independent of the true data distribution and only depends on m and N. Without loss of generalization, all the results in the rest of this paper will be computed for the biased case in (4), and accordingly, comparisons will be made with the biased estimator of Shannon's entropy in (5).

III. RENYI'S ENTROPY FOR GENERALIZED EXPONENTIAL DISTRIBUTIONS

The probability density function $p_\beta(x)$ for a general Gaussian distribution can be written as

$$p_\beta(x) = C_1(\beta, C_2)e^{-C_2|x|^\beta} \tag{6}$$

where $1 \le \beta \le \infty$ defines the mode of the distribution. The Laplacian distribution ($\beta = 1$) and Gaussian distribution ($\beta = 2$) are members of the generalized Gaussian family of distributions. Using the definition of (1), Renyi's entropy on $p_\beta(x)$ can be written as,

$$H_\alpha(P_\beta) = \frac{1}{(1-\alpha)}\log\int_{-\infty}^{\infty}C_1^\alpha e^{-\alpha C_2|x|^\beta}dx \tag{7}$$

We know from (5) that $C_1^{-1} = \int e^{-C_2|x|^\beta}$, and therefore incorporate it in (6) and proceed to find an analytical solution for Renyi's entropy on generalized exponential distributions

$$H_\alpha^{true}(p_\beta) = -\log C_1 - \frac{\log\alpha}{\beta(1-\alpha)} \tag{8}$$

Note that for Shannon's entropy ($\alpha = 1$), we use the L'Hospital's rule as,

$$H_1^{true}(p_\beta) = -\log C_1 - \frac{1}{\beta}\left(\lim_{\alpha\to 1}\frac{\log\alpha}{(1-\alpha)}\right) \tag{9}$$

IV. SIMULATIONS AND RESULTS

In this section, we compare the accuracy of the generalized Renyi's entropy estimator (for higher orders of α) and the Shannon's entropy estimator ($\alpha = 1$), using a number of Monte-Carlo simulations, carried over a range of α and β values. To illustrate its practical utility, we present an example application in which we use the Renyi's entropy estimator to estimate mutual information in linearly and non-linearly coupled systems.

***Experiment A*:** In this simulation, we illustrate the importance of having m>1 while determining the Renyi's entropy using sample spacing estimates. Fig 1. plots the true entropy against the estimated entropy. For a fixed order of $\alpha = 2$, the entropy is estimated for some standard distributions such as Laplacian, Gaussian and the Uniform distributions. Fig. 1 shows that the estimator converges to the true entropy with increasing m. The variance is also seen to reduce with increasing m. With the sample-size set to N=10^4, this example provides enough justification for our rationale in setting the sample spacings to $m = \sqrt{N}$.

Experiment B: Our second Monte-Carlo simulation consists of estimating the Renyi's entropy for data distributions at different values of β and at various orders of α (0.5, 1, 2, and 3). The sample data generated for each of the distributions had zero mean and unit-variance. For $N = 10^4$ data samples, the true entropy values of the generated sample data were determined analytically and plotted against the estimated entropy values, as shown in Fig. 2. One can observe that the entropy estimates for $\alpha = 2$ and $\alpha = 3$ are closer to their corresponding true values, regardless of the order of the distribution β . This emphasizes that regardless of the bias, entropy can be more accurately estimated at higher order values of α (>1), as opposed to estimating using Shannon's entropy (α =1).

Experiment C: Our third simulation attempts to investigate the effect of sample-size (*N*) in entropy estimates. Fig 3. plots both the true entropy and the entropy estimates as a function of the parameter α , at various sample sizes (*N*), for a Gaussian~N(0,1) distributed data. It is easy to observe the higher accuracy achieved in estimating the entropy, with increasing *N*. The variance was also observed to decrease with increasing *N*. Fig 3. also indicates that the entropy estimates closely approximate the true entropy for α , in the range of [1,4]. Entropy estimates outside this range are seen to have larger bias and variance and therefore not recommended.

Experiment D: Estimating mutual information using Renyi's sample spacing entropy estimator. We now demonstrate a practical usage of this estimator by applying it in estimating the mutual information between two time-series. Mutual information is a widely used statistical measure, for detecting non-linear coupling/dependencies in multivariate dynamic and statistical structures. It is a symmetric measure and utilizes the complete structure of the data while estimating their functional relationships. Renyi's Mutual Information is defined as [6]

$$I_{R_\alpha}(x) = \frac{1}{(\alpha-1)} \log \int_{-\infty}^{\infty} \frac{f_X(x)^\alpha}{\prod_{i=1}^{M} f_{X_i}(x_i)^{(\alpha-1)}} dx. \quad (10)$$

An approximation to this quantity, inspired by the entropy expansion of Shannon's mutual information is the sum of Renyi's marginal entropies minus the joint entropy, which can be written as

$$\sum_{i=1}^{M} H_{R_\alpha}(x_i) - H_{R_\alpha}(x) = \frac{1}{(\alpha-1)} \log \frac{\int f_X(x)^\alpha dx}{\int \prod_{i=1}^{M} f_{X_i}(x_i)^\alpha dx_i} \quad (11)$$

Our experiment setup consists of two uniformly distributed random signals ε(-1,1), $\mathbf{x}=\{x_1(t), x_2(t), t \in T\}$ linearly coupled through a rotation matrix $\mathbf{R}$,

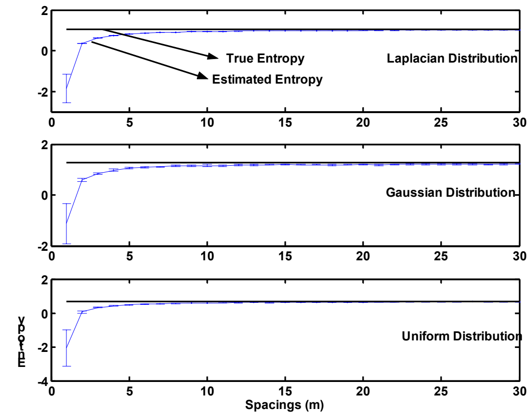

Figure 1. Illustrating the need to have higher sample spacings (*m*). Renyi's entropy was estimated for fixed $\alpha = 2$ and sample size of $N = 10^4$ samples.

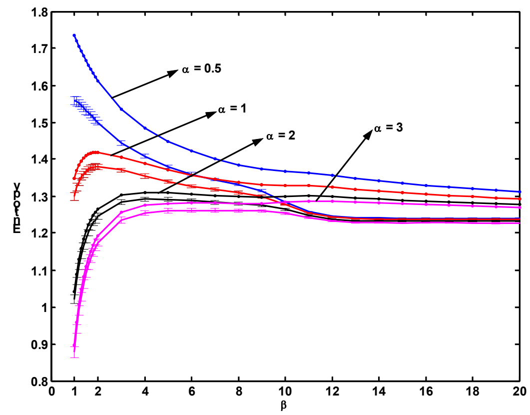

Figure 2. Entropy estimates on family of exponential distributions, determined by β (x-axis) and computed for various orders of α . Sample-size was fixed at $N = 10^4$ samples. It is evident that higher orders of $1 < \alpha < 4$ are seen to have better accuracies in terms of bias and variance.

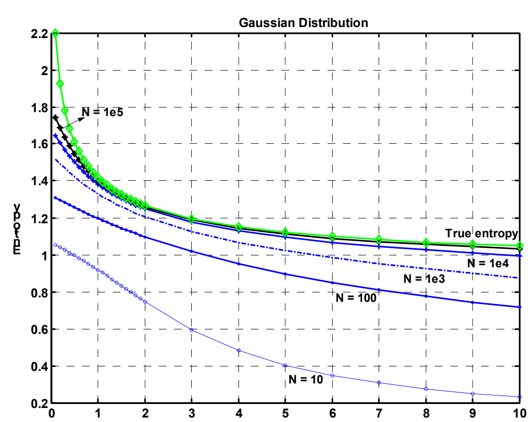

Figure 3. Illustrating the effect of sample size (*N*) in obtaining accurate estimates of the entropy of a data set. The plot also reinforces the need to have $1 < \alpha < 4$.

$$\mathbf{R} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}, \ \theta \in [0, \pi/4] \tag{12}$$

to yield

$$\mathbf{y} = \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \mathbf{R} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \mathbf{Rx} \tag{13}$$

The objective is to quantitatively index the amount of coupling between y_1 and y_2 by determining Renyi's mutual information $I_{R_\alpha}(\mathbf{y})$ as approximated by (11), i.e.,

$$I_{R_\alpha}(\mathbf{y}) \approx \sum_{i=1}^{2} H_{R_\alpha}(y_i) - H_{R_\alpha}(\mathbf{y}) \tag{14}$$

Since the components of $\boldsymbol{y}$ are formed from a linear combination of $\mathbf{x}$, their marginal distribution involves convolution of two uniform pdfs that usually has the form of a trapezoid or a triangle. The marginal entropies in question are analytically determined to be

$$H_{R_\alpha}(y_i) = -\log\left(\frac{-2\sqrt{2}A^2}{\begin{array}{c} 3(\cos\theta_1 - \sin\theta_1) \\ + A^2(2\sqrt{2})\sin\theta_1 \end{array}} \right) \tag{15}$$

where $\theta_1 = {\pi}/{4} - \theta$.

The joint entropy of the components of $\mathbf{y}$ is the same as joint entropy of the components of $\mathbf{x}$. This can be shown as follows:

$$H_{R_\alpha}(\mathbf{y}) = H_{R_\alpha}(\mathbf{x}) - \log(\det(\mathbf{R})) = H_{R_\alpha}(\mathbf{x}) \tag{16}$$

since $\log(\det(\mathbf{R})) = 0$. Knowing that the sources x_1, x_2 are independent, their joint entropy is equal to the sum of their marginals. Eventually, the theoretical Renyi's mutual information can be written as,

$$I_{R_\alpha}(\mathbf{y}) = \sum_{i=1}^{2} H_{R_\alpha}(y_i) - H_{R_\alpha}(x_1) - H_{R_\alpha}(x_2) \tag{17}$$

However, for all practical purposes, we do not have the source and the mixing information. While it is possible to compute the marginal entropies of the $\mathbf{y}$ components, the joint entropy needs some extra effort. The problem arises due to the fact that our definitions for Renyi's entropy estimation using sample-spacings are only limited to estimating marginal entropies. In order to compute the joint entropy, we propose to treat it as an ICA problem where the observations can be construed to be formed from a linear or non-linear combination of independent sources. By estimating the demixing matrix and hence the independent sources, the problem reduces to finding the marginal entropy of the estimated sources. Mathematically, we can describe it as

$$\hat{H}_{R_\alpha}(\mathbf{y}) = H_{R_\alpha}(\hat{\mathbf{x}}) - \log(\det(\mathbf{W}^{-1})) \tag{18}$$

where $\mathbf{W}^{-1}$ is the estimated demixing matrix and $\hat{\mathbf{x}}$ are the estimated independent components. If we have a perfect estimation of the ICA solution, then

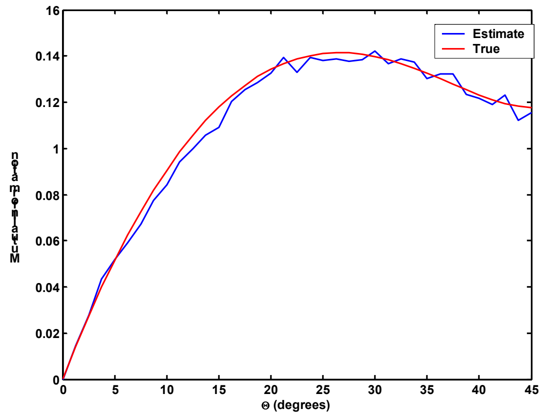

Figure 4. Mutual information between two synthetically mixed uniform sources, as a function of θ (in degrees). The estimated mutual information is in close agreement with the theoretical value for all rotation angles.

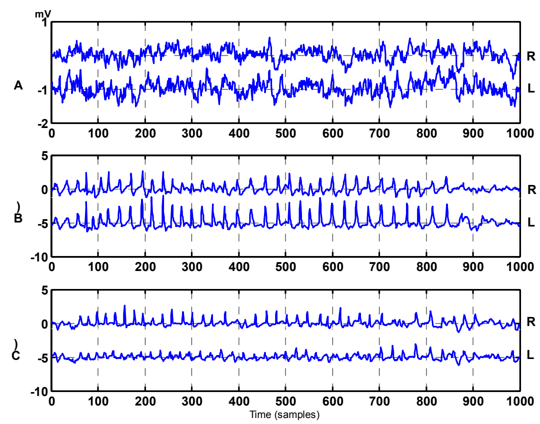

Figure 5. Rat EEG signals recorded from left and right intracranial electrodes. Left electrode signals are slightly offset for visualization purpose.

$\hat{H}_{R_\alpha}(\mathbf{y}) = H_{R_\alpha}(\mathbf{y})$, which implies $\mathbf{R} = \mathbf{W}^{-1}$ and $H_{R_\alpha}(\hat{\mathbf{x}}) = H_{R_\alpha}(\mathbf{x})$.

We applied the GED-ICA algorithm [4] for source-separation and then estimated the mutual-information at various degrees of mixing determined by $\mathbf{R}$. Fig 4. shows the average estimated and the true mutual information curves, superimposed on each other, as a function of the rotation angle (θ). Averaging was performed over 100 Monte-Carlo trials, the sample size was fixed at $N = 10^6$ and α was set equal to 2. Notice that the estimated mutual information is consistent with the actual (true) mutual information. As with most other estimation techniques, there is a small statistical error associated with estimation. As noted earlier, the bias problem surfaced in the estimation and it was compensated by adding a constant term that made the mutual information zero for statistically independent case (θ = 0). The same constant was then added to the mutual information estimates computed at other rotations as well. Overall, this experiment encourages us to use Renyi's entropy estimator for estimating dependencies in coupled systems.

Experiment E: Application to EEG data. One application where mutual information is commonly used is in quantifying the spatio-temporal dependencies of multivariate EEG data. As a start, we analyze the dependencies between two channels in 3 sets of male rat EEG data [5]. The data length was 5 seconds and was recorded from the left and right channels on the frontal cortex, corresponding to different stages of epileptic seizure (Fig. 5). Visually, the sets B and C consist of a lot of spike discharges and seem to be highly synchronized. Set A is a normal EEG and it is not easy to observe how well the right and left channels are synchronized with each other. Mutual information computed on all the three sets of data revealed that the synchronization was more pronounced in B, compared to A and C (the order was B > A > C, as seen in Table 1). In [5], it was shown that, with the exception of mutual information computed using box method, all the other measures give qualitatively equivalent results. The non-linear measures [5] were however, shown to be more sensitive and explained the possible nonlinear dependency between the signals. Comparisons of our proposed technique with few other measures, as shown in Table 1., indicates that our measure provides results that are consistent with the results obtained from other measures.

Examples	**I(R;L)**	Similarity-Index N(R|L)	N(L|R)	Coherence (at 9 Hz)
Set A	**0.3176**	0.46	0.42	0.88
Set B	**0.5021**	0.63	0.69	0.86
Set C	**0.1865**	0.24	0.32	0.40

Table1. Coupling strength evaluated for the 3 sets of Male Rat EEG data using different measures: I(R;L) is the estimated mutual information using our technique, N(R|L) and N(L|R) are the similarity-index coefficients indicating dependencies between Right (R) and Left (L) channels in each direction. Note the discrepancy with cross-coherence where the sets A and B are almost equally coupled.

V. DISCUSSION

In this study, we proposed a sample-spacing based estimator for Renyi's entropy, as a generalization of the exisintg estimator for Shannon's entropy. Based on order statistics, this estimator shows high estimation accuracy in terms of bias/variance balance and experimentally is observed to be most sample-efficient for entropy orders of $1<\alpha<4$. In terms of sample spacings, we demonstrate that using more intervals in the order statistics ensures better approximation of the true entropy. Consequently, it is recommended that in practice $m=\sqrt{N}$ or a similar monotonic function of the sample size is utilized. While this method is computationally very efficient for single-dimensional entropy estimation, it is not suitable for adaptation and its generalization to high dimensionality data is problematic (some approaches have been studied for the Sahannon case in the literature already). In this paper, we proposed using ICA as a tool to decompose the high dimensional entropy estimation problem into multiple single-dimensional entropy estimation problems. We demonstrated this principle in estimating mutual information, assuming linear ICA, an assumption which need not be generally true as most dynamical systems in real world have nonlinear interactions. Nonlinear ICA algorithms [7] might be more suitable for this purpose and will be studied in this context in a future publication.

ACKNOWLEDGMENTS

This paper/research was supported by contract number DAAD-16-03-C-0054 funded through DARPA and the U.S. Army Natick Soldier Center.

REFERENCES

[1] E.G. Miller, J.W.Fisher, "ICA Using Spacings Estimates Of Entropy," Proceedings of the Fourth International Symposium on ICA and Blind Signal Separation, 2003.

[2] O.Vasicek, "A Test for Normality Based on Sample Entropy," Journal of the Royal Statistical Society Series B, vol. 38, no. 1, pp. 54-59, 1976.

[3] T.M. Cover, J.A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[4] L. Parra, P. Sajda, "Blind Source Separation via Generalized Eigenvalue Decomposition," Journal of Machine Learning Research, vol 4. pp. 1261-1269, 2003.

[5] R.Q. Quiroga, A. Kraskow, T. Kreuz, P. Grassberger, "On the Performance of Different Synchronization Measure in Real Data: A Case Study on EEG Signals," Physical Review E, vol. 65, 2002.

[6] K. Hild, D. Erdogmus, J.C. Principe, "Blind Source Separation Using Renyi's Mutual Information," IEEE Signal Processing Letters, vol. 8, no. 6, pp. 174-176, 2001.

[7] C. Jutten, J. Karhunen, "Advances in Blind Source Separation (BSS) and Independent Component Analysis (ICA) for Nonlinear Mixtures," International Journal of Neural Systems, vol. 14, no. 15, pp. 267-292, 2004.