ELSEVIER

Neurocomputing 49 (2002) 25–38

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Blind source separation using Renyi's α-marginal entropies

Deniz Erdogmus*, Kenneth E. Hild II, Jose C. Principe

Computational NeuroEngineering Laboratory, University of Florida, Gainesville, FL 32611, USA

Received 13 March 2001; accepted 22 October 2001

Abstract

We have recently suggested the minimization of a nonparametric estimator of Renyi's mutual information as a criterion for blind source separation. Using a two-stage topology, consisting of spatial whitening and a series of Givens rotations, the cost function reduces to the sum of marginal entropies, just like in the Shannon's entropy case. Since we use a Parzen window density estimator and eliminate the joint entropy by employing an orthonormal demixing matrix, the problems of probability density function inaccuracy due to truncation of series expansion and the estimation of joint pdfs in high-dimensional spaces (given the typical paucity of data) are avoided, respectively. In our previous formulation, the algorithm was restricted to Renyi's second-order entropy and Gaussian kernels for the Parzen window estimator. The present work extends the previous results by formulating a new estimation methodology for Renyi's entropy, which allows the designer to choose any order of entropy and any suitable kernel function. Simulations illustrate that the proposed method compares favorably to Hyvarinen's FastICA, Bell and Sejnowski's Infomax and Common's minimum of mutual information. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Blind source separation; Renyi's entropy; Parzen estimation

1. Introduction

A typical blind source separation (BSS) system operates on observations that are obtained by passing unknown independent signals through an unknown mixing matrix. The block diagram of the BSS scheme we assume is given in Fig. 1. An observation vector $\mathbf{z} = \mathbf{H}^{\mathrm{T}}\mathbf{s}$ is obtained from the sources. First, a spatial whitening is applied to the observed data, $\mathbf{x} = \mathbf{W}^{\mathrm{T}}\mathbf{z}$, where the whitening transform $\mathbf{W}$ is evaluated from the

* Corresponding author.
E-mail address: deniz@cnel.ufl.edu (D. Erdogmus).

PII: S0925-2312(02)00526-X

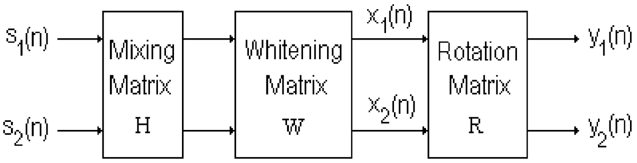

Fig. 1. System block diagram for two-source/two-measurement scheme.

autocorrelation matrix of $\mathbf{z}$ in the usual manner, namely, $\mathbf{W}=\mathbf{\Phi}\mathbf{\Lambda}^{-1/2}$, where $\mathbf{\Phi}$ is the matrix of eigenvectors of the autocorrelation matrix of $\mathbf{z}$, and $\mathbf{\Lambda}$ is the corresponding eigenvalue matrix. The adaptive part of the topology is the rotation matrix which produces an output $\mathbf{y}=\mathbf{R}(\theta)\mathbf{x}$.

In the BSS literature, the minimization of the mutual information (MMI) between outputs is considered to be the natural information theoretic criterion [3,4,8]. However, two of the most well-known methods for BSS [2] and [9] use, respectively, the maximization of output entropy and fourth-order cumulants. Shannon's mutual information can be written as the sum of Shannon's marginal entropies minus the joint entropy. One difficulty in using Shannon's MMI is the estimation of the marginal entropies. In order to estimate the marginal entropy, Comon and others approximate the output marginal probability density functions (pdf) with truncated polynomial expansions [4,5,8], which naturally introduces error in the estimation procedure. There are also parametric approaches to BSS, where the designer assumes a specific parametric model for the source distributions based on previous knowledge in the problem [4]. A well-known result from statistical signal processing theory is that if the designer chooses an accurate parametric model for the problem, it will outperform any nonparametric approach, as the ones proposed in this paper. However, it is also well known that the penalty for model mismatch is also high, so there is an intrinsic compromise on the use of parametric modeling. An algorithm proposed by Xu et al. [14] avoids the polynomial expansion by employing the nonparametric Parzen windowing to estimate directly Renyi's joint entropy at the output of the mapper [10]. Unfortunately, Xu's method requires estimation of the n-dimensional joint entropy (n number of sources), and nonparametric pdf estimation using Parzen windows is ill posed in high-dimensional spaces [14]. Our recently proposed algorithm avoids this shortcoming and has proved to be superior to many commonly accepted methods because it requires much less data to achieve the same performance level as shown in [7]. The algorithm in [7] was restricted to Renyi's quadratic mutual information and Gaussian kernels in Parzen windowing because the analytic formulation in [10] only applies to Gaussian kernels and quadratic mutual information. Recently, we overcame these limitations by introducing a new estimator for Renyi's entropy, which again utilizes Parzen windowing in the pdf estimation phase [6]. In this paper we combine these advances for blind source separation yielding the minimization of Renyi's mutual information (MRMI) algorithm.

The organization of this paper is as follows. First, we derive the cost function for BSS starting from Renyi's mutual information. In Section 3, we demonstrate how to estimate Renyi's entropy nonparametrically, i.e. directly from the samples, and in

Section 4, we define the information potential field and the information force and we demonstrate their role in adaptation. Section 5 is devoted to the derivation of the gradient of the cost function with respect to the rotation angles for use in the steepest descent algorithm. Finally, in Sections 6 and 7, we present separation results and conclude with a discussion of these results.

2. Derivation of the cost function

Recall the following equality for Shannon's definitions of mutual information, marginal and joint entropies for two random variables x and y:

$$I_S(x,y) = H_S(x) + H_S(y) - H_S(x,y). \tag{1}$$

The same equality is not valid for Renyi's definitions of these quantities because Renyi's entropy lacks the recursivity property of Shannon's entropy. Nevertheless, we will show that we can slightly modify Renyi's mutual information expression such that it preserves the global minimum of the mutual information. Renyi's mutual information for an n-dimensional random variable y is defined as [12]

$$I_{R_\alpha}(y) = \frac{1}{\alpha-1}\log\int_{-\infty}^{\infty}\frac{f_Y(y)^\alpha}{\prod_{o=1}^{n} f_o(y^o)^{\alpha-1}}\,\mathrm{d}y. \tag{2}$$

However, the sum of Renyi's marginal entropies minus the joint entropy is

$$\sum_{o=1}^{n} H_{R_\alpha}(y^o) - H_{R_\alpha}(y) = \frac{1}{\alpha-1}\log\frac{\int_{-\infty}^{\infty} f_Y(y)^\alpha\,\mathrm{d}y}{\int_{-\infty}^{\infty}\prod_{o=1}^{n} f_o(y^o)^\alpha\,\mathrm{d}y}. \tag{3}$$

Although this is not identical to (2), it is very similar in structure. In addition, (2) and (3) are both nonnegative and they both evaluate to zero if and only if the joint pdf can be written as the product of the marginal densities, i.e. when the output signals are statistically independent. This can be seen easily by letting the joint density, which is the integrand in the numerator, to be equal to the product of marginal densities, which is the integrand in the denominator. In that case, the argument of the logarithm in (3) becomes unity, hence the minimum value of zero is achieved, and thus the sources are separated. On the other hand, if the right-hand side in (3) becomes zero, the argument of the logarithm becomes unity, thus the numerator is equal to the denominator. For this to occur between a joint distribution and its marginals, it is necessary for the marginal random variables to be independent. Having proved that the expression in (3) is a valid criterion for measuring independence, we adopt it as the cost function instead of the actual mutual information, given in (2).

We now assume the two-stage demixing process proposed in [5], which consists of spatial whitening (sphering) and then rotation in n-dimensions. Only the rotation matrix is adapted to minimize the quantity in (3). Now, using the fact that Renyi's joint entropy is invariant to rotations [7], we can remove this term and reduce the cost function to

$$J = \sum_{o=1}^{n} H_{\mathrm{R}_\alpha}(y^o), \tag{4}$$

which mimics the cost function of [5,15] with Renyi's entropy substituted for Shannon's.

The observations are first whitened to eliminate correlation, and then rotated to a proper angle to restore independence, as shown in Fig. 1. The parameter vector θ of the rotation matrix is adapted to minimize the cost function in (4). This vector consists of $n(n-1)/2$ parameters θ_{ij}, $j > i$, where each parameter represents the amount of Givens rotation in the i–j plane. The overall rotation matrix is the product of the individual in-plane rotation matrices:

$$R(\theta) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} R_{ij}(\theta_{ij}). \tag{5}$$

In (5), all products are performed sequentially from the right (or left). The important point is to perform these operations in the same order and from the same side when evaluating the gradient expression. The Givens rotation in the i–j plane is defined as an identity matrix whose (i,i)th, (i,j)th, (j,i)th, and (j,j)th entries are modified to read $\cos\theta_{ij}, -\sin\theta_{ij}, \sin\theta_{ij}$, and $\cos\theta_{ij}$, respectively.

3. Estimating Renyi's entropy

Renyi's entropy with parameter α for a random variable y^o with pdf $f_o(.)$ is defined as (we will drop the $\pm\infty$ limits from the equations from now on)

$$H_\alpha(y^o) = \frac{1}{1-\alpha} \log \int_{-\infty}^{\infty} f_o^\alpha(y^o)\,\mathrm{d}y^o = \frac{1}{1-\alpha} \log E[f_o^{\alpha-1}(y^o)]. \tag{6}$$

The Parzen window pdf estimate of a random variable y^o for which only the samples $\{y_1^o, \dots, y_N^o\}$ are given, is defined by

$$\hat{f}_o(y^o) = \frac{1}{N} \sum_{i=1}^{N} \kappa_\sigma(y^o - y_i^o), \tag{7}$$

where the kernel function is $\kappa_\sigma(.)$, whose size is specified by the parameter σ, as defined in the following section [10].

When we substitute the sample mean for the expected value operator in (6), and then replace the actual pdf with its Parzen window estimate in (7) evaluated at the corresponding sample, we obtain our new estimator for Renyi's entropy of order α. Notice that in (8), the index j runs for the sample mean, and the index i runs for Parzen windowing.

$$\hat{H}_\alpha(y^o) = \frac{1}{1-\alpha} \log \left[\frac{1}{N^\alpha} \sum_j \left(\sum_i \kappa_\sigma(y_j^o - y_i^o) \right)^{\alpha-1} \right]. \tag{8}$$

The argument of the log in (6) is called the information potential [11], and will be denoted by $V_\alpha(y^o)$. Hence, a nonparametric, biased estimator for the information potential is given by

$$\hat{V}_\alpha(y^o) = \frac{1}{N^\alpha}\sum_j \left(\sum_i \kappa_\sigma(y_j^o - y_i^o)\right)^{\alpha-1}. \tag{9}$$

The information potential will be used later in the gradient computation. We have shown in [6] that, a symmetric differentiable kernel with a negative-definite Hessian at the origin guarantees that the minimization of the nonparametrically estimated entropy and the actual entropy occurs at the same point; hence, this should be the necessary guideline to choose a valid kernel function. In that work, we also proved that there is a correspondence with the kernel function in this estimator for information potential and the smoothing function in the global optimization method of convolution smoothing [6,13]. The same relationship is valid in this case, and it is trivial to modify that proof to the cost function we use here.

The following question comes into one's mind immediately. How good is the sample mean approximation of the expected value in the definition of the information potential? From the central limit theorem, we know that the distribution of the sample mean is asymptotically Gaussian. It is also known that the sample mean is an unbiased and asymptotically consistent estimator for the expected value operator. For these reasons, the sample mean is widely used and we also use it here. Moreover, the combination of the sample mean with the Parzen window estimator of Renyi's entropy possesses a very special property that was noted upon careful investigation of the relationship between this new estimator and the previously defined quadratic entropy estimator [11]. The previous quadratic information potential estimator using Gaussian kernels was defined as

$$\begin{aligned}
V_2(y^o) &= \int_{-\infty}^{\infty} f_o^2(y^o)\,\mathrm{d}y^o \cong \int_{-\infty}^{\infty}\left(\frac{1}{N}\sum_i G_\sigma(y^o - y_i^o)\right)^2 \mathrm{d}y^o \\
&= \int_{-\infty}^{\infty}\left(\frac{1}{N^2}\sum_i\sum_j G_\sigma(y^o - y_i^o)G_\sigma(y^o - y_j^o)\right)\mathrm{d}y^o \\
&= \frac{1}{N^2}\sum_i\sum_j \int_{-\infty}^{\infty} G_\sigma(y^o - y_i^o)G_\sigma(y^o - y_j^o)\,\mathrm{d}y^o \\
&= \frac{1}{N^2}\sum_i\sum_j G_{\sigma\sqrt{2}}(y_j^o - y_i^o).
\end{aligned} \tag{10}$$

This derivation makes use of the fact that the integral of the product of two Gaussian functions is another Gaussian function with twice the variance. Notice that there is no approximation in (10) apart from the implicit Parzen window pdf estimation. Now let us look at what our new estimator with the sample mean gives for quadratic entropy with the same choice of kernel function. We get the results by direct substitution of

$\alpha = 2$ and $\kappa = G_\sigma$ in (9):

$$\hat{V}_2(y^o) = \frac{1}{N^2} \sum_i \sum_j G_\sigma(y_j^o - y_i^o). \tag{11}$$

We conclude that the estimator in (11) has exactly the same form but a larger variance than (10) since it effectively uses a smaller kernel size in the Parzen window estimation. Remarkably, the sample mean approximation followed by Parzen window estimation with Gaussian kernels can be compensated by simply choosing a larger kernel size in (11) (specifically $\sqrt{2}$ times the original kernel size). For quadratic entropy calculations with Parzen estimation, any choice of kernel in (10) can be mimicked by letting the kernel in (9) equal

$$\kappa_{\text{new}}(x_j - x_i) = \int_{-\infty}^{\infty} \kappa_{\text{old}}(x - x_i)\kappa_{\text{old}}(x - x_j)\,\mathrm{d}x. \tag{12}$$

Gaussian kernels, however, are very special because by simply rescaling the kernel function to the appropriate size we can achieve this equality between the new and old estimators. This analysis shows that the sample mean estimation of the α-information potential is effectively a productive approach.

4. Information potential field and information forces

These quantities were first defined in [11] and analyzed in the context of BSS. At that time, only an estimator for Renyi's quadratic entropy ($\alpha = 2$) was available. Since we can now estimate any order of entropy with (8), we can extend the definition of information potential and information force, and furthermore, explore their relationships with their quadratic counterparts. One of the appeals of the information potential is its analogy to physics, where the samples become *information particles* and the kernels define the interaction laws. Thus, in an information potential field, it becomes possible to discuss information forces that these particles exert on each other. From (9), we have the following information potential (energy) estimator:

$$\hat{V}_{\alpha,\sigma}(y^o) = \frac{1}{N^\alpha} \sum_j \left(\sum_i \kappa_\sigma(y_j^o - y_i^o) \right)^{\alpha-1}, \tag{13}$$

where the one-dimensional size-σ kernel can be written in terms of the unit-size kernel according to

$$\kappa_\sigma(x) = \frac{1}{\sigma}\,\kappa(x/\sigma). \tag{14}$$

Notice that in this one-dimensional case, the standard deviation acts as a natural scaling factor for the Gaussian kernel.

Since potential energy of a set of particles is the sum of the individual potential energies, we can instantly write the potential energy of an information particle (sample)

y_j^o from the above expression as

$$\hat{V}_{\alpha,\sigma}(y_j^o) = \frac{1}{N^\alpha}\left(\sum_i \kappa_\sigma(y_j^o - y_i^o)\right)^{\alpha-1}. \tag{15}$$

Now, we can define and compute the information force acting on this particle as

$$F_\alpha(y_j^o) = \frac{\partial \hat{V}_\alpha(y_j^o)}{\partial y_j^o} = \frac{(\alpha-1)}{N^\alpha}\left(\sum_i \kappa_\sigma(y_j^o - y_i^o)\right)^{\alpha-2}\left(\sum_{i\neq j} \kappa'_\sigma(y_j^o - y_i^o)\right). \tag{16}$$

In this form, the information force expression is not very informative. We can understand the nature of the order-α force better when we write it as a function of the force created by the information potential for $\alpha = 2$:

$$F_\alpha(y_j^o) = (\alpha-1)\hat{f}_o^{\alpha-2}(y_j^o)F_2(y_j^o), \tag{17}$$

where the pdf estimation is performed by Parzen windowing as in (7) and the quadratic force is defined as

$$F_2(y_j^o) = \frac{1}{N^2}\left(\sum_{i\neq j} \kappa'_\sigma(y_j^o - y_i^o)\right). \tag{18}$$

This quadratic force expression reduces to the exact same definition in [11] when Gaussian kernels are assumed. With this formulation, it also becomes possible to define the force exerted on a particular sample by another sample. The force on y_j^o due to y_i^o is given by

$$\begin{aligned} F_\alpha(y_j^o; y_i^o) &= (\alpha-1)\hat{f}_o^{\alpha-2}(y_j^o)F_2(y_j^o; y_i^o), \\ F_2(y_j^o; y_i^o) &= \frac{1}{N^2}\kappa'_\sigma(y_j^o - y_i^o). \end{aligned} \tag{19}$$

We have now completed the formulation of information forces and set up the link between the order-α force and the quadratic force. Basically, the quadratic force can be regarded as the foundation of information forces of all orders. Forces of any order can be written as a scaled version of the quadratic force, where the scaling factor is a power of the probability density of the particle that the force acts on. We see that the order-α information force is scaled from the quadratic force by a power of the probability density for that specific particle. For $\alpha > 2$, this scale factor is larger for particles with greater probability densities, and for $\alpha < 2$, it is larger for particles with smaller probability densities. Thus, for $\alpha > 2$, larger forces will act on the concentrated regions of the data to spread them apart.

Since the relation between entropy and information potential is the logarithm, the information force at the output of the mapper is all what matters to train an adaptive system with a gradient-based method to maximize or minimize entropy. The gradient of the cost function we have introduced in (4) with respect to the parameters of the rotation matrix can be written in terms of the information forces in the following manner:

$$\frac{\partial J}{\partial \theta} = \sum_o \frac{\partial \hat{H}_\alpha(y^o)}{\partial \theta} = \sum_o \sum_j \frac{1}{1-\alpha} \frac{\partial \hat{V}_\alpha(y_j^o)/\partial y_j^o}{\hat{V}_\alpha(y^o)} \frac{\partial y_j^o}{\partial \theta}$$

$$= \sum_o \sum_j \frac{1}{1-\alpha} \frac{F_\alpha(y_j^o)}{\hat{V}_\alpha(y^o)} S_\theta(y_j^o), \tag{20}$$

where o is the index running over the output channels, and j the index running over the samples. The derivative of an output sample with respect to the weights is termed the sensitivity; hence, the overall gradient is a function of the information forces, the information potentials, and the sensitivities of the information particles.

In order to understand the role of the information forces in the adaptation process, it is helpful to study (20) in detail. Observe that the gradient of the cost function with respect to the weights consists of a sum of gradients generated by each information particle y_j^o. Each of these components is directly proportional to the information force on the corresponding particle and inversely proportional to the potential of that output channel.

5. The gradient vector

Recall the gradient expression given in (20). The forces and information potentials can be evaluated exactly as shown in Section 4. The sensitivity is simply the gradient of the corresponding output with respect to the parameters. It can be calculated using the Givens rotation matrices as follows:

$$y_j^o = R^o x_j,$$

$$S_{\theta_{ij}}(y_j^o) = \frac{\partial y_j^o}{\partial \theta_{ij}} = \frac{\partial R^o}{\partial \theta_{ij}} x_j = \left(\frac{\partial R}{\partial \theta_{ij}}\right)^o x_j, \tag{21}$$

where R^o is the oth row of R, and x_j the jth whitened sample vector. The partial derivative of the overall rotation matrix R with respect to the parameter θ_{ij} can be evaluated easily from

$$\frac{\partial R}{\partial \theta_{ij}} = \left(\prod_{p=1}^{i-1} \prod_{q=p+1}^{n} R_{pq}\right) \left(\prod_{q=i}^{j-1} R_{iq}\right) R'_{ij} \left(\prod_{q=j+1}^{n} R_{iq}\right) \left(\prod_{p=i+1}^{n} \prod_{q=p+1}^{n} R_{pq}\right). \tag{22}$$

The derivative of R_{ij} is simply a sparse matrix with the obvious entries being the derivatives of the corresponding sin and cos functions of θ_{ij}. The rotation angles then can be updated using a steepest descent algorithm with the use of this gradient.

6. Simulations

The whitening-rotation scheme has a very significant advantage. When this topology is used with a large number of samples, oftentimes, there are no local minima in the cost function. Consider the two-source separation problem. The rotation matrix consists of a single parameter, which can assume values in the interval $[0, 2\pi)$. As far as separation is concerned, there are four equivalent solutions, which correspond to two permutations of each source and the two possible signs for each source. The value of the cost function is periodic with $\pi/2$ over the rotation angle θ, and most often is a very smooth function (sinusoidal like), which is easy to search.

Numerous simulations were performed with this new BSS algorithm using different α and kernels on synthetic data and audio instantaneous mixtures. In order to compare the results from different algorithms, we used a signal-to-distortion ratio, which is defined as

$$SDR = \frac{1}{n}\sum_{i=1}^{n} 10\log_{10}\left(\frac{(\max q_i)^2}{q_i q_i^{\mathrm{T}} - (\max q_i)^2}\right), \tag{23}$$

where $\mathbf{q} = \mathbf{R}\mathbf{W}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}$, and q_i is the ith row of $\mathbf{q}$. This criterion effectively measures the distance of $\mathbf{q}$ from an identity matrix and is invariant to permutations and scaling.

We first start with an investigation of the effect of α on the separation of instantaneous mixtures, when the source kurtosis spans reasonable values (the so-called super- and sub-Gaussian signals). Although our nonparametric method in principle can separate signals independent of its kurtosis (since the pdf is estimated at the output directly), a question of paramount importance is ‘what value of entropy order should one use for different source densities in order to achieve optimal performance?’ In search of the answer to this question, a series of Monte Carlo simulations are performed, using source distributions of different kurtosis values. In all these simulations, the two sources are assumed to have the same generalized Gaussian density, which is given by $G_\upsilon(x) = C\exp(-|x|^\upsilon/(\upsilon E[|x|^\upsilon]))$. The parameter υ controls the kurtosis of the density and this family includes distributions ranging from Laplacian ($\upsilon = 1$) to uniform ($\upsilon \to \infty$). Gaussian distribution is a special case corresponding to ($\upsilon = 2$), which leads to the classification of densities as super- and sub-Gaussian for ($\upsilon < 2$) and ($\upsilon > 2$), respectively. For a given kurtosis value, the training data set is generated from the corresponding generalized Gaussian density and a random mixing matrix is selected. Then the separation is performed using various entropy orders (tracing the interval from 1.2 to 8 in steps of 0.4) and Gaussian kernels. The Gaussian kernel size was set at 0.25, and the adaptation using MRMI was run to achieve a convergence of the SDR within 0.1 dB (although in practice this cannot be used as the stopping criterion), which usually occurred in < 50 iterations with a step size of 0.2.

According to these simulations, the optimal entropy orders for the corresponding kurtosis value of the source densities are presented in Table 1. These results indicate that, for super-Gaussian sources, entropy orders $\geqslant 2$ should be preferred, whereas for sub-Gaussian sources, entropy orders smaller than 2, perhaps closer to 1 or even smaller than 1, should be preferred. These results are in conformity with our expectations from the analysis of the information forces in Section 4. As we have seen in that analysis, entropy orders larger than 2 emphasize samples in concentrated regions of data, whereas smaller orders emphasize the samples in sparse regions of data. If the mixtures belong to different kurtosis classes, then the quadratic entropy can be employed as it puts

Table 1
Optimal entropy order versus source density kurtosis

	Kurtosis of sources	Optimal entropy order
Super-Gaussian sources	0.8 ($\upsilon = 1$)	6.4
	0.5 ($\upsilon = 1.2$)	5.2
	0.2 ($\upsilon = 1.5$)	2
Sub-Gaussian sources	−0.8 ($\upsilon = 4$)	1.2
	−0.9 ($\upsilon = 5$)	1.6
	−1.0 ($\upsilon = 6$)	1.2

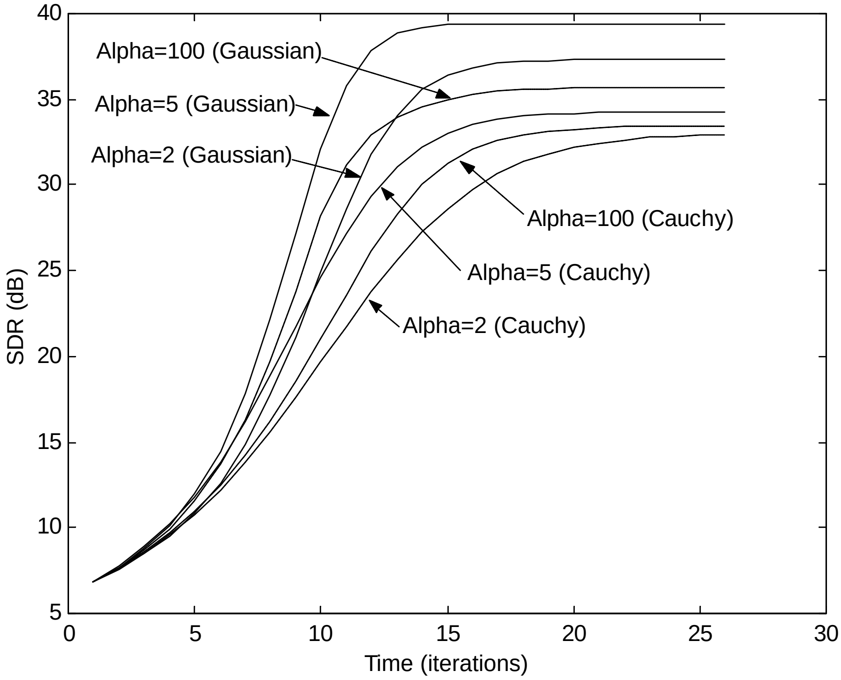

Fig. 2. Evolution of the signal-to-distortion ratio during iterations for two sources.

equal emphasis on all data points regardless of their probability density. This effect is very different from some of the available algorithms where the BSS algorithms diverge if the kurtosis of the sources are misestimated [8,9]. Another interesting aspect of this simulation is that it seems to imply that Shannon information definition ($\alpha \to 1$) is not particularly appropriate to separate super-Gaussian sources, although it may be useful for sub-Gaussian sources.

The next question addresses the performance of the MRMI algorithm for a realistic source such as speech. Fig. 2 shows the evolution of the SDR values as a function of

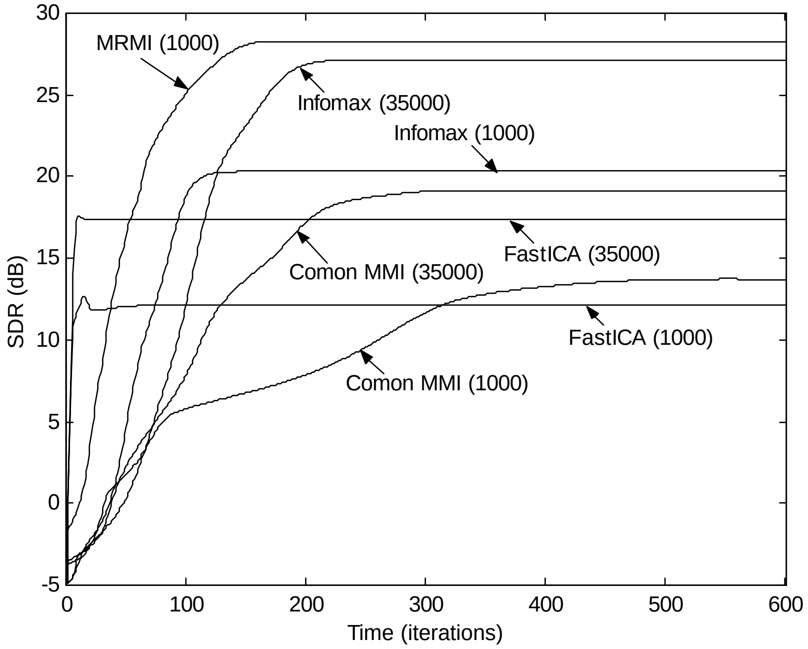

Fig. 3. SDR versus iterations for MRMI, Infomax, and FICA using the designated number of samples.

number of iterations in a two-audio source problem for different choices of the kernel function and the parameter α. These plots clearly demonstrate that for both kernels, a better separation is achieved when $\alpha = 5$ is utilized. Also, we observe that the solutions generated using the Gaussian kernel are better than those generated with the Cauchy kernel, which hints at the possibility of determining an optimal kernel choice for a given data set. How to do this, however, is still an open question, but we know that the optimal kernel will be a function of the source distributions. There is no impressive deterioration of performance when α changes from 5 to 2, since both provide SDRs larger than 30 dBs (20 dB is considered an acceptable separation performance). We also see that for very large α values performance deteriorates as we can expect due to the smoothing effect in the kernel.

For a final comparison, we show SDR plots in Fig. 3 for our MRMI method with $\alpha = 2$ and Gaussian kernels, the FastICA (FICA) [9] with the symmetric approach and the cubic nonlinearity, Infomax [2] with Amari's natural gradient [1], and Comon's MMI using an instantaneous mixture of 10 audio sources. The sources consist of one music source, four female and five male speakers. Spatial prewhitening is used for each method, and the mixing matrix entries were chosen from a uniform density on $[-1, 1]$. The numbers in parentheses are the number of data samples used to train each algorithm. It is clearly seen from the figure that the MRMI method achieves better performance, although it uses a smaller data set. The improved data efficiency of the MRMI method is discussed in greater detail in [7], but we attribute it to the fact that we are directly estimating entropy at the output, and our method seems to capture better the information contained in the samples. Although the MRMI method converges in fewer iterations than the others, keep in mind that it has $O(N^2)$ computational complexity per update as compared to $O(N)$ for the other two methods. Yang and Amari's MMI algorithm was also applied to this problem, however, we were never able to achieve an acceptable separation level; therefore, the corresponding results are not included here.

7. Conclusions

We have extended our previous work on BSS with Renyi's entropy by removing the limitations on the choice of entropy order and kernel function. The proposed method, MRMI, employs Renyi's mutual information to arrive at a simple and performance-wise advantageous BSS algorithm, when compared to many well-known algorithms such as Bell and Sejnowski's Infomax, Comon's MMI and Hyvarinen's FastICA. We have shown that this new algorithm uses the data efficiently and is therefore able to achieve a better performance for a given small set of samples, or in other words, it requires less data to achieve the same performance as the other competing algorithms. This property is vital in tracking a changing environment.

The parameter α in Renyi's definition of entropy stands as a design parameter whose effect on the performance and convergence properties of the algorithm is yet to be explored in more detail. We have demonstrated how this parameter changes the information forces acting on an information particle (sample) compared to the quadratic case. A weighting factor that is proportional to the estimated probability density of the particle of interest scales the quadratic force, hence the larger the probability density of the particle, the larger the force that acts on it. This will cause the natural clusters of the data samples to disintegrate and spread faster since the density estimates for a group of particles that are closely spaced will be higher than isolated particles in the perimeter of the space. Thus, by adjusting the parameter α, one can control how fast this spreading of the data clusters will occur. Analytical analysis and simulations had shown that for super-Gaussian sources an entropy order $\geqslant 2$, and for sub-Gaussian sources, an entropy order smaller than 2 should be used.

Another advantage of the MRMI is that it takes advantage of the large literature and methods for training nonlinear systems (such as the backpropagation algorithm) by substituting the injected error by the information force. This coupled with the superior estimation capabilities of the Parzen estimator leads us to predict that MRMI can be applied even in the case of nonlinear mixtures. As a final comment, it is noted that the proposed method achieves better separation for a small number of data points, because the other algorithms either require the estimation of a joint entropy or the fourth order cumulants, both of which require a larger number of samples due to the dimensionality and the sensitivity to outliers, respectively.

In this paper, no form of optimization of the kernel function is employed. Future work on how to optimize the kernel function choice and especially the kernel size must be conducted. This, the authors believe, will improve both the final performance and the convergence speed of the algorithm.

Acknowledgements

This work is partially supported by NSF Grant ECS-9900394.

References

[1] S. Amari, Neural learning in structured parameter spaces—natural Riemannian gradient, in: Proceedings of the NIPS'96, 1996, pp. 127–133.

[2] A. Bell, T. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (1995) 1129–1159.

[3] J. Cardoso, Blind signal separation: statistical principles, Proc. IEEE 86 (10) (1998) 2009–2025.

[4] S. Choi, A Cichocki, S. Amari, Flexible independent component analysis, J. VLSI Signal Process. 26 (2000) 25–38.

[5] P. Comon, Independent component analysis, a new concept? Signal Process. 36 (1994) 287–314.

[6] D. Erdogmus, J.C. Principe, Generalized information potential for training adaptive systems, IEEE Trans. Neural Networks, 2001, to appear.

[7] K.E. Hild II, D. Erdogmus, J.C. Principe, Blind source separation using Renyi's mutual information, IEEE Signal Process. Lett. 8 (2001) 174–176.

[8] A. Hyvarinen, Survey on independent component analysis, Neural Comput. Surveys 2 (1999) 94–128.

[9] A. Hyvarinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE Trans. Neural Networks 10 (1999) 626–634.

[10] E. Parzen, On estimation of a probability density function and mode, in: Time Series Analysis Papers, Holden-Day, Inc., CA, 1967.

[11] J.C. Principe, D. Xu, J. Fisher, Information theoretic learning, in: S. Haykin (Ed.), Unsupervised Adaptive Filtering, Wiley, New York, 2000, pp. 265–319.

[12] A. Renyi, Probability Theory, North-Holland, Amsterdam, 1970.

[13] R.Y. Rubinstein, Simulation and the Monte Carlo Method, Wiley, New York, 1981.

[14] D. Xu, J.C. Principe, J. Fisher, H. Wu, A novel measure for independent component analysis (ICA), to appear in IEEE Transactions, 2002.

[15] H. Yang, S. Amari, Adaptive online learning algorithms for blind separation: maximum entropy and minimum mutual information, Neural Comput. 9 (1997) 1457–1482.

Deniz Erdogmus received his B.S. in Electrical & Electronics Engineering and B.S. in Mathematics in 1997, and his M.S. in Electrical & Electronics Engineering, with emphasis on systems and control, in 1999, all from the Middle East Technical University, Ankara, Turkey. From 1997 to 1999, he worked as a research engineer at the Defense Industries Research and Development Institute (SAGE) under The Scientific and Technical Research Council of Turkey (TUBITAK). Since 1999, he has been working towards his Ph.D. at the Electrical and Computer Engineering Department at University of Florida, under the supervision of Jose C. Principe. His current research interests include information theory and its applications to adaptive systems and adaptive systems for signal processing, communications and control.

Kenneth E. Hild II received the B.S. degree in Electrical Engineering in 1992 and the M.S. degree in Electrical Engineering in 1996, both from The University of Oklahoma. The emphasis during his graduate work at OU was in the areas of communications, signal processing and controls. His thesis research topic was non-linear equalization of partial response channels. From 1995 to 1999 he was employed full-time with Seagate Technologies, Inc. where he served as an Advisory Development Engineer in the Advanced Concepts Group. Kenneth returned to academia in 1998, where he is presently in his fourth year of the Ph.D. program at the University of Florida. His current interests include blind source separation and information theoretic learning.

Jose C. Principe is Professor of Electrical and Computer Engineering and Biomedical Engineering at the University of Florida where he teaches advanced signal processing, machine learning and artificial neural networks (ANNs) modeling. He is BellSouth Professor and the Founder and Director of the University of Florida Computational NeuroEngineering Laboratory (CNEL). His primary area of interest is processing of time varying signals with adaptive neural models. The CNEL Lab has been studying signal and pattern recognition principles based on information theoretic criteria (entropy and mutual information).Dr. Principe is an IEEE Fellow. He is the Chair of the Technical Committee on Neural Networks of the IEEE Signal Processing Society, Member of the Board of Governors of the International Neural Network Society, and Editor in Chief of the IEEE Transactions on Biomedical Engineering. He is a member of the Advisory Board of the University of Florida Brain Institute. Dr. Principe has more than 70 publications in refereed journals, 10 book chapters, and 160 conference papers. He directed 35 Ph.D. dissertations and 45 Master theses. He recently wrote an interactive electronic book entitled ”Neural and Adaptive Systems: Fundamentals Through Simulation” published by John Wiley and Sons.