Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Signal Processing 86 (2006) 182–194

SIGNAL PROCESSING

www.elsevier.com/locate/sigpro

An analysis of entropy estimators for blind source separation

Kenneth E. Hild II[a], Deniz Erdogmus[b], Jose C. Principe[c,*]

[a]*Department of Radiology, The University of California at San Francisco, San Francisco, CA 94122, USA*
[b]*Departments of Computer Science and Engineering and Biomedical Engineering, Oregon Health & Science University, Beaverton, OR 97006, USA*
[c]*Department of Electrical and Computer Engineering, The University of Florida, Gainesville, FL 32611, USA*

Received 1 December 2003; received in revised form 10 January 2005; accepted 27 April 2005
Available online 20 July 2005

Abstract

An extensive analysis of a non-parametric, information-theoretic method for instantaneous blind source separation (BSS) is presented. As a result a modified stochastic information gradient estimator is proposed to reduce the computational complexity and to allow the separation of sub-Gaussian sources. Interestingly, the modification enables the method to simultaneously exploit spatial and spectral diversity of the sources. Consequently, the new algorithm is able to separate i.i.d. sources, which requires higher-order spatial statistics, and it is also able to separate temporally correlated Gaussian sources, which requires temporal statistics. Three reasons are given why Renyi's entropy estimators for Information-Theoretic Learning (ITL), on which the proposed method is based, is to be preferred over Shannon's entropy estimators for ITL. Also contained herein is an extensive comparison of the proposed method with JADE, Infomax, Comon's MI, FastICA, and a non-parametric, information-theoretic method that is based on Shannon's entropy. Performance comparisons are shown as a function of the data length, source kurtosis, number of sources, and stationarity/correlation of the sources.

Keywords: Blind source separation; Information theoretic learning; Renyi's quadratic entropy; Kurtosis; Independent component analysis

1. Introduction

Blind source separation (BSS) is a method of extracting one or more desired signals from an observed mixture of signals. Strictly speaking the term 'blind' denotes that nothing is known about either the sources, including the source statistics, or the mixing process. Suppose that N samples of a

*Corresponding author. Tel.: +1 352 392 2682; fax: +1 352 392 0044.
E-mail addresses: k.hild@ieee.org (K.E. Hild II), derdogmus@ieee.org (D. Erdogmus), principe@cnel.ufl.edu (J.C. Principe).

0165-1684/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.sigpro.2005.04.015

set of M zero-mean desired signals, $s_m(n)$, for $m = \{1, 2, \ldots, M\}$ and $n = \{1, 2, \ldots, N\}$, are available and are combined into an $(M \times N)$ matrix, $\boldsymbol{S}$. Suppose further that the linear, instantaneous mixing matrix is denoted as $\boldsymbol{A}$ and that M observations, $x_m(n)$, are available and are combined into an $(M \times N)$ matrix $\boldsymbol{X}$. The observations can then be represented mathematically as, $\boldsymbol{X} = \boldsymbol{AS}$. Demixing is attempted by linearly combining the observations. This produces M outputs at each time instant, $y_m(n)$, which together form an $(M \times N)$ matrix $\boldsymbol{Y}$ that can be expressed as, $\boldsymbol{Y} = \boldsymbol{WX} = \boldsymbol{WAS}$. For $\boldsymbol{A}$ full rank the most obvious solution is $\boldsymbol{W} = \boldsymbol{A}^{-1}$, in which case $\boldsymbol{Y} = \boldsymbol{S}$ as desired. The BSS problem can therefore be stated as follows: given a set of observations $\boldsymbol{X}$, find the $\boldsymbol{W}$ such that $\boldsymbol{Y}$ is the best approximation of $\boldsymbol{S}$. For additional details on BSS, see papers by Cardoso [1] and Hyvarinen [2].

The canonical contrast for BSS is to minimize the mutual information (MI) between the outputs [1]. If the observations are sphered prior to demixing and the demixing matrix is constrained to be a pure rotation, then minimizing MI is equivalent to minimizing the sum of marginal entropies [3]. Hence, this class of criteria for BSS entails the selection of a definition of entropy and a method to estimate the entropy from samples. Herein, the definition of entropy is restricted to the family of entropies formulated by Alfred Renyi [4] and the three methods used to estimate entropy from data are all based on Parzen window probability density function (pdf) estimation using Gaussian kernels [3,5]. Renyi's definition of entropy allows for a fairly comprehensive examination of this class of criteria since it represents, as a function of a single user-defined parameter α, a family of entropies that encompasses Shannon's definition [6,7] in the limit as α approaches 1. The entropy estimator is also a function of a single user-defined parameter, σ, which represents the width of the Gaussian kernel. Consequently, the class of BSS algorithms that consist of minimizing a sum of marginal entropies can be studied by observing the effect of jointly selecting α and σ for each of the three entropy estimators.

This particular approach to BSS falls under the general framework of Information Theoretic Learning (ITL), a term coined in a paper by Principe et al. [8] to denote a class of optimization algorithms that replace the conventional mean square error (MSE) criterion in the adaptation of linear and nonlinear systems. More specifically, ITL methods are concerned with the extremization of criteria based on a formulation of either (Renyi's) entropy or a quadratic measure of divergence that may be computed directly from samples. This paradigm represents a general optimization procedure that unifies supervised and unsupervised learning and has been used for function approximation, feature extraction, clustering, and for BSS. With respect to BSS, $\alpha = 2$ is used in the original paper by Hild et al. [3], while the extension to any value of α (except $\alpha = 1$) is covered in a paper by Erdogmus et al. [9]. The present paper provides a systematic study of the joint effect of Renyi's entropy order, α, and the kernel size used in the entropy estimation, σ, for three entropy estimators with special emphasis placed on the separation of sub-Gaussian sources. The results of this discussion suggest slight modifications to the originally proposed criterion and provides new insight as to why Renyi's quadratic entropy ($\alpha = 2$) is preferred over both Shannon's entropy ($\alpha = 1$) and kurtosis-based methods.

2. Renyi's entropy for BSS

The criterion under consideration is the sum of Renyi's marginal entropies, which is expressed as

$$J_\alpha(Y) = \sum_{m=1}^{M} H_\alpha(Y_m) \quad (\text{for } \alpha > 0), \tag{1}$$

where $H_\alpha(Y_m)$ is Renyi's marginal entropy of order α for the mth output. The discussion is initially limited to theoretical entropies and is then followed by a discussion of three sample-based entropy estimators.

2.1. Renyi's theoretical entropy

The Central Limit Theorem states that the pdf of a summation of independent random variables

tends toward the Gaussian distribution. Therefore, the goal of BSS is to force the pdf of each output, $f_{Y_m}(y_m)$, to be as far from Gaussian as possible. The nice feature of Shannon's entropy is that, for a fixed variance, it is maximized for Gaussian distributions. This is ideal for BSS when a sphering/rotation architecture is used because separation can be achieved simply by minimizing the sum of (Shannon's) entropies. This is true independent of whether the source distributions are sub-Gaussian or super-Gaussian as demonstrated in Fig. 1. This figure shows a plot of both Renyi's quadratic and Shannon's (theoretical) entropies as a function of β, where β is the parameter of a generalized Gaussian pdf

$$f_{Y_m}(y_m) = B_m e^{-C_m|y_m|^\beta}$$

and where $\beta<2$ refers to the super-Gaussian region, $\beta>2$ refers to the sub-Gaussian region, and B_m and C_m are functions of β that ensure the pdf integrates to 1 and that yield a pdf corresponding to a unit-variance random variable. The values in Fig. 1 are numerical estimates of the theoretical entropies. The asterisks indicate analytically computed values of Renyi's entropies for $\beta = 1$, 2, and infinity, which correspond to a Laplacian, Gaussian, and uniform pdf, respectively. The values of Renyi's quadratic entropy and Shannon's entropy for a uniform random variable are identical (this is true for all $\alpha>0$). This figure shows that, for the generalized Gaussian family where $1\leqslant\beta\leqslant10$, Shannon's entropy is maximized for $\beta = 2$, as expected, and Renyi's quadratic entropy is maximized for β equal to 4.

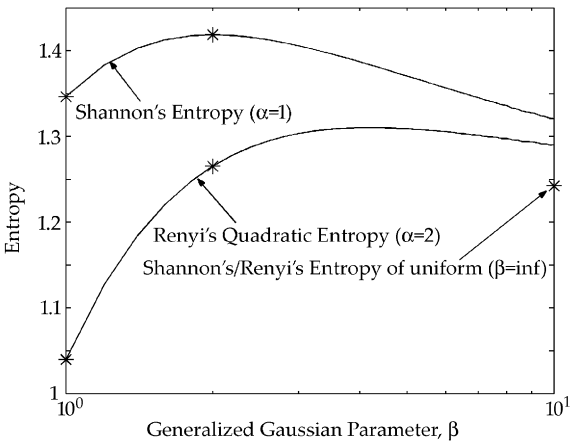

Fig. 1. Renyi's theoretical entropy, for $\alpha = 1$ and $\alpha = 2$, versus β.

In Fig. 2 the entropy of a mixture of 2 Laplacian sources is plotted as a function of rotation angle, θ, for values of $\beta = 1$, 2.7, 5, and 10, where $k\pi/2$ radians corresponds to separation for k any integer and the (2×2) matrix representing the product of the mixing and demixing matrices is given by

$$WA = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}.$$

The results for both Renyi's quadratic and Shannon's entropies in Fig. 2 were scaled for visualization purposes so that the minimum value is 0 and the maximum value is 1. For Laplacian sources, shown in the upper left subplot, the two results are virtually identical. In fact, although it is not shown, there is very little difference in the space of the demixing coefficients between Renyi's quadratic and Shannon's entropies for all super-Gaussian pdfs of the generalized Gaussian family. The remaining subplots show that the behavior of Renyi's quadratic entropy for sub-Gaussian data is much more complex than Shannon's entropy.

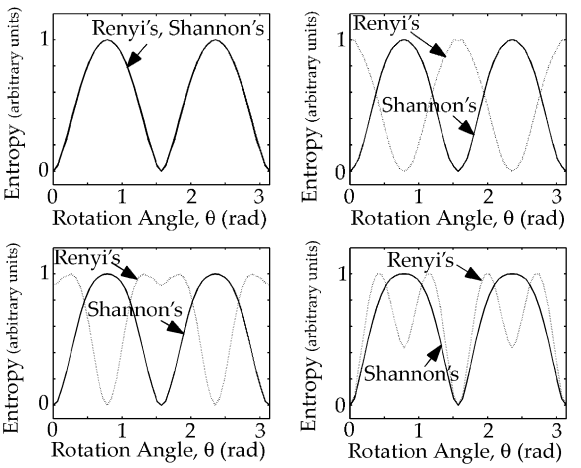

Fig. 2. Renyi's quadratic theoretical entropy versus rotation angle ($k\pi/2$ rad is solution). Upper left, upper right, lower left, lower right subplots are for $\beta = 1$, 2.7, 5, 10, respectively.

Notice that a local minimum always occurs at the solution (there is a minimum at the solution in the upper right subplot of Fig. 2 although it is difficult to see). However, it is neither the only minimum nor the global minimum for $2<\beta<8$. The fact that Renyi's quadratic entropy has an acceptable performance surface for super-Gaussian but not sub-Gaussian signals is related to the monotonicity of Renyi's quadratic entropy as a function of β. In order for Eq. (1) to be a suitable criterion for BSS, it is necessary that the entropy used in the definition is monotonically increasing for $1\leqslant\beta<2$ and monotonically decreasing for $2<\beta\leqslant$ infinity. If, on the other hand, the entropy is only guaranteed to be monotonic (increasing or decreasing) for both of these regions then it is a simple matter to include the appropriate change of sign in Eq. (1) so that its minimization leads to separation. This relaxation of the constraints for separation requires extra information, which can and must be estimated from the data. In particular, the modified criterion must determine whether each output is sub-Gaussian or super-Gaussian. The result is that Eq. (1), which consists of *theoretical* entropies, is an ideal demixing criterion for BSS without regard for the Gaussianity of the sources only when $\alpha=1$ (Shannon's entropy). Likewise, $\alpha<0.6$, $\alpha=1$, $\alpha>4$ are the only values of α are appropriate for Eq. (1) if the appropriate sign change (based on the Gaussianity) is included as shown in the first author's dissertation [10].

2.2. Renyi's empirical entropy

The preceding discussion is limited to theoretical quantities as opposed to using an entropy estimator, which produces values based on a finite number of samples. For the case that the pdfs are estimated using Parzen windows [5] with Gaussian kernels, the entropy estimator is [9]

$$\hat{H}_\alpha(Y_m,\sigma) = \frac{1}{1-\alpha}\log\frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{N}\sum_{k=1}^{N}G(y_m(n)-y_m(k),2\sigma^2)\right)^{\alpha-1} \quad (\text{for } \alpha>0,\ \alpha\neq 1), \tag{2}$$

where $G(y_m(n),\ \sigma^2)$ is the value of a Gaussian kernel evaluated at $y_m(n)$ and σ is a user-defined parameter referred to as the kernel size. Details of the derivations of this equation and Eq. (4) below are not provided here as they may be found in papers by Erdogmus et al. [9,11]. Eq. (2) is not valid for $\alpha=1$ since it results in the indeterminant value 0/0. There are two ways in which an analogous expression can be found for an estimator for $\alpha=1$. In the first method the gradient of Eq. (2), which is needed for gradient-based adaptation anyway, is found and then α is set to 1 in the gradient expression. The second method is derived by noticing that Shannon's (theoretical) entropy can be written in terms of an expectation as follows, $H_1(Y_m)=-E\ [\log(f_{Ym}(Y_m))]$. Approximating the expectation with the sample mean and using Parzen window estimation for the pdf produces the following estimator for Shannon's entropy:

$$\hat{H}_1(Y_m,\sigma) = \frac{-1}{N}\sum_{n=1}^{N}\log\frac{1}{N}\sum_{k=1}^{N}G(y_m(n)-y_m(k),\sigma^2) \quad (\text{for } \alpha=1) \tag{3}$$

which is similar to that used previously by Viola et al. [12] for processing magnetic resonance images. It is simple to show that the gradient of Eq. (3) is identically the gradient of Eq. (2) with $\alpha=1$.

Both Eqs. (2) and (3) have $O(N^2)$ computational complexity. An $O(N)$ estimator may be obtained with the help of the Stochastic Information Gradient (SIG) [11]. This involves the removal of one of the summations in the entropy estimator of Eq. (2), producing a third entropy estimator

$$\hat{H}_2(Y_m,\sigma) = -\log\frac{1}{N}\sum_{k=1}^{N}G(y_m(k)-y_m(k-p),2\sigma^2) \quad (\text{for } \alpha=2), \tag{4}$$

where the difference in time between the outputs, p, is user-defined. The recommended value is $p=1$, which is particularly suitable for applications requiring online entropy manipulation [13]. Notice that dropping either summation of Eq. (2) results in essentially the same entropy estimator as Eq. (4). That is, the distribution that maximizes/minimizes one entropy estimator will maximize/minimize the

other [10]. Eq. (4) can also be derived from Eq. (3) by removing the outer summation.

The O(N) entropy estimator of Eq. (4) is not a function of α and it can be derived from Equation (2) for any $\alpha>0$. However, it is of practical importance to know that the only O(N^2) entropy estimator that Eq. (4) approximates well is Eq. (2) with $\alpha=2$, otherwise known as Renyi's quadratic entropy. The reason Eq. (4) approximates Eq. (2) only for $\alpha=2$ is that the inner summation term of Eq. (2) is raised to the power of 1 only for this value of α. Good approximation can be guaranteed if either one of the following two conditions is met:

- Multiple entropy estimates are averaged, where the (time) indices of the data are uniformly randomized for each estimate [11].
- The data is i.i.d. and N is sufficiently large, e.g. $N>1000$ requires only a single estimate [10].

Unlike theoretical entropy, entropy estimators require the selection of the kernel size, σ, in addition to selecting the entropy order, α. For minimizing/maximizing an entropic cost function experience shows that σ should be restricted to be between 0.1 and 2 for unit-variance signals. Fig. 3 shows Renyi's O(N) entropy estimator, given by Eq. (4), for four values of kernel size as a function of the generalized Gaussian parameter, β. The measurements were made using $N=1000$ i.i.d. data samples and they demonstrate that the O(N) entropy estimator produces a good approximation of the theoretical value of Renyi's quadratic entropy when N is sufficiently large and σ is small (relative to the standard deviation of the data). Notice that, up to $\sigma=0.5$, a bias is introduced that is largely independent of β. The fact that it is independent of β implies that it has no effect on minimization or maximization of the entropy estimator. As the kernel size increases further to $\sigma=1$ the bias is no longer independent of β and the shape no longer resembles the theoretical entropy curve for $\alpha=2$. Interestingly, this is beneficial in the context of BSS since it makes the plot of entropy versus β monotonic in the sub-Gaussian region.

Fig. 4 shows the plot of entropy versus β when the entropy estimators of Eq. (2) and (3) are used. Upon close observation of Fig. 4 it can be seen that increasing σ has a tendency to upward bias the estimates for large β. Several important conclusions can be drawn from this. For $\alpha=1$ and $\sigma>0.75$, the plot no longer peaks at $\beta=2$. This implies that ITL algorithms based on Shannon's entropy estimator need to incorporate appropriate sign change(s) for $\sigma>0.75$. For $\alpha=2$ the entropy plot is monotonic in the super-Gaussian region for all four values of σ while the

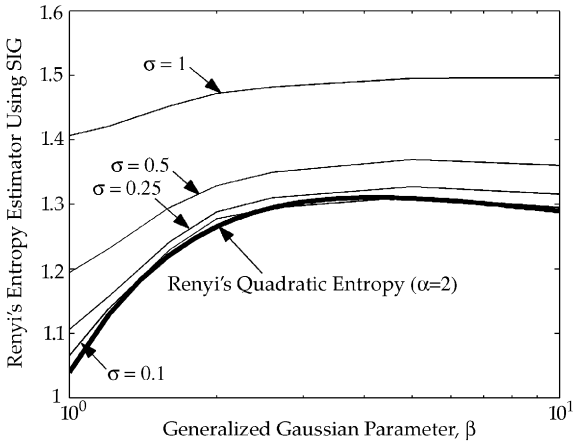

Fig. 3. Renyi's entropy estimator using SIG for four values of kernel size ($\sigma=0.1$, 0.25, 0.5, and 1) versus β. The thick line is the result for Renyi's theoretical entropy.

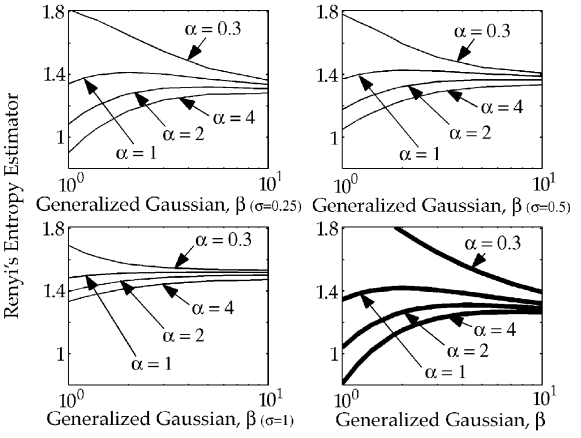

Fig. 4. Renyi's entropy estimators versus β. Upper left, upper right, and lower left subplots are for $\sigma=0.25$, 0.5, and 1, respectively. The lower right subplot shows, for sake of comparison, the results for Renyi's theoretical entropies.

sub-Gaussian region becomes monotonic only for $\sigma > 1$. Therefore, with the appropriate value of σ and as long as the criterion uses the appropriate sign for each marginal entropy, Renyi's quadratic sample-based entropy is suitable as a criterion for BSS. This was not the case for the theoretical entropy with $\alpha = 2$. It is also inferred from Fig. 4 that $\alpha < 0.3$ and $\alpha > 4$ is appropriate for BSS for all interesting values of σ.

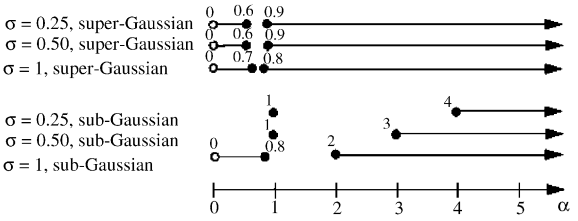

Fig. 5. Monotonicity requirements.

3. Selection of Renyi's entropy order, α, for BSS

To assist with the selection of α the statistical properties of the entropy estimation are quantified. It should be noted that the entropy estimation used in ITL has different objectives than the entropy estimation normally considered in coding or channel capacity. The main difference is that ITL involves the *extremization* of, e.g., an entropy-based criterion. Consequently, the performance does not necessarily depend on how well the entropy is estimated, but on how accurately the coefficients can be found that minimize or maximize entropy. The desire is to select α such that the resulting criterion produces an estimate of the rotation angle(s) that is both unbiased and has small variance. Since asymptotic analyses of the bias and variance do not favor one value of α over another and since a closed-form expression for the case of finite N and non-zero σ is not known to exist, the following evaluation is necessarily heuristic.

3.1. Statistical analysis

In an ITL context one necessary requirement for the choice of α (and σ) is that the curves of Fig. 4, which represent mean values, are monotonic for $1 \leqslant \beta < 2$ and $2 \leqslant \beta$ infinity. The requirements for monotonicity are given in Fig. 5 as a function of α for three of the more interesting values of σ. Aside from monotonicity it is tempting to select α based on Fig. 4 by choosing the value whose curves have the largest slope in both the super-Gaussian and sub-Gaussian regions since this implies the maximum discriminability, hence robustness, with respect to estimation. This would be a valid approach except that the variance of the entropy estimator also varies with β. Therefore, evaluation of different values of α is performed using the following metric, which takes into account both types of information.

$$\rho(\alpha, \sigma, \beta) = \frac{\sqrt{E[\hat{H}_\alpha(X_\beta, \sigma)^2]}}{|E[\hat{H}_\alpha(X_1, \sigma)] - E[\hat{H}_\alpha(X_2, \sigma)]|}.$$

In this equation X_β represents a random variable having a generalized Gaussian distribution with parameter β. The numerator is the standard deviation of the entropy estimate and the denominator, which accounts for the slope of the curve, is the absolute difference in the mean values for $\beta = 1$ and 2. With this definition inferences for $\beta < 2$ are more accurate than for those made for $\beta > 2$. This compromise was made since experience indicates that the largest differences in the performance of the different BSS algorithms occur for super-Gaussian sources. A small value of $\rho(\alpha,\sigma,\beta)$ is indicative of a good estimator.

Fig. 6 shows the normalized standard deviation of the entropy estimate versus β for $\alpha = 1$ and 2, where 100 Monte Carlo trials were used, $N = 1000$ samples, and $\sigma = 0.25$. Also shown, for sake of perspective, is the standard deviation of several moment estimators (defined in the same manner as ρ), which are constructed using the sample mean of the data raised to the appropriate power [14]. Several interesting conclusions may be drawn from this figure. Shannon's entropy estimator is the least desirable of those shown for sub-Gaussian distributions. Renyi's quadratic entropy estimator for super-Gaussian data outperforms the estimator for fourth-order moments, which is commonly used as a criterion for BSS. Also, some methods

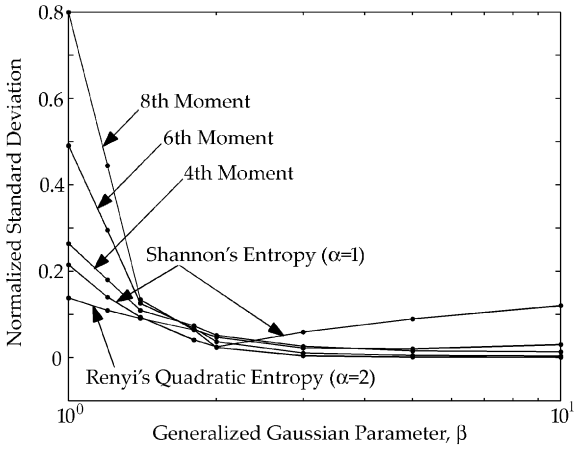

Fig. 6. Normalized standard deviation of several different entropy and moment estimators versus β ($N = 1000$, $\sigma = 0.25$).

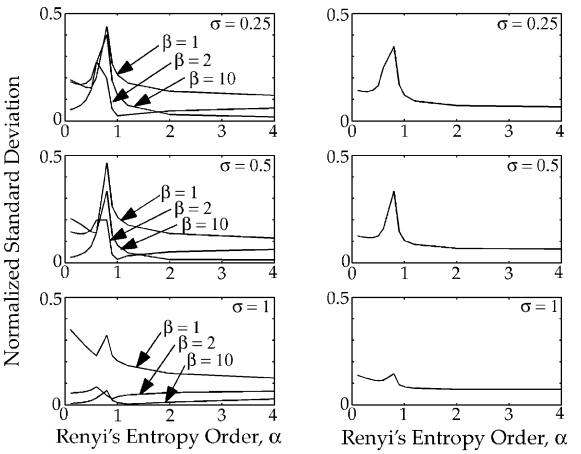

Fig. 7. Normalized standard deviation versus α. The left column shows results for $\beta = 1$, 2, and 10. The right column is the mean of the results for $\beta = 1$, 2, and 10.

are not robust in that they have a low ρ for some values of β and a large ρ for other values of β. For example, the 8th-order moment estimator has the lowest ρ (of the methods considered) for uniformly distributed data, but it also has the highest ρ for Laplacian-distributed data.

A more complete picture of the effect of α on ρ may be found in Fig. 7, which shows results for $\sigma = 0.25$, 0.5 and 1. When the data is super-Gaussian the tendency is for the results to improve as α increases (the peak is due to the mean entropies of Laplacian and Gaussian distributions being similar for $\alpha = 0.73$). This is easily understood since (1) Laplacian data is heavy-tailed and (2) decreasing values of α emphasize the tails of the distribution [15]. This is also the reason that kurtosis-based methods do not perform well for super-Gaussian data [2]. When the data is sub-Gaussian the value of α that produces the smallest ρ decreases from 4 to 1.2 as σ increases from 0.25 to 1. Interestingly, when the data is Gaussian-distributed there is a minimum in ρ at or near $\alpha = 1$, which corresponds to Shannon's entropy. Notice that small values of α perform poorly for super-Gaussian data, but they perform well for sub-Gaussian data (for $\sigma = 1$). The right column of Fig. 7 shows the mean results averaged over the three different source distributions (the conclusions are unchanged if Gaussian-distributed sources are left out of the average). The combination of the mean results from Fig. 7 and the monotonicity requirements shown in Fig. 5 does not yield a single value of α that is noticeably superior to all other values. However, it is clear that one should select $\alpha > 1$. Additionally, one may expect more robust adaptation by choosing $\alpha \geqslant 2$, which excludes the obvious choice of α corresponding to Shannon's entropy.

3.2. *Arguments for selecting* $\alpha = 2$

Since the statistical properties of the estimator for the densities in the exponential family are insufficient to select a single preferred value of α, other characteristics are considered. An obvious choice is to use $\alpha = 1$ since it does not require explicit determination of sub/super-Gaussianity (for small σ) when Parzen estimation is used. However, there are three good arguments for selecting $\alpha = 2$,

- The normalized standard deviation is at or near the minimum value for $\alpha = 2$
- Unlike any other value of α, there exists an entropy estimator for $\alpha = 2$ that reduces the complexity of the entropy estimation from $O(N^2)$ to $O(N)$
- Unlike any other value of α, the $O(N)$ entropy estimator for $\alpha = 2$ allows the criterion to

exploit spectral diversity in addition to spatial diversity

The last two items are a direct consequence of using the SIG approximation to estimate entropy. The second item was discussed in Section 2.2 and the third property stems from the nonlinear transformation of $(y_m(k)-y_m(k-p))^2$, which includes information contained in the autocorrelation at lag p. The proof that a criterion based on Eq. (4) can make use of spectral diversity, as well as the conditions required for the proof, is given in the first author's dissertation [10]. The essential conditions are that the correlations at lag p must be positive, as commonly occurs for natural signals when p is small, and the auto-correlations of the sources at lag p must be distinct, as expected. The ability to make use of spectral diversity is especially useful if the sources are Gaussian-distributed. Spectral information has been used in numerous BSS methods that are based on second-order statistics and has only rarely been discussed for use in information-theoretic methods [16,17]. To exploit spectral diversity it is important that Eq. (4) is used without randomizing the time indices since randomization destroys the temporal structure of the data. This gives an important advantage of the SIG estimator over Eqs. (2) and (3), for any value of α, since their computation involves an average over all possible pair-wise permutations of the time indices and cannot, therefore, make use of spectral diversity.

4. General purpose MRMI and MRMI-SIG algorithms

Presented below are three practical algorithms for BSS. These algorithms are computed directly from samples and are appropriate irrespective of the Gaussianity of the sources. They are found by replacing the theoretical entropies in Eq. (1) with the entropy estimators of Eqs. (2)–(4) and including a change of sign as needed. The determination of sub/super-Gaussianity is estimated using the sign of the kurtosis, for which super-Gaussian sources (generally) have a positive value and sub-Gaussian sources (generally) have a negative value. While this approximation works well in practice, there are known counterexamples, e.g., super-Gaussian sources having zero kurtosis [18]. The first criterion uses the entropy estimator of Eq. (2)

$$\hat{J}_\alpha(Y) = \frac{1}{1-\alpha}\sum_{m=1}^{M}\operatorname{sign}\left(\sum_{k=1}^{N}(y_m^4(k)-3y_m^2(k))\right) \times \log\frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{N}\sum_{k=1}^{N}G(y_m(n)-y_m(k),2\sigma_m^2)\right)^{\alpha-1} \quad (\text{for } \alpha>0,\ \alpha\neq 1) \tag{5}$$

the second uses Eq. (3)

$$\hat{J}_1(Y) = -\frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N}\log\frac{1}{N}\sum_{k=1}^{N}G(y_m(n)-y_m(k),\sigma^2) \quad (\text{for } \alpha=1) \tag{6}$$

and the third employs the SIG approximation of Eq. (4),

$$\hat{J}_2(Y) = -\sum_{m=1}^{M}\operatorname{sign}\left(\sum_{k=1}^{N}(y_m^4(k)-3y_m^2(k))\right) \times \log\frac{1}{N}\sum_{k=1}^{N}G(y_m(k)-y_m(k-p),2\sigma_m^2) \quad (\text{for } \alpha=2), \tag{7}$$

where σ should be chosen in accordance with Fig. 5. These three criteria will be referred to as the (modified) Minimum Renyi's Mutual Information (MRMI), Minimum Shannon Mutual Information (MSMI), and (modified) MRMI-SIG criteria, respectively. Notice that the kurtosis estimation is performed on each output. As a result no a priori information concerning each source kurtosis is required. Also, Eq. (6) does not include a sign-change term as it assumes a small kernel size. If it is desired to use Shannon's entropy with a large kernel size one can always use the criterion of Eq. (5) with α near to 1, e.g. $\alpha = 0.95$ or 1.05.

Whenever σ is small and N is large, the entropy estimators on which these three criteria are built provide a good approximation of their respective theoretical entropy. Hence, these sample-based criteria are appropriate for BSS, with small σ, whenever the respective theoretical entropy is appropriate for BSS. A list of the range of α appropriate for BSS when the

criterion is based on a summation of Renyi's theoretical entropies is given in Section 2.1. The preferred criterion, MRMI-SIG, estimates Renyi's (theoretical) quadratic entropy when σ is small and approximates kurtosis when σ is large, as can easily be shown with a Taylor series approximation [10] (for i.i.d. data or if the time indices are randomized). This is ideal since MRMI and MRMI-SIG are able to take advantage of the improved statistical properties of Renyi's quadratic entropy for super-Gaussian sources as previously shown in Fig. 6 and to benefit from the monotonicity of the kurtosis for sub-Gaussian sources, simply by changing σ. The recommendation is to use a single small value of σ for all outputs that are positively kurtotic and a single large value for outputs that are negatively kurtotic. In addition, if the sources have spectral diversity then MRMI-SIG is able to use this information simply by not randomizing the time indices.

5. Comparisons

This section consists of a detailed comparison of the suggested criterion, MRMI-SIG, with MSMI from Eq. (6), JADE [19], FastICA [20], Comon's MI [21], and Infomax [22]. Two additional methods were also tried, which included an MI method by Pham [23] and one by Yang and Amari [24]. The method by Pham appears to be inappropriate for sub-Gaussian sources and preliminary results from the Yang and Amari method were disappointing, so results from these two methods are not reported here. All the methods under consideration use higher-order statistics. Methods that use only second-order statistics were not included since most of the separation tasks in the comparison are for i.i.d. sources, for which second-order statistics is insufficient. The comparisons assume an off-line implementation, which implies that the data may be re-used for any number of epochs. The performance is measured using the signal-to-interference ratio (SIR), which is given by

$$\mathrm{SIR} = \arg\max_{\boldsymbol{k}} \frac{1}{M}\sum_{m=1}^{M} 10\log 10\left(\frac{P_{k_i}}{P_i - P_{k_i}}\right),$$

where k_i, for $i = 1, 2, \ldots, M$, is an element of $\{1, 2, \ldots, M\}$, k_i not equal to k_j for i not equal to j, P_{k_i} is the power of source k_i in output i, and P_i is the total power of output i. The set of k_i terms, $\boldsymbol{k}$, are determined by assuming a particular permutation of the output signals. Due to the permutation indeterminacy inherent in BSS, the permutation that maximizes the summation above is the one of interest. The SIR is a measure of mean separation performance across channels where larger values represent better performance and values above 20 dB correspond to inaudible interference when audio sources are used. A total of ten Monte Carlo runs are performed for each separation task. Each of the Monte Carlo runs uses a different mixing matrix, whose entries are chosen uniformly in $[-1, +1]$.

In order to take advantage of any spectral diversity, randomization is not used for MRMI-SIG whenever the outputs are such that the mean (across channels) of the normalized correlation coefficient at lag p exceeds 0.4, a value which was experimentally determined. The kernel sizes for MRMI-SIG are chosen to be 0.25 and 1 for positively and negatively kurtotic outputs, respectively. While it is possible to fine tune the kernel sizes in order to avoid local minima [15], this was not done. The kernel size for MSMI is chosen to be 0.5 in order that the maximum entropy pdf remains the Gaussian and, due to the $O(N^2)$ complexity, MSMI uses a maximum of 500 randomly selected data points. For all synthetically created i.i.d. data the nonlinearities of the Infomax algorithm were selected to be the cumulative distribution functions (cdf's) of the sources, in which case the Infomax algorithm becomes a maximum-likelihood method [25]. This prevents the need to adapt the nonlinearities and represents a best-case scenario for the Infomax algorithm since knowledge of the source distributions is not normally available in the context of BSS. Some results are also included for speech data, for which a sigmoid nonlinearity is a decent approximation of the cdf. In this latter case the sigmoid nonlinearity is used. In all cases the tap weight update for Infomax uses the natural gradient [26], which is also known as the relative gradient [27].

The first separation task is to separate $M = 5$ sources for different combinations of β and N. Six different exponentially increasing values of β are

used. These values are 1, 1.2, 1.7, 2.7, 5, and 10 (for each test, all five sources have the same β). An exponential increase was used since $\beta = 2$ is the logical choice for the midpoint. In addition, seven different values were used for the data length. These values are 100, 200, 500, 1000, 2000, 5000, and 10,000. This made a total of 42 different combinations for each method. The samples for this task are drawn in an i.i.d. fashion so that the data is stationary and temporally independent.

Fig. 8 shows the results for each method averaged over ten Monte Carlo trials. In this figure each subplot represents a different value of β. Aside from some initial differences in data efficiency (i.e. for small values of N) notice that the different methods perform almost identically as the distribution of the sources become increasingly uniform.

Fig. 9 shows the results averaged over β. This figure indicates that MRMI-SIG is the most data-efficient method. Despite the exponentially greater computational complexity and implicit determination of sub/super-Gaussianity, MSMI performs worse than both MRMI-SIG and JADE at all values of N. The performance for this method is flat above $N = 500$ due to the imposed data-length restriction. Interestingly, Infomax performs worse than both MRMI-SIG and JADE at all values of

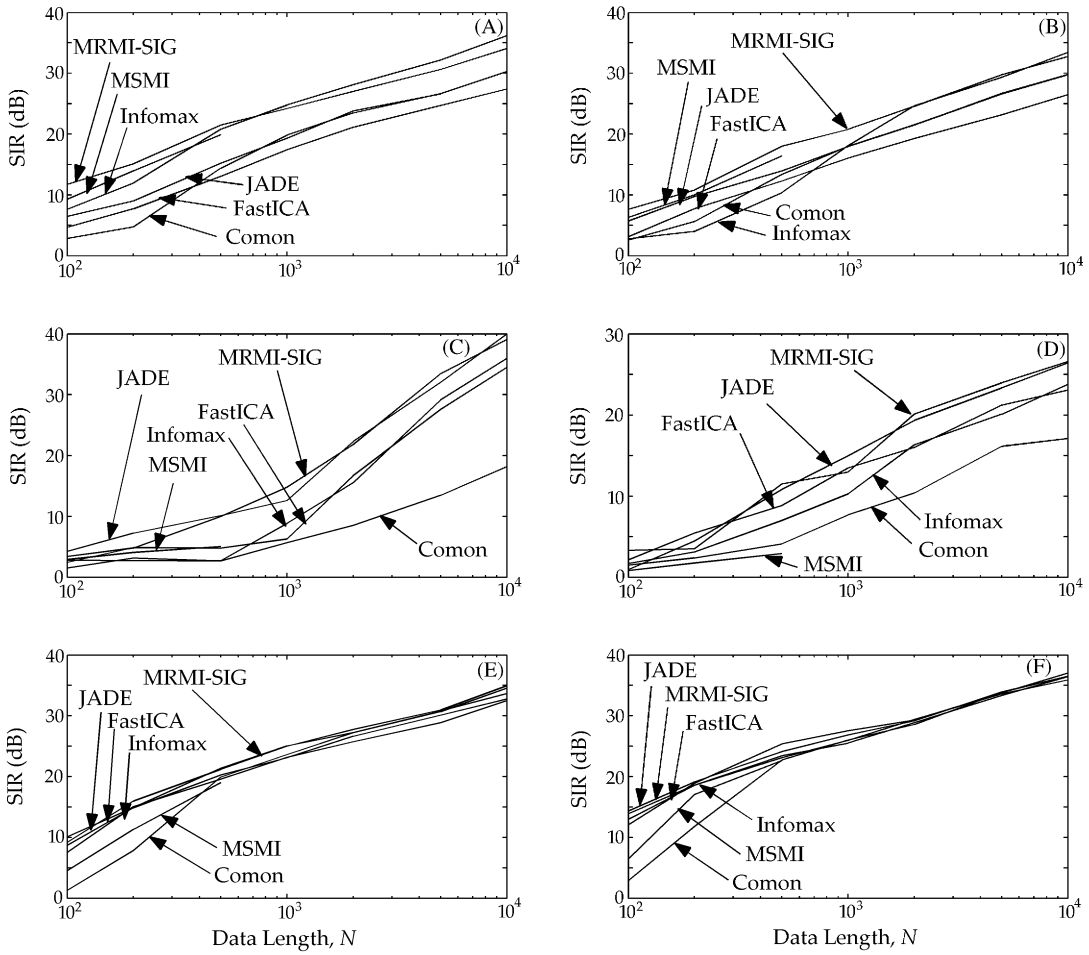

Fig. 8. SIR versus N for the competing BSS methods for i.i.d. sources. (A) $\beta = 1$, (B) $\beta = 1.2$, (C) $\beta = 1.7$, (D) $\beta = 2.7$, (E) $\beta = 5$, (F) $\beta = 10$.

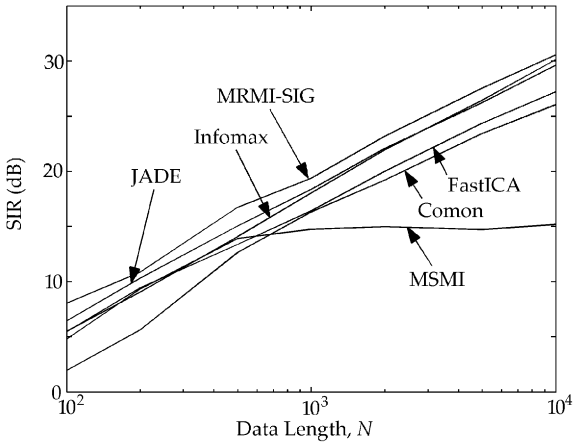

Fig. 9. Mean SIR as a function of N, averaged over $\beta = 1$, 1.2, 1.7, 2.7, 5, and 10, for i.i.d. sources.

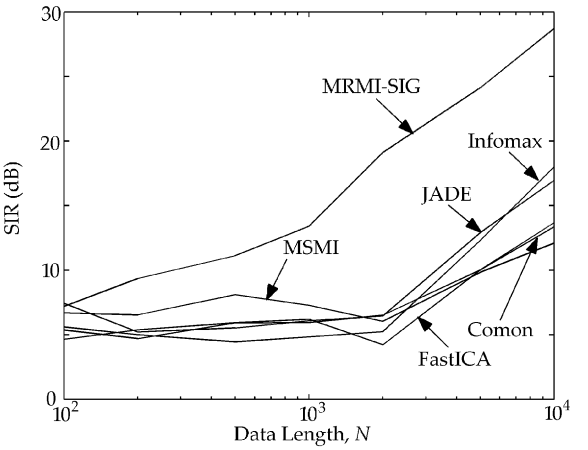

Fig. 10. SIR as a function of N data samples, for $M = 5$ audio sources.

N even though knowledge of the true source cdfs was used to construct the nonlinearities. FastICA performs the worst when there is little data and Comon's MI method performs the worst for $N > 1000$ (neglecting MSMI).

The next two separation tasks use artificially mixed audio sources. There are a total of 50 sources, of which 24 were speech (approximately Laplacian-distributed) and 26 were music (most of which were slightly super-Gaussian). Each Monte Carlo trial uses M randomly selected sources. One task is to separate $M = 5$ sources as a function of N, while the second task varied M and used a constant data length of $N = 10,000$. These results are shown in Figs. 10 and 11, respectively. Notice that the performance of all the methods are reduced from that of the i.i.d. sources due to the reduction of available statistical information caused by the time-correlation of the audio sources. However, MRMI-SIG is reduced much less than the others. Unlike before, MSMI is able to improve as N increases above 500, as shown in Fig. 10, since the 500 randomly selected points become less likely to be temporally correlated as N increases. Infomax performs quite well in this case even though a sigmoid nonlinearity is used, which is not perfectly tuned to the cdf of the sources. In fact, it surpasses the performance of JADE which had outperformed it for i.i.d. data. Fig. 10 shows that MSMI is better than all methods except

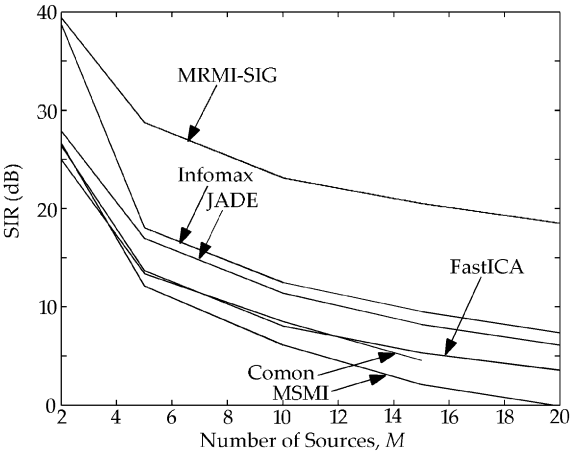

Fig. 11. SIR as a function of M audio sources, for $N = 10,000$ data samples.

MRMI-SIG for $N < 500$. Notice that the rank of performance is consistent with the findings of the statistical properties of the entropy and moment estimators for super-Gaussian distributions, as shown previously in Fig. 6. The performance advantage of MSMI is not seen in Fig. 11 because of the restriction on the amount of data used. Recall that MSMI is limited to 500 data points while all other methods use more than a magnitude of order more data. Also, it appears that the performance for all the methods, except

Table 1
Adaptation time (in relative magnitudes of order)

	$M=5$ $N=100$	$M=5$ $N=1000$	$M=20$ $N=10{,}000$
MRMI-SIG	3	4	7
JADE	1	1	2
Infomax	3	4	6
Comon's MI	3	4	8
FastICA	1	1	3
MSMI	4	5 ($N=500$)	6 ($N=500$)

MRMI-SIG, is flat up to and including $N=2000$, with the slope across all methods basically identical for $N>2000$. The better performance of MRMI-SIG for $N<2000$ is attributed to the extraction of temporal dependencies. In Fig. 11 the data point corresponding to $M=20$ for Comon's MI method is unavailable because the time of adaptation for large M, an order of magnitude longer than the other gradient-based methods, became unbearably long.

This comparison was for an off-line BSS implementation, therefore the time of adaptation is considered unimportant (except in extreme cases). However, the amount of time required for each is listed in Table 1. Keep in mind that, had the comparison fixed the adaptation time, the gradient-based algorithms would have traded performance for time. Nevertheless, it is quite impressive that JADE and FastICA perform as well as they do, and yet require very little training time as compared to the other algorithms.

6. Conclusions

This paper presents a detailed study on the use of Renyi's entropy for blind source separation. In this context Renyi's entropy has very different properties than Shannon's entropy. The fundamental difficulty of Renyi's (theoretical) entropy is that it peaks for the Gaussian distribution only when $\alpha=1$. Consequently, this is the only value of α appropriate for Eq. (1) (which ignores the sub/super-Gaussianity of the sources). This paper presents a method to counteract this limitation for sample-based entropy estimators by taking advantage of the combined effect of Renyi's entropy parameter, α, and the kernel size of the Parzen window estimator, σ. It should be mentioned that the arguments in this paper are only made for exponential distributions and cannot be guaranteed to generalize to other source distributions. However, extensive experience with these methods, as well as the results of the audio mixtures, indicates the validity of this approach.

The findings suggest that the previously published MRMI-SIG criterion [11] should be modified to (1) use a large kernel size for sub-Gaussian sources, (2) select the sign of each marginal entropy in the sum based on the kurtosis of the associated source estimate, and (3) refrain from randomizing the time indices when the sources are highly temporally correlated. Likewise, the MRMI criterion [9] should implement the first two of the three changes above (the third is not applicable). The need for these changes passed unnoticed in the paper by Erdogmus et al. [15] when optimizing α because the changes are not needed for super-Gaussian sources and, even without the changes as Fig. 2 shows for $\beta=10$, there is roughly a 50% probability of obtaining the global minimum for sub-Gaussian sources. While Parzen windows may be applied to create a non-parametric BSS algorithm for any positive value of α, three reasons are given why $\alpha=2$ (corresponding to MRMI-SIG) is preferred over all other values of α including $\alpha=1$, which corresponds to Shannon's entropy. They are as follows: (1) nearly minimal normalized standard deviation, (2) the ability to exploit spectral diversity, and (3) exponentially reduced computational complexity.

Acknowledgments

This work was partially supported by NSF ECS #0300340.

References

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

[2] A. Hyvarinen, Survey on independent component analysis, Neural Comput. Surv. 2, (1999) 94–128, http://www.cse.ucsc.edu/NCS/.
[3] K. Hild II, et al., Blind source separation using Renyi's mutual information, IEEE Signal Processing Lett. 8(6) (June 2001) 174–176.
[4] A. Renyi, Probability Theory, North-Holland Publishing Company, Amsterdam, Netherlands, 1970 (Chapter 9).
[5] E. Parzen, On estimation of a probability function and mode, Ann. Math. Statist. 33 (3) (1962) 1065–1076.
[6] T. Cover, J. Thomas, Elements of Information Theory, Wiley, New York, NY, 1991, pp. 20–21 (Chapter 2).
[7] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. (July 1948) 379–423; (October 1948) 623–656.
[8] J.C. Principe, et al., Learning from examples with information theoretic criteria, J. VLSI Signal Processing Systems 26(1/2) (August 2000) 61–77.
[9] D. Erdogmus, et al., Blind source separation using Renyi's alpha-marginal entropies, Neurocomputing 49(1) (December 2002) 25–38.
[10] K. Hild II, Blind separation of convolutive mixtures using Renyi's divergence, Ph.D. Dissertation, University of Florida, 2003, pp. 3–39, 148–159.
[11] D. Erdogmus et al., On-line entropy manipulation: stochastic information gradient, IEEE Signal Processing Lett. 10(8) (August 2003) 242–245.
[12] P. Viola, et al., Empirical entropy manipulation for real-world problems, Advances in Neural Information Processing Systems, Denver, CO, 27–30 November 1995, pp. 851–857.
[13] K. Hild II, et al., Blind source separation of time-varying, instantaneous mixtures using an on-line algorithm, International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, 13–17 May 2002, pp. 993–996.
[14] C. Bourin, P. Bondon, Efficiency of high-order moment estimators, IEEE Trans. Signal Process. 46(1) (January 1998) 255–258.
[15] D. Erdogmus, J.C. Principe, Generalized information potential criterion for adaptive system training, IEEE Trans. Neural Networks 13(5) (September 2002) 1035–1044.
[16] D. Pham, Contrast functions for blind separation and deconvolution of sources, International Workshop on Independent Component Analysis and Signal Separation, San Diego, CA, 9–12 December 2001, pp. 37–42
[17] B.A. Pearlmutter, L.C. Parra, Maximum likelihood blind source separation: a context-sensitive generalization of ICA, Advances in Neural Information Processing Systems, vol. 9, MIT Press, Cambridge, MA, December 1996, pp. 613–619.
[18] C.D. Giurcaneanu, I. Tabus, On the sign of kurtosis, International Conference on Acoustics, Speech, and Signal Processing, Helsinki, Finland, 19–22 June 2000, pp. 499–502.
[19] J.F. Cardoso, A. Souloumiac, Blind beamforming for non-Gaussian signals, Radar Signal Processing IEE Proc F 140(6) (December 1993) 362–370.
[20] A. Hyvarinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE Trans Neural Networks 10(3) (May 1999) 626–634.
[21] P. Comon, Independent component analysis, a new concept?, Signal Processing 36(3) (April 1994) 287–314.
[22] A. Bell and T. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7(6) (November 1995) 1129–1159.
[23] D. Pham, Fast algorithm for estimating mutual information, entropies, and score functions, International Workshop on Independent Component Analysis and Signal Separation, Nara, Japan, 1–4 April 2003, pp. 17–22.
[24] H. Yang, S. Amari, Adaptive online learning algorithms for blind separation: maximum entropy and minimum mutual information, Neural Comput. 9(7) (October 1997) 1457–1482.
[25] J.F. Cardoso, Infomax and maximum likelihood for blind source separation, IEEE Signal Processing Lett. 4(4) (April 1997) 112–114.
[26] S. Amari, "Neural learning in structured parameter spaces—natural Riemannian gradient", Advances in Neural Information Proc. Systems, Denver, CO, 2–5 December 1996, pp. 127–133.
[27] J.F. Cardoso, B. Laheld, Equivariant adaptive source separation, IEEE Trans. Signal Processing 44(12) (December 1996) 3017–3030.