Blind Source Separation Using Renyi's Mutual Information

Kenneth E. Hild, II, Deniz Erdogmus, and José Príncipe

***Abstract*—A blind source separation algorithm is proposed that is based on minimizing Renyi's mutual information by means of nonparametric probability density function (PDF) estimation. The two-stqge process consists of spatial whitening and a series of Givens rotations and produces a cost function consisting only of marginal entropies. This formulation avoids the problems of PDF inaccuracy due to truncation of series expansion and the estimation of joint PDFs in high-dimensional spaces given the typical paucity of data. Simulations illustrate the superior efficiency, in terms of data length, of the proposed method compared to fast independent component analysis (FastICA), Comon's minimum mutual information, and Bell and Sejnowski's Infomax.**

***Index Terms*—Blind source separation, component analysis, Givens rotations, mutual information, Renyi's entropy, Renyi's information.**

I. Introduction

MINIMIZATION of the mutual information (MMI) between outputs is considered the ideal information theoretic criteria for blind source separation (BSS) [1]–[3]. The (Shannon) mutual information can be written as the sum of the (Shannon) marginal entropies minus the joint entropy. One of the difficulties of using 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 (PDFs) with truncated polynomial expansions [2]–[4], a process that inherently introduces error in the estimation procedure. Another MMI method proposed by Xu *et al.* [5] avoids the polynomial expansion by approximating the Kullback–Leibler divergence with the Cauchy–Schwartz inequality and estimates Renyi's entropy nonparametrically by Parzen windows. The second method requires estimation of the N-dimensional joint entropy, and nonparametric PDF estimation using Parzen windows is ill-posed in high-diemsional spaces [5]. We propose below a new algorithm based on Renyi's mutual information that requires only marginal entropy estimation and avoids both polynomial expansions and estimation of Renyi's joint entropy.

II. Cost Function

Renyi's mutual information is defined as [6]

$$I_{R_a}(y)=\frac{1}{\alpha-1}\log\int_{-\infty}^{\infty}\frac{f_Y(y)^\alpha}{\prod_{i=1}^{N}f_{Y_i}(y_i)^{\alpha-1}}dy. \tag{1}$$

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

$$\sum_{i=1}^{N}H_{R_a}(y_i)-H_{R_a}(y)=\frac{1}{\alpha-1}\log\frac{\int_{-\infty}^{\infty}f_Y(y)^\alpha\,dy}{\int_{-\infty}^{\infty}\prod_{i=1}^{N}f_{Y_i}(y_i)^\alpha\,dy_i} \tag{2}$$

which differs from (1). However, both (1) and (2) are nonnegative [6], and both evaluate to zero when and only when the joint PDF can be written as a product of the marginals, which occurs identically when the statistically independent sources are separated. Therefore, minimizing Renyi's MI can be accomplished by, minimizing the sum of (Renyi's) marginal entropies minus the joint entropy, just as with Shannon's entropy. Notice that (2) is not a reformulation of MI, but is preferred since Renyi's entropy can be estimated nonparametrically as proposed in [9].

In order to produce an algorithm that scales favorably as the number of sources increases, the estimation of the joint entropy is avoided by the two-step parameterization proposed in [4]. The first stage performs spatial whitening, and the second stage performs a rotation in N-dimensions. The rotation, denoted as matrix R, is adapted to minimize the cost function given by (2). In fact, since Renyi's joint entropy is invariant to rotations

$$\begin{aligned}H_{R_a}(y)&=\frac{1}{1-\alpha}\log\int_{-\infty}^{\infty}f_Y(y)\,dy\\&=\frac{1}{1-\alpha}\log\int_{-\infty}^{\infty}\frac{f_X(x)^\alpha|\det R|}{|\det R|^\alpha}dx=H_{R_a}(x)\end{aligned} \tag{3}$$

the joint entropy may be discarded from the adaptation process, leaving only the marginal entropies. Hence, the cost function becomes simply

$$J=\sum_{i=1}^{N}H_{R_a}(y_i) \tag{4}$$

which mimics the cost function of [3], [4], but with Renyi's entropy substituted for Shannon's entropy. Using the nonparametric Renyi's entropy estimator proposed in [9], we arrive at a new algorithm for BSS.

Manuscript received November 20, 2000. This work was supported in part by National Science Foundation Contract ECS 9900394. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. G. Ramponi.

The authors are with Computational NeuroEngineering Laboratory, Electrical and Computer Engineering Department, University of Florida, Gainesville, FL 32611 USA.

Publisher Item Identifier S 1070-9908(01)03293-X.

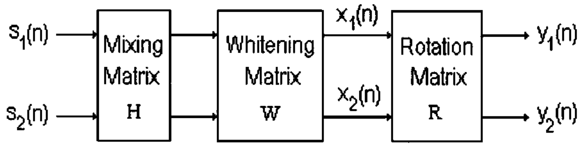

Fig. 1. System block diagram for two sources/observations.

III. System Description

The overall BSS block diagram for two inputs/observations is given in Fig. 1. The equations for a system having N sources/observations are given by $\mathbf{x} = \mathbf{W}^T\mathbf{H}^T\mathbf{s}$ and $\mathbf{y} = \mathbf{R}^T\mathbf{x}$, where $\mathbf{W} = \Phi\Lambda^{-1/2}$, Φ is the matrix of eigenvectors of the autocorrelation of $\mathbf{H}^T\mathbf{s}$, and Λ is the corresponding eigenvalue matrix. Notice that $E[\mathbf{x}\mathbf{x}^T] = \mathbf{I}_N$ is the $(N \times N)$ identity matrix, due to the spatial whitening. The rotation matrix $\mathbf{R}$ is constructed from the product of $N(N-1)/2$. Givens rotation matrices[1] are $\mathbf{R}_{ij}$, where $\mathbf{R}_{ij}$ equals $\mathbf{I}_N$ with elements $\mathbf{I}_N(i,i)$, $\mathbf{I}_N(i,j)$, $\mathbf{I}_N(j,i)$, and $\mathbf{I}_N(j,j)$ modified to $\cos\theta_{ij}$, $\sin\theta_{ij}$, $-\sin\theta_{ij}$, and $\cos\theta_{ij}$, respectively, where $\mathbf{I}_N(i,j)$ is the element of $\mathbf{I}_N$ located at the ith row and jth column. There is one rotation angle for each $\mathbf{R}_{ij}$, whose purpose is to perform a rotation in the output space in the plane specified by the ith and jth (orthogonal) basis vectors. The gradient of $\mathbf{R}$ with respect to θ_{ij} is also needed for the algorithm development and it is denoted as $\nabla\mathbf{R}_{ij}$.

IV. Stochastic Gradient Descent Algorithm

When Parzen windowing is used with a Gaussian kernel, the estimate for Renyi's quadratic marginal entropy simplifies to [9]

$$H_{R_2}(y_i) = -\log\frac{1}{L^2}\sum_{j=1}^{L}\sum_{k=1}^{L}G(y_i(j)-y_i(k),2\sigma^2) \quad (5)$$

where $G(x,2\sigma^2)$ is a Gaussian PDF, and $y_i(j)$ is the jth sample of output y_i. Notice that the infinite-limit integral disappears and that there are no approximations involved (aside from the implicit PDF estimation using Parzen windows). Substituting (5) into (4) and taking the derivative with respect to θ_{ij} produces (6), shown at the bottom of the page, where $(\nabla_{ij}\mathbf{R})_i$ is the ith column of $\nabla_{ij}\mathbf{R}$, and $\mathbf{x}(m)$ is the vector of $\mathbf{x}$ at time m. The overall update equation for stochastic gradient descent is then $\theta(n+1) = \theta(n) - \eta\Delta\theta(n)$, where $\theta(n)$ and $\Delta\theta(n)$ are vectors of angles, and η is the step size.

V. Results

The proposed method is compared to FastICA [7], Bell and Sejnowski's Infomax [8], and Comon's MMI method [4] using an instantaneous mixture of ten audio sources consisting of one music source and speech from five male and four female speakers. Spatial prewhitening is used for each method, the values of the mixing coefficients are chosen uniformly in $[-1, 1]$ and the performance criterion is the signal-to-distortion ratio, defined as

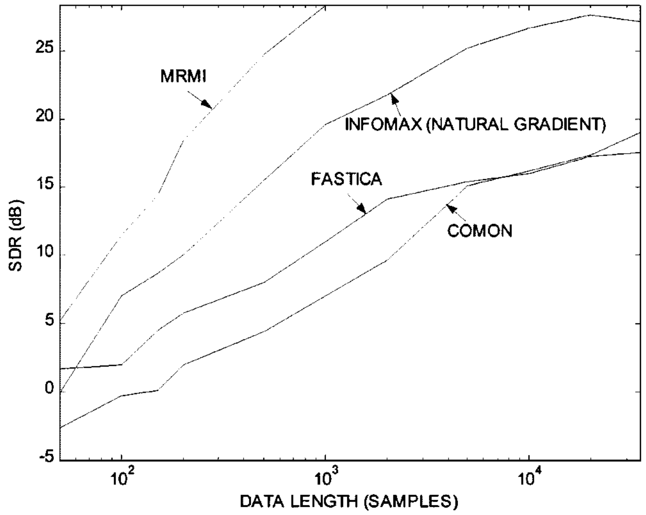

Fig. 2. Signal-to-distortion plots for ten sources/observations. Notice that the x-axis is logarithmic.

$$\text{SDR} = \frac{1}{N}\sum_{i=1}^{N}10\log_{10}\left(\frac{(\max q_i)^2}{q_i q_i^T - (\max q_i)^2}\right) \quad (7)$$

where $\mathbf{q} = \mathbf{R}^T\mathbf{W}^T\mathbf{H}^T$, $\mathbf{q}_i$ is the ith row of $\mathbf{q}$, and $\max(\mathbf{q}_i)$ equals the maximum of the argument. This criterion effectively measures the distance of $\mathbf{q}$ from the product of a permutation matrix with a diagonal matrix.

Fig. 2 shows the SDR for each method as a function of L, the data length in number of samples. FastICA uses the symmetric approach and the cubic nonlinearity, Infomax uses Amari's natural gradient [11], Comon's method uses a fourth-order PDF expansion (requiring estimates of third and fourth-order cumulants) and our algorithm, MRMI, uses a kernel size σ^2 of 0.5 ($L \leq$ 100) or 0.25 ($L >$ 100). In addition, small step sizes were used to maximize SDR. As can be seen from the figure, the MRMI method requires much less data to yield a given performance level (good separation is achieved at an SDR of 20 dB). This is imperative in a number of applications such as channel equalization and in nonstationary environments. In hearing aid applications for example, Torkkola explains [10] that "Any system which attempts to ... apply some means of inverse filtering would have to be adaptable on almost a frame-by-frame basis to be effective." The results for MRMI are not shown beyond L = 1000 due to the computational

$$\Delta\theta_{ij} = \frac{\partial}{\partial\theta_{ij}}\sum_{k=1}^{N}H_{R_2}(y_k) = -\sum_{k=1}^{N}\frac{\sum_{m=1}^{L}\sum_{n=1}^{L}G(y_k(m)-y_k(n),2\sigma^2)(y_k(m)-y_k(n))(\nabla_{ij}R)_i^T(x(m)-x(n))}{\sum_{m=1}^{L}\sum_{n=1}^{L}G(y_k(m)-y_k(n),2\sigma^2)} \quad (6)$$

[1] This is the number of unique pairwise combinations of N orthogonal basis vectors.

complexity, which is $O(L^2)$ compared to $O(L)$ for the other three methods.

VI. Conclusions

A new BSS algorithm has been developed that is based on the minimization of Renyi's mutual information. Unlike the Comon approach, which uses Shannon's entropy and requires truncation of a PDF series expansion, there are no approximations in the proposed method due to the utilization of Renyi's quadratic entropy. While this algorithm is computationally more complex than its predecessors, it has been shown to be much more efficient in terms of the amount of data samples required, an item that is paramount in tracking a rapidly changing environment.

References

[1] J. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, 1998.

[2] A. Hyvarinen, "Survey on independent component analysis," *Neural Comput. Surv.*, pp. 94–128, 1999.

[3] H. Yang and S. Amari, "Adaptive online learning algorithms for blind separation: Maximum entropy and minimum mutual information," *Neural Comput.*, vol. 9, pp. 1457–1482, 1997.

[4] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, pp. 287–314, 1994.

[5] D. Xu, J. Principe, J. Fisher, and H. Wu, "A novel measure for independent component analysis (ICA)," in *Proc. Int. Circuits Syst. Symp.'98*, vol. II, pp. 1161–1164.

[6] A. Rényi, *Probability Theory*. Amsterdam, The Netherlands: North-Holland, 1970.

[7] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks*, vol. 10, pp. 626–634, 1999.

[8] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.

[9] J. Principe, D. Xu, and J. Fisher, "Information theoretic learning," in *Unsupervised Adaptive Filtering*, S. Haykin, Ed. New York: Wiley, 2000, pp. 265–319.

[10] K. Torkkola, "Blind separation for audio signals—Are we there yet?," in *Proc. 1st Int. Workshop Independent Component Analysis and Signal Separation'99*, Aussois, France, Jan. 1999, pp. 239–244.

[11] S. Amari, "Neural learning in structured parameter spaces—Natural Riemannian gradient," in *Proc. Neural Information Processing Systems*. MIT, Cambridge, MA, 1997, pp. 127–133.