INDEPENDENT COMPONENT ANALYSIS USING JAYNES' MAXIMUM ENTROPY PRINCIPLE

Deniz Erdogmus, Kenneth E. Hild II, Yadunandana N. Rao, Jose C. Principe

CNEL, ECE Department, University of Florida, Gainesville, FL 32611, USA

ABSTRACT

ICA deals with finding linear projections in the input space along which the data shows most independence. Therefore, mutual information between the projected outputs, which are usually called the separated outputs due to links with blind source separation (BSS), is considered to be a natural criterion for ICA. Minimization of the mutual information requires primarily the estimation of this quantity from the samples, and then adaptation of the separation matrix parameters using a suitable optimization approach. In this paper, we present a numerical procedure to estimate an upper bound for the mutual information based on density estimates motivated by Jaynes' maximum entropy principle. The gradient of the mutual information with respect to the adaptive parameters, then turns out to be extremely simple.

1. INTRODUCTION

Independent components analysis (ICA) is the problem of finding directions in the data space such that the independence of the projections is maximized [1]. It is a special case of the more general problem of blind source separation (BSS), which deals with obtaining estimates of the unknown source signals using a number of mixed observations, where the mixing process is also unknown [2]. In order to be able to solve this problem, some assumptions regarding the statistical properties of the sources must be made. One commonly used assumption is the independence of the original sources. In the instantaneous linear mixture case of BSS, which is identical to the ICA problem, the relationship between the observation vector **x** and the source vector **s** is

$$\mathbf{x} = \mathbf{H}\mathbf{s} \tag{1}$$

In this expression, **H** represents the unknown mixing matrix. Limiting ourselves to the square mixture case where the number of observations equals the number of sources, the square matrix **H** is assumed to be invertible. The separation is achieved by finding a matrix **A** that optimizes some criterion that measures the independence of the outputs, which are given by $\mathbf{y}=\mathbf{A}\mathbf{x}$.

A natural criterion for measuring independence is Shannon's mutual information [3]. For *n* random variables $Y_1,\ldots,Y_n$, whose joint pdf is $f_{\mathbf{Y}}(\mathbf{y})$ and marginal pdfs are $f_1(y^1),\ldots, f_n(y^n)$, respectively, this mutual information is defined as follows [4].

$$I(\mathbf{Y}) = \int_{-\infty}^{\infty} f_{\mathbf{Y}}(\mathbf{y}) \log\left(f_{\mathbf{Y}}(\mathbf{y}) \Big/ \prod_{o=1}^{n} f_o(y^o) \right) d\mathbf{y} \tag{2}$$

where $\mathbf{y}=[y^1,\ldots,y^n]^T$. Equivalently, Shannon's mutual information can be written in terms of the marginal and joint entropies [4].

$$I(\mathbf{Y}) = \sum_{o=1}^{n} H(Y_o) - H(\mathbf{Y}) \tag{3}$$

where Shannon's entropy is defined as

$$H(Y_o) = -\int_{-\infty}^{\infty} f_o(y^o) \log f_o(y^o) dy^o \tag{4}$$

$$H(\mathbf{Y}) = -\int_{-\infty}^{\infty} f_{\mathbf{Y}}(\mathbf{y}) \log f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} \tag{5}$$

for marginal and joint entropies, respectively. Although minimization of the output mutual information is considered to be a natural criterion, due to difficulties associated with estimating it in a robust manner, two of the most popular ICA algorithms, namely Infomax [5] and Fast ICA [6] use other criteria. Specifically, Infomax maximizes the joint output entropy of the nonlinearly transformed data and Fast ICA uses fourth-order statistics.

The difficulty in using information theoretic measures lies in estimating the density of the underlying data. Successful algorithms must use robust probability density function (pdf) estimators in order to improve convergence to asymptotic performance and to minimize sensitivity to outliers. A common approach is polynomial expansions [7,8,9]. Truncation of these expansions introduce inherent errors in the estimation of the information theoretic criteria. Alternative density estimation approaches include Parzen windowing [10], and orthonormal basis functions [11]. Kernel estimates are used by several researchers in ICA [12,13,14].

In this paper, we will undertake the minimum mutual information approach in training the whitening-rotation

topology as previously done by many other researchers [7,8,12]. The estimate of mutual information, however, will be based on maximum-entropy density estimates obtained in accordance with Jaynes' maximum entropy principle [15]. This principle basically states that the probability density that best fits the already available experimental data, yet that makes minimal commitment regarding any *unseen* measurements should be adopted. Therefore, the data density is obtained by solving a constrained entropy maximization problem, where the constraints guarantee consistency with available data and the maximization of entropy (uncertainty) represents the minimization of commitment to unseen data.

In this approach, since the density estimates will be based on the maximum entropy principle, the estimated value of the criterion will constitute an upper bound for its actual value. Due to this minimization of an upper bound (the maximum value) the presented algorithm is, in a sense, a *minimax* approach. Therefore, this criterion and the associated learning algorithm will be referred to as Minimax ICA. We expect Minimax ICA to exhibit the robustness properties of maximum entropy methods.

2. THE TOPOLOGY AND THE OBJECTIVE

The algorithm will use the common whitening-rotation scheme in which the separation is performed in two steps. The whitening matrix generates unit-variance uncorrelated signals, which are then transformed by a coordinate rotation in order to maximize independence. Assuming the measurements are (made) zero-mean, the whitening matrix is determined from the eigendecomposition of the measurement covariance matrix. Specifically, $\mathbf{W}=\mathbf{\Lambda}^{-1/2}\mathbf{\Phi}^T$, where $\mathbf{\Lambda}$ is the diagonal eigenvalue matrix and $\mathbf{\Phi}$ is the corresponding orthonormal eigenvector matrix of the measurement covariance matrix $\mathbf{\Sigma}=E[\mathbf{x}\mathbf{x}^T]$. The whitened signals are obtained by $\mathbf{z}=\mathbf{W}\mathbf{x}$. The coordinate rotation is achieved by optimizing an orthonormal matrix $\mathbf{R}$, which is parameterized using Givens rotations [16], using the mutual information criterion.

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

The Givens parameterization for a rotation matrix involves the multiplication of in-plane rotation matrices as shown in (6). Here, the vector $\mathbf{\Theta}$ contains the Givens angles θ_{ij}, i=1,…,n-1 and j=i+1,…,n. Each of the matrices $\mathbf{R}^{ij}(\theta_{ij})$ are the so-called in-plane rotations and are given by an nxn identity matrix whose four entries are modified as follows: $(i,i)^{\text{th}}$, $(i,j)^{\text{th}}$, $(j,i)^{\text{th}}$, and $(j,j)^{\text{th}}$ entries are modified to read $\cos\theta_{ij}$, $-\sin\theta_{ij}$, $\sin\theta_{ij}$, and $\cos\theta_{ij}$, respectively. There are a total of $n(n-1)/2$ Givens angles to be optimized.

Considering the mutual information criterion given in (2), and noticing that for outputs obtained from $\mathbf{y}=\mathbf{R}\mathbf{z}$, the joint entropy remains constant for changing rotations, the criterion reduces to the sum of marginal output entropies.

$$J(\mathbf{\Theta}) = \sum_{o=1}^{n} H(Y_o) \tag{7}$$

The elimination of the requirement for estimating the joint entropy is a very significant gain in terms of algorithmic robustness, since the estimation of high-dimensional densities requires an exponentially increasing number of samples.

3. THE MAXIMUM ENTROPY PROBLEM

Given a set of samples $\{x_1,\ldots,x_N\}$, the maximum entropy density estimate for X can be obtained by solving the following constrained optimization problem.

$$\max_{p_X(x)} H = -\int_{-\infty}^{\infty} p_X(x)\log p_X(x)dx$$
$$\text{subject to } E[f_k(X)] = \alpha_k \quad k = 1,\ldots,m \tag{8}$$

where the constraint functions $f_k(.)$ are selected *a priori* by the designer and the constants α_k are calculated using the given samples with sample mean approximation.

Using calculus of variations and the Lagrange multipliers method, the solution to this constrained maximization problem is determined as

$$p_X(x) = C(\boldsymbol{\lambda})\exp\left(\sum_{k=1}^{m}\lambda_k f_k(x)\right) \tag{9}$$

where $\boldsymbol{\lambda} = [\lambda_1 \ \ldots \ \lambda_m]^T$ is the Lagrange multiplier vector and $C(\boldsymbol{\lambda})$ is the normalization constant. The Lagrange multipliers need to be solved simultaneously from the constraints. In the continuous random variable case, however, this is not an easy task, since these equations involve expectation integrals. Using integration by parts, and using the assumption that the actual distribution is close to the maximum entropy distribution, it is possible to derive a much simpler formula to solve for the Lagrange multipliers. Consider

$$\alpha_i = E[f_i(X)] = \int_{-\infty}^{\infty} f_i(x)p_X(x)dx \tag{10}$$

Applying integration by parts with the following definitions,

$$u = p_X(x) \qquad v = \int f_i(x)dx \overset{\Delta}{=} F_i(x)$$
$$du = \left(\sum_{k=1}^{m}\lambda_k f_k'(x)\right)p_X(x) \quad dv = f_i(x)dx \tag{11}$$

where $f_k'(.)$ denotes the derivative of the constraint function, and $F_k(.)$ denotes its integral, we obtain

$$\alpha_i = F_i(x)p_X(x)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} F_i(x)\left(\sum_{k=1}^{m}\lambda_k f_k'(x)\right)p_X(x)dx \quad (12)$$

It is important that the constraint functions are carefully selected such that their analytical integrals are known. In addition, if $p_X(x)$ decays faster than $F_i(x)$, then the first term in (12) goes to zero. This yields

$$\begin{aligned}\alpha_i &= -\sum_{k=1}^{m}\lambda_k \int_{-\infty}^{\infty} F_i(x)f_k'(x)p_X(x)dx \\ &= -\sum_{k=1}^{m}\lambda_k E\left[F_i(X)f_k'(X)\right] \stackrel{\Delta}{=} -\sum_{k=1}^{m}\lambda_k\beta_{ik}\end{aligned} \quad (13)$$

Under the assumption that the actual and maximum entropy distributions are close to each other (in the sense that these expectations are approximately equal) β_{ik} can be estimated from the samples of *X*. Finally, defining the vector $\boldsymbol{\alpha} = [\alpha_1 \ \ldots \ \alpha_m]^T$ and the matrix $\boldsymbol{\beta} = [\beta_{ik}]$, the Lagrange multipliers are determined by the following linear system of equations: $\boldsymbol{\lambda} = -\boldsymbol{\beta}^{-1}\boldsymbol{\alpha}$.

4. MINIMAX ICA LEARNING RULE

The minimax ICA approach is based on minimizing the criterion in (7) by adapting the rotation matrix of the whitening-rotation scheme described in Section 2. Since in this paper, we parameterize the rotation matrix using Givens angles, the optimization algorithm will update these parameters. If the steepest descent approach is employed, the updates become simple. It can easily be shown that the derivative of Shannon's (marginal entropy) of a random variable, denoted by $H(Y_o)$, with respect to a Givens angle, denoted by θ_{pq}, is given by

$$\frac{\partial H(Y_o)}{\partial\theta_{pq}} = -\sum_{k=1}^{m}\lambda_k^o\frac{\partial\alpha_k^o}{\partial\theta_{pq}} \quad (14)$$

In (14), $\boldsymbol{\lambda}^o$ is the Lagrange multiplier associated with the pdf of the o^{th} output signal and α_k^o is the k^{th} constraint for the pdf of the o^{th} output. The Lagrange multipliers are obtained easily based on the previous discussion. The derivative of the constraint constant with respect to the Givens angle is determined from the output samples. Since by definition

$$\alpha_k^o = \frac{1}{N}\sum_{j=1}^{N} f_k(y_{o,j}) \quad (15)$$

where $y_{o,j}$ is the j^{th} sample at the o^{th} output for the current angles, the derivative in (14) is

$$\begin{aligned}\frac{\partial\alpha_k^o}{\partial\theta_{pq}} &= \frac{1}{N}\sum_{j=1}^{N} f_k'(y_{o,j})\frac{\partial y_{o,j}}{\partial\theta_{pq}} \\ &= \frac{1}{N}\sum_{j=1}^{N} f_k'(y_{o,j})\left(\frac{\partial y_{o,j}}{\partial\mathbf{R}_{o:}}\right)^T\left(\frac{\partial\mathbf{R}_{o:}}{\partial\theta_{pq}}\right)^T \\ &= \frac{1}{N}\sum_{j=1}^{N} f_k'(y_{o,j})\mathbf{x}_j^T\left(\frac{\partial\mathbf{R}}{\partial\theta_{pq}}\right)_{o:}^T\end{aligned} \quad (16)$$

In (16), the subscript in $\mathbf{R}_{o:}$ and $\left(\partial\mathbf{R}/\partial\theta_{pq}\right)_{o:}$ denote the o^{th} row of the corresponding matrix. The derivative of the rotation matrix with respect to θ_{pq} is

$$\begin{aligned}\frac{\partial\mathbf{R}}{\partial\theta_{pq}} &= \left(\prod_{i=1}^{p-1}\prod_{j=i+1}^{n}\mathbf{R}^{ij}(\theta_{ij})\right)\left(\prod_{j=p+1}^{q-1}\mathbf{R}^{pj}(\theta_{pj})\right) \\ &\quad \frac{\partial\mathbf{R}^{pq}(\theta_{pq})}{\partial\theta_{pq}}\left(\prod_{j=q+1}^{n}\mathbf{R}^{pj}(\theta_{pj})\right)\left(\prod_{i=p+1}^{n-1}\prod_{j=i+1}^{n}\mathbf{R}^{ij}(\theta_{ij})\right)\end{aligned} \quad (17)$$

Once the gradient is constructed from the individual derivatives with respect to the Givens angles, these parameters can be updated by

$$\boldsymbol{\Theta}_{t+1} = \boldsymbol{\Theta}_t - \eta\sum_{o=1}^{n}\frac{\partial H(Y_o)}{\partial\boldsymbol{\Theta}} \quad (18)$$

where η is a small step size.

5. SIMULATIONS

In this section, we demonstrate the effect of user-determined parameters on the performance of Minimax ICA and present comparisons with benchmark ICA algorithms. All algorithms are operated in batch training mode in all the simulations. For performance evaluation, we will use signal-to-interference ratio (SIR) as the measure, which is made possible by the fact that the actual mixing matrix **H** is known in the simulation environment. The SIR is defined as [12]

$$SIR\ (dB) = \frac{1}{n}\sum_{o=1}^{n} 10\log_{10}\frac{\max_k(\mathbf{O}_{ok}^2)}{\mathbf{O}_{o:}\mathbf{O}_{o:}^T - \max_k(\mathbf{O}_{ok}^2)} \quad (19)$$

where **O**=**RWH**. This measure is the average ratio in decibels (dB) of the main signal power in each output channel to the total power of the interfering signals.

In the first set of Monte Carlo simulations, we investigate the affect of the number of samples and the number of constraints on the performance of Minimax ICA. For simplicity, we use a 2x2 instantaneous linear mixture case, where the entries of the mixing matrix **H** is selected randomly from a uniform distribution in [-1,1] for each one of the 100 runs. The two independent source

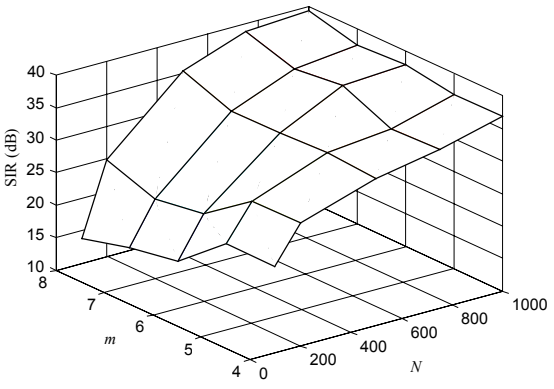

Figure 1. Average SIR of Minimax ICA versus number of moment constraints and number of samples.

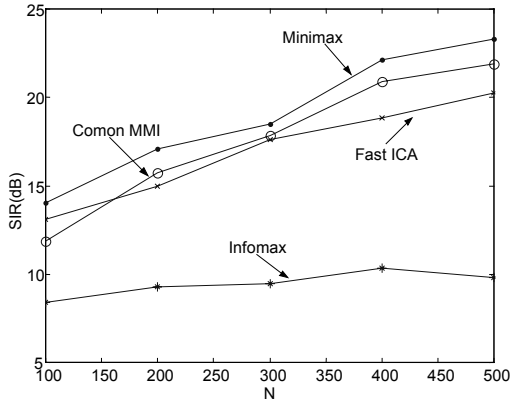

Figure 2. Average SIR of Minimax ICA, Comon's

distributions are uniform and Gaussian. For constraint functions, we use $f_k(x)=x^k$, which correspond to the moments of the random variable X. Repeating these experiments for the number of constraints m=4,5,6,7,8 and the number of samples N=100,200,500,750,1000, the average SIR surface shown in Fig. 1 is obtained. As expected, regardless of the number of constraints being used, performance increases with an increasing number of samples. On the other hand, for a given number of samples, the number of constraints (moments) can be increased up to a certain point to improve performance. After a critical value we expect the performance to start decreasing due to the fact that higher order moments require higher number of samples for accurate estimation. In Fig. 1, for small number of samples, we observe that the performance does not increase when the order of moments increase. The fluctuations are most likely due to the insufficient number of Monte Carlo simulations. If the moment order is larger than the number of samples can tolerate, the performance of the algorithm would degrade. This means, as more samples are available, the performance of Minimax ICA can be improved by using more constraints.

In our second case study, we will conduct Monte Carlo simulations to compare the performance of four algorithms: Minimax ICA (with the first four moments as constraints), Comon's minimum mutual information (MMI) algorithm [7], Fast ICA [6], and Infomax [5]. We performed some simulations using Yang's minimum mutual information algorithm [8], however, these preliminary results were not comparable to the performance of the other algorithms. Therefore, we do not present those results here. In each run, N samples of a source vector composed of one Gaussian, one Laplacian (super-Gaussian), and one uniformly (sub-Gaussian) distributed entry were generated. The 3x3 mixing matrix **H**, whose entries are randomly selected from the interval [-1,1], is also generated. The mixed signals are then fed into the four algorithms (for Fast ICA and Infomax pre-whitening is applied to speed-up convergence). The average SIR of all algorithms obtained over 100 Monte Carlo runs are shown in Fig. 2. In these simulations, Infomax was not very successful, because generic sigmoid nonlinearities were used at its output layer that did not match the exact cdfs of the sources. Although the performance of Infomax would increase if the nonlinearities were matched to the source cdfs, that would result in an unfair comparison by providing additional information to the algorithm regarding the statistical structure of the source signals.

Clearly, the results in Fig. 2 demonstrate that for the same sample set, Minimax ICA achieves better separation. Since these simulations used only 4 moments as constraints, it is possible to improve the performance of Minimax ICA significantly, especially for large data sets, by using additional higher order moments. Hence, the average SIR curve for Minimax ICA shown in Fig. 2 could be regarded as a lower bound for its performance.

6. CONCLUSIONS

Independent component analysis is a problem where information-theoretic optimization criteria finds natural application. Minimum output mutual information is one such information-theoretic criterion, which has been successfully applied in the ICA context. Different approaches usually focus on employing various pdf estimation techniques, which can then be substituted in the mutual information definition to obtain a sample-estimate for this quantity.

In this paper, we have once again exploited this well appreciated information-theoretic criterion, along with the widely accepted whitening-rotation topology, in order to solve the ICA problem. As for the pdf estimation, we have utilized the maximum entropy density estimates, whose robustness is supported by Jaynes' principle. We have

proposed a method for determining the Lagrange multipliers of the constrained entropy maximization problem, which leads to the aforementioned density estimates. Since ICA involves continuous random variables (continuous-valued signals), determining these parameters is especially difficult due to the reasons discussed in the text. The proposed approach for determining the Lagrange multipliers numerically is based on the assumption that certain moments of the actual density and the corresponding maximum entropy density are close to each other. The solution then reduces to solving a system of linear equations, from which these parameters are determined. This assumption, in an ICA setting where the signal densities are not extremely spiky (like a delta-train), is usually approximately satisfied and the density estimates are generally useful. In a setting with spiky signal distributions, such as digital communications, where the signal values are selected from a finite alphabet, the assumption would become invalid.

We have demonstrated with Monte Carlo simulations that, when the data moments are used as the constraints, an increase in the number of constraints (i.e., inclusion of higher order moments) produces an increased performance (if the data length is sufficient). In addition, comparisons with three other benchmark algorithms (Comon's MMI, Fast ICA, and Infomax) indicates that the Minimax algorithm performs well.

The computational load of Minimax ICA in batch mode is slightly higher than the other algorithms considered here. To give a rough idea, using the first four moments as constraints, one update using Minimax ICA requires the evaluation of sample moments of all the outputs up to order 8, solving a 4x4 linear system of equations, and taking the derivative of all output samples with respect to the Givens angles.

Minimax ICA, since it is based on Shannon's entropy, does not require adjustments according to the sub/super-Gaussianity of the sources. It utilizes the presented training data to efficiently extract information from the given set. Since the maximum entropy density estimate *commits minimally to unseen data*, Minimax ICA has good generalization capabilities (as also demonstrated by the SIR measure used as the figure of merit).

Future studies will be directed to optimizing the constraint functions to maximize performance for a given data set. In addition, a low-complexity recursive version of Minimax ICA for on-line separation will be pursued.

Acknowledgments: This work is partially supported by NSF grant ECS-9900394.

7. REFERENCES

[1] A. Hyvarinen, J. Karhunen, E. Oja, *Independent Component Analysis*, Wiley, New York, 2001.

[2] A. Cichocki, S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, Wiley, New York, 2002.

[3] J.F. Cardoso, A. Souloumiac, "Blind Beamforming for Non-Gaussian Signals," *IEE Proc. F Radar and Signal Processing*, vol. 140, no. 6, pp. 362-370, 1993.

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

[5] A. Bell, T. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," *Neural Computation*, vol. 7, pp. 1129-1159, 1995.

[6] A. Hyvarinen, "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis," *IEEE Transactions on Neural Networks*, vol. 10, pp. 626-634, 1999.

[7] P. Comon, "Independent Component Analysis, a New Concept?" *Signal Processing*, vol. 36, no. (3), pp. 287-314, 1994.

[8] H.H. Yang, S.I. Amari, "Adaptive Online Learning Algorithms for Blind Separation: Maximum Entropy and Minimum Mutual Information," *Neural Computation*, vol. 9, pp. 1457-1482, 1997.

[9] D. Erdogmus, K.E. Hild II, J.C. Principe, "Independent Component Analysis Using Renyi's Mutual Information and Legendre Density Estimation," *Proc. IJCNN'01*, pp. 2762-2767, Washington, DC, 2001.

[10] E. Parzen, "On Estimation of a Probability Density Function and Mode," *Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065-1076, 1962.

[11] M. Girolami, "Orthogonal Series Density Estimation and the Kernel Eigenvalue Problem," *Neural Computation*, vol. 14, no. 3, pp. 669-688, 2002.

[12] K.E. Hild II, D. Erdogmus, J.C. Principe, "Blind Source Separation Using Renyi's Mutual Information," *IEEE Signal Processing Letters*, vol. 8, pp. 174-176, 2001.

[13] D. Xu, J.C. Principe, J. Fisher, H.C. Wu, "A Novel Measure for Independent Component Analysis," *Proc. ICASSP'98*, pp. 1161-1164, Seattle, Washington, 1998.

[14] D.T. Pham. Blind Separation of Instantaneous Mixture of Sources via the Gaussian Mutual Information Criterion," *Signal Processing*, vol. 81, pp. 855-870, 2001.

[15] E.T. Jaynes, "Information Theory and Statistical Mechanics," *Physical Review*, vol. 106, no. 4, pp. 620-630, 1957.

[16] G. Golub, C.V. Loan, *Matrix Computation*, John Hopkins University Press, Baltimore, MD, 1993.