INDEPENDENT COMPONENTS ANALYSIS USING RENYI'S MUTUAL INFORMATION AND LEGENDRE DENSITY ESTIMATION

Deniz Erdogmus, Kenneth E. Hild II, Jose C. Principe
Computational NeuroEngineering Lab (CNEL), University of Florida, Gainesville, FL 32611
[deniz,hildk,principe]@cnel.ufl.edu

ABSTRACT

We have previously proposed the use of quadratic Renyi's mutual information, estimated using Parzen windowing, as an ICA criterion and showed that it utilizes data more efficiently than classical algorithms like InfoMax and FastICA. We suggested the use of Renyi's definition of information theoretic quantities rather than Shannon's definitions since the Shannon's definitions are already included in Renyi's as special cases. In the estimation of probability densities using kernel methods, the choice of the kernel width is an important issue that affects the overall performance of the system, and there is no known way of determining the optimal value. Legendre polynomial expansion of a probability distribution, on the other hand, has two advantages. Hardware implementation is trivial and it does not require the choice of any parameter except for the point of truncation of the series. The rule for this assignment is simple: the longer the series, the more accurate the density estimation becomes. Thus, we combine these two schemes, namely Renyi's entropy and Legendre polynomial expansion for probability density function estimation to obtain a simple ICA algorithm. This algorithm is then tested on blind source separation, time-series analysis, and data reduction.

I. INTRODUCTION

Independent components analysis (ICA) has become one of the crucial topics of research in signal processing recently. The classical InfoMax algorithm by Bell and Sejnowski [1] and the FastICA algorithm by Hyvarinen [2] are only two of the important achievements in this field. However, in our previous studies, we showed that our ICA criterion, involving Renyi's mutual information, uses data more efficiently than both of these classical algorithms [3]. The probability density function (pdf) estimator we used was Parzen windowing, which required an assignment of the kernel width. There is no mathematically rigorous method of assigning kernel widths optimally for kernel methods; therefore, we were inclined to find a pdf estimator that did not require such parameter assignment tasks. Polynomial methods are suitable for this purpose. Comon and others used truncated polynomial expansions for pdf estimations [4-6], and typically Edgeworth or Gram-Charlier expansions were utilized, which became intractable and extremely complex even for very small orders of truncation. Legendre polynomials, which were used successfully by Friedman in the projection pursuit context [7], however, are recursive in nature; therefore, the higher order terms in the expansion can easily be computed. This is also an important factor when an easy evaluation of the gradient of the cost function is necessary.

In this paper, we propose a new ICA criterion that employs Renyi's entropy definition and Legendre polynomial expansion to estimate it. First, we briefly describe the cost function. Next we derive the gradient of the cost function with respect to the weight vector of the system. In the case studies section, we present three applications where the algorithm may be used. These include blind source separation (BSS), time-series analysis, and data reduction / feature extraction. Finally, we finish with the conclusions.

II. ICA COST FUNCTION

Mutual information is considered the natural cost function for ICA [5,6,8]. Shannon's definition of mutual information is extensively exploited in the ICA and BSS literature for that reason. Alternatively, we propose the use of Renyi's definition of mutual information as the criterion, since it already contains Shannon's definition as a special case; hence, it is more general. Renyi's mutual information of order α between n random variables with the joint pdf $f_Y(y)$ is given by [9]

$$I_{R_\alpha}(y) = \frac{1}{\alpha - 1} \log \int_{-\infty}^{\infty} \frac{f_Y(y)^\alpha}{\prod_{i=1}^{n} f_{Y_i}(y_i)^{\alpha-1}} dy \qquad (1)$$

For Renyi's definition, the mutual information is not identically equal to the sum of Renyi's marginal entropies minus the joint entropy, which is

$$\sum_{i=1}^{n} H_{R_\alpha}(y_i) - H_{R_\alpha}(\mathrm{y}) = \frac{1}{\alpha - 1} \log \frac{\int_{-\infty}^{\infty} f_Y(\mathrm{y})^\alpha \, d\mathrm{y}}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} f_{Y_i}(y_i)^\alpha \, dy_i} \quad (2)$$

Notice that both (1) and (2) achieve their minimum values when the n random variables are independent, i.e. when the joint pdf is equal to the product of the marginals. For this reason, we can employ (2) as an ICA criterion instead of (1). Recall that, for Shannon's definitions, i.e. as $\alpha \to 1$, the expressions in (1) and (2) become identically equal. If we utilize a two-step parameterization that involves spatial whitening followed by a rotation, as proposed by Comon [4], we can further simplify the cost function. The proposed architecture is shown below in Fig. 1.

Figure 1. ICA System Block Diagram

Since in this case we will only adapt the angles of the Given's rotation matrix, we can drop off the joint entropy from the cost function, because we know that Renyi's joint entropy, like Shannon's, is invariant under rotation [3]. Thus, the cost function becomes

$$J = \sum_{i=1}^{n} H_{R_\alpha}(y_i) \quad (3)$$

In order to evaluate the cost function, we need to obtain a nonparametric estimate of the entropy of each of the outputs. Renyi's entropy of order α for a random variable y_i is defined as [9]

$$\begin{aligned} H_{R_\alpha}(\mathrm{y_i}) &= \frac{1}{1-\alpha} \log \int_{-\infty}^{\infty} f_{Y_i}(\mathrm{y_i})^\alpha \, d\mathrm{y_i} \\ &= \frac{1}{1-\alpha} \log E\left[f_{Y_i}^{\alpha-1}(\mathrm{y_i}) \right] \end{aligned} \quad (4)$$

To obtain the nonparametric estimator, we approximate the expectation by the sample mean, and we substitute the Legendre polynomial expansion for the pdf, which is given by [7]

$$f_X(x) = \sum_{j=0}^{\infty} a_j P_j(x) \quad (5)$$

where the coefficients and the polynomials are given by (6) below. In the nonparametric estimator, the coefficients are also evaluated using the sample mean in place of the expected value.

$$\begin{aligned} a_j &= \frac{2j+1}{2} E\left[P_j(x) \right] \\ P_j(x) &= \begin{cases} 1, & j = 0 \\ x, & j = 1 \\ \frac{1}{j}\left[(2j-1) x P_{j-1}(x) - (j-1) P_{j-2}(x) \right], & j \geq 2 \end{cases} \end{aligned} \quad (6)$$

In solving for the independent components, (3) is evaluated by substituting (4)-(6) and minimized with respect to the angles of the rotation matrix R which is composed of a product of rotations in individual planes.

$$R = \prod_{m=1}^{n-1} \prod_{l=m+1}^{n} R_{ml}(\theta_{ml}) \quad (7)$$

It is important to note that, it is crucial to preserve the ordering of the matrix multiplications when evaluating the cost function and the gradient in training.

III. TRAINING THE SYSTEM WITH STEEPEST DESCENT

Steepest descent is a simple and fast converging algorithm, which is extensively utilized in adaptive systems, although, in general, there is a probability of being stuck in a local optimum. For the proposed cost function, however, this does not impose any problems, because all optima are global [19], and the cost function is a smooth periodic surface in the weight space. The gradient of (3) with respect to the angle θ_{ml} is given by

$$\begin{aligned} \frac{\partial J}{\partial \theta_{ml}} &= \sum_{o=1}^{n} \frac{1}{1-\alpha} \frac{\partial V_\alpha(y^o) / \partial \theta_{ml}}{V_\alpha(y^o)} \\ \frac{\partial V_\alpha(y^o)}{\partial \theta_{ml}} &= \frac{(\alpha-1)}{N} \sum_{i=1}^{N} f_{Y_o}^{\alpha-2}(y_i^o) \cdot \frac{\partial f(y_i^o)}{\partial \theta_{ml}} \\ \frac{\partial f(y_i^o)}{\partial \theta_{ml}} &= \sum_{j=0}^{\infty} \left[\frac{\partial a_j}{\partial \theta_{ml}} P_j(y_i^o) + a_j \frac{\partial P_j(y_i^o)}{\partial \theta_{ml}} \right] \end{aligned} \quad (8)$$

The coefficients a_j depend on the weights since they are estimated from the samples; therefore, in the gradient computation this must be taken into account. The gradient of the polynomials with respect to the angles can be computed by recursive equations that arise from the recursion in (6), and the expression that relates the output samples to the angles in (7). Note that the output vector y is related to the observation vector z

at the output of the spatial whitening by $y = Rz$. Thus, the o[th] output channel depends on the o[th] row of the rotation matrix R.

IV. STOCHASTIC APPROXIMATION OF MUTUAL INFORMATION AND ITS STOCHASTIC GRADIENT

It is rather simple to obtain a stochastic gradient algorithm that would implement the ICA procedure outlined above on a sample-by-sample basis. For this, obvious stochastic version, we just need to follow Widrow's example [10], by dropping all the expectation operators and replacing them with the instantaneous values of their arguments. With these simplifications, the instantaneous entropy estimator for (4) becomes

$$H_{R_\alpha}(\mathrm{y_i}) = \frac{1}{1-\alpha}\log f_{Y_i}^{\alpha-1}(\mathrm{y_i}) \tag{9}$$

where the instantaneous estimate of the pdf using stochastic coefficients is given by

$$f_X(x) = \sum_{j=0}^{\infty} \frac{2j+1}{2} P_j^2(x) \tag{10}$$

The stochastic gradient then becomes

$$\begin{aligned} \frac{\partial J}{\partial \theta_{ml}} &= -\sum_{o=1}^{n} f_{Y_o}^{\alpha-2}(y_i^o) \cdot \frac{\partial f(y_i^o)}{\partial \theta_{ml}} \\ \frac{\partial f(y_i^o)}{\partial \theta_{ml}} &= \sum_{j=0}^{\infty} (2j+1) P_j(y_i^o) \frac{\partial P_j(y_i^o)}{\partial \theta_{ml}} \end{aligned} \tag{11}$$

where the gradient of the polynomial with respect to the rotation angle can be computed similarly to the batch adaptation case given in the previous section.

It turns out that this obvious stochastic gradient estimator for the ICA cost function in (3) is not as robust as Widrow's LMS is for MSE. Many simulations with BSS examples showed that the stochastic gradient could not identify the independent speakers unless only one of them was speaking and all the other sources were silent (except for ambient noise).

V. CASE STUDIES

In this section, we present the results of three problems where ICA methods can be applied, and in particular, we will apply the algorithm sketched in the previous section.

Blind Source Separation

In instantaneous BSS, our aim is to recover the independent source signals from a vector observed signals, which were formed by an instantaneous mixture of the sources. If the number of observations is greater than or equal to the number of independent sources, and the mixing matrix is invertible, then the solution can be found by minimizing the criterion in (3) using the gradient given in (8).

Results are demonstrated here for a mixture of two different sentences by the same female speaker. The observations are obtained by synthetically mixing the two sentences with the mixing matrix H. We want the overall matrix RWH to be a permutation of a diagonal matrix. The mixing matrix and the overall matrix after training with 500 randomly chosen samples are given below. The Legendre polynomial expansion was truncated at the 10[th] order.

$$H = \begin{bmatrix} 0.66 & 0.16 \\ 0.72 & 0.63 \end{bmatrix}, \quad RWH = \begin{bmatrix} 8.65 & 0.14 \\ 0.60 & 7.82 \end{bmatrix} \tag{12}$$

Fig. 2 demonstrates how the deformed signal space after the mixing operation is restored by the ICA algorithm. The recovered source signals are almost exactly the same as the originals, except for the scaling factor introduced in the separation process, which is unavoidable.

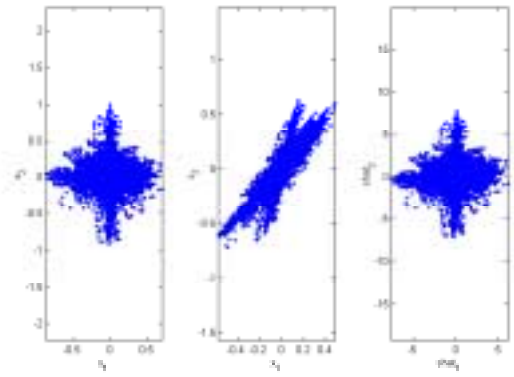

Figure 2. Sources, Observations and Recovered Signals

We have applied this algorithm to the separation of randomly chosen and randomly mixed sources. The separation results for various choices of truncation order (as low as 5) and number of training samples (as low as 50) showed that the algorithm successfully recovers the independent components in the observation vector.

Time-Series Analysis

When training an adaptive FIR filter with the MSE criterion, a well-known procedure to eliminate the effect of eigenvalue-spread is to first employ principal components analysis (PCA) to the input vectors, thus obtaining input vectors that have uncorrelated entries. Then it becomes possible to adapt the weights of the FIR filters with different stepsizes, which are determined from the eigenvalues of the auto-correlation matrix, hence avoid the slow-convergence problem

associated with large eigenspread. Another possible use of this method is to obtain the maximum variance preserving components in an input space having higher dimension than the length of the filter. Then the dimension of the higher order input space could be reduced to be equal to the filter order that is being used, thus obtaining a longer memory depth for the FIR filter, by keeping only the higher energy components.

A similar strategy may be employed, but using information theoretical criteria and ICA. Suppose we are going to train an adaptive, length-L FIR filter as a single step predictor. We can use the L most recent signal values as the input vector, or we can first consider a longer input vector, extending back in time even more, and then apply ICA to this longer input vector and choose the L independent components that possess maximum entropy. This way, we preserve the signal characteristics that have maximum average information. Then we can train the FIR filter to minimize the error entropy, where the error signal is the difference between the filter output and the desired signal. For more information about error-entropy minimization see [11-13].

The intuition behind this approach can be heuristically stated as follows. We know that minimizing the error entropy maximizes the mutual information between the actual filter output and the desired output [14]. By choosing the ICA solutions that possess the most entropy (though it must be noted that some kind of normalization scheme must be employed, which we did not do in order to avoid mistaking dynamic range with entropy), we try to maximize the information transferred through the FIR filter to its output, thus helping maximize the mutual information between that and the desired output, assuming that there exists some type of mechanism (which we wish to approximate) that controls the information transfer from the input to the desired output.

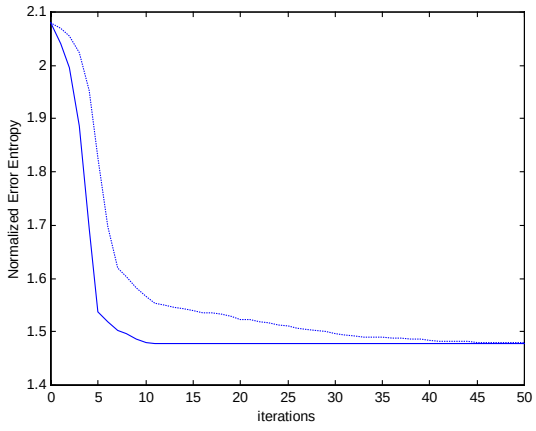

Figure 3. Error entropy vs iterations in FIR training: using best two ICA solutions with maximum entropy (solid), and two most recent samples in signal (dashed)

We have applied the ICA algorithm to extract the 'most informative 2 components' from the 3 dimensional delay vectors formed from the signal. The minimum-error-entropy (MEE) criterion was then employed in training the FIR filters with two taps, one with the input vectors formed from the ICA solutions, and the other with the input vectors formed from the signal values themselves. In numerous experiments, it was consistently observed that the filter that was trained with the ICA components at the input converged faster than the competing filter. However, it did not necessarily achieve smaller error entropy. The reason for this may be the lack of normalization procedures in the choice of the components or the use of different entropy estimators in ICA and MEE algorithms. Fig. 3 shows the convergence of the error entropies for the two filters for one such training scenario. The Legendre polynomial expansion is truncated at 10^{th} order.

Data Reduction

ICA can also be effectively utilized in feature extraction for efficient classifier design. Torkkola had shown that, using a mutual information estimator defined by Principe *et. al.*, it is possible to extract features that lead to the design of classifiers that perform much better than those which are designed based on features that are extracted by PCA like algorithms, which consider merely the variance in the data rather than the information content [15,18]. Torkkola's approach was to maximize the mutual information between the reduced-dimension data and the original data. Alternatively, we propose the use of ICA and again the maximum entropy criterion to select the independent components from the feature vectors that possess maximum entropy. Normalization procedures to avoid mistaking dynamic range with entropy are again crucial in this method.

We briefly remark that the intuition lying behind this approach depends on Fano's bound, and the generalized bounds that we have recently introduced, which contain Fano's bound as a special case [16,17]. These bounds state that increasing the entropy of the input space of the classifier can decrease the classification error probability, as well as increasing the mutual information between the input and the output spaces. Thus, we wish to construct feature vectors of a fixed dimension such that the entropy is maximized. However, to avoid duplicating the information in entries of the feature vectors, we choose the components such that the mutual information between them is minimized.

We used the letter recognition data set acquired from the UCI Machine Learning Laboratory web page

(www.ics.uci.edu/AI/ML/Machine-Learning.html) for testing purposes. As an example, the 6th to 9th features of the feature vectors of samples from the letters Z and M are input to the ICA algorithm (since the algorithm runs in batch mode, it requires more data and much more time, $O(N^2)$ with respect to the data, to converge; therefore, we used a portion of the full feature vector which was 16-dimensional). Once the ICA algorithm converged using 100 training samples (50 for each letter) and a Legendre polynomial approximation of order 10, the two components with the larger entropies were chosen.

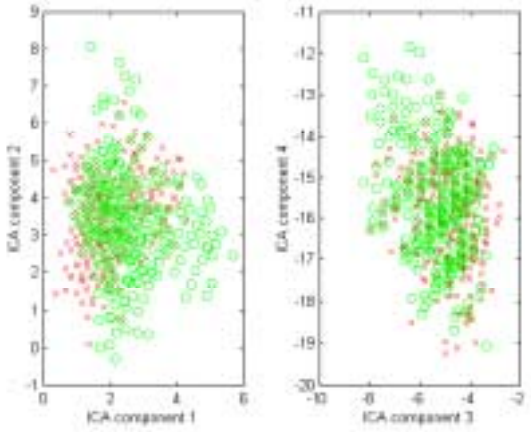

Figure 4. ICA components plotted for Z (x) and M (o) a) Maximum b) Minimum entropy components

Fig. 4 shows the best two and worst two features that resulted as the solution of the ICA algorithm. Clearly, the maximum entropy components on the left are more helpful in distinguishing between the two classes compared with the minimum entropy components on the right.

VI. CONCLUSIONS

In this paper, we have proposed an alternative estimator for an ICA criterion, which had been successfully applied previously to blind source separation. This new criterion, based on Renyi's definition of entropy and Legendre polynomials for density estimation, was applied to three basic problems where the application of ICA methods is promising.

We have provided the gradient expression for the topology/cost-function pair proposed and also investigated the possibility of obtaining a stochastic gradient estimator that would allow sample-by-sample processing of the data to extract independent components. It was found that the obvious stochastic gradient inspired by Widrow's LMS was not sufficiently robust.

The first case example was blind source separation. For this example, the algorithm successfully found the inverse of the mixing matrix. Secondly, the algorithm was applied to time-series analysis in order to determine the most informative independent components of a high dimensional delay-line profile of the signal. The reduced dimensional input vectors consisting of the most informative components were then utilized to train an FIR filter as a single-step predictor. It was observed that the input vectors composed of the more informative components resulted in faster convergence, an expected effect that is similar to what happens in the PCA-MSE case. Finally, we have employed the ICA algorithm to extract the most informative reduced dimensional features from a larger dimensional feature vector. This is an important issue in data visualization and classifier design when the high dimensionality of the feature vector is a problem. The algorithm was successful in identifying the components that would yield less classification error.

When we compared the performances of the Legendre polynomial based estimator and the Parzen window based estimator for the ICA cost function utilized in this paper, we found that the Parzen window estimator performed much better; however, we did not present any results related to this comparison since it is out of the scope of this work, and will be presented elsewhere. This, we believe, results from the pair-wise interaction that the kernels in Parzen windowing impose on the 'information particles' (see [18] for a thorough treatment of information particles in the BSS context). On the other hand, polynomial methods lack this property of pair-wise interaction between training samples, thus performing worse. There is one other item to keep in mind. Preliminary results indicate that the Legendre polynomial expansion requires a higher order than the Edgeworth/Gram-Charlier expansions for a similar performance level.

Nevertheless, Legendre polynomials are well-suited for hardware implementation since they are evaluated recursively from a difference equation. This provides a substantial advantage in speed of computation over the kernel methods. In addition, the polynomial can be expanded to much higher orders than the Edgeworth and Gram-Charlier expansions, which become computationally infeasible at orders greater than four.

Future studies may be conducted on comparing the estimation accuracies and performances of kernel methods, polynomial methods, and other methods in entropy evaluation and in problems of interest to the ICA field. We stress that, as long as the estimators preserve the location of the extreme points of the ideal cost functions, in terms of finding the solution, they are equally useful. Hence the question becomes, how does the estimator formulation affect the dynamics of adaptation and the profile of the performance surface.

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

REFERENCES

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

[2] A. Hyvarinen, “Fast and robust fixed-point algorithms for independent component analysis,” *IEEE Transactions on Neural Networks*, vol. 10, pp. 626-634, 1999.

[3] K.E. Hild II, D. Erdogmus, J.C. Principe, "Blind Source Separation Using Renyi's Mutual Information", to appear in IEEE Signal Processing Letters, June 2001.

[4] A. Hyvarinen, “Survey on Independent Component Analysis,” *Neural Computing Surveys 2*, pp. 94-128, 1999.

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

[6] P. Comon, “Independent Component Analysis, a New Concept?” *Signal Processing*, vol. 36, pp. 287-314, 1994.

[7] J. H. Friedman, “Exploratory projection pursuit,” *Journal of the American Statistical Association*, vol. 82, pp. 249-266, March 1987.

[8] J. Cardoso, “Blind Signal Separation: Statistical Principles,” *Proc. of the IEEE*, vol. 86, 1998.

[9] A. Renyi, *Probability Theory*, American Elsevier Publishing Company Inc., New York, 1970.

[10] B. Widrow, S.D.Stearns, *Adaptive Signal Processing*, Prentice Hall, NJ, 1985.

[11] D. Erdogmus, J.C. Principe, “Comparison of Entropy and Mean Square Error Criteria in Adaptive System Training Using Higher Order Statistics”, Proc. ICA 2000, Helsinki.

[12] D. Erdogmus, J.C. Principe, “Convergence Analysis of the Information Potential Criterion in FIR Filter Training,” submitted to Neural Networks for Signal Processing 2001.

[13] D. Erdogmus, J.C. Principe, “Entropy Minimization Algorithm for Multilayer Perceptrons,” to appear in the Proc. of Int. Joint Conf. on Neural Networks (IJCNN) 2001, Washington, D.C., July 2001.

[14] D. Erdogmus, J.C. Principe, "An Entropy Minimization Algorithm for Short Term Prediction of Chaotic Time Series", submitted to IEEE Transactions on Signal Processing.

[15] Kari Torkkola, "Visualizing class structure in data using mutual information," Proc. of Neural Networks for Signal Proc. (NNSP) 2000, Sydney, Australia, December 11-13, 2000.

[16] D. Erdogmus, J.C. Principe, "Information Theoretic Lower and Upper Bounds for Error Probability of Classifiers", submitted to IEEE Trans. on Pattern Analysis and Machine Intelligence.

[17] D. Erdogmus, J.C. Principe, “Information Transfer Through Classifiers and its Relation to Probability of Error,” to appear in the Proc. of Int. Joint Conf. on Neural Networks (IJCNN) 2001, Washington, D.C., July 2001.

[18] J.C. Principe, D. Xu, J. Fisher, “Information Theoretic Learning,” in *Unsupervised Adaptive Filtering*, vol I, Simon Haykin Editor, 265-319, Wiley, 2000.

[19] D. Erdogmus, K.E. Hild II, J.C. Principe, "Blind Source Separation Using Renyi's Marginal Entropy", submitted to Neurocomputing Special Issue on Blind Source Separation and Independent Component Analysis, Feb 2001.