A Novel Measure for Independent Component Analysis (ICA)

Dongxin Xu, Jose C. Principe, John Fisher III, Hsiao-Chun Wu

Computational NeuroEngineering Laboratory
NEB 486, Electrical and Computer Engineering Department
University of Florida, Gainesville, FL 32611, USA
(xu, principe, fisher, wu)@cnel.ufl.edu

ABSTRACT

Measures of independence (and dependence) are fundamental in many areas of engineering and signal processing. Shannon introduced the idea of Information Entropy which has a sound theoretical foundation but sometimes is not easy to implement in engineering applications. In this paper, Renyi's Entropy is used and a novel independence measure is proposed. When integrated with a nonparametric estimator of the probability density function (Parzen Window), the measure can be related to the "potential energy of the samples" which is easy to understand and implement. The experimental results on Blind Source Separation confirm the theory. Although the work is preliminary, the "potential energy" method is rather general and will have many applications.

1. INTRODUCTION

Information theory is a powerful tool in communication, signal processing, and even machine learning. The parallel between information and energy is well known and here their measures will be also linked. This paper shows that our proposed information measure is connected to "potential energy" in the scenario of "learning from examples". In this case, a machine learns from the interaction with its environment through examples, where each example is a point in a problem space (e.g. feature space, input-output space, etc) and can be regarded as a "particle" in the "potential field" formed by all other examples. As a result, machine learning can be thought of as the interaction between the examples driven by the "forces" among them.

Early in 1928, Hartley proposed a logarithmic measure of information [1]. Later in 1948, Shannon pointed out that Hartley's measure is valid only if all events are equiprobable [2]. Further he coined the term "Information Entropy" which is the mathematical expectation of Hartley's measure. In 1960, Renyi generalized Shannon's idea by using an exponential function rather than a linear function to calculate the mean [3], [4]. Later on, other forms of information entropy appeared (e.g. Havrda and Charvat's measure, Kapur's measure) [5]. Although Shannon's entropy is the only one which possesses all the postulated properties for information measure, the other forms such as Renyi's, Havrda and Charvat's are equivalent with regards to entropy maximization [5]. In a real problem, which form to use depends upon other requirements such as ease of implementation.

The motivation of this work is to find a direct and general measure of information for discrete samples. The fact that all the above measures are functions of the probability density function (pdf) make pdf estimation inevitable. In the "learning from examples" scenario, information is provided in terms of examples and the pdf on the output space of a learning machine can be estimated using a nonparametric estimator (i.e. Parzen Window [7]). Although nonparametric estimation of the pdf from samples in high dimensional spaces is ill-posed [13], here the dimension of the output space is under user control. Entropy maximization for a learning machine with finite dynamic range outputs is achieved by the criterion of mean squared difference between the output pdf estimated by Parzen Window and the uniform distribution [8]. Obviously, such a measure is crucial for the development of "learning from examples" algorithms because it allows direct interaction between samples.

It turns out that Renyi's entropy with order 2 (also called quadratic entropy [6]) can be gracefully integrated with the Parzen Window method resulting in a "potential energy" measure of information for discrete samples. However, instead of entropy maximization, optimization of mutual information (both maximization and minimization) is usually more desired in a learning process [8], [14]. Unfortunately, there is no such graceful integration of the above entropy measures with respect to optimiza-

tion of the mutual information. Based on the Cauchy-Schwartz inequality, a novel measure is proposed which has a quadratic form and results in a "potential energy" representation. Although strict justification has not yet been obtained that the measure is appropriate for dependence (maximization of mutual information), it is evidently a measure of independence (minimization of mutual information). As a direct consequence, the measure is applied to blind source separation which is an important practical example of independent component analysis (ICA) [9], [10], [11].

2. Renyi's Entropy and Potential Energy

Let $a_i \in R^m, i = 1, \ldots, N$, be a set of samples from a random variable $Y \in R^m$ in m-dimensional space. One interesting question is what will be the entropy associated with this set of data points. One answer lies in the estimation of the data pdf by the Parzen Window method using a Gaussian kernel:

$$f_Y(y) = \frac{1}{N}\sum_{i=1}^{N} G(y-a_i, \sigma^2 I) \qquad \textbf{(EQ 1)}$$

where $G(\ ,\ )$ is the Gaussian function, σ^2 is the variance, and $I \in R^{m\times m}$ is the identity matrix. When Shannon's entropy is used along with this pdf estimation, the measure may become very complex. Fortunately, Renyi's entropy with order 2 may lead to a simpler form. Generally, Renyi's entropy with order α (differential entropy for continuous random variable) is as eq. 2:

$$R_\alpha(Y) = \frac{1}{1-\alpha}\log\int f_Y(y)^\alpha dy, \ \alpha > 0, \alpha \neq 1 \qquad \textbf{(EQ 2)}$$

If $S(Y)$ is the Shannon's entropy (differential) for Y, then $\lim_{\alpha\to 1} R_\alpha(Y) = S(Y)$, and $R_\beta(Y) \geq S(Y) \geq R_\gamma(Y)$, for $0 < \beta < 1$ and $1 < \gamma$. So, Shannon's entropy can be viewed as one member of Renyi's entropy family. When $\alpha = 2$, Renyi's entropy $R_2(Y)$ is also called quadratic entropy. Combining eq.1 and eq.2 and using the relation in the Appendix, the entropy measure for a set of discrete data points $H(\{a_i\})$ becomes:

$$\begin{cases} H(\{a_i\}) = R_2(Y|\{a_i\}) = -\log P(\{a_i\}) \\ P(\{a_i\}) = \dfrac{1}{N^2}\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{N} G(a_i-a_j, 2\sigma^2 I) \end{cases} \qquad \textbf{(EQ 3)}$$

Making the analogy between data points and "particles", $P(\{a_i\})$ can be regarded as the overall potential energy since $G(a_i-a_j, 2\sigma^2 I)$ can be taken as the potential energy of "particle" a_i in the potential field of "particle" a_j, or vice versa. So, maximizing entropy in this case is equivalent to minimizing potential energy. If all the data points can be freely moved in a certain region of the space, then the forces between each pair of points: $\partial G(a_i-a_j, 2\sigma^2 I)/\partial a_i$ or $\partial G(a_i-a_j, 2\sigma^2 I)/\partial a_j$ will drive all the points to a state with minimum potential energy, at least locally. The interaction among data samples can also be thought of as an error that needs to be minimized to achieve maximum entropy [8].

Suppose that the problem is one of finding a mapping: $R^n \to R^m$: $Y = q(X, \theta)$ where θ is a set of unknown parameters such that the entropy in the output space $H(\{a_i\})$ is maximized. If we consider the points a_i as the outputs of the mapping $a_i = q(b_i, \theta)$ when the given input data are $b_i \in R^n$, $i = 1, \ldots, N$ the problem is to find θ such that the potential energy in the output space is minimized. In this case, following the idea of "error back-propagation" [12], the forces will be back-propagated to each parameter in θ according to the chain rule:

$$\frac{\partial}{\partial\theta}P(\{a_i(\theta)\}) = \sum_{i=1}^{N}\frac{\partial P}{\partial a_i}\frac{\partial a_i}{\partial\theta} \qquad \textbf{(EQ 4)}$$

i.e. we obtain a general, nonparametric, and sample-by sample methodology to adapt arbitrary nonlinear (smooth) mappings based on entropy maximization.

3. Cauchy-Schwartz Independence Measure

As pointed out above, mutual information is a more general idea than entropy, and it is sometimes more desirable [14]. The mutual information between two random variables can be measured with the Kullback-Leibler divergence $K(f, g) = \int f(x)\log(f(x)/g(x))dx$ where $f(x)$ and $g(x)$ are two pdfs. The corresponding Renyi's measure of divergence between two pdfs with order α is: $R_\alpha(f, g) = \log\left(\int (f(x)^\alpha / g(x)^{\alpha-1})dx\right)/(\alpha-1)$. We can see that neither of them can be integrated with Parzen Window in eq.1 to produce a simple result. Based on the Cauchy-Schwartz inequality, the following symmetric measure is proposed:

$$C(f,g) = \log \frac{\left(\int f(x)^2 dx\right)\left(\int g(x)^2 dx\right)}{\left(\int f(x)g(x)dx\right)^2} \quad \text{(EQ 5)}$$

It is easy to show that $C(f,g) \geq 0$ and the equality holds true if and only if $f(x) = g(x)$. For two random variables Y_1 and Y_2 (with marginal pdfs $f_{Y_1}(y_1)$, $f_{Y_2}(y_2)$ and joint pdf $f_{Y_1Y_2}(y_1, y_2)$), the independence measure can be written as $C(Y_1, Y_2) =$

$$\log \frac{\left(\iint f_{Y_1Y_2}(y_1,y_2)^2 dy_1 dy_2\right)\left(\iint f_{Y_1}(y_1)^2 f_{Y_2}(y_2)^2 dy_1 dy_2\right)}{\left(\iint f_{Y_1Y_2}(y_1,y_2) f_{Y_1}(y_1) f_{Y_2}(y_2) dy_1 dy_2\right)^2}$$

because the non-negative $C(Y_1, Y_2)$ equals 0 if and only if Y_1 and Y_2 are statistical independent. If a set of data: $\{a_i, i = 1, \ldots, N\}$ in the joint space of Y_1 and Y_2, (i.e. $a_i = (a_{i1}, a_{i2})^T$ is in the same space as $(Y_1, Y_2)^T$) is given or observed, the pdf estimation in eq.1. can be used and the independence measure between Y_1 and Y_2 based on the set of data becomes:

$$C((Y_1, Y_2)|\{a_i\}) = \log \frac{P(\{a_i\})P_1(\{a_{i1}\})P_2(\{a_{i2}\})}{P_c(\{a_i\})^2} \quad \text{(EQ 6)}$$

where $P(\{a_i\})$ is the potential in the whole space, $P_l(j, \{a_i\}) = \frac{1}{N}\sum_{i=1}^{N} G(a_{jl} - a_{il}, 2\sigma^2 I_l)$ is the partial marginal potential ($l = 1, 2$), $P_l(\{a_i\}) = \frac{1}{N}\sum_{j=1}^{N} P_l(j, \{a_i\})$ is the marginal potential, and $P_c(\{a_i\}) = \frac{1}{N}\sum_{j=1}^{N} P_1(j, \{a_i\})P_2(j, \{a_i\})$ is the cross-potential. The independence of two variables requires both small joint potential, small marginal potentials and large cross potential. It should be noted that Y_1 or Y_2 can be either scalars or vectors and that eq. 6 can be easily extended to more than two variables.

4. Application to ICA & Blind Source Separation

Given a input data set $\{b_i\}$ ICA seeks to find a set of parameters θ in a parametric mapping $Y = q(X, \theta)$ so that all the components of Y are statistically independent. The Cauchy-Schwartz independence measure can naturally be used in this problem. The forces associated with different potentials described above can be calculated and back-propagation can then adjust the parameters θ.

Blind Separation is a specific case of ICA, where the observed data $X = AS$ is a linear mixture ($A \in R^{m \times m}$) of independent source signals ($S = (S_1, \ldots, S_m)^T$, S_i independent with each other). The problem is to find a projection $W \in R^{m \times m}$, $Y = WX$ so that $Y = S$ up to a permutation and scaling. Below, we present the results of a linear de-mixing system trained with the proposed method

5. Experiments

For ease of illustration, only 2-source-2-sensor problem is tested. There are two experiments presented here:

Experiment 1 tests the performance of the method on a very sparse data set. Two different colored Gaussian noise segments are used as sources, with 30 data points for each segment. The data distribution for source signals, mixed signals and recovered signals are plotted in figure 1. Figure 2 is the training curve which shows how the SNR of de-mixing-mixing product matrix (WA) changes with iteration (SNR approaches to 36.73dB). Both figures show that the method works well.

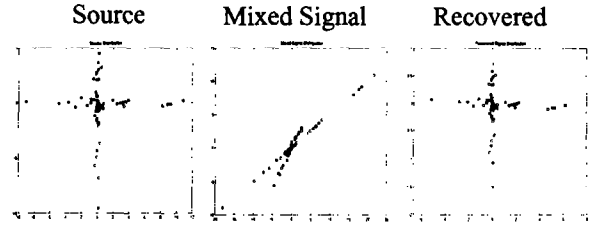

Figure 1. Data Distribution

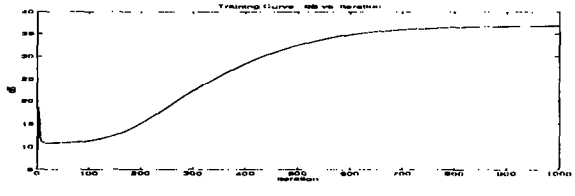

Figure 2. Training Curve for Experiment 1

Experiment 2 uses two speech signals from the TIMIT database as source signals (Figure 3). The mixing matrix is [1, 3.5; 0.8, 2.6] where two mixing direction [1, 3.5] and [0.8, 2.6] are similar. Whitening is first done on mixed sig-

nals. An on-line implementation is tried in this experiment, in which a short-time window slides over the speech data. In each window position, speech data within the window are used to calculate potentials, related forces and back-propagated forces to adjust the de-mixing matrix. As the window slides, all speech data will make contribution to the de-mixing and the contributions are accumulated. The training curve (SNR vs. sliding index, SNR approaches to 49.15dB) is shown in figure 4 which tells us that the method converges fast and works very well. We can even say that it can track the slow change of mixing. Although whitening is done before the "potential energy" method, we believe that whitening process can also be incorporated into this method.

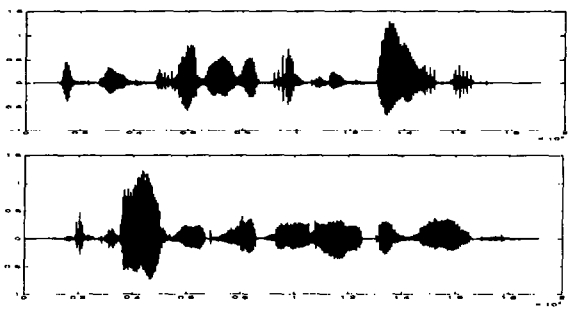

Figure 3. Two speech signals

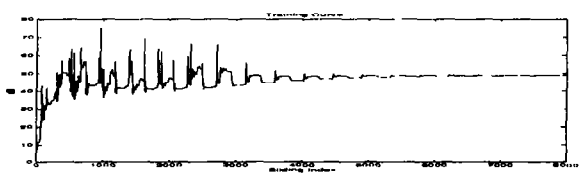

Figure 4. Training Curve

6. CONCLUSION & DISCUSSION

This paper gives a direct measure for entropy and independence of discrete data sets. This is significant because entropy and independence can be related to potentials and the optimization can be computed directly by the interaction between samples. However, the kernel size is a function of the data and the way to choose it and how it may affect the performance remains to be investigated. In the Cauchy-Schwartz independence measure, instead of the ratio, a difference can also be used but it may produce performance surfaces that are more difficult to search. Similarly, the mean squared difference between joint pdf and factorized marginal pdfs can also be used which are essentially the same as the Cauchy-Schwartz measure.

APPENDIX

Let $G(x-a_i, \Sigma_1)$ and $G(x-a_j, \Sigma_2)$ be two Gaussian function with mean a_i and a_j, covariance matrix Σ_1 and Σ_2 respectively, where $x, a_i, a_j \in R^m$, $\Sigma_1, \Sigma_2 \in R^{m \times m}$, then there is following relation:

$$\int_{-\infty}^{+\infty} G(x-a_i, \Sigma_1) G(x-a_j, \Sigma_2) dx = G(a_i - a_j, (\Sigma_1 + \Sigma_2))$$

Acknowledgments: This work was partially supported by DARPA grant F33615-97-1-1019.

References

[1] Hartley, R.V. "Transmission of information". Bell System Technical Journal, 7, 1928.
[2] Shannon, C.E. "A mathematical theory of communication". Bell Sys. Tech. J. 27, 1948, pp379-423, 623-653
[3] Renyi, A. "Some Fundamental Questions of Information Theory". Selected Papers of Alfred Renyi, Vol 2.
[4] Renyi, A. "On Measures of Entropy and Information". Selected Papers of Alfred Renyi, Vol. 2. pp 565-580
[5] Kapur, J.N. "Measures of Information and Their Applications". John Wiley & Sons. 1994
[6] Jumarie, G. "Relative Information, theory and applications", Springer-Verlag. 1990
[7] Parzen, E. "On the estimation of a probability density function and the mode", Ann. Math. Stat. 33, 1962, p1065
[8] Fisher, J. III. "Non-linear extensions to the minimum average correlation energy filter" Ph.D dissertation, Dept. of Electrical Engineering, University of Florida. 1997.
[9] Bell, A.J. and Sejnowski, T.J. "An information-maximization approach to blind separation and blind deconvolution", Neural Comput., Vol.7, no.6, pp1004-1034
[10] Yang, H.H. and Amari S. "Adaptive on-line learning algorithms for blind separation -- maximum entropy and minimum mutual information", to appear in Neural Computation, 1997
[11] Cardoso, J.-F. "Infomax and Maximum Likelihood for Blind Source Separation", IEEE Signal Processing Letters, Vol. 4. No. 4. April 1997, pp112-114
[12] Rumelhart, D.E., Hinton, G.E. and Williams, J.R. "Learning representations by back-propagating errors", Nature (London), 323, 1986, pp533-536.
[13] Vapnik, V.N. "The Nature of Statistical Learning Theory", Springer, 1995.
[14] Linsker R. "An application of the principle of maximum information preservation to linear systems", in Advances in Neural Information Processing Systems 1, Touretzky D.S. (ed), Morgan-Kauffman