SUPERVISED NEURAL NETWORK TRAINING USING THE MINIMUM ERROR ENTROPY CRITERION WITH VARIABLE-SIZE AND FINITE-SUPPORT KERNEL ESTIMATES

Umut Ozertem, Deniz Erdogmus
CSEE Department, OGI, Oregon Health & Science University, Portland, Oregon, USA

Abstract. The insufficiency of mere second-order statistics in many application areas have been discovered and more advanced concepts including higher-order statistics, especially those stemming from information theory like error entropy minimization are now being studied and applied in many contexts by researchers in machine learning and signal processing. The main drawback of using minimization of output error entropy for adaptive system training is the computational load when fixed-size kernel estimates are employed. Entropy estimators based on sample spacing, on the other hand, have lower computational cost, however they are not differentiable, which makes them unsuitable for adaptive learning. In this paper, a nonparametric entropy estimator that blends the desirable properties of both techniques in a variable-size finite-support kernel estimation methodology. This yields an estimator suitable for adaptation, yet has computational complexity similar to sample spacing techniques. The estimator is illustrated in supervised adaptive system training using the minimum error entropy criterion.

I. INTODUCTION

Since the earlier work of Wiener on adaptive filtering mean square error (MSE) has been used as a widely accepted criterion for adaptive system training [1,2,3]. The main reasons behind this choice are the assumption that the real life phenomena can be sufficiently described using second order statistics and the analytical and computational simplicities of this method. Under Gaussianity assumption, MSE, which solely constrains second-order statistics, would be capable of extracting all possible information from a signal whose characteristics are solely defined by its mean and variance. In the absence of Gaussianity assumption, especially for non-linear signal processing, a more suitable approach would be to constrain the information content of the signals rather than simply their energy.

Although Gaussianity assumption has proven to provide successful solutions for many practical problems, it is evident that this approach needs to be refined while dealing with non-linear systems. Moreover, the insufficiency of mere second-order statistics in many application areas have been discovered and more advanced concepts including higher-order statistics, especially those stemming from information theory are now being studied and applied in many contexts in machine learning and signal processing [4,5].

Entropy is introduced by Shannon as a measure of the average information in a given probability distribution function [6,7]. Entropy, being a functional of the probability density function itself, includes all the higher order statistical properties defined in probability density function. Hence, entropy is superior to MSE as an optimality criterion due to the fact that minimizing the error entropy constrains all moments of error pdf, not only the first and second moments. Besides, using error entropy as a criterion for adaptive system training is conceptually straightforward. Given the samples of the input and output, the entropy of the output error evaluated over the training set has to be minimized. Minimizing the output error entropy is equivalent to minimizing the distance between the probability density functions of the desired and output sequences [8]. Specifically, in the case of Shannon entropy this corresponds to minimizing the Kullback-Leibler divergence.

Since analytical data distributions are not available in many practical situations, in the plug-in approach to nonparametric entropy estimation [9], an estimate of the probability density function of the signal is substituted into the sample mean approximation for the expectation. The non-parametric kernel density estimator (KDE) is typically the estimator of choice for this purpose. In KDE, the probability density is approximated by a sum of kernels whose centers are translated to the sample locations. A suitable and commonly accepted kernel function is the Gaussian, which is attractive for adaptation purposes, because it is continuously differentiable and it leads to continuously differentiable density estimates. KDE is a consistent estimator and is also proven to have a good asymptotic behavior. However, nonparametric entropy estimation using KDE results in $O(N^2)$ complexity, where N is the number of samples [8]. Therefore, KDE based adaptation is computationally prohibitive for large training sets. On the other hand, the density estimators based on sample spacing have lower computational cost, $O(Nm)$, where m is the neighborhood size [10,11]. However, the results provided by these estimators are not differentiable, and not suitable for adaptive learning.

In this paper we propose a continuously differentiable entropy estimation technique based on a variable-size finite-support kernel entropy estimator that blends the desirable properties of both techniques: differentiability and continuity of the kernel estimators and the computational simplicity of the sample spacing estimators. The derivation of the estimator is motivated by a kernel estimate interpretation of the standard sample spacing estimates and relaxing the rectangular kernel to a smooth finite-support polynomial kernel. In this paper, we have utilized a fourth order polynomial as the kernel on the support in order to have the least exponential order while maintaining smoothness of the derivatives up to order two as well as keeping the piecewise nature of the kernel to a minimum of pieces. With the reduced computational cost, the usage of entropy-based adaptation criteria becomes more applicable for large data sets. We illustrate the estimator's suitability in minimum error entropy training of adaptive systems.

II. Entropy Estimation

Kernel Estimates: The fixed-size kernel estimates rely on the plug-in estimation methodology and the use of Parzen windowing (also called KDE). For a suitable kernel function $K(.)$, the pdf estimate for a random variable e with samples $\{e_1,\ldots,e_N\}$ is [12,13]

$$\hat{p}(e) = \frac{1}{N}\sum_{i=1}^{N} K(e-e_i) \tag{1}$$

Consequently, the Renyi's order-α entropy estimate of e is found to be [14]

$$\hat{H}_\alpha(e) = \frac{1}{1-\alpha}\log\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N} K(e_j - e_i)\right)^{\alpha-1} \tag{2}$$

Sample Spacing Estimates: The *order statistics* of a sample of a random variable is known to be simply the elements of the sample rearranged in a non-decreasing order. Consider a random variable e, whose samples are denoted by $\{e_1<e_2<\ldots<e_N\}$, which are labeled in non-decreasing order. The 1-spacing estimator of the Shannon entropy of e is given by [10,11]

$$H(e) = \frac{1}{N}\sum_{i=1}^{N-m}\log\left(\frac{N}{m}\left(e_{i+1}-e_i\right)\right) \tag{3}$$

It can easily be seen that the sample spacing entropy estimator can be derived from a plug-in estimation perspective by assuming the following density estimator for the order statistics of e:

$$\hat{p}(e) = \begin{cases} \dfrac{1}{(N+1)(e_{i+1}-e_i)} & \text{if} \quad e_i \le e < e_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

This estimate is generated such that the expected value of the probability mass of the two successive elements of a sample of a random variable is $1/(N+1)$ and corresponds to assuming uniform kernels with variable size for each sample interval.

The variance of this estimator can be decreased by a factor of m considering successive m-sample intervals instead of using each successive sample pair. This is known as the m-spacing estimator and it is shown to be a consistent estimate of entropy provided that m increases with sample size. The latter estimator has a better asymptotic behavior compared to the 1-spacing estimator.

This approach can be interpreted as a summation of uniform density kernels of finite support, where the kernels are located at the successive error samples and the height of each kernel is determined by the sample spacing as in (4), accordingly. Using the above interpretation, we can replace each uniform kernel with a continuously differentiable kernel to develop a continuously differentiable density estimate, which is suitable for adaptive learning.

Variable-Size Finite-Support Kernel Estimates: Defining the midpoints of each sample pair $\{e_i,e_{i+m}\}$ as $\tilde{e}_i$, we obtain the set of kernel center locations as $\{\tilde{e}_1,\tilde{e}_2,\cdots,\tilde{e}_{N-m}\}$:

$$\tilde{e}_i = \frac{e_{i+m}+e_i}{2} \tag{5}$$

By construction, the successive samples of (5) are also in a non-decreasing order. Using (5), the probability density given in (4) is rewritten as

$$p(e) = \sum_{i=1}^{N-m} K_{\sigma_i}(e-\tilde{e}_i) \tag{6}$$

where K_{σ_i} should be selected for each $\tilde{e}_i$ such that it is nonzero in the interval $[e_i,e_{i+m}]$ and zero otherwise. While the nonzero segment of the kernel can be selected to be any suitable smooth function satisfying the required boundary conditions, a fourth order polynomial is the minimum order and simplest kernel choice that also has enough degrees of freedom to meet the boundary and smoothness constraints. Hence, we proceed with this choice to illustrate the technique:

$$K_{\sigma_i} = \begin{cases} A_i\,(e-\sigma_i)^2(e+\sigma_i)^2 & e\in[-\sigma_i,\sigma_i] \\ 0 & \textit{otherwise} \end{cases} \tag{7}$$

In (7), σ_i are selected as described above

$$\sigma_i = \frac{e_{i+m}-e_i}{2} \tag{8}$$

and A_i, the normalization constant can be easily evaluated to be

$$A_i = \frac{15\sigma_i^{-5}}{16(N-m)} \tag{9}$$

in this case. This kernel is continuous up to its fourth order derivative with respect to its argument e, which is more than enough for the typical first and second order iterative learning rules, its width is automatically determined from the data eliminating the need to seek an optimal kernel size, which is a drawback in fixed-size kernel methods, and it has finite support on $[e_i,e_{i+m}]$, which will lead to the computationally efficient entropy estimator that we present next.

Given the continuous and differentiable density estimate in (6) for e, one can easily write the entropy of for this random variable by plugging this estimator in the entropy definition, and approximating the expected value with a sample mean. To illustrate, we consider Renyi's quadratic entropy:

$$\begin{aligned} H_2(e) &= -\log \int p^2(e)de \\ &= -\log E[p(e)] \\ &\approx -\log \frac{1}{N}\sum_{j=1}^{N}\frac{1}{N-m}\sum_{i=1}^{N-m} K_{\sigma_i}(e_j - \tilde{e}_i) \qquad (10) \\ &= -\log\left(\frac{1}{(N)(N-m)}\sum_{i=1}^{N-m}\sum_{j=i+1}^{i+m-1} K_{\sigma_i}(e_j - \tilde{e}_i)\right) \end{aligned}$$

In (10), the last step incorporates the finite-support nature of the kernels by removing unnecessary kernel evaluations in one of the summations, thus reduces the computational requirement by m/N-fold. Substituting (5) for $\tilde{e}_i$, (7) for K_{σ_i}, (8) for σ_i, and (9) for A_i, we get

$$H_2(e) \approx -\log\left(\sum_{i=1}^{N-m}\sum_{j=i+1}^{i+m-1}\frac{30(e_j - e_{i+m})^2(e_j - e_i)^2}{N(N-m)^2(e_{i+m} - e_i)^5}\right) \quad (11)$$

where the entropy is solely defined in terms of the available samples with no parameters to adjust except m.

In the KDE method, in general an infinite support kernel such as a Gaussian or a Laplacian density is used, since fixed-size finite-support kernels result in a poor asymptotic behavior. In this case, however, we started from the order-statistics based sample spacing estimator and integrated the kernel concept into this well-known entropy estimator to rewrite the corresponding density estimator as a sum of finite-support kernels, where the kernel size is locally adjusted to the data spread through the m-neighborhood relation. Selecting m, the number of samples in each neighborhood must be based on the total number of samples to guarantee asymptotic consistency. Typically, in the literature $m = \sqrt{N}$ is recommended and we also employ this formula [11]. This selection satisfies $lim_{N\to\infty} m=\infty$ for asymptotic unbiasedness besides the asymptotic consistency due to $lim_{N\to\infty} m/N=0$. The parameter m controls the trade-off between the computational complexity and the asymptotic behaviour as shown above. The computational complexity, which is $O(Nm)$, becomes $O(N^{3/2})$ for the recommended selection. For appropriate cases with some a priori knowledge about data, one can even choose m as a constant value and get $O(N)$ complexity with an inconsistent, but computationally more efficient entropy estimator.

Initialize weights.
Repeat the following until convergence

- *Select artificial noise power from schedule*
- *Add noise to the clean desired output*
- *Calculate training set errors from noisy data*
- *Sort error samples in ascending order and sort the corresponding input and hidden layer activations*
- *Evaluate gradient as shown in (12)*
- *Perform line search to adjust the step size*
- *Update weights*

Table 1. Outline of the algorithm for training MLPs.

III. Minimum Error Entropy Criterion for Adaptive Learning

Minimization of output error entropy has been shown to be superior to methods based on second-order statistics as mentioned before [8,14]. Employing the entropy estimate in (11) for adaptive learning, one can overcome the main drawback of this procedure: high computational complexity.

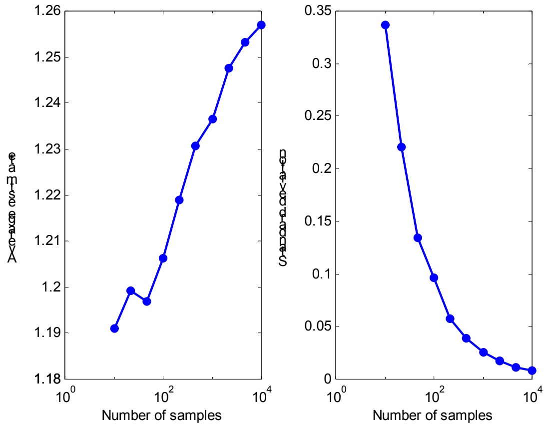
Given an adaptive system with weight vector **w** and a training set consisting of input-output pairs, typically gradient descent algorithm is utilized, while other alternatives such as the Newton method exist. The gradient of the error entropy with respect to the weights can be obtained by letting e_i denote the training error samples in (11) and taking the derivative with respect to the system parameters. This yields:

$$\frac{\partial V_2(e)}{\partial w} = \frac{30}{N(N-m)}\sum_{i=1}^{N-m}\sum_{j=i+1}^{i+m-1}\begin{pmatrix}(e_{i+m}-e_i)^{-6} \\ (e_j - e_{i+m}) \\ (e_j - e_i)\end{pmatrix}\cdot \Delta_{ij}$$

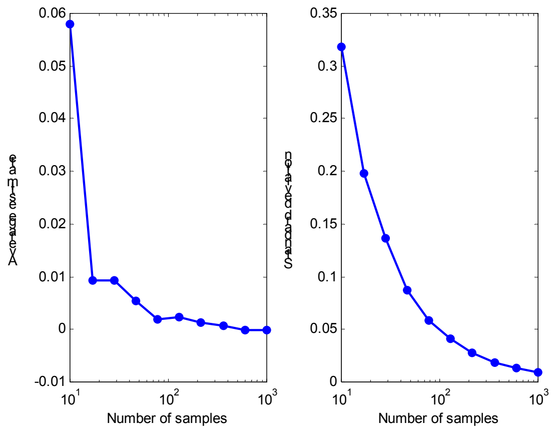
$$\Delta_{ij} = \begin{pmatrix} 5\left(\frac{\partial y_{i+m}}{\partial w} - \frac{\partial y_i}{\partial w}\right)(e_j - e_{i+m})(e_j - e_i) \\ -(e_{i+m} - e_i)\left(\frac{\partial y_j}{\partial w} - \frac{\partial y_{i+m}}{\partial w}\right)(e_j - e_i) \\ -(e_{i+m} - e_i)(e_j - e_{i+m})\left(\frac{\partial y_j}{\partial w} - \frac{\partial y_i}{\partial w}\right)\end{pmatrix} \quad (12)$$

where $V_2(e)$ is the argument of the logarithm in (10), which is also called the quadratic information potential [14], and $\partial y_i/\partial \mathbf{w}$ denotes the gradient of the adaptive system output for the i^{th} sample, which can be evaluated for linear filters and standard neural networks easily as known from the relevant literature on backpropagation [2].

Inspecting (11) and (12), one can observe that in the ideal case where all error samples approach zero, both the objective function H, and its gradient approach $-\infty$. While, in practice, perfect zero errors are never achieved, and both functions will attain finite values at the optimal solution, the performance surface around this solution might resemble a *funnel* resulting in a numerically

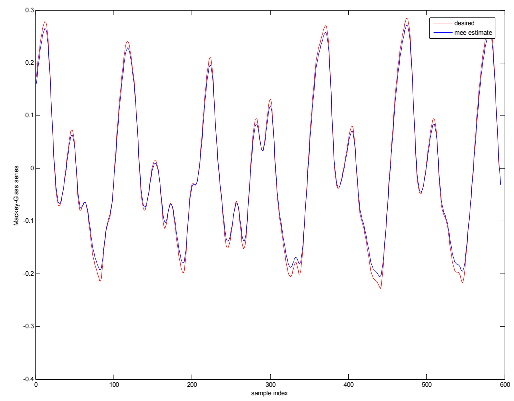


(a)

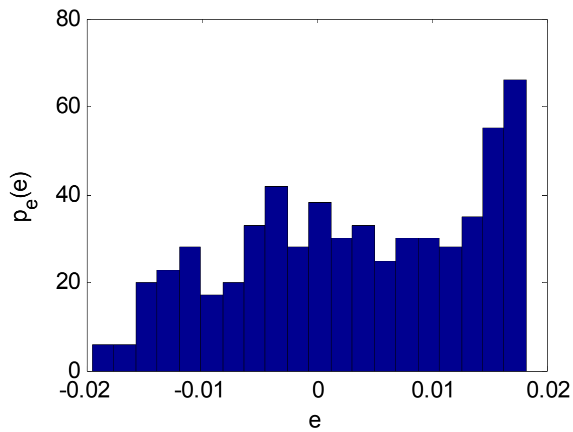


(b)

Figure 1. Average estimates and standard deviations of the quadratic entropy estimator in (11) versus sample size for (a) Gaussian distributed data and (b) uniformly distributed data.



(a)



(b)

Figure 2. (a) Mackey-Glass chaotic time series prediction results (b) estimation error histogram for the testing phase

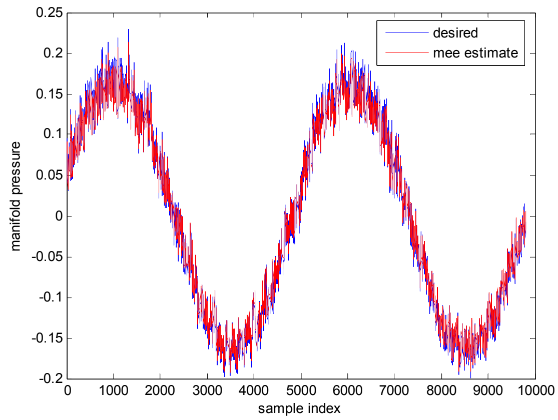
unstable gradient descent learning algorithm near the solution. This numerical issue can be easily resolved by superimposing an artificial measurement noise onto the desired output, whose power controls the minimum error one can attain. During learning, both the learning rate and the artificial noise power can be reduced gradually in order to guide the weights to the global optimum smoothly. The outline of the algorithm is given in Table 1.

IV. Experimental EResults

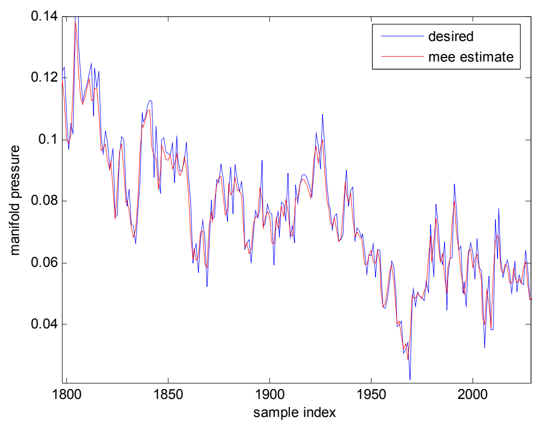
The Entropy Estimator: We demonstrate the performance of the entropy estimator on synthetic data generated according to unit-variance Gaussian and unit-support uniform distributions. The estimator in (11) is applied to datasets of various sizes drawn from these two distributions using the recommended m value. The averages and standard deviations versus sample size of the estimates over 1000 Monte Carlo simulations are shown in Fig. 1. The true entropy of the Gaussian distribution is about 1.26 and that of the uniform distribution is 0. As expected, the systematic bias and variance both decrease asymptotically as the number of samples increases in both cases. Also note that the standard deviation is typically orders of magnitude larger than the bias, therefore, the bias does not contribute significantly to the overall error of the estimator.

Minimum Error Entropy Training: To demonstrate the performance of the minimum error entropy criterion in (11), we utilize two multilayer perceptron (MLP) training examples.

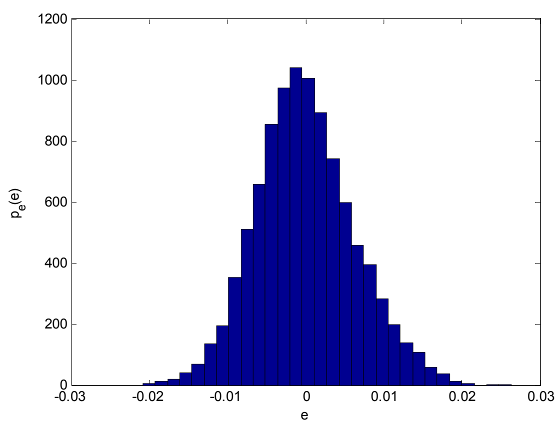
The first example is the short-term prediction of the Mackey-Glass chaotic time series [15] with parameter $\tau = 30$. A TDNN topology used for this purpose has a 3-tap input, 5 neurons in the hidden layer and a single linear output neuron. The embedding dimension suggested by Taken's Embedding Theorem is 5 for the Mackey-Glass series [16]. Choosing the embedding dimension as 3, constrains the reconstruction space to



(a)



(b)



(c)

Figure 3. (a) MEE results for engine manifold dynamics for 10000 samples (b) MEE results close up (c) estimation error histogram for the testing phase

be three-dimensional and increases the difficulty of the prediction problem. The data set is generated to be a 800 sample Mackey-Glass time series and first 200 samples used for training and the remaining for testing.

Results obtained with MEE for the given short-term prediction is presented in Fig. 2 along with the error probability density function for the testing set. Similar results are obtained for the training set, however, the MSE for testing set is slightly higher than that of training set, which can be interpreted as an indication that no over-fitting occurs in the training phase. This is also seen by the test error distribution in Fig. 2b.

The second example is chosen to be a realistic system, namely, identification of the realistic nonlinear engine manifold dynamics in a car engine [17,18]. The engine manifold model assumes the manifold pressure and manifold temperature as the states, and the pressure as the system output. The input is the angle of the throttle that controls the amount of air flowing into the manifold. Using an ARMA model for the desired output here, the last two recent values of the desired response and the input is used for predicting the desired system output.

Results corresponding to nonlinear engine manifold dynamics are presented in Fig. 3, along with the estimation error pdf, which mostly corresponds to the Gaussian measurement noise in the data. A close up figure is also presented for this example to underline the accuracy of the MEE criterion, since the simulation results for over the whole set in Fig. 3a are not representative for presenting the accuracy in estimating high frequency characteristics of the desired output. As in the first data set, the MSE for the testing set is slightly higher than the MSE evaluated over the training set, hence, no over-fitting occurs in the training.

V. Conclusions

In this paper, we proposed a hybrid entropy estimator that exploits the variable-size kernel density estimators based on the plug-in estimation principle. In order to increase computational efficiency, inspired by the sample-spacing entropy estimation technique, the kernels are piecewise defined to be finite-support, but maintain various orders of continuity and differentiability. The latter property is required for applicability in adaptive system training. In this paper, the nonzero segment of the kernel is selected to be a fourth order polynomial, since this is the lowest order polynomial that satisfies continuity and differentiability at the boundaries while requiring a single analytical expression over the whole support. Lower order polynomials (e.g., piecewise linear or quadratic kernels) would require multiple segments over the support, while higher order polynomials would introduce more (perhaps undesirable) flexibility.

The proposed entropy estimator is tested on uniform and Gaussian distributed data and is shown to be asymptotically unbiased and consistent. It has also been utilized in supervised neural network training in the minimum error entropy training framework in order to illustrate its utility. The MEE framework was proposed earlier and has been demonstrated to outperform standard squared error criteria in supervised training. Previous work, however, utilized fixed-size infinite-support kernels, which has high computational demands. The estimator proposed in this paper has been motivated by the need to reduce the computational complexity in these estimators while maintaining smoothness, and the need to have a parameter-free completely nonparametric estimator where the kernel size is set locally and automatically based on the samples.

REFERENCES

[1] B. Widrow, S.D. Stearns, *Adaptive Signal Processing*, Prentice Hall, New Jersey, 1985.

[2] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed., Prentice Hall, Upper Saddle River, New Jersey, 1999.

[3] A.H. Sayed, *Fundamentals of Adaptive Filtering*, Wiley, New York, 2003.

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

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

[6] C.E. Shannon, W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, 1964.

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

[8] D. Erdogmus, J.C. Principe, "An Error-Entropy Minimization Algorithm for Supervised Training of Nonlinear Adaptive Systems," IEEE Transactions on Signal Processing, vol. 50, no. 7, pp. 1780-1786, 2002.

[9] J. Beirlant, E.J. Dudewicz, L. Gyorfi, E.C. van der Meulen, "Nonparametric Entropy Estimation: An Overview," International Journal of Mathematical and Statistical Sciences, vol. 6, no. 1, pp. 17-39, 1997.

[10] O.Vasicek, "A test for normality based on sample entropy," Journal of the Royal Statistical Society Series B, vol. 38, no. 1, pp. 54-59, 1976.

[11] E.G. Miller, J.W.Fisher, "ICA Using Spacings Estimates Of Entropy," Proceedings of the Fourth International Symposium on ICA and Blind Signal Separation, 2003.

[12] E. Parzen, "On Estimation of a Probability Density Function and Mode", in *Time Series Analysis Papers*, Holden-Day, Inc., San Diego, California, 1967.

[13] L. Devroye, G. Lugosi, *Combinatorial Methods in Density Estimation*, Springer, New York, 2001.

[14] D. Erdogmus, J.C. Principe, "Generalized Information Potential Criterion for Adaptive System Training," IEEE Transactions on Neural Networks, vol. 13, no. 5, pp. 1035-1044, 2002.

[15] Kaplan, D., Glass, L., *Understanding Nonlinear Dynamics*, Springer-Verlag, NY, 1995.

[16] Kuo, J.M., *Nonlinear Dynamic Modeling With Artificial Neural Networks*, Ph.D. Dissertation, University of Florida, 1993.

[17] J.D. Powell, N.P. Fekete, C-F. Chang, "Observer-Based Air-Fuel Ratio Control," IEEE Control Systems Magazine, vol. 18, no. 5, pp. 72-83, 1998.

[18] D. Erdogmus, A.U. Genc, J.C. Principe, "A Neural Network Perspective to Extended Luenberger Observers," Institute of Measurement and Control, vol. 35, pp. 10-16, 2002.