Minimum Error Entropy Luenberger Observer

Jian-Wu Xu[1], Deniz Erdogmus[2], Jose C. Principe[1]

[1]CNEL, Dept. of Electrical and Computer Engineering, University of Florida, USA
[2]Dept. of Computer Science and Engineering, Oregon Graduate Institute, OHSU, USA

Abstract - **In this paper, we apply the information-theoretic learning (ITL) technique to the extended Luenberger observer. Instead of prespecifying the globally stable observer gains for nonlinear dynamic systems, we propose minimizing the entropy of the error between the measurement and the estimated output to update the observer gains. A stochastic gradient-based algorithm is presented and the performance of the entropy observer is demonstrated on linear and nonlinear dynamic systems. We also point out that this approach leads to the introduction of kernel methods into state estimation.**

I. INTRODUCTION

State estimation drew an intense attention in control and signal processing community following Kalman's seminal paper [1]. Modern approaches on state estimation are based on Luenberger's design [2,3]. The classical Luenberger observer deals with linear systems. Recently there has been work on the stable observer design for nonlinear and time-varying systems [4,5]. These extensions to nonlinear systems were focused on analytical design techniques. The main method is to set the observer gains such that the overall linearized error dynamics matrix, composed of the gain vector, the Jacobians of the state dynamics and the output mapping, has stable eigenvalues over a closed subset of the state space. Convergence can be proven and stability can be guaranteed for those state trajectories residing in this subset under these conditions [4,5]. This analytical extended Luenberger observer design has been successfully applied to realistic nonlinear system models.

In contrast to this analytical design, recently we proposed an adaptive extended Luenberger observer using mean-square-error (MSE), where the observer gains are continuously updated during state estimation [6]. MSE can extract all the information in the data provided that the dynamic system is linear and the noise is Gaussian distributed. However, when the system becomes nonlinear and the noise distribution is non-Gaussian, MSE fails to capture all the information in the error sequences.

In order to extend Luenberger observer to nonlinear dynamic system with non-Gaussian-distributed noise, an alternative criterion is needed in order to achieve optimality. Entropy is a natural extension beyond MSE since entropy is a function of probability density function (pdf), which considers all high order statistics [7].

Information theoretic approaches have been previously proposed as a natural extension of Kalman filtering to nonlinear and non-Gaussian systems and signals [8]. This early work focused on the theoretical aspects of entropy-based state estimation without providing a feasible practical algorithm. The contribution of this paper is a practical algorithm that implements error-entropy-based state estimation that has low computational complexity for a feasible real-time implementation as well.

II. EXTENDED LUENBERGER OBSERVER

Consider a linear time-invariant (LTI) dynamic system described by the equations

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k \\ y_k &= Cx_k + Du_k \\ x_0 &= x_0 \end{aligned} \tag{1}$$

where

- $x_k \in \Re^n$ is the state vector,
- $u_k \in \Re^m$ is the system input,
- $y_k \in \Re^t$ is the system output,
- x_0 are the initial conditions (probably unknown).

It is often necessary to construct estimates of the state vectors that are not available through direct measurement in control design. A state estimator can be specified as $\hat{x}_k = \Im(u_k, y_k)$, where $\Im$ denotes an operator and $\hat{x}_k$ is called the *state estimate*. It is desired that the estimation error, defined by $\tilde{x}_k = x_k - \hat{x}_k$, be small in some sense.

Luenberger proposed state observers for multivariable dynamic systems [2], which deal with state estimation for deterministic systems. The estimation error is fed back through a proportional term so that the closes-loop state estimator is stable and the estimation error will approach to zero asymptotically provided that the original system is observable. The Luenberger observer is given by

$$\begin{aligned} \hat{x}_{k+1} &= A\hat{x}_k + Bu_k + L(y_k - \hat{y}_k) \\ \hat{y}_k &= C\hat{x}_k + Du_k \end{aligned} \tag{2}$$

This work was supported by NSF grant ECS-0300340.

If the observer gain vector L is set to a value such that the estimator error dynamics given by

$$\tilde{x}_{k+1} = x_{k+1} - \hat{x}_{k+1} = (A - LC)\tilde{x}_k \tag{3}$$

has stable eigenvalues, then the global asymptotic stability of the observer is guaranteed [9].

It is straightforward to extend the Luenberger observer to nonlinear system. Consider a nonlinear time-varying dynamic system given by

$$\begin{aligned} x_{k+1} &= f(x_k, u_k, k) \\ y_k &= g(x_k, u_k, k) \end{aligned} \tag{4}$$

the extended Luenberger observer is specified by

$$\begin{aligned} \hat{x}_{k+1} &= f(\hat{x}_k, u_k, k) + L(y_k - \hat{y}_k) \\ \hat{y}_k &= g(\hat{x}_k, u_k, k) \end{aligned} \tag{5}$$

The extended Luenberger observer (5) is intended to deal with deterministic nonlinear dynamic system, it is also possible to apply the formulation to stochastic systems. The issue of stochastic state estimation arises when the noise and disturbance acting on state transition equations and measurements are considerable and cannot be satisfactorily filtered out [10]. Consider a nonlinear, time-varying, stochastic dynamic system described by equations

$$\begin{aligned} x_{k+1} &= f(x_k, u_k, k) + v_k \\ y_k &= g(x_k, u_k, k) + w_k \end{aligned} \tag{6}$$

where v_k, w_k are zero-mean white noise for state transition and measurement respectively. The restriction to zero-mean noise is not a loss of generality. We can always add one more dimension to state transition equation and measurement to take care of a nonzero mean noise. A proportional extended Luenberger observer is defined as in (5). In the stochastic state estimation scenario, zero mean and small variance are typical desired characteristics of the estimation error.

For the Luenberger observer (2), there is a solid analytical method to select the observer gain L such that the observer will behave according to control design requirements. But such a method is not yet available for the extended Luenberger observer (5). Instead of prespecifying the observer gain, we apply the information theoretic learning technique to update the observer gain during the course of state estimation so that the entropy of the error between the measurement and estimated output is minimized at each step.

III. INFORMATION-THEORETIC LEARNING

Information theoretic learning (ITL) is a signal processing technique that combines information theory and adaptive systems. ITL utilizes information theory as a criterion to update the structure of adaptive system in order to achieve a certain performance [11]. Traditionally, mean-square-error (MSE) is the optimality criterion to perform supervised training of adaptive systems. The main reason for the wide use of MSE resides in the fact that quadratic criteria combined with linear systems result in analytically tractable mathematics and lead to solutions like the Wiener-Hopf equation [12]. For linear systems and Gaussian distributed signals, second-order statistics are able to extract all the information present in the data, thus yield optimal training solutions in an information theoretic perspective. For example, the well-known Kalman filter, using the MSE criterion, from adaptive signal processing point of view, is the optimal filter in the information theoretic sense, since it deals with linear systems corrupted with white Gaussian noise [8].

However, many contemporary signal processing problems extend beyond the linearity and Gaussianity assumptions, therefore to achieve optimality in an information theoretic framework, one has to go beyond second-order statistics as optimality criteria. To this end, we need to consider the higher-order statistics of the signals since arbitrary distributions, unlike the Gaussian, are not only characterized by their 2nd-order statistics.

Information theoretic criteria provide natural and intuitive means of dealing with higher-order statistics of the signals, since they are derived based on particular postulates such as additivity [7]. Entropy, which measures the average information content in a random variable with a particular probability distribution was previously proposed as a criterion for supervised adaptive filter training and it was shown to provide better neural network generalization compared to MSE [13].

Given a random variable X with probability distribution function (pdf) $f_X(x)$, Shannon's entropy is defined by [14]

$$H_s(X) = -\int f_X(x) \log f_X(x) dx \tag{7}$$

One drawback of using Shannon's entropy as a cost function in adaptive signal processing is that it is difficult to estimate the quantity directly from data samples. In fact, Renyi's entropy, which includes Shannon's entropy as a special case, leads to a practical estimator for entropy directly from data when combined with a nonparametric estimator. Renyi's entropy of order-α is given by [15]

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \int f_X^\alpha(x) dx = \frac{1}{1-\alpha} \log E[f_X^{\alpha-1}(x)] \tag{8}$$

As can be shown, using L'Hopital's rule, the limit of Renyi's entropy, as α approaches to 1, yields Shannon's entropy. In order to estimate Renyi's entropy directly from data samples $\{x_1, x_2, \ldots x_N\}$, Parzen windowing with kernel function $\kappa_\sigma(.)$ is employed [16]. Approximating the expectation operator with sample mean, we obtain the following estimator for Renyi's entropy [13]

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left[\frac{1}{N^\alpha} \sum_{j=1}^{N} \left(\sum_{i=1}^{N} \kappa_\sigma(x_j - x_i) \right)^{\alpha-1} \right] \tag{9}$$

The parametric definition of Renyi's entropy provides extra freedom to the designer. Most commonly used and easy to evaluate is the quadratic entropy, which is given by

$$H_2(X) = -\log\left[\frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\kappa_\sigma\left(x_j - x_i\right)\right] \qquad (10)$$

When Renyi's entropy is utilized as a cost function that encompasses all the information lies in sample data in supervised learning, it is straightforward and intuitive to minimize the cost function with respect to system structure. Because the "log" function is monotonically increasing, minimizing Renyi's entropy is equivalent to maximizing the argument of logarithm function, which is called the *information potential*.

To apply information-theoretic learning to state estimation, we aim to minimize Renyi's quadratic entropy (10) or equivalently maximize the quadratic information potential of the error, defined as the difference between measurements y_k and estimated output $\hat{y}_k$, with respect to observer gain *L*. To this end, a gradient-based learning algorithm is developed to update observer gain *L* during the course of state estimation so that the estimated state will approach to the true state asymptotically in a statistical sense.

IV. STOCHASTIC GRADIENT ALGORITHM

Given the data sample set $\{u_k, y_k\}_{k=1}^{N}$ up to time step *N*, the entropy observer will generate the estimated state x_N such that Renyi's quadratic entropy of error will be minimized. When we directly apply Renyi's quadratic entropy in (10) to state estimation, the algorithm suffers from *O(N²)* computational complexity since quadratic entropy (10) is a batch method which needs all the previous data samples. To reduce the computational complexity, we derive the stochastic information gradient (SIG) for Renyi's quadratic entropy so that the algorithm can handle online, instantaneous computation for state estimation.

Dropping the expectation and evaluating its argument at the most recent sample of a random variable *X*, we obtain the stochastic gradient for quadratic entropy [17],

$$H_2(X) = -\log E[f_X(x)] \approx -\log f_X(x_k) \qquad (11)$$

x_k denotes the most recent data. Since the probability density function (pdf) of *X* is unknown in practice, we use Parzen windowing to estimate it. In a nonstationary scenario, the online pdf estimator can be obtained using a sliding window. Assuming a length-*W* window of data samples, the stochastic pdf estimator evaluated at x_k is

$$\hat{f}_X(x_k) = \frac{1}{W}\sum_{i=k-W}^{k-1}\kappa_\sigma(x_i - x_k) \qquad (12)$$

Hence, the stochastic quadratic entropy at time step *k* is given by

$$\hat{H}_2(X) = -\log\frac{1}{W}\sum_{i=k-W}^{k-1}\kappa_\sigma(x_i - x_k) \qquad (13)$$

Define the instantaneous error $e_k = y_k - \hat{y}_k$, then the cost function used to update the extended Luenberger observer gain *L* would be the stochastic quadratic information potential of error signal, i.e.

$$J = \frac{1}{W}\sum_{i=k-W}^{k-1}\kappa_\sigma(e_i - e_k) \qquad (14)$$

Applying the chain rule and taking the derivative of *J* with respect to the observer gain *L*, we get the update rule for the observer gain *L* for every time step *k*, thus the observer is updated during the course of state estimation.

Suppose the dynamic system is multi-input and multi-output (MIMO), *L* will be a matrix of dimension *n-by-t*, we can derive the stochastic gradient with respect to each column of *L*, denoted by $L_{:,j}$.

$$\frac{\partial J}{\partial L_{:,j}} = \frac{1}{W}\sum_{i=k-W}^{k-1}\kappa'_\sigma(e_i - e_k)\left[g'(\hat{x}_k, u_k, k)\frac{\partial \hat{x}_k}{\partial L_{:,j}} - g'(\hat{x}_i, u_i, i)\frac{\partial \hat{x}_i}{\partial L_{:,j}}\right]$$

$$\frac{\partial \hat{x}_k}{\partial L_{:,j}} = \left[f'(\hat{x}_{k-1}, u_{k-1}, k-1) - L \cdot g'(\hat{x}_{k-1}, u_{k-1}, k-1)\right] \cdot \frac{\partial \hat{x}_{k-1}}{\partial L_{:,j}} + \left(y_{k-1} - \hat{y}_{k-1}\right)_j \cdot I_{n\times n}$$

$$g'_j(\hat{x}_k, u_k, k) = \left[\frac{\partial g(\hat{x}_k, u_k, k)}{\partial \hat{x}_k}\right]_j$$

$$L_{:,j} \leftarrow L_{:,j} + \eta \cdot \frac{\partial J}{\partial L_{:,j}} \qquad (15)$$

where η is step size for adaptation. Throughout the paper, a Gaussian kernel is used, although other choices are possible and will be investigated in the future.

It can be shown that the stochastic gradient will update the observer gain *L* in the mean to the minimum error entropy (MEE) optimal values [17].

V. SIMULATIONS

In this section, we demonstrate the performance of the proposed minimum error entropy observer on state estimation for a linear time-invariant system and also for the Van der Pol oscillator. In order to illustrate the performance, we compare the results of entropy observer with the extended Luenberger observer using MSE [6].

A. Linear time-invariant (LTI) system

We first study a linear time-invariant system, here a single-input single-output excited by white noise of

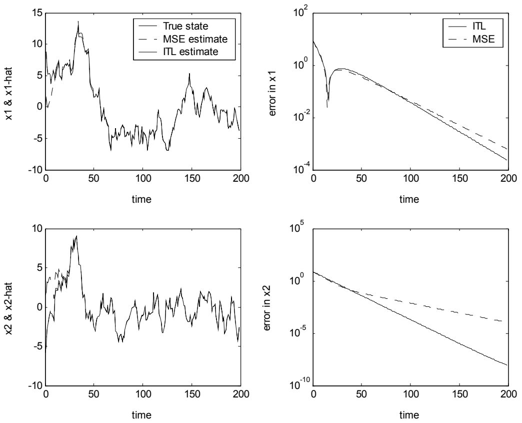

Fig. 1(a) LTI system states, estimates and estimation error without measurement noise

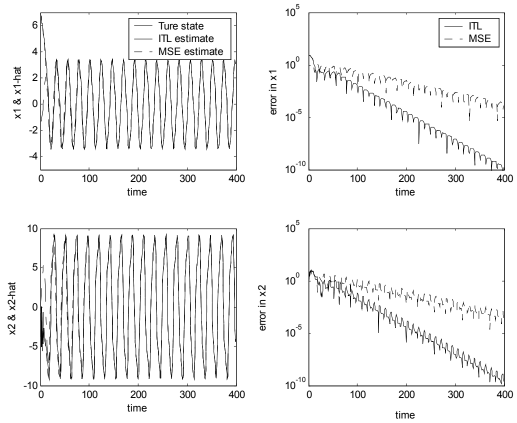

Fig. 2(a) Van del Pol system states, estimates and estimation error without measurement noise

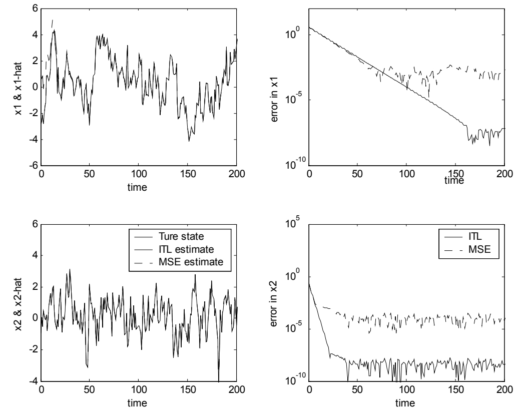

Fig. 1(b) LTI system states, estimates and estimation error with measurement noise

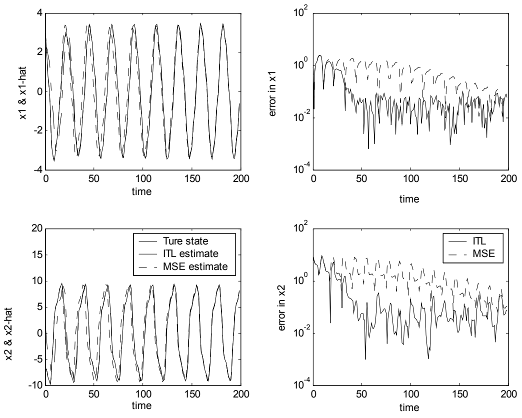

Fig. 2(b) Van del Pol system states, estimates and estimation error with measurement noise

exponential distribution. The system dynamic is given by

$$x_{k+1} = \begin{bmatrix} 0.9 & 0.1 \\ 0 & 0.5 \end{bmatrix} x_k + \begin{bmatrix} 1 \\ -0.9 \end{bmatrix} u_k \qquad (16)$$
$$y_k = \begin{bmatrix} 1 & 0 \end{bmatrix} x_k + w_k$$

where w_k is zero-mean white noise with any arbitrary distribution.

Two simulation results are presented in Fig. 1. The first one is without measurement noise, i.e. w_k =0. The left two subplots in each figure present the true state and estimated states using ITL and MSE, and the right two subplots give the state estimation error. Notice that in Fig. 1(a), the absolute value of the state estimation errors using ITL decay exponentially much faster than the one using MSE, which means entropy observer converges to the true state much faster than the one trained by MSE. Whereas, in Fig. 1(b), the measurements are corrupted with zero-mean white noise of uniform distribution at 15dB signal-to-noise ratio (SNR), the simulation results suggest that the state estimation errors from the entropy observer not only decay faster, but also achieve a much lower level, up to 10^{-5} difference, than the one by MSE. Since the noise is not Gaussian, ITL extract more information from the error sequence than MSE.

B. Van der Pol Oscillator

Next, we apply the minimum error entropy Luenberger observer to the discretized (first-order difference) Van der Pol oscillator dynamic system. The system is characterized by the following equations.

$$x_{1,k+1} = x_{1,k} + T \cdot x_{2,k}$$
$$x_{2,k+1} = x_{2,k} - 9T \cdot x_{1,k} + \mu \cdot T\left(1 - x_{1,k}^2\right) x_{2,k} \qquad (17)$$
$$y_k = x_{1,k} + w_k$$

where T is the sampling time used in discretization and the smaller T the better approximation. Note that though the continuous Van der Pol oscillator is globally stable, the first-order discretization leads to instability in parts of state space. In the case that the state trajectory goes through this unstable region, neither the entropy observer nor the MSE extended Luenberger observer can follow the diverging trajectory. But, as long as the state trajectory remains in the stable region, both observers converge smoothly.

In simulations, we used a sampling time of T=0.1 and the oscillator parameter is μ=0.5. The state estimation errors exhibit similar behavior to that observed in the LTI system. The first simulation corresponds to noiseless measurements. As can be seen from the results, the entropy observer converges faster than the extended Luenberger observer using MSE. Under the noisy measurement case, where the white noise of uniform distribution is 15dB, the state estimation error from entropy observer yields a lower bound than that of extended Luenberger observer by MSE. The bound is affected by the measurement noise.

Case studies from linear time-invariant system and nonlinear dynamic system suggest that entropy observer outperforms the extended Luenberger observer using MSE in terms of faster convergence and smaller steady-state estimation error.

VI. ALGORITHM DISCUSSIONS

A. Computational Complexity

It is important that the developed algorithm for state estimation problems be on-line. By introducing the Stochastic Information Gradient (SIG), we transformed a batch algorithm to an online one and reduced the computational complexity from $O(N^2)$ to $O(W)$, where N is the total data up to time step N and W is the window length used in SIG. The designer considers a trade-off in choosing the window length W, since smaller W results in less computational complexity, while larger W is required to improve estimation accuracy and reduce misadjustment.

B. Mean-invariance of Entropy

Entropy is mean-invariant, which means its value remains constant even if the mean has been shifted. Let $x' = x + \xi$, where ξ is a replacement, then

$$\begin{aligned} H_\alpha(X') &= \frac{1}{1-\alpha}\log\int f_{X'}^{\alpha}(x+\xi)dx' \\ &= \frac{1}{1-\alpha}\log\int f_{X}^{\alpha}(x)dx = H_\alpha(X) \end{aligned} \tag{18}$$

Thus training with entropy will lead to a set of optimal weights. This is the reason that we have to properly modify the output system bias to yield zero mean error over the training data set [13]. However, we don't need any extra processing about mean-invariance in the proposed entropy observer. The reason is that the bias is feedback to the observer and it will decay exponentially to zero as long as the entropy observer is stable.

C. Optimization issue

Training with information-theoretic learning may exhibit some local minima [13]. From the extensive simulations above, we notice that gradient descent algorithm sometimes failed to reach the global minimum. The kernel size σ in the Parzen window $\kappa_\sigma(.)$ controls the smoothness of the error information potential. A kernel annealing approach is proposed to update the kernel size during adaptation in order to achieve the global minimum [18]. One problem with kernel annealing is how to set the annealing rate (a common unsolved problem in all variants of stochastic annealing). This issue will be addressed in a future paper, and in this preliminary report, we use a constant kernel size for simplicity.

VII. KERNEL METHODS

Kernel methods have become a hot research topic in the machine learning community since the introduction of Support Vector Machines (SVM) [19]. Kernel-based algorithms are nonlinear versions of linear algorithms where the data has been nonlinearly transformed to a high dimensional feature space where we only need to compute the inner product via the kernel function. Kernel methods have been successfully used in classification, regression and data analysis [19]. In essence, the eigendecomposition of a positive function (the kernel) is utilized to define the following inner product for the transformation space:

$$\kappa_\sigma(x-x') = \sum_{k=1}^{\infty}\lambda_k\varphi_k(x)\varphi_k(x') = \langle\Phi(x),\Phi(x')\rangle \tag{19}$$

Recently we formulated information-theoretic learning based on Parzen window density estimators as a kernel method, which allows us to address the kernel methods from an information processing point view [20]. Revisiting the information potential for order-2 Renyi's entropy in (10), we notice that

$$\begin{aligned} V(X) &= \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\kappa_\sigma(x_j - x_i) = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\langle\Phi(x_j),\Phi(x_i)\rangle \\ &= \left\langle\frac{1}{N}\sum_{i=1}^{N}\Phi(x_i),\frac{1}{N}\sum_{j=1}^{N}\Phi(x_j)\right\rangle = \langle\mathbf{m}^\Phi,\mathbf{m}^\Phi\rangle \\ &= \left\|\mathbf{m}^\Phi\right\|^2 \end{aligned} \tag{20}$$

where m^Φ is the mean vector of the transformed data. Thus, the quadratic information potential turns out to be the inner product of the mean vector of the nonlinearly transformed data in the Hilbert kernel space. By applying information-theoretic learning technique to the extended Luenberger observer for nonlinear dynamic systems, we are essentially transforming the input data to a high dimensional feature space so that a second-order, linear algorithm is performed.

The authors hypothesize that a kernel Kalman filter can be developed where the Kalman filter equations can be applied to the nonlinearly transformed dynamic system in the high dimensional Hilbert kernel space. The mapping back to original input space enables us to deal with nonlinear, non-Gaussian state estimation problems. In order to achieve this, future work will involve merging dynamical state estimation with kernel methods.

VIII. CONCLUSIONS

In this paper, we developed an error-entropy-based observer to deal with nonlinear state estimation problems. We constructed an adaptive Luenberger observer based on information theoretic learning rather than trying to pre specify a globally stable observer gain vector. A stochastic gradient-based algorithm was developed for feasible real-time implementation.

The performance of the proposed observer is evaluated in linear time-invariant and Van der Pol oscillator systems by comparing with a similar adaptive Luenberger observer trained using the mean-square-error criterion. The simulation results suggested that the entropy observer converges faster to the true state and has lower state estimation error than its square-error counterpart.

We also pointed out that the proposed entropy observer forms a link between the popular kernel methods in machine learning and state estimation. Future work will involve developing a "Kernel Kalman Filter" to address nonlinear non-Gaussian state estimation.

REFERENCES

[1] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," Trans. of ASME, Journal of Basic Engineering, 82(Series D), pp. 35-45, 1960.

[2] D. G. Luenberger, "Observers for Multivariable Systems," IEEE Trans. Automatic Control, vol. 11, no. 2, pp. 190-197, 1966.

[3] D. G. Luenberger, "An Introduction to Observers," IEEE Trans. Automatic Control, vol. AC-16, no. 6, pp. 596-602, 1971.

[4] C. Elmas, H. Zelaya de la Parra, "Application of a full-order Luenberger Observer for a Position Sensorless Operation of a Switched Reluctance Motor Drive," IEE proceedings on Control Theory Application, Vol. 143, no. 5, 1996.

[5] T. Du, M.A. Brdys, " Implementation of Extended Luenberger Observer for Joint State and Parameter Estimation of PWM Induction Motor Drive," Proceedings of the 5th European Conference on Power Electronic Applications, vol. 4, pp. 439-444, 1993.

[6] D. Erdogmus, A. U. Genc, J.C. Principe, "A Neural Network Perspective to Extended Luenberger Observers," Measurement and Control, Special Feature on Recent Advances in Neural Networks, Vol.35, No 1, Feb., 2002.

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

[8] X. Feng, K.A. Loparo, Y. Fang, "Optimal State Estimation with Active Probing for Stochastic Systems: An Information Theoretic Approach," IEEE Transaction on Automatic Control, vol. 42, no. 6, pp. 771-785, 1997

[9] T. Kailath, *Linear Systems,* Prentice-Hall, Englewood Cliffs, NJ, 1985.

[10] T. P. McGarty, *Stochastic Systems and State Estimation*, John Wiley & Sons, NY, 1974.

[11] J.C. Principe, D. Xu, J. Fisher, "Information Theoretic Learning," in *Unsupervised Adaptive Filtering*, (Ed. S. Haykin), Wiley, NY, 2000, pp. 265-319.

[12] B. Widrow and S. D. Stearns, *Adaptive Signal Processing,* Prentice-Hall, Englewood Cliffs, NJ, 1985.

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

[14] C. E. Shannon, "A Mathematical Theory of Communication," Bell Sys. Tech. J., vol. 27, pp. 379-423, 623-653, 1948.

[15] A. Renyi, *Probability Theory,* Elsevier, NY, 1970.

[16] E. Parzen, "On Estimation of a Probability Density Function and Mode," in *Time Series Analysis Papers*, Holden-Day, CA, 1967.

[17] D. Erdogmus, J. C. Principe, K. E. Hild II, "On-Line Entropy Manipulation: Stochastic Information Gradient," Signal Processing Letter, Vol. 10, No. 8, pp. 242-245, 2003.

[18] D. Erdogmus, J.C. Principe, "Convergence Properties and Data Efficiency of the Minimum Error Entropy Criterion in ADALINE Training," IEEE Trans. on Signal Processing, vol. 51, no. 7, pp. 1966-1978, Jul 2003.

[19] B. Scholkopf, A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, & Beyond*, MIT, MA, 2001.

[20] R. Jenssen, D. Erdogmus, J. C. Principe, T. Eltoft, "Towards a unification of information theoretic learning and kernel methods," in proceedings of Machine Learning for Signal Processing, São Luís, Brazil, 2004.