DETECTION OF NONLINEARLY DISTORTED SIGNALS USING MUTUAL INFORMATION

Umut Ozertem[1], Deniz Erdogmus[1], Ignacio Santamaria[2]
[1] CSEE Department, Oregon Health & Science University, Portland, Oregon, USA
[2] GTAS, Department of Communications, University of Cantabria, Santander, Spain

ABSTRACT

The traditional matched filter is the optimal signal detector under quite restrictive conditions, such as linearity and Gaussianity. In this paper, a nonlinear filter topology based on mutual information is proposed to exploit higher order statistics instead of linear second-order measures. Results demonstrate superior performance in nonlinear amplitude and temporal signal distortion situations.

1. INTRODUCTION

Detection of a *known* waveform in noise is an important fundamental problem having a wide range of applications, communications, radar, and biomedical engineering to name just a few. Under the additive white Gaussian noise (AWGN) and linear channel assumptions, optimal detection is achieved by the conventional *matched filter*. However, if the noise distribution is non-Gaussian or the waveform suffers a nonlinear distortion, the matched filter becomes sub-optimal in signal detection performance, since it relies on correlation. Besides, the matched filter method, by definition, assumes that the exact form of signal that is to be detected is known and time invariant. To overcome these shortcomings of the matched filter, we propose a nonlinear filter topology based on a mutual information (MI) criterion.

In earlier work, it is demonstrated that MI-based methods are superior to the second order statistical measures in nonlinear signal processing [1]. Hence, in detection of nonlinearly distorted signals in noise, a suitable criterion is the mutual information (MI) between the filter output and the class label (throughout the paper class refers to the two cases of the signel being present or not in the received signal and the class label for these cases are 1 and 0, respectively). This choice is motivated by lower and upper bounds in information theory that relate this quantity to probability of classification error. In principle, MI measures nonlinear dependencies between a set of random variables taking higher order statistical structures existing in the data into account, as opposed to linear and second-order statistical measures such as correlation and covariance [2].

In this paper, we propose a method for determining an optimal nonlinear filter that maximizes the Shannon mutual information between the filter output and the class label, hence improves the signal detection and false alarm performance for nonlinear channels and non-Gaussian noise distributions. Illustrative performance comparisons are carried out using typical nonlinearities and noise distributions assumed in communication channel models.

2. THEORETICAL BACKGROUND

In pattern recognition, it is well known that the average probability of classification error is related to the MI between the feature vectors and the class labels. Specifically, Fano's and Hellman & Raviv's bounds demonstrate that probability of error is bounded from below and above by quantities that depend on the Shannon MI between these variables [3,4]. Maximizing this MI reduces both bounds, therefore, forces the probability of error to decrease.

Although Shannon's MI is traditionally used as a measure of shared information, fundamentally it is a measure of divergence from independence for two random variables. In this nonlinear filtering approach, we are interested in the MI between the continuous-valued y, which is the nonlinear filter output, and the discrete-valued class label c. Shannon MI between y and c is defined in terms of the entropies of the overall data and the individual classes as [2]:

$$I_S(y;c) = H_S(y) - \sum_c p_c H_S(y \mid c) \tag{1}$$

where p_c are the prior class probabilities. The Shannon entropy is given by

$$\begin{aligned} H_S(y) &= -\int p(y) \log p(y) dy \\ H_S(y \mid c) &= -\int p(y \mid c) \log p(y \mid c) dy \end{aligned} \tag{2}$$

where $p(y|c)$ are the class conditional distributions and the overall data distribution is

$$p(y) = \sum_c p_c \, p(y \mid c) \tag{3}$$

Under this framework of *nonlinear filtering that maximizes mutual information of the output with class labels*, the adaptive learning procedure to find these optimal projections follows the block diagram shown in Fig. 1. First, the input samples are shifted through a delay line. The length of the delay line is equal to the length of the signal to be detected, assuming the signal length is known and constant. The obtained input vector is fed into the nonlinear filter, which generates the test statistic to be thresholded for signal detection. The filter contains a weight vector **v** that needs to be optimized to maximize the MI criterion [5,6,7].

The conditional class entropies and the overall data entropy have to be estimated in order to approximate MI. In our system, a KDE-based plug-in estimator [8,9,10] is used for this purpose. Given a set of independent and identically distributed (iid) samples $\{y_1,\ldots y_N\}$, which can be partitioned into subsets corresponding to each class as $\{y_1^c,\ldots,y_{N_c}^c\}$, the entropies in (1) can be estimated by [9]:

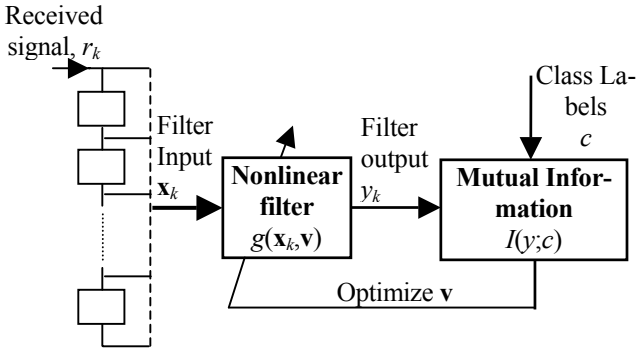

Figure 1. Detecting signal in noise using mutual information.

$$H_S(\mathbf{y}) = -\frac{1}{N}\sum_{j=1}^{N}\log\frac{1}{N}\sum_{i=1}^{N}K(\mathbf{y}_j-\mathbf{y}_i)$$
$$H_S(\mathbf{y}\mid c) = -\frac{1}{N_c}\sum_{j=1}^{N_c}\log\frac{1}{N_c}\sum_{i=1}^{N_c}K(\mathbf{y}_j^c-\mathbf{y}_i^c) \quad (4)$$

3. NONLINEAR ADAPTIVE FILTERING

Given the received signal r_k, the time-delay vector samples $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}$ are constructed from the shifted samples of r_k. Each vector is also associated with a class label (0 or 1) yielding $\{c_1,c_2,\ldots,c_N\}$. We are interested in finding a nonlinear transformation $y=g(\mathbf{x})$ such that the Shannon MI between the projection and the class label is maximized.

According to the theory of reproducing kernels for Hilbert spaces (RKHS), the eigenfunctions $\{\overline{\varphi}_1(\mathbf{x}),\overline{\varphi}_2(\mathbf{x}),...\}$ collected in vector notation as $\overline{\boldsymbol{\varphi}}(\mathbf{x})$ of a kernel function K that satisfy the Mercer conditions [11] form a basis for the Hilbert space of finite-power nonlinear functions [12,13].[1] Therefore, every finite-L_2-norm nonlinear transformation $g(\mathbf{x})$ can be expressed as a linear combination of these bases:

$$y = g(\mathbf{x}) = \mathbf{v}^T\overline{\boldsymbol{\varphi}}(\mathbf{x}) \quad (5)$$

As we will show next, such linear combinations of nonlinear basis functions arise naturally from the KDE-based nonparametric estimates of mutual information in the context of feature subspace projections.

Consider the Shannon mutual information between the high-dimensional filter input vector and the class label.

$$\begin{aligned} I_S(\mathbf{x};c) &= \sum_c\int p_{\mathbf{x}c}(\mathbf{x},c)\log\frac{p_{\mathbf{x}c}(\mathbf{x},c)}{p_{\mathbf{x}}(\mathbf{x})p_c}d\mathbf{x} \\ &= \sum_c p_c\int p_{\mathbf{x}|c}(\mathbf{x}\mid c)\log\frac{p_{\mathbf{x}|c}(\mathbf{x}\mid c)}{p_{\mathbf{x}}(\mathbf{x})}d\mathbf{x} \\ &= \sum_c p_c E_{\mathbf{x}|c}\left[\log\frac{p_{\mathbf{x}|c}(\mathbf{x}\mid c)}{p_{\mathbf{x}}(\mathbf{x})}\right] \end{aligned} \quad (6)$$

The pdfs $p_{\mathbf{x}|c}$ and $p_{\mathbf{x}}$ in (6) are estimated using KDE with $K(.)$ as the kernel. The conditional expectation can be approximated by a sample mean over the appropriate samples. This leads to

$$I_S(\mathbf{x};c) \approx \sum_c\frac{p_c}{N_c}\sum_{j=1}^{N_c}\log\frac{(1/N_c)\sum_{i=1}^{N_c}K(\mathbf{x}_j^c-\mathbf{x}_i^c)}{(1/N)\sum_{i=1}^{N}K(\mathbf{x}_j^c-\mathbf{x}_i)} \quad (7)$$

Assuming that K is a Mercer kernel (with some abuse of notation) we can write $K(\mathbf{x}-\mathbf{x}')=\overline{\boldsymbol{\varphi}}^T(\mathbf{x})\overline{\boldsymbol{\Lambda}}\overline{\boldsymbol{\varphi}}(\mathbf{x}')$. Consequently, the mutual information estimate becomes

$$I_S(\mathbf{x};c) \approx \sum_c\frac{p_c}{N_c}\sum_{j=1}^{N_c}\log\left|\frac{N\overline{\boldsymbol{\varphi}}^T(\mathbf{x}_j^c)\overline{\boldsymbol{\Lambda}\boldsymbol{\Phi}}_{\mathbf{x}}\mathbf{m}_c}{N_c\overline{\boldsymbol{\varphi}}^T(\mathbf{x}_j^c)\overline{\boldsymbol{\Lambda}\boldsymbol{\Phi}}_{\mathbf{x}}\mathbf{1}}\right| \quad (8)$$

where we define the membership vector $\mathbf{m}_c$ for each class c, such that $\mathbf{m}_{ci}=1$ if $c_i=c$, 0 otherwise, and the vectors $\mathbf{e}_i$ whose i^{th} entry is 1 and all others are zeros, as well as a vector of ones, denoted by $\mathbf{1}$. In addition, we introduced the matrix $\overline{\boldsymbol{\Phi}}_{\mathbf{x}}=[\overline{\boldsymbol{\varphi}}(\mathbf{x}_1)\cdots\overline{\boldsymbol{\varphi}}(\mathbf{x}_N)]$, where $N=N_0+N_1$. Defining the average vectors of the transformed features for each class and for the whole training set as $\overline{\boldsymbol{\mu}}_c=(1/N_c)\overline{\boldsymbol{\Phi}}_{\mathbf{x}}\mathbf{m}_c$ (for the feature vectors from class c) and $\overline{\boldsymbol{\mu}}=(1/N)\overline{\boldsymbol{\Phi}}_{\mathbf{x}}\mathbf{1}$ (for the whole data set), we equivalently obtain:

$$I_S(\mathbf{x};c) \approx \sum_c\frac{p_c}{N_c}\sum_{j=1}^{N_c}\log\left|\frac{\overline{\boldsymbol{\varphi}}^T(\mathbf{x}_j)\overline{\boldsymbol{\Lambda}}\overline{\boldsymbol{\mu}}_c}{\overline{\boldsymbol{\varphi}}^T(\mathbf{x}_j)\overline{\boldsymbol{\Lambda}}\overline{\boldsymbol{\mu}}}\right| \quad (9)$$

Note that so far we have only utilized the true eigenfunctions and the eigenvectors of the kernel function. According to our projection model in (5), the projection is accomplished in the kernel-induced $\boldsymbol{\varphi}$-space, and the best L_2-orthogonal approximation for $\overline{\boldsymbol{\varphi}}(\mathbf{x})$ is given by

$$\overline{\boldsymbol{\varphi}}(\mathbf{y}) = \mathbf{v}\mathbf{v}^T\overline{\boldsymbol{\varphi}}(\mathbf{x}) \quad (10)$$

This leads to the following cost function that needs to be maximized by optimizing the weight vector $\mathbf{v}$:

$$J(\mathbf{v}) = \sum_c\frac{p_c}{N_c}\sum_{j=1}^{N_c}\log\left|\frac{\overline{\boldsymbol{\varphi}}^T(\mathbf{x}_j)\mathbf{v}\mathbf{v}^T\overline{\boldsymbol{\Lambda}}\mathbf{v}\mathbf{v}^T\overline{\boldsymbol{\mu}}_c}{\overline{\boldsymbol{\varphi}}^T(\mathbf{x}_j)\mathbf{v}\mathbf{v}^T\overline{\boldsymbol{\Lambda}}\mathbf{v}\mathbf{v}^T\overline{\boldsymbol{\mu}}}\right| \quad (11)$$

In practice, analytical expressions for the (infinitely many) eigenfunctions of the kernel function are not available. However, suitable approximates can be obtained using the available training samples. Spectral methods provide the necessary tools to achieve this. Following the common procedure in spectral methods, using all training samples in pairs as $\mathbf{K}_{ij}=K(\mathbf{x}_i-\mathbf{x}_j)$, we define the affinity matrix. The matrix $\mathbf{K}$ can be decomposed into its eigenvalues and eigenvectors as $\mathbf{K}=\boldsymbol{\Phi}_{\mathbf{x}}^T\boldsymbol{\Lambda}\boldsymbol{\Phi}_{\mathbf{x}}$, which are essentially approximations of the sought eigenfunctions and eigenvalues of the kernel function. Specifically, according to the Nystrom routine [14], the eigenfunctions can be approximated using the eigendecomposition of the affinity matrix $\mathbf{K}$ as follows:

$$\boldsymbol{\varphi}(\mathbf{x}) = \sqrt{N}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Phi}_{\mathbf{x}}\mathbf{k}(\mathbf{x}) \quad (12)$$

where $\mathbf{k}(\mathbf{x})=[K(\mathbf{x}-\mathbf{x}_1),\ldots,K(\mathbf{x}-\mathbf{x}_N)]^T$.

With this substitution, the nonlinear feature transformations become $\mathbf{y}=\mathbf{v}^T\boldsymbol{\varphi}(\mathbf{x})$. Using this nonparametric approximation, the estimate for the criterion in (11) becomes

[1] The bar denotes the true eigenfunctions/values of the kernel.

$$J(\mathbf{v}) = \sum_c p_c \log \left| \frac{\mathbf{v}^T \boldsymbol{\mu}_c}{\mathbf{v}^T \boldsymbol{\mu}} \right| \tag{13}$$

where $\boldsymbol{\mu}_c = (1/N_c)\boldsymbol{\Phi}_{\mathbf{x}}\mathbf{m}_c$ and $\boldsymbol{\mu} = (1/N)\boldsymbol{\Phi}_{\mathbf{x}}\mathbf{1}$ are the class and overall mean vectors of the data in the $\boldsymbol{\varphi}$-space. It is important to note that the class priors p_c are estimated from the training data by N_c/N. Imposing the constraint $\mathbf{v}^T\mathbf{v}=1$, we need to maximize (13). A very important observation is that these mean vectors are orthogonal to each other with their individual norms equal to $p_c^{-1/2}$, p_c being the class prior probability. This is due to the fact that the data transformations are calculated using (12) for both training and testing data. To see this consider the following:

$$\boldsymbol{\mu}_c = \frac{1}{N_c}\sum_{j=1}^{N_c}\sqrt{N}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Phi}_{\mathbf{x}}\mathbf{k}(\mathbf{x}_j^c) = \ldots \approx \frac{\sqrt{N}}{N_c}\boldsymbol{\Phi}_{\mathbf{x}}\mathbf{m}_c \tag{14}$$

Now the inner product between two mean vectors is:

$$\boldsymbol{\mu}_c^T\boldsymbol{\mu}_d = \frac{N}{N_c N_d}\mathbf{m}_c^T\mathbf{m}_d = \begin{cases} N/N_c & \text{if } c = d \\ 0 & \text{if } c \neq d \end{cases} \tag{15}$$

Thus, the mean vectors of each class in the $\boldsymbol{\varphi}$-space create an orthogonal (but not normal) basis for the space in which our optimization variable $\mathbf{v}$ lies in. Defining a basis matrix $\mathbf{M} = [\boldsymbol{\mu}_0\ \boldsymbol{\mu}_1]$, which satisfies $\mathbf{M}^T\mathbf{M}=\mathbf{P}^{-1}$, where $\mathbf{P}=diag(p_1,\ldots,p_C)$, we can express $\mathbf{v}$ as

$$\mathbf{v} = \mathbf{M}\mathbf{P}^{1/2}\boldsymbol{\alpha} \tag{16}$$

where $\boldsymbol{\alpha}^T\boldsymbol{\alpha}=1$. Using (16), and the identities $\boldsymbol{\mu} = \mathbf{M}\mathbf{p}$ and $\mathbf{M}^T\boldsymbol{\mu}_c = p_c^{-1}\mathbf{e}_c$, where $\mathbf{p}$ is the vector of class priors and $\mathbf{e}_c$ is the canonical unit vector in direction c as defined earlier, the maximization problem in (13) can be converted to a problem in terms $\boldsymbol{\alpha}$ subject to $\boldsymbol{\alpha}^T\boldsymbol{\alpha}=1$ as:

$$\max_{\boldsymbol{\alpha}} J(\boldsymbol{\alpha}) = \sum_{c=0}^{1} p_c \log \frac{|\alpha_c| p_c^{1/2}}{\left| \sum_{d=1}^{C} \alpha_d p_d^{1/2} \right|} - \sum_{c=0}^{1} p_c \log p_c \tag{17}$$

Notice that, due to the constraint $\boldsymbol{\alpha}^T\boldsymbol{\alpha}=1$, we can express all feasible solutions of $\boldsymbol{\alpha}$ in terms of rotations of a unit norm vector. For convenience, consider rotations of the form $\boldsymbol{\alpha}=\mathbf{R}\mathbf{q}$, where $\mathbf{q}$ is a vector consisting of entries $q_c=p_c^{1/2}$. With some more manipulations, the solution to the constrained optimisation problem in (17) is found to be $\boldsymbol{\alpha}=[-p_1^{1/2}, p_0^{1/2}]^T$. An extension of the formulation presented here to more than two classes is possible and will be treated in a future publication. In the signal detection scenario (or any other hypothesis testing problem), however, the number of classes is always two.

A crucial issue in the success of the proposed nonlinear detection filter is the suitable selection of the kernel function. A practical consideration in selecting the kernel function in all spectral methods is the selection of the functional form of the kernel as well as the width of the kernel. Typically, this problem is tackled by trying to optimize the parameters for a family of kernels of some specific type. The connection to density estimation, presented in (7), clearly indicates that the kernel function should be selected to match the distribution

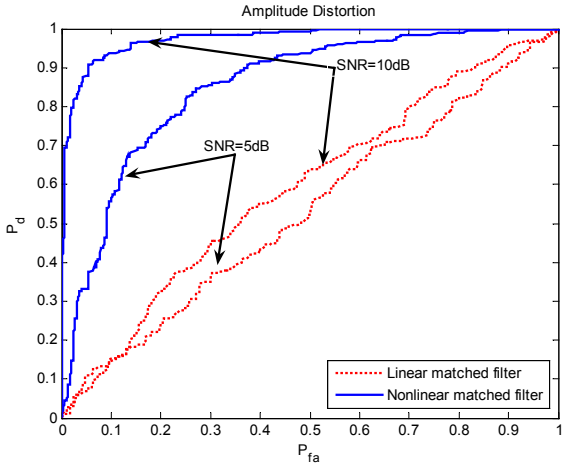

Figure 2. Performance comparison for signal detection in AWGN with nonlinear amplitude distortion.

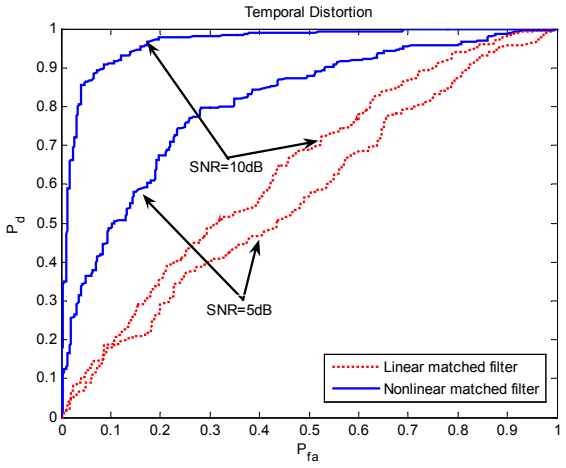

Figure 3. Performance comparison for signal detection in AWGN with temporal distortion.

of the data as much as possible. For simplicity, in the following experiments, a circular Gaussian kernel is assumed and its width parameter (variance) is determined utilizing the rule of thumb by Silverman that gives the *optimal* kernel size for the data set assuming that a Gaussian distribution underlies [15]:

$$\sigma^2 = \frac{1}{n} tr(\boldsymbol{\Sigma}_{\mathbf{x}}) \left(\frac{4}{(2n+1)N} \right)^{2/(n+4)} \tag{20}$$

where n is the dimensionality of the data $\mathbf{x}$, N is the number of samples, and $\boldsymbol{\Sigma}_{\mathbf{x}}$ is the sample covariance of the training set.

4. EXPERIMENTAL RESULTS

The matched filter is known to be optimal under AGN and linear channel assumptions, in which case the proposed nonlinear matched filter has identical receiver characteristics with the matched filter. Additionally, the matched filter structure, by definition, assumes that the signal structure is

known. However, the matched filter becomes sub-optimal for the cases where these assumptions are not valid or there is a perturbation in the received signal from the predetermined one. We demonstrate the performance gain obtained by using the proposed nonlinear matched filter in the following two examples.

Nonlinear Channel Distortion: The matched filter performance declines drastically if the channel introduces nonlinear distortions to the transmitted signal. The detection and false alarm performance of the nonlinear filter is much better as compared to the conventional linear method as shown in Fig. 2. In consistency with the literature on digital communications, the channel nonlinearity in this example is taken to be third order polynomial [16]. The received signal is simply $r_k=h(s_k)+n_k$, where h is the nonlinear distortion and n is AWGN.

Temporal Deviation from the Original Signal: In some biomedical signal detection problems (such has heart-beats and neuronal spiking activity), the target waveform does not always follow the same temporal shape exactly. In such situations, correlating the received signal with a standard template will result in reduced detection performance. To illustrate the performance of the nonlinear matched filter in such situations, here we utilize a sinusoid with *unknown* frequency. The matched filter uses a sinusoid with a slightly perturbed frequency as the template in order to simulate the *rough* approximation effect of the template to the target signal. The results shown on Fig. 3 clearly demonstrate the superior performance of the nonlinear filter.

Upon observing the performance of the nonlinear filter in both experiments, we can also conclude that the performance of this filter does not require the knowledge of the true signal waveform. The filter can easily be modified to accommodate for multiple target signal scenarios by allocating a separate class for each target.

5. CONCLUSIONS

In this paper we proposed a nonlinear matched filter based on mutual information. Although the conventional linear matched filter has optimal performance under linearity and Gaussianity conditions, it loses its optimality in the absence of these requirements. The experimental results showed that the proposed filter is superior to the linear matched filter in the presence of nonlinear distortions, which is an expected consequence of using mutual information as opposed to linear and second-order statistical measures.

REFERENCES

[1] 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.

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

[3] R.M. Fano, *Transmission of Information: A Statistical Theory of Communications*, MIT Press, New York, 1961.

[4] M.E. Hellman, J. Raviv, "Probability of Error, Equivocation and the Chernoff Bound," IEEE Transactions on Information Theory, vol. 16, pp. 368-372, 1970.

[5] J.C. Principe, J.W. Fisher, D. Xu, "Information Theoretic Learning," in *Unsupervised Adaptive Filtering*, S. Haykin Editor, John Wiley & Sons, New York, pp.265-319, 2000,.

[6] K. Torkkola, "Feature Extraction by Non-Parametric Mutual Information Maximization," Journal of Machine Learning Research, vol. 3, pp. 1415-1438, 2003.

[7] K.D. Bollacker, J. Ghosh, "Linear Feature Extractors Based on Mutual Information," *Proceedings of International Conference on Pattern Recognition*, pp. 720-724, Vienna, Austria, 1996.

[8] D. Erdogmus, *Information Theoretic Learning: Renyi's Entropy and its Applications to Adaptive System Training*, PhD Dissertation, University of Florida, Gainesville, Florida, 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] 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.

[11] J. Mercer, "Functions of Positive and Negative Type, and Their Connection with the Theory of Integral Equations," Transactions of the London Philosophical Society A, vol. 209, pp. 415-446, 1909.

[12] G. Wahba, *Spline Models for Observational Data*, SIAM, Philedelphia, Pennsylvania, 1990.

[13] H. Weinert (ed.), *Reproducing Kernel Hilbert Spaces: Applications in Statistical Signal Processing*, Hutchinson Ross Pub. Co., Stroudsburg, Pennsylvania, 1982.

[14] C. Fowlkes, S. Belongie, F. Chung, J. Malik, Spectral Grouping Using the Nystrom Method," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, pp. 298-305, 2004.

[15] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.

[16] X.N. Fernando, A.B. Sesay, "Nonlinear Channel Estimation Using Correlation Properties of PN Sequences," *Proc. IEEE Canadian Conference on Electrical and Computer Engineering*, pp. 469-474, 2001.