IDENTIFICATION AND PREDICTION OF NONLINEAR SYSTEMS USING ORTHONORMAL FUNCTIONS

Iain Scott and Bernard Mulgrew

Dept. of Electrical Eng., The University of Edinburgh, Edinburgh EH9 3JL,
Scotland, U.K., Tel/Fax: +44 [131] 650 5660 / 650 6554. E-Mail: is@ee.ed.ac.uk

ABSTRACT

In a recent paper Mulgrew [1] proposed a nonlinear filtering structure which utilises a set of orthonormal expansions to model nonlinear dynamical systems. Provisional results were presented for a simple 1–dimensional system. In this paper we extend the analysis of this structure to multi–dimensional filtering, and examine the application of the orthonormal structure for nonlinear system identification and communications channel equalisation. The link between the choice of Fourier basis functions and popular kernel probability density estimation techniques is examined.

1 Introduction

Nonlinear adaptive filtering, whether it be for prediction, identification or equalisation, has been of increasing interest during the past decade. Several nonlinear structures have become common, amongst which the multi–layer perceptron, radial basis function network, and the functional–link adaptive neural network have been of particular interest. All these networks share a common philosophy - the input space is expanded via a distributed nonlinearity into a higher dimensional space in which the representation is enhanced. This paper considers a similar structure - in this case an orthonormal set of basis functions are derived for the task.

The fundamentals of this work are drawn from Wiener's work on nonlinear system modeling [2] - a given nonlinear system is modeled by a series of orthogonal nonlinear functions - Wiener chose polynomial series derived from the Volterra series. In this paper we consider a model based on the multi–dimensional Fourier series. This choice of nonlinearity yields several notable advantages:

a) the calculation of the functions is computationally simple - the structure consists of fixed linear combiners and fixed nonlinearities or lookup tables.

b) convergence of gradient descent algorithms should be consistent and optimum.

c) after convergence, the contribution of each nonlinear function to the model will be directly related to the magnitude of the corresponding coefficient in the linear combiner.

The last of these points is the most appealing - this provides a simple method for removing unimportant terms, and thus deriving parsimonious system models.

2 Orthonormal nonlinear filter

The two step nature of the filtering structure is depicted in Figure 1. A N–dimensional input vector $\boldsymbol{x}(k)$ is expanded through the fixed nonlinear expansions $\{g_n(\boldsymbol{x})\}$. These functions are the multi–dimensional Fourier set

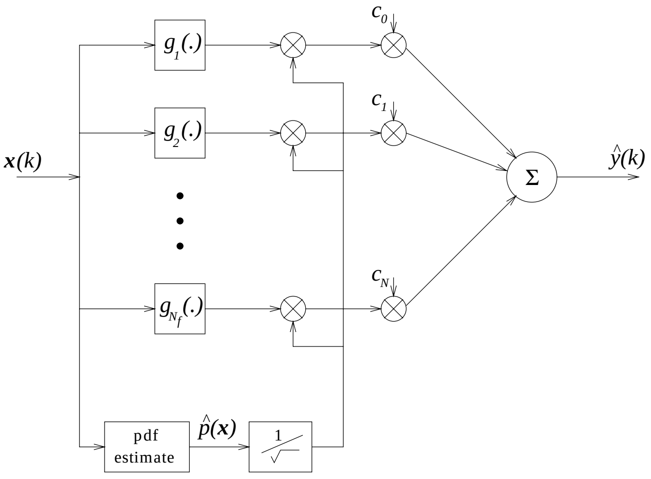

Figure 1: Modified Wiener structure showing signal dependent scaling.

$$g_n(\boldsymbol{x}) \quad = \quad \exp\left\{j\,\boldsymbol{\omega}_n^T \boldsymbol{x}\right\}, \qquad (1)$$

where $\boldsymbol{\omega}_n$ is chosen from a regular lattice in N–dimensional space, and $j = \sqrt{-1}$. These functions themselves form an orthonormal basis in N–dimensional space, however when scaled by the probability density function (pdf) $p(\boldsymbol{x})$, they form an orthonormal basis set for the nonlinear adaptive filter. The combination of a fixed signal–independent nonlinear expansion and a signal–dependent normalisation forms a set of signal–dependent Wiener functions, which satisfy the following orthonormality criterion

$$\int_U f_i(\boldsymbol{x}) f_j(\boldsymbol{x})\, d\boldsymbol{x} \quad = \quad \delta_{ij}, \tag{2}$$

where U is a hypercube in N–dimensional space, and δ_{ij} is the dirac function, and

$$f_i(\boldsymbol{x}) \quad = \quad \frac{1}{\sqrt{p(\boldsymbol{x})}}\, g_i(\boldsymbol{x}). \tag{3}$$

On first inspection, estimating the vector pdf $p(\boldsymbol{x})$ may seem a complex problem, however the choice of a Fourier basis set can lead to a simple estimate of the vector pdf through the characteristic function. The multi–dimensional characteristic function $\boldsymbol{\phi}(\boldsymbol{\omega})$ is defined as follows

$$\boldsymbol{\phi}(\boldsymbol{\omega}) \quad = \quad \int \exp\left\{j\boldsymbol{\omega}^T\boldsymbol{x}\right\}\, p(\boldsymbol{x})\, d\boldsymbol{x}. \tag{4}$$

This can be estimated through the time average of $g_n(\boldsymbol{x})$ as

$$\hat{\boldsymbol{\phi}}(\boldsymbol{\omega}_n) \quad = \quad \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} g_n(\boldsymbol{x}). \tag{5}$$

Given this estimate of the characteristic function we may form an estimate of the pdf $\hat{p}(\boldsymbol{x})$, through the inverse Fourier transform

$$\hat{p}(\boldsymbol{x}) \quad = \quad C\sum_n \hat{\boldsymbol{\phi}}(\boldsymbol{\omega}_n)\, g_n^*(\boldsymbol{x}), \tag{6}$$

where $*$ denotes complex conjugate, and C is some scaling constant. The possibility of (6) producing a complex result is avoided by taking the magnitude of the right hand side as the pdf estimate. The network output is then formed as

$$\hat{y}(k) \quad = \quad \sum_{n=1}^{N_f} c_n f_n(\boldsymbol{x}(k)), \tag{7}$$

where N_f is the number of basis functions.

3 Discussion

There are several advantages to using Fourier series representations, one of which is the computational modularity which they allow. One could envisage each of the basis functions $g_n(\boldsymbol{x})$ being computed by means of a look–up table. Figure 2 illustrates a simple example for an embedding dimension of 3. The elements of the frequency vector $\boldsymbol{\omega}_n$ form the weights in a fixed linear combiner, the output of which is passed to a sine look–up table. Another advantage of Fourier series representations is the ease with which they may be extended to any arbitrary multi–dimensional problem. A further benefit is the similarity which exists between Fourier series density estimators and kernel estimators, another class of commonly used density estimators [3, pp. 36].

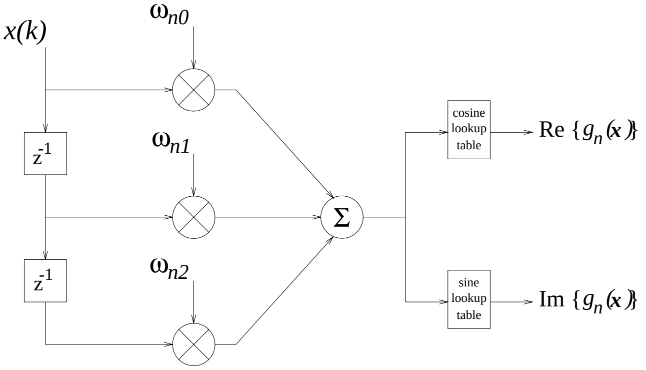

Figure 2: Implementation of $N = 3$ sine lookup table for computing the multi–dimensional Fourier series.

In their simplest form, kernel estimators perform density estimation by centering a scaled function, called the kernel, at each observation of the available data. Suppose we call this kernel $\mathcal{K}$, it is chosen to satisfy

$$\int \mathcal{K}(\boldsymbol{x})\, d\boldsymbol{x} \quad = \quad 1. \tag{8}$$

Usually $\mathcal{K}$ is chosen to be a unimodal probability density function, for example the common normal density function $\mathcal{N}\left(0,\sigma^2\right)$, although kernels which are not densities can also be used. The value of the kernel estimate at some $\boldsymbol{x}$ is simply the average of the n kernel ordinates

$$\hat{p}(\boldsymbol{x}) \quad = \quad K^{-1}\left|\boldsymbol{h}\right|^{-1/2}\sum_{k=1}^{K}\mathcal{K}\left(\boldsymbol{h}^{-T/2}(\boldsymbol{x}-\boldsymbol{x}_k)\right) \tag{9}$$

where the $\boldsymbol{x}_k$ are the previous observations of the random process, and $\boldsymbol{h}$ is called the bandwidth or smoothing parameter. Although not immediately obvious, a kernel estimator can be expressed over any finite interval as a Fourier series. This arises because we may express the kernel function $\mathcal{K}(\boldsymbol{x})$ as the multi–dimensional Fourier series

$$\mathcal{K}(\boldsymbol{x}) \quad = \quad \sum_n b_n \exp\left\{j\boldsymbol{\omega}_n^T\boldsymbol{x}\right\}, \tag{10}$$

in which $b_n = b_{-n}$ are the Fourier coefficients of $\mathcal{K}(\boldsymbol{x})$, and $\boldsymbol{\omega}_n$ is the 'frequency' vector associated with the n–th basis function. Note the limits of the summation are unspecified and may extend to infinity. These types of series–kernel techniques have substantial advantages over parametric techniques which use closed form representations. Now we note that the generalised form of the Fourier series estimator in (6) can be written as

$$\hat{p}(\boldsymbol{x}) \quad = \quad \sum_n b_n\hat{B}_n \exp\left\{j\boldsymbol{\omega}_n^T\boldsymbol{x}\right\}, \tag{11}$$

in which $\hat{B}_n$, $n = 0, \pm1, \pm2, \ldots$ are the sample Fourier series coefficients given by

$$\hat{B}_n \quad = \quad K^{-1} \sum_{k=1}^{K} \exp\left\{-j\boldsymbol{\omega}_n^T \boldsymbol{x}_k\right\}. \tag{12}$$

Note the similarity which exists between this expression and that for the characteristic function of (4). In (11) the $\{b_n\}$ can be thought of as a set of multipliers. Substituting the Fourier series representation of the kernel function (10) in the expression for the kernel estimator (9), for the case $\boldsymbol{h} = \mathbf{1}$

$$\hat{p}\left(\boldsymbol{x}\right) = K^{-1} \sum_{k=1}^{K} \left[\sum_{n} b_n \exp\left\{j\boldsymbol{\omega}_n^T \left(\boldsymbol{x} - \boldsymbol{x}_k\right)\right\} \right], \tag{13}$$

we see that this has the same functional form as the Fourier series estimator of (6) and (11). This similarity between the Fourier series approach taken in the previous section and the general class of kernel estimators is opportune, since it implies that the Fourier series technique of (6) will share many of the properties of kernel estimators, most importantly their ability to generalise to many problems. Having decided upon the Fourier basis set, the most important parameter is the number of basis functions n. In order to attain good estimation of the pdf a large number of basis functions would be desirable, however it is important to avoid the problems of parameter explosion which may result, especially with high embedding dimensions.

4 Examples

A. Communications channel equalisation

In this example a binary sequence $s\left(k\right)$ is transmitted through a dispersive channel then corrupted by additive noise, as illustrated in Figure 3. The transmitted symbol $s\left(k\right)$ is assumed to be an independent sequence of value either +1 or -1 with equal probability. The channel is modeled as a finite impulse response filter with the following transfer function

$$H\left(z\right) \quad = \quad \sum_{i=0}^{N} h_i \, z^{-i}. \tag{14}$$

The additive noise $e\left(k\right)$ is assumed to be a white Gaussian sequence. The task of the equaliser is to reconstruct the original signals $s\left(k\right)$ using the observations of the channel state

$$\boldsymbol{x}\left(k\right) \quad = \quad \left[x\left(k\right) \;\cdots\; x\left(k-m+1\right) \right]^T, \tag{15}$$

where the integer m is known as the order of the equaliser, and the sample delay $\tau = 1$. It is common to introduce a delay Δ to the equaliser so that at the sample instant k the equaliser estimates the input symbol $s\left(k - \Delta\right)$. The symbol decision is taken as follows

$$\hat{s}\left(k - \Delta\right) \quad = \quad \operatorname{sgn}\Big(f\left(\boldsymbol{x}\left(k\right)\right)\Big), \tag{16}$$

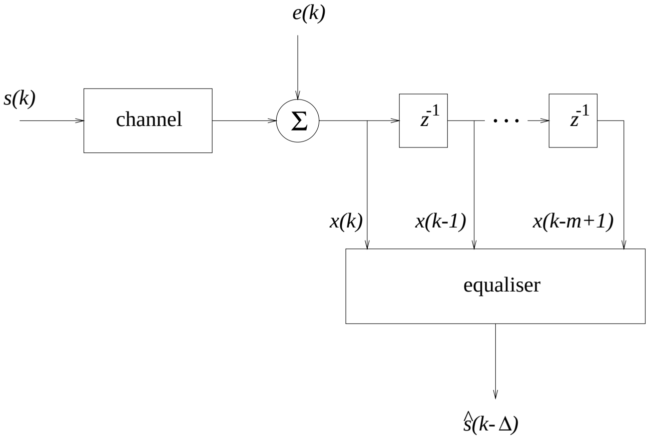

Figure 3: Schematic of a digital data transmission system.

where $f\left(\boldsymbol{x}\left(k\right)\right)$ is the orthonormal equaliser output, and

$$\operatorname{sgn}\left(y\right) \quad = \quad \begin{cases} 1, & y \geq 0 \\ -1, & y < 0 \end{cases} \tag{17}$$

represents a slicer. This operation defines decision regions in the channel state space, and the set of points $\boldsymbol{x}$ that satisfy

$$f\left(\boldsymbol{x}\right) \quad = \quad 0, \tag{18}$$

is termed the decision boundary. This boundary divides the m–dimensional space into two sets X_1 and X_{-1}, into which we partition incoming channel data as either resulting from $\hat{s}\left(k - \Delta\right) = 1$, or $\hat{s}\left(k - \Delta\right) = -1$. An example channel is given by the following transfer function

$$H\left(z\right) \quad = \quad 0.3482 \; + \; 0.8704z^{-1} \; + \; 0.3482z^{-2}(19)$$

An equaliser of order $m = 4$ and delay $\Delta = 1$ was selected. An orthonormal equaliser was formed using $5^4 = 625$ functions and was trained using the LMS algorithm over 2000 data symbols. A linear equaliser of order 4, and a RBF network of 64 centres trained using the orthogonal least squares (OLS) algorithm [4] and 500 samples and block least squares learning were also constructed, and the error probability was computed for a range of signal/noise ratios, the results of which are shown in Figure 4. This problem is part of a group of channel equalisation problems for which the optimal decision boundary is nonlinear, and it is therefore little surprise that the linear equaliser has the poorest performance of the three equaliser structures. Both of the nonlinear structures exhibit similar performance, on average around 4 dB better than the linear filter.

From this example it appears that the orthonormal filter can be made to perform equally with the RBF network, although with a considerably larger number of functions. However, this neglects the training overhead of a RBF network - centres must be chosen either by a forward regression technique like the OLS algorithm used here, or by clustering of the input data.

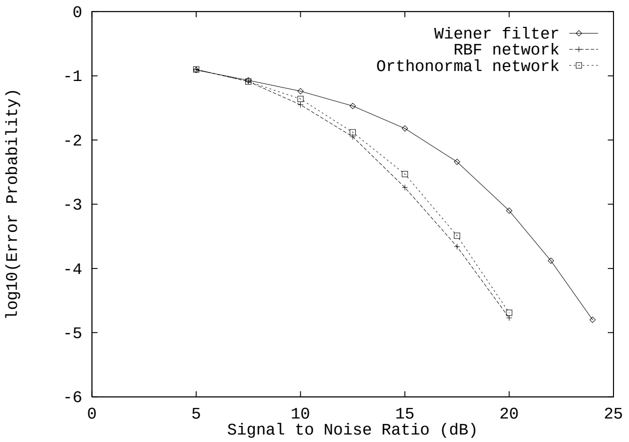

Figure 4: Performance comparison of different equaliser structures, for stationary channel $H(z) = 0.3482 + 0.8704z^{-1} + 0.3482z^{-2}$, $m = 4$ and $\Delta = 1$.

B. Nonlinear system identification

The problem can be stated as that of finding a model which approximates some nonlinear functional expansion $f_s(\cdot)$ of lagged inputs and outputs. The functional form of $f_s(\cdot)$ in a practical system is generally very complicated and is rarely available. A model has to be constructed based on some known simpler function (or functions) which can be used to represent the more complex system. A one–step ahead prediction of system can be derived by defining the input vector $\boldsymbol{x}(k)$ as

$$\begin{aligned}\boldsymbol{x}(k) \;&=\; \Big[y(k-1), \ldots, y(k-n_y), \\ &\qquad u(k-1), \ldots, u(k-n_u)\Big]^T, \qquad (20)\end{aligned}$$

where $y(k)$, $u(k)$ are respectively the system output and input, and n_y and n_u are the maximum lags in the output and input respectively. The identification involves determining values for the network weights based on the input–output observations $\{u(k), y(k)\}$. The example chosen was the following Volterra system

$$\begin{aligned}y(k) \;&=\; 0.4\,y(k-1)\Big(1.0 \;-\; u(k-2)\Big) \\ &\quad + u(k-2)\,u(k-3) \;+\; e(k), \qquad (21)\end{aligned}$$

for which $n_y = 1$ and $n_u = 3$, and $u(k)$, $e(k)$ are white Gaussian noise sequences with variances 1 and 0.001 respectively. $e(k)$ represents additive noise. Orthonormal, RBF and linear filters were constructed, each with an input vector of dimension $N = n_y + n_u = 4$. A total of $81 (= 3^4)$ basis functions were chosen for the orthonormal network, and 30 centres were selected for the RBF network using the OLS algorithm. A training set of 500 points was selected and the normalised mean squared error was computed over a separate test sequence of 500 points. The NMSE of the orthonormal network model was $\bar{\zeta}_o = -21.88$ dB, whereas for the RBF network $\bar{\zeta}_r = -21.84$ dB, over the test sequence. The linear model performed poorly, having a NMSE of $\bar{\zeta}_l = -0.46$ dB.

The autocorrelations of the orthonormal and RBF network error sequences are plotted in Figure 5. If the model found is good, then the residual error should be uncorrelated from sample to sample. It is apparent from Figure 5 that the autocorrelations are all within a 95% confidence band.

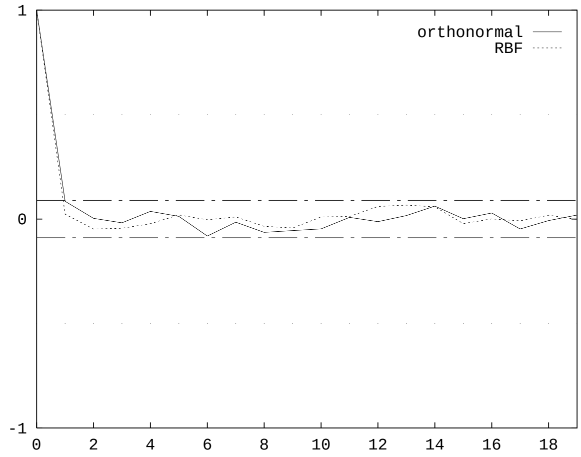

Figure 5: Autocorrelations of the residual sequence for the orthonormal and RBF predictors. $- \cdot -$ indicates 95% confidence band.

5 Conclusions

This paper has presented a systematic approach to nonlinear adaptive filtering of signals. The technique uses multi–dimensional Fourier basis functions to form Wiener functions through an estimate of the input pdf. A simple method based on the characteristic function was described for forming an estimate of the pdf. The performance of the filter was examined for the equalisation of a digital communication channel, and for the identification of a nonlinear system, and was found to be close to that of popular RBF networks.

References

[1] B. Mulgrew, "Orthonormal functions for nonlinear signal processing and adaptive filtering", in *ICASSP-94*, Adelaide, Australia, April 1994, pp. III–541 – III–544.

[2] M. Schetzen, "Nonlinear System Modeling Based on the Wiener Theory", *Proceedings of the IEEE*, vol. 69, no. 12, pp. 1557–1573, December 1981.

[3] M. E. Tarter and M. D. Lock, *Model–Free Curve Estimation*, Number 56 in Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1993.

[4] S. Chen, C. F. N. Cowan, and P. M. Grant, "Orthogonal least squares learning algorithm for radial basis function networks", *IEEE Trans. Signal Processing*, vol. 2, no. 2, pp. 302–309, March 1991.