15th Triennial World Congress, Barcelona, Spain

A NEW SOLUTION TO VOLTERRA SERIES ESTIMATION

Tony J. Dodd and Robert F. Harrison

Department of Automatic Control and Systems Engineering
The University of Sheffield, Sheffield S1 3JD, UK
e-mail: {t.j.dodd, r.f.harrison}@shef.ac.uk

Abstract: Volterra series expansions represent an important model for the representation, analysis and synthesis of nonlinear dynamical systems. However, a significant problem with this approach to system identification is that the number of terms required to be estimated grows exponentially with the order of the expansion. In practice, therefore, the Volterra series is typically truncated to consist of, at most, second degree terms only. In this paper it is shown how the ideas of reproducing kernel Hilbert spaces (RKHS) can be applied to provide a practicable solution to the problem of estimating Volterra series. The approach is based on solving for the Volterra series in a linearised feature space (corresponding to the Volterra series) which leads to a more parsimonious estimation problem.

Keywords: Volterra series, Hilbert spaces, system identification, time series analysis, identification algorithms

1. INTRODUCTION

Volterra series expansions represent an important model for the representation, analysis and synthesis of nonlinear dynamical systems. The idea of the Volterra series expansion is to form a model for the output of the system as a polynomial in the delayed inputs (Priestley, 1988). Such a model has been shown to provide a good representation for a wide class of nonlinear systems (Boyd and Chua, 1985). It is particularly attractive given that the unknown parameters enter linearly and therefore in the minimum mean square error case the parameters can, at least in principle, be determined exactly (Koh and Powers, 1985). However, the number of terms increases exponentially with the order of the expansion. Therefore, in practical terms it is usually necessary to use severely truncated series or employ particular reduced order structures (Ling and Rivera, 1998).

We explain how the ideas of reproducing kernel Hilbert spaces (RKHS) (Aronszajin, 1950; Wahba, 1990) can be applied to provide a more practicable solution to the estimation of Volterra kernels. In particular we are interested in the case of Volterra series of orders higher than two. It is these models which present significant difficulty when attempting to estimate the Volterra kernels.

The main idea behind the approach is to use a particular reproducing, or Mercer, kernel to summarise the complete Volterra series. This is achieved using a mapping into a feature space which is a RKHS (Vapnik, 1995). This feature space corresponds to the space formed by the Volterra series. However, it is not necessary to use the Volterra series terms themselves and instead we use inner products between the terms. This leads to a considerable simplification of the estimation problem. It is this alternative, computable approach to Volterra series estimation which is the novel contribution of this paper.

In the next section we introduce the discrete Volterra series for representing nonlinear discrete-time input-output models. The problems with this approach are discussed as motivation for the new approach. In Section 3 a RKHS corresponding to the Volterra series is constructed. The problem and solution of approximation in RKHS is described and a simple form of the Volterra kernel are described in Sections 4 and 5.

Finally, an example of the new approach to Volterra series estimation is described.

2. VOLTERRA SERIES EXPANSIONS

Consider now the nonlinear model consisting of observable input and output processes $u(t), u(t-1), \ldots$ and $y(t), y(t-1), \ldots$ respectively. A general (non-anticipative) model for $y(t)$ takes the form

$$y(t) = f(u(t), u(t-1), \ldots). \tag{1}$$

Suppose that f is sufficiently well behaved so that we can expand it in a Taylor series about some fixed point to give (Priestley, 1988)

$$\begin{aligned} y(t) = h_0 &+ \sum_{m_1=0}^{\infty} h_1(m_1)u(t-m_1) \\ &+ \sum_{m_1=0}^{\infty}\sum_{m_2=0}^{\infty} h_2(m_1,m_2)u(t-m_1)u(t-m_2) \\ &+ \cdots \end{aligned} \tag{2}$$

where the Volterra kernels (coefficients [1]) are formally given by

$$h_0 = f(\bar{u}), \quad h_1(m_i) = \left(\frac{\partial f}{\partial u(t-m_i)}\right)_{\bar{u}},$$
$$h_2(m_i,m_j) = \left(\frac{\partial^2 f}{\partial u(t-m_i)\partial u(t-m_j)}\right)_{\bar{u}}, \ldots$$

with $\bar{u}$ the fixed point about which the expansion is taken. It is normally assumed that the coefficients $h_k(m_1, m_2, \ldots, m_k)$ are symmetric with respect to permutations of $m_1, m_2, \ldots, m_k$. Such a model has been shown to provide a good representation for a wide class of nonlinear systems (Boyd and Chua, 1985).

We can form the truncated version of Eq. 2 giving the Volterra model of *degree*, L, and *memory length*, M, thus

$$\begin{aligned} y(t) = h_0 + \sum_{n=1}^{L} \Bigg\{ &\sum_{m_1=0}^{M-1}\sum_{m_2=0}^{M-1}\cdots\sum_{m_n=0}^{M-1} h_n(m_1, m_2, \\ &\ldots, m_n)\prod_{i=1}^{n} u(t-m_j) \Bigg\} \end{aligned} \tag{3}$$

This model consists of multidimensional convolutions between the Volterra coefficients and the input terms. The output is linear with respect to the coefficients and, therefore, under the assumption of stationarity, if we solve for the coefficients with respect to a minimum mean square error criterion this will have a single global minimum. The coefficients can then, in principle, be found using the calculus of variations or orthogonal projections (Koh and Powers, 1985). However, the computational complexity is found to increase exponentially with the order of the model. For example with $M = 10, L = 10$, taking account of the symmetry in the coefficients, we are still required to estimate approximately $184,756$ parameters. Notwithstanding the computational burden, to have confidence in the estimates large quantities of data would be required.

To limit the number of parameters the model is often truncated to 2nd or 3rd degree. However, the number of parameters can still pose a problem and 2nd order models only describe the system nonlinearity in a very limited operating range. To include higher degree nonlinearity without introducing too many model parameters it is therefore necessary to seek parsiminious, reduced-order alternatives by imposing additional structure. Examples include cascade structures composed of linear filters in series with memoryless nonlinearities (Korenberg, 1991) or linear combinations of tensor products of simple basis vectors (Nowak and Van Veen, 1996). Both of these approaches require nonlinear optimisation - of the parameters in the former and of the structure in the latter.

An alternative is to construct a sparse representation by searching for the most significant terms (Yao, 1999). Again, however, this is a nonlinear optimisation problem solved, for example, using genetic algorithms. Efficient frequency domain methods have also been reported which reduce significantly the number of computations as compared to the standard time domain methods (Im and Powers, 1996; Reed and Hawksford, 2000). These approaches, though, still scale exponentially with the degree and memory of the filter. In this paper we take an alternative approach in which no approximation to the model structure is necessary but which leads to a significant reduction in the number of parameters to be estimated.

3. REPRODUCING KERNEL HILBERT SPACE OF VOLTERRA SERIES

In order to find a simple solution to Volterra series it will be useful to construct a Hilbert space of functions corresponding to the Volterra series. This Hilbert space, which will be shown to be a RKHS, will allow for a particularly simple solution which is readily computable.

First we define a variable, x, the components of which are the delayed input samples, i.e. $x_1 = u(t), x_2 = u(t-1), \ldots, x_i = u(t-i+1), \ldots$. We assume that the maximum delay of interest is $M-1$ such that $x \in \mathbb{R}^M$. The Volterra series, Eq. 3, can then be written as

$$\begin{aligned} y(t) = h_0 + \sum_{n=1}^{L} \Bigg\{ &\sum_{m_1=0}^{M-1}\sum_{m_2=0}^{M-1}\cdots\sum_{m_n=0}^{M-1} h_n(m_1, m_2, \\ &\ldots, m_n)\prod_{i=1}^{n} x_i \Bigg\}. \end{aligned} \tag{4}$$

[1] We will use coefficients in the remainder to differentiate these from the reproducing kernels which will be introduced subsequently.

This is equivalent to expanding the input x into a nonlinear feature space consisting of all possible polynomials in the x_i up to, and including, degree L. For example if we consider the case of $L = 2$ and $M = 2$ we have the following feature expansion

$$\phi(x) = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \end{pmatrix}. \tag{5}$$

We denote the number of terms in this feature expansion by $l = \frac{(L+M)!}{L!M!}$.

The Volterra series is then expressed, using the abuse of notation $y(t) = y(x) = y(x(t))$, in terms of this feature space as

$$y(t) = y(x) = \langle w, \phi(x) \rangle. \tag{6}$$

where the vector w is an appropriate one-to-one mapping of the Volterra coefficients $h_i(\cdot)$. The feature mapping $\phi(x) : \mathbb{R}^n \to \mathcal{H}$ maps low dimensional inputs into the (typically) high dimensional space $\mathcal{H}$. In the previous example we see that the mapping is from $\mathbb{R}^2$ to a six dimensional space of features.

We now need to show that this feature space corresponds to a Hilbert space. A Hilbert space is a linear space, upon which is defined an inner product, and which is also complete with respect to the metric defined by the inner product (the space is complete if every Cauchy sequence of points converges such that the limit is also a point in the space).

We take as the Hilbert space the set of functions of the form

$$y(x) = \sum_{i=0}^{l} w_i \phi_i(x) \tag{7}$$

for any $w_i \in \mathbb{R}$ and where the upper limit may be infinite. We define the inner product in the space to be

$$\left\langle \sum_{i=0}^{l} v_i \phi_i(x), \sum_{i=0}^{l} w_i \phi_i(x) \right\rangle_{\mathcal{H}} = \sum_{i=0}^{l} \frac{v_i w_i}{\lambda_i} \tag{8}$$

where the λ_i will be defined shortly but for now can be considered simply as a sequence of positive numbers. The associated norm then has the form

$$||y||^2_{\mathcal{H}} = \sum_{i=0}^{l} \frac{w_i^2}{\lambda_i}. \tag{9}$$

The linear combination of terms, Eq. 7, together with the inner product, Eq. 8, is then a Hilbert space, $\mathcal{H}$ [2].

Additionally we define the function

$$k(x, x') = \sum_{i=0}^{l} \lambda_i \phi_i(x) \phi_i(x'). \tag{10}$$

This function will therefore correspond to a dot product in l_2 such that

$$k(x, x') = \left\langle \sum_{i=0}^{l} \sqrt{\lambda_i} \phi_i(x), \sum_{i=0}^{l} \sqrt{\lambda_i} \phi_i(x') \right\rangle_{l_2}. \tag{11}$$

The function, $k(x, x')$, thus defined, has the two important properties that:

(1) For a fixed x, $k(x, \cdot)$ belongs to the Hilbert space $\mathcal{H}$ since

$$k(x, \cdot) = \sum_{i=0}^{l} \{\lambda_i \phi_i(x)\} \phi_i(\cdot) \tag{12}$$

and, for fixed x, $\phi_i(x)$ are a set of numbers, therefore defining the new set of weights $w'_i = \lambda_i \phi_i(x)$

$$k(x, \cdot) = \sum_{i=0}^{l} w'_i \phi_i(\cdot) \tag{13}$$

which is of the general form, Eq. 7.

(2) For every $y \in \mathcal{H} : y = \sum_{i=1}^{l} w_i \phi_i$ we have the reproducing property

$$\begin{aligned} \langle y, k(x, \cdot) \rangle_{\mathcal{H}} &= \sum_{i=0}^{l} \frac{w_i w'_i}{\lambda_i} = \sum_{i=0}^{l} \frac{w_i \lambda_i \phi_i(x)}{\lambda_i} \\ &= \sum_{i=0}^{l} w_i \phi_i(x) = y(x) \end{aligned} \tag{14}$$

i.e. point evaluations of y at x are equal to the inner product of y with $k(x, \cdot)$.

The function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ where $x, x' \in \mathcal{X}$, with these properties, is known as the reproducing kernel of the Hilbert space $\mathcal{H}$, which is called a reproducing kernel Hilbert space (RKHS).

Formally a RKHS is a Hilbert space of functions on some parameter set $\mathcal{X}$ with the property that, for each $x \in \mathcal{X}$, the evaluation functional L_x, which associates f with $f(x)$, $L_x f \to f(x)$, is a bounded linear functional (Wahba, 1990). The boundedness means that there exists a scalar $M = M_x$ such that

$$|L_x f| = |f(x)| \leq M||f|| \text{ for all } f \text{ in the RKHS}$$

where $||\cdot||$ is the norm in the Hilbert space.

The original definition, given by properties 1 and 2, then follows from this formal definition by the Riesz representation theorem. The characterisation of the kernel is encompassed in the Moore-Aronszajin theorem (Wahba, 1990).

Theorem 3.1. To every RKHS there corresponds a unique positive-definite function (the reproducing kernel) and conversely given a positive-definite function k on $\mathcal{X} \times \mathcal{X}$ we can construct a unique RKHS of real-valued functions on $\mathcal{X}$ with k as its reproducing kernel.

Given the kernel, $k(x, \cdot)$, functions in the RKHS corresponding to the original expansion, Eq. 7, can now be expressed in terms of the kernel instead as

[2] The completeness of this space to form a Hilbert space can be proven (Wahba, 1990).

$$y(x) = \sum_i a_i k(x, x^i) \tag{15}$$

for $a_i \in \mathbb{R}$. A well defined inner product is then (Wahba, 1990)

$$\left\langle \sum_i a_i k(x^i, \cdot), \sum_j b_j k(x^j, \cdot) \right\rangle_{\mathcal{F}} = \sum_{i,j} a_i b_j \langle k(x^i, \cdot), k(x^j, \cdot) \rangle_{\mathcal{F}} = \sum_{i,j} a_i b_j k(x^i . x^j).$$

4. ESTIMATION OF THE VOLTERRA SERIES IN RKHS

In the general theory of Hilbert spaces we consider functions as points in $\mathcal{H}$ and it is therefore not possible to look at the value of a function at a point. However, if $\mathcal{H}$ is a RKHS then we can express the value of the function y at the point x^i (we use the superscript notation, x^i, to signify different inputs as opposed to the different components of x which we denote x_i) as the inner product

$$y(x^i) = \langle y, k(x^i, \cdot) \rangle_{\mathcal{H}} \tag{16}$$

using the reproducing property of the RKHS. The significant advantage of RKHS then is that we can approximate functions using a finite series of point evaluations (observations). The reproducing property, Eq. 16, defines a linear sampling operator which we denote by L, i.e.

$$z_i = L_i y \tag{17}$$

where we use z_i to denote the observation of y at the point x^i. The sampling operator is a linear evaluation functional, defined on $\mathcal{H}$, which associates real numbers to the function y.

Suppose we have N such observations at a set of distinct values of x and denoting the complete set of observations by $z^N = [z_1, \ldots, z_N]^T$ then

$$z^N = Ly = \sum_{i=1}^{N} (L_i y) e_i \tag{18}$$

where $e_i \in \mathbb{R}^N$ is the ith standard basis vector.

The Volterra approximation problem can now be stated as follows: given the Hilbert space of functions $\mathcal{H}$, the set of functions $\{k(x^i, \cdot)\}_{i=1}^N \subset \mathcal{H}$ and the observations $\{z_i\}_{i=1}^N$, find a function $y \in \mathcal{H}$ such that Eq. 18 is satisfied.

It can be shown that a solution always exists provided the $k(x_i, \cdot)$ are linearly independent (which is satisfied if the x^i are distinct) (Bertero *et al.*, 1985), it is not unique however. A unique solution can be found, though, which has minimal norm, the so-called normal solution. Interestingly the computation of this normal solution is always well-posed in the strict mathematical sense, i.e. the solution depends continuously on the data. However, it can be strongly ill-conditioned and therefore exhibit numerical instability. To avoid this instability we therefore seek a solution to the regularisation problem (which leads to a particularly simple solution): find $y \in \mathcal{H}$ such that

$$\hat{y}(x) = \arg\min_{y \in \mathcal{H}} \sum_{i=1}^{N} l(z_i - y(x^i)) + \frac{\rho}{2} \|y\|^2_{\mathcal{H}} \tag{19}$$

where $l(\cdot, \cdot)$ is a convex loss function and $\rho \geq 0$ is the regularisation parameter.

Substituting Eqs. 7 and 9 we obtain, for a RKHS,

$$\hat{y}(x) = \arg\min_{y \in \mathcal{H}} \sum_{i=1}^{N} l\left(z_i - \sum_{j=0}^{l} w_j \phi_j(x^i)\right) + \frac{\rho}{2} \sum_{j=0}^{l} \frac{w_j^2}{\lambda_j}. \tag{20}$$

Minimising this expression with respect to the coefficients w_j and equating to zero we obtain

$$-\sum_{i=1}^{N} l'\left(z_i - \sum_{j=0}^{l} w_j \phi_j(x^i)\right) \phi_j(x^i) + \rho \frac{w_j}{\lambda_j} = 0 \tag{21}$$

where l' is the derivative of the loss function. Defining a new set of coefficients

$$a_i = \frac{1}{\rho} l'(z_i - y(x^i)) \tag{22}$$

then, in terms of these, we have

$$w_j = \lambda_j \sum_{i=1}^{N} a_i \phi_j(x^i). \tag{23}$$

The solution of the variational approximation problem is therefore given by

$$\hat{y}(x) = \sum_{j=0}^{l} w_j \phi_j(x) = \sum_{j=0}^{l} \lambda_j \sum_{i=1}^{N} a_i \phi_j(x^i) \phi_j(x) = \sum_{i=1}^{N} a_i \sum_{j=0}^{l} \lambda_j \phi_j(x^i) \phi_j(x) = \sum_{i=1}^{N} a_i k(x, x^i). \tag{24}$$

There are two points of interest: (i) even though we are considering mappings into very high (possibly infinite dimensional spaces) the computation remains finite and directly proportional to the size of the available data. We are therefore able to solve for Volterra series with arbitrarily large numbers of terms with a finite computation. (ii) the form of the solution is independent of the loss function l and is always a linear superposition of the kernel functions.

In the particular case of $l(\cdot, \cdot) = (\cdot, \cdot)^2$ we have the standard regularised least-squares solution. Substituting Eq. 24 for the output into Eq. 22

$$a_i = \frac{1}{\rho} l'(z_i - y(x^i)) = \frac{1}{\rho} \left(z_i - \sum_{j=1}^{N} a_j k(x^i, x^j)\right). \tag{25}$$

In matrix form

$$a = \frac{1}{\rho}(z^N - Ka) \tag{26}$$

where K is the kernel (Gram) matrix defined as $K_{ij} = k(x^i, x^j)$. The coefficients are then given by the solution of

$$(K + \rho I)a = z^N. \tag{27}$$

5. THE VOLTERRA KERNEL

The Volterra reproducing kernel was expressed previously as the series expansion

$$k(x, x') = \sum_{i=0}^{l} \lambda_i \phi_i(x) \phi_i(x') \tag{28}$$

where the $\phi_i(x)$ correspond to polynomials in x. It was seen in the previous section that the estimation of the Volterra filter could be reduced to a problem which scales with the number of data. However, the reproducing kernel still involves l terms.

Consider instead the form of the kernel

$$k(x, x') = (1 + \langle x \cdot x' \rangle)^L \tag{29}$$

which corresponds to a mapping into the space of all possible polynomials of degree $\leq L$. It is known that this kernel has an expansion of the form Eq. 28 (Vapnik, 1995) and can therefore be used instead.

As an example consider the case of $L = 2, M = 2$, i.e. $x \in \mathbb{R}^2$ for which

$$\begin{aligned} k(x, x') &= (1 + \langle x \cdot x' \rangle)^2 = (1 + x_1 x_1' + x_2 x_2')^2 \\ &= 1 + 2x_1 x_1' + 2x_2 x_2' + 2x_1 x_1' x_2 x_2' \\ &\quad + (x_1 x_1')^2 + (x_2 x_2')^2 \end{aligned}$$

which is equivalent to considering the feature mapping $\sqrt{\lambda_i}\phi_i(x)$ given by

$$\{\sqrt{\lambda_i}\phi_i(x)\}_{i=1}^{6} = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2)$$

which can be proven by evaluating (in l_2)

$$k(x, x') = \begin{pmatrix} 1 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ \sqrt{2}x_1 x_2 \\ x_1^2 \\ x_2^2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ \sqrt{2}x_1' \\ \sqrt{2}x_2' \\ \sqrt{2}x_1' x_2' \\ (x_1')^2 \\ (x_2')^2 \end{pmatrix}. \tag{30}$$

We now see the role of the λ_i's, which ensure the equivalence of the forms, Eqs. 10 and 29.

6. EXAMPLE

As an example of the application of the RKHS approach to Volterra series estimation consider the discrete-time nonlinear dynamical system described by the following equation (Billings and Voon, 1986)

$$\begin{aligned} y(t) &= 0.5y(t-1) + 0.3y(t-1)u(t-1) \\ &\quad + 0.2u(t-1) + 0.05y^2(t-1) + 0.6u^2(t-1) \end{aligned}$$

with the observations generated as

$$z(t) = y(t) + \varepsilon(t) \tag{31}$$

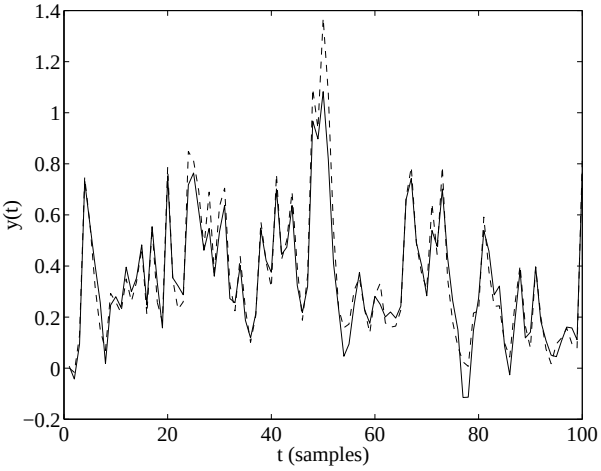

Fig. 1. Typical predicted output ('–') for a Volterra RKHS model with $L = 5, M = 2$ and actual noise free true output ('- -').

where $\varepsilon(t) \sim N(0, 0.1)$ (note that this is a very noisy signal with a signal-to-noise ratio of approximately 30%). The system includes both delayed inputs and outputs and therefore we would expect a Volterra model (based only on delayed inputs) to be of a high order to provide good predictive performance. In identifying the system the data were generated from an initial condition of $y(1) = 0.1$ and the control input was sampled as $u(t) \sim N(0.2, 0.1)$. Various models were considered with varying memory lengths and polynomial degrees as shown in Table 1. In all cases the quadratic loss function is used for which the solution is given by Eq. 27. The kernel matrix is constructed using the kernel defined by Eq. 29 with the appropriate value of L and $x^i = [u(t-i+1), \ldots, u(t-M+1)]^T$.

For the models considered, the value of ρ was first estimated using a set of 500 data samples for training and 200 independent samples for validation. The value of ρ was estimated as corresponding to the minimum of the mean-squared error on this validation set. Given the estimated value of ρ each model was then trained and tested for 10 different training and testing data sets of 500 samples each. The average over these runs of the mean-squared error is shown in Table 1. An example prediction over the first 100 samples of one of the test sets is shown in Figure 1.

The purpose of these results is simply to demonstrate the applicability of the new technique and not as an exhaustive investigation of how to find good models. It can be seen from the results that good prediction performance is achievable and that large Volterra models can be estimated. The reason that the model with

Table 1. Comparison of the average mean squared error for six different example Volterra RKHS models.

L	M	ρ	Average mse
2	2	0.05	0.0254
5	1	0.1	0.0159
5	2	3.5	0.0045
5	3	3.8	0.0056
5	4	13.0	0.0062
10	10	500.0	0.4214

$L = 10, M = 10$ performed so poorly is probably due to insufficient data and/or overfitting of the model to the training data. However, we see that the "optimum" model ($L = 5, M = 2$) is considerably better than the simple $L = 2, M = 2$ case. For this "optimum" case the average mean-squared error of 0.0045 compares very favourably to the noise variance of 0.01.

Using Eq. 23 it is possible to convert the kernel model back into the equivalent Volterra series. An example of this for the case $L = 5, M = 2$ is

$$\begin{aligned} y(t) = {} & 0.1215 - 0.0826u(t) + 0.0836u(t-1) \\ & -0.0002u(t)u(t-1) + 0.0395u^2(t) \\ & +0.2637u^2(t-1) - 0.0704u^2(t)u(t-1) \\ & +0.0124u(t)u^2(t-1) + 0.0649u^3(t) \\ & +0.0256u^3(t-1) + 0.0988u^3(t)u(t-1) \\ & +0.0580u^2(t)u^2(t-1) - 0.0379u(t)u^3(t-1) \\ & -0.0156u^4(t) - 0.0553u^4(t-1) \\ & -0.0270u^4(t)u(t-1) + 0.0798u^3(t)u^2(t-1) \\ & +0.0286u^2(t)u^3(t-1) - 0.0109u(t)u^4(t-1) \\ & -0.0097u^5(t) - 0.0485u^5(t-1). \end{aligned}$$

7. CONCLUSIONS

A computationally efficient approach to the estimation of large scale Volterra series has been presented. The approach makes use of a Hilbert space corresponding to the Volterra series which was shown to be a RKHS. The solution to approximation in the Volterra RKHS with respect to a large class of loss functions was shown to be simply a linear combination of a set of kernel functions. The main reason for using the RKHS approach is that the number of coefficients which needs to be estimated is only proportional to the number of data. This can therefore represent a significant reduction over the standard Volterra series case (for which an arbitrarily large number coefficients may be present). Finally, the approach was demonstrated on a highly nonlinear benchmark system.

ACKNOWLEDGEMENTS

The authors would like to thank the referees for their comments and the UK EPSRC for their financial support under Grant No. GR/R15726/01.

8. REFERENCES

Aronszajin, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society* **68**, 337–404.

Bertero, M., C. De Mol and E.R. Pike (1985). Linear inverse problems with discrete data. I: General formulation and singular system analysis. *Inverse Problems* **1**, 301–330.

Billings, S.A. and W.S.F. Voon (1986). Correlation-based model validity tests for non-linear models. *International Journal of Control* **44**, 235–244.

Boyd, Stephen and Leon O. Chua (1985). Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Transactions on Circuits and Systems* **32**(11), 1150–1161.

Im, Sungbin and Edward J. Powers (1996). A fast method of discrete third-order Volterra filtering. *IEEE Transactions on Signal Processing* **44**(9), 2195–2208.

Koh, Taiho and Edward J. Powers (1985). Second-order Volterra filtering and its application to nonlinear system identification. *IEEE Transactions on Acoustics, Speech and Signal Processing* **33**(6), 1445–1455.

Korenberg, M.J. (1991). Orthogonal approaches to time-series analysis and system identification. *IEEE Signal Processing Magazine* **8**, 29–43.

Ling, Wei-Ming and Daniel E. Rivera (1998). Control relevant model reduction of Volterra series models. *Journal of Process Control* **8**(2), 79–88.

Nowak, Robert D. and Barry D. Van Veen (1996). Tensor product basis approximations for Volterra filters. *IEEE Transactions on Signal Processing* **44**(1), 36–50.

Priestley, M.B. (1988). *Non-Linear and Non-Stationary Time Series Analysis*. Academic Press.

Reed, M.J. and M.O.J. Hawksford (2000). Efficient implementation of the Volterra filter. *IEE Proceedings - Vision, Image and Signal Processing* **147**(2), 109–114.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag. New York.

Wahba, G. (1990). *Spline Models for Observational Data*. Vol. 50 of *Series in Applied Mathematics*. SIAM. Philadelphia.

Yao, Leehter (1999). Genetic algorithm based identification of nonlinear systems by sparse Volterra filters. *IEEE Transactions on Signal Processing* **47**(12), 3433–3435.