Monograph:
Dodd, T.J. and Harrison, R.F. (2001) Volterra Series Estimation via Reproducing Kernel Hilbert Spaces (1). UNSPECIFIED. ACSE Research Report 804 . Department of Automatic Control and Systems Engineering

Volterra Series Estimation via Reproducing Kernel Hilbert Spaces[1]

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

Research Report No. 804
October 2001

[1]Research conducted under EPSRC Grant No. GR/R15726/01

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 order 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 estimation of 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.

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 theory, 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 in estimating the Volterra kernels.

The main idea behind the approach is to use a particular reproducing or Mercer kernel to essentially 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 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 Hilbert space corresponding to the Volterra series is constructed. This forms the basis for the subsequent treatment. The idea of RKHS is introduced in Section 4 and it is shown that the Hilbert space of Volterra series is a RKHS. The problem and solution of approximation in RKHS is described. Finally, an example of the new approach to Volterra series estimation is described and some conclusions drawn.

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) 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

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

This model consists of multidimensional convolutions between the Volterra coefficients and 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 theory, 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 $M = 15, L = 3$, taking account of the symmetry in the coefficients, we are still required to estimate approximately 815 parameters. In any case we would require large amounts of data in order to be confident in the estimates.

To limit the number of parameters the model is often truncated to 2nd or 3rd order. However, the number of parameters can still pose a problem and 2nd order models only describe the system nonlinearity in a very limited operation range. In order to include higher order nonlinearity without introducing too many model parameters it is therefore necessary to seek parsiminious,

reduced-order alternatives by imposing additional structure. The Hammerstein and Wiener strutures are examples of these (Ling and Rivera 1998). Other approaches include the use of orthogonal basis function expansions such as Laguerre series. However, 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 solved.

3 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.

Firstly we define a variable, x, the components of which are the delayed inputs, 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

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

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 possible feature expansion

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

which is not unique (Vapnik 1995).

The Volterra series is then expressed in terms of this feature space as

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

where the vector w is a 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

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

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

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

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

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

The linear combination of terms, Eq. 7, together with the inner product, Eq. 8, is then a Hilbert space [1]. More specifically, the space of functions corresponds to a RKHS which we now describe in detail.

4 Reproducing Kernel Hilbert Spaces

4.1 Formal Definition

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 $f(x)$, $L_x f \rightarrow f(x)$, is a bounded linear functional (Wahba 1990). The boundedness means that there exists a scalar $M = M_x$ such that

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

where $||\cdot||$ is the norm in the Hilbert space. An important class of Hilbert spaces where this does not hold are the L_2 spaces on subsets of $\mathbb{R}^d$.

By the Riesz representation theorem we then arrive at the more generally quoted definition of a RKHS which is based on a symmetric, positive-definite function $k(\cdot, \cdot)$.

Definition 4.1 *(Parzen 1961) A Hilbert space $\mathcal{H}$ is said to be a reproducing kernel Hilbert space, with reproducing kernel k, if the members of $\mathcal{H}$ are functions on some set $\mathcal{X}$, and if there is a kernel k on $\mathcal{X} \times \mathcal{X}$ having the following two properties; for every $x \in \mathcal{X}$ (where $k(\cdot, x')$ is the function defined on $\mathcal{X}$, with value at x' in $\mathcal{X}$ equal to $k(x, x')$):*

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

1. $k(\cdot, x') \in \mathcal{H}$; *and*

2. $\langle f, k(\cdot, x') \rangle_{\mathcal{H}} = f(x')$

for every f *in* $\mathcal{H}$.

The characterisation of the kernel is encompassed in the Moore-Aronszajin theorem (Wahba 1990).

Theorem 4.1 *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.*

Suppose that $k(x, x')$ is continuous and

$$\int_{\mathcal{X}} \int_{\mathcal{X}} k^2(x, x') dx dx' < \infty \tag{10}$$

then there exists an orthonormal sequence of continuous eigenfunctions $\{\phi_i\}_{i=1}^{\infty}$ in $L_2[\mathcal{X}]$ with associated eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ such that (Wahba 1990)

$$\int_{\mathcal{X}} k(x, x') \phi_i(x') dx' = \lambda_i \phi_i(x), \quad i = 1, 2, \ldots \tag{11}$$

It can then be shown that if we let

$$f_i = \int_{\mathcal{X}} f(x) \phi_i(x) dx, \tag{12}$$

then $f \in \mathcal{H}$ if and only if (Wahba 1990)

$$\sum_{i=1}^{\infty} \frac{f_i^2}{\lambda_i} < \infty \tag{13}$$

and

$$\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{\infty} \frac{f_i^2}{\lambda_i}. \tag{14}$$

Condition 13 requires that the generalised Fourier coefficients f_i go to zero not slower that the eigenvalues of k which is equivalent to a smoothness condition on the functions f.

Expanding the function f in a Fourier series with respect to the ϕ_i we have

$$f(x) = \sum_i f_i \phi_i(x) \tag{15}$$

and comparing Eqs. 15 and 14 with Eqs. 7 and 9 we see that for $f = y(t)$ then $f_i = w_i$.

4.2 Volterra Series in RKHS

It can also be shown that a valid (in fact the unique) reproducing kernel satisfies (Wahba 1990)

$$k(x,x') = \sum_i \lambda_i \phi_i(x)\phi_i(x') \tag{16}$$

which is valid under the assumption of Eq. 10 (known as the Mercer theorem hence the term Mercer kernel often used for $k(\cdot,\cdot)$). The kernel will therefore correspond to a dot product in l_2 such that

$$k(x,x') = \langle \sum_i \sqrt{\lambda_i}\phi_i(x), \sum_i \sqrt{\lambda_i}\phi_i(x') \rangle. \tag{17}$$

We consider now the particular instance of the nonlinear feature mapping consisting of polynomials to order L. The polynomial kernel is defined as

$$k(x,x') = (1 + \langle x \cdot x' \rangle)^L \tag{18}$$

which corresponds to a mapping into the space of all possible polynomials of degree $\leq L$. Now it is known that this kernel has an expansion of the form (Vapnik 1995)

$$k(x,x') = \sum_{i=1}^{\dim(\mathcal{H})} \lambda_i \phi_i(x)\phi_i(x') \tag{19}$$

where the ϕ_i are the polynomials of degree up to L, which constitutes a basis in the set of polynomials of degree L, λ_i are a sequence of positive numbers and $\dim(\mathcal{H})$ is the dimension of the associated Hilbert space (equal to the total number of terms in the polynomial expansion).

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

$$\begin{aligned} k(x,x') &= (1 + \langle x \cdot x' \rangle)^2 = (1 + x_1x_1' + x_2x_2')^2 \\ &= 1 + 2x_1x_1' + 2x_2x_2' + 2x_1x_1'x_2x_2' \\ &\quad + (x_1x_1')^2 + (x_2x_2')^2 \end{aligned}$$

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

$$\{\sqrt{\lambda_i}\phi_i(x)\}_{i=1}^6 = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2)$$

which can be proven by evaluating (in l_2)

$$k(x,x') = \begin{pmatrix} 1 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ \sqrt{2}x_1x_2 \\ x_1^2 \\ x_2^2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ \sqrt{2}x_1' \\ \sqrt{2}x_2' \\ \sqrt{2}x_1'x_2' \\ (x_1')^2 \\ (x_2')^2 \end{pmatrix}. \tag{20}$$

We therefore see that the Volterra series expansion corresponds to a RKHS with a particular choice of kernel, Eq. 18. The so-called features are not strictly the ones we considered originally as they are now scaled according to the square roots of the eigenvalues. The effects of these will simply be a rescaling of the coefficients w_i in the original expansion. However, we are now still faced with the prospect of solving for a potentially infinite sequence of coefficients. The reason for using RKHS will be demonstrated in the next section.

4.3 Approximation 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 if we define the set of functions $\{\psi_i(x)\}_{i=1}^N$ as

$$\psi_i(x) = k(x, x^i), \quad i = 1, \ldots, N, \tag{21}$$

then we can express the value of the function f 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 (Vijayakumar and Ogawa 1999)

$$f(x^i) = \langle f, \psi_i \rangle. \tag{22}$$

We are then able to define a linear sampling operator S which relates the complete set of observations to the underlying function

$$z = Sf + \varepsilon \tag{23}$$

where z and ε are now vectors $\in \mathbb{R}^N$. The sampling operator is that which satisfies

$$z = \sum_{i=1}^{N} \langle f, \psi_i \rangle e_i \tag{24}$$

where e_i is the ith unit vector i.e. consisting of zero elements except the ith element equal to 1. The learning problem is then that of obtaining an estimate, $\hat{f}$, of f given z, which can be considered as the inverse problem of obtaining the operator $\hat{S}^{-1}$

$$\hat{f} = \hat{S}^{-1} z. \tag{25}$$

It can be shown that a solution to Eq. 25 always exists provided the ψ_i are linearly independent (Bertero, De Mol, and Pike 1985). However it is not unique. 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 problem: find $f \in \mathcal{H}$ such that

$$\hat{f}(x) = \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} l(z_i, f(x^i)) + \rho \|f\|_{\mathcal{H}}^2 \tag{26}$$

where the norm is taken over the RKHS of interest and $\rho \geq 0$ is the so-called regularisation parameter. The solution is then given by:

Theorem 4.2 *(Wahba 1999) Let $l(\cdot, \cdot)$ be convex then any solution to Eq. 26 has a representation of the form*

$$\hat{f}(x) = \sum_{i=1}^{N} a_i k(x, x^i). \tag{27}$$

The coefficients a_i are determined by the specific loss function l used. Two cases of specific interest are possible: the ϵ-insensitive function

$$l(e) = |e|_\epsilon = \begin{cases} 0 & \text{if } |e| < \epsilon \\ |e| - \epsilon & \text{otherwise} \end{cases} \tag{28}$$

whereby the solution of Eq. 26 is the support vector machine approximator and the coefficients must be determined using quadratic programming (Smola 1998). The second case, and that which we are primarily interested in is the quadratic loss

$$l(e) = e^2 \tag{29}$$

for which the coefficients can be determined in closed form, thus

$$a = (K + \rho I)^{-1} z, \quad [K]_{ij} = k(x_i, x_j). \tag{30}$$

The main point of interest is that, 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. The penalty is that we lose the transparency that the original expansion in polynomial terms gives us.

5 Example

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

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

with the observations generated as

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

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%). Note that 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 orders as shown in Table 1. In all cases the quadratic loss function is used for which the solution is given by Eq. 30.

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 ten 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.

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

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 $L = 10, M = 10$ 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 favourably to the noise variance of 0.01.

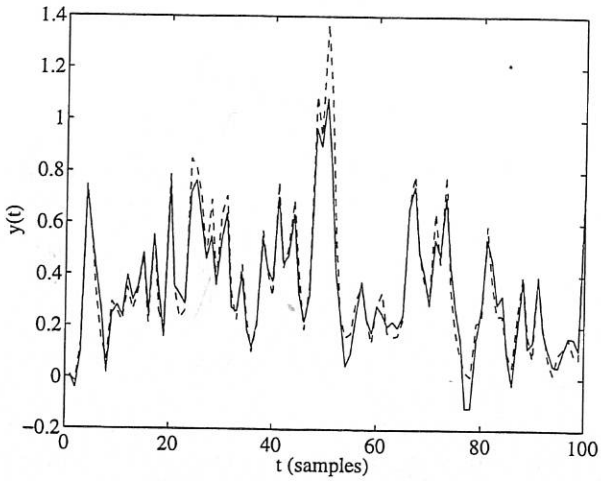

Figure 1: Typical predicted output ('–') for a Volterra RKHS model with $L = 5, M = 2$ and actual noise free true output ('- -').

6 Conclusions

Volterra series provide a good general representation for a wide class of nonlinear systems. The series is attractive in that the output is linear with respect to the unknown coefficients. Therefore, under the assumption of stationarity, with respect to the minimum mean square error criterion a single global minimum exists for the coefficients which can be solved for. However, the number of terms required to be estimated grows exponentially with the order of the expansion. In practise, therefore, the Volterra series is typically truncated to a lower order model.

In this paper an alternative framework for Volterra series has been described based on constructing a corresponding Hilbert space. Such a Hilbert space was shown to be a RKHS which has certain nice properties. The solution to approximation in RKHS with respect to a large class of loss functions is simple and well known in the case of the quadratic and ε-insensitive cases. The latter leads to a support vector machine. The main reason for using the RKHS approach is that the number of coefficients which need 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 for which an infinite dimensional Volterra series is required. This is because the system output is dependent on both delayed outputs and inputs. The RKHS approach was shown to provide good predictive performance. A more detailed version of this paper is currently in preparation which will include computational aspects for large data sets, additional examples and a Bayesian

analysis.

References

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

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

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

Boyd, S. and L. 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.

Koh, T. and E. 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.

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

Parzen, E. (1961). An approach to time series analysis. *Annals of Mathematical Statistics 32*, 951–989.

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

Smola, A. J. (1998). *Learning with Kernels.* Ph. D. thesis, Informatik der Technischen Universitat Berlin.

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

Vijayakumar, S. and H. Ogawa (1999). RKHS-based functional analysis for exact incremental learning. *Neurocomputing 29*, 85–113.

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

Wahba, G. (1999). Support vector machines, reproducing kernel hilbert spaces and randomized GACV. In B. Schölkopf, C. J. Burges, and A. J. Smola (Eds.), *Advances in Kernel Methods: Support Vector Learning*, pp. 69–88. The MIT Press.

Vapnik, V. (1998). *Statistical Learning Theory.* Adaptive and Learning Systems for Signal Processing, Communications and Control. John Wiley & Sons.

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

Weiner, H. (1965). The gradient iteration in time series analysis. *Journal of the Society for Industrial and Applied Mathematics 13*(4), 1096–1101.

Williams, C. (1999). Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M. Jordan (Ed.), *Learning in Graphical Models*, pp. 599–621. The MIT Press.

A Proofs Relating to Adjoint Operators in RKHS

Consider the operator $L : \mathcal{F} \to \mathcal{Z}$ where $\mathcal{Z}$ is the N dimensional Euclidean space with inner product $\langle g, h \rangle_{\mathcal{Z}} = \sum_{i=1}^{N} g_i h_i$, for $g, h \in \mathcal{Z}$, then, for $z^N \in \mathcal{Z}, f \in \mathcal{F}$ the adjoint operator L^* is defined by

$$\langle Lf, z^N \rangle_{\mathcal{Z}} = \langle f, L^* z^N \rangle_{\mathcal{F}} \tag{31}$$

and transforms the observation vector z^N into an element of $\mathcal{F}$ or more precisely the finite dimensional subspace $\mathcal{F}_N$. In a RKHS the operator L acting on f has the form $Lf = \sum_{i=1}^{N} \langle f, k(x_i, \cdot) \rangle_{\mathcal{F}} \cdot e_i$, where $e_i \in \mathbb{R}^N$ is the ith standard basis vector. The following results apply to the operator L and its adjoint L^*.

Theorem A.1 *Given the operator L and its adjoint L^* defined by*

$$\langle Lf, z^N \rangle_{\mathcal{Z}} = \langle f, L^* z^N \rangle_{\mathcal{F}} \tag{32}$$

then in a RKHS with $Lf = \sum_{i=1}^{N} \langle f, k(x_i, \cdot) \rangle_{\mathcal{F}} \cdot e_i$ the adjoint L^* is given by

$$L^* z^N = \sum_{i=1}^{N} z_i k(x_i, \cdot). \tag{33}$$

Proof. Solving for the LHS of Eq. 32

$$\langle Lf, z^N \rangle_{\mathcal{Z}} = \sum_{i=1}^{N} \langle f, k(x_i, \cdot) \rangle_{\mathcal{F}} \cdot z_i = \sum_{i=1}^{N} f(x_i) z_i. \tag{34}$$

By assumption we set $L^* z^N = \sum_{i=1}^{N} z_i k(x_i, \cdot)$ and solving for the RHS of Eq. 32

$$\langle f, L^* z^N \rangle_{\mathcal{F}} = \left\langle f, \sum_{i=1}^{N} z_i k(x_i, \cdot) \right\rangle_{\mathcal{F}} = \sum_{i=1}^{N} z_i \langle f, k(x_i, \cdot) \rangle_{\mathcal{F}} \tag{35}$$

the latter due to the linearity property of the inner product. But this is simply equal to $\sum_{i=1}^{N} z_i k(x_i, \cdot)$. □

Theorem A.2 *The operator LL^* is equal to the kernel (Gram) matrix K, i.e.*

$$LL^* = \sum_{j=1}^{N} \sum_{i=1}^{N} k(x_i, x_j) e_j e_i^T. \tag{36}$$

Proof. The operator LL^* acting on z^N can be expressed, using the previous results, as follows:

$$\begin{aligned} LL^* z^N &= L\left(\sum_{i=1}^{N} z_i k(x_i, \cdot)\right) \\ &= \sum_{j=1}^{N} \left\langle \sum_{i=1}^{N} z_i k(x_i, \cdot), k(x_j, \cdot) \right\rangle_{\mathcal{F}} \cdot e_j \\ &= \sum_{j=1}^{N} \sum_{i=1}^{N} z_i \langle k(x_i, \cdot), k(x_j, \cdot) \rangle_{\mathcal{F}} \cdot e_j \\ &= \sum_{j=1}^{N} \sum_{i=1}^{N} z_i k(x_i, x_j) e_j \end{aligned}$$

and therefore the operator $LL^* = \sum_{j=1}^{N} \sum_{i=1}^{N} k(x_i, x_j) e_j e_i^T$. □