Available online at www.sciencedirect.com
SCIENCE DIRECT®
automatica
Automatica 41 (2005) 1263–1272
www.elsevier.com/locate/automatica

Brief paper

Identification of MIMO Hammerstein models using least squares support vector machines☆

Ivan Goethals*, Kristiaan Pelckmans, Johan A.K. Suykens, Bart De Moor

K.U.Leuven, ESAT-SCD-SISTA, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

Received 1 April 2004; received in revised form 5 October 2004; accepted 17 February 2005
Available online 18 April 2005

Abstract

This paper studies a method for the identification of Hammerstein models based on least squares support vector machines (LS-SVMs). The technique allows for the determination of the memoryless static nonlinearity as well as the estimation of the model parameters of the dynamic ARX part. This is done by applying the equivalent of Bai's overparameterization method for identification of Hammerstein systems in an LS-SVM context. The SISO as well as the MIMO identification cases are elaborated. The technique can lead to significant improvements with respect to classical overparameterization methods as illustrated in a number of examples. Another important advantage is that no stringent assumptions on the nature of the nonlinearity need to be imposed except for a certain degree of smoothness.

Keywords: Hammerstein models; ARX; LS-SVM; MIMO systems; Kernel methods

1. Introduction

Throughout the last few decades, the field of linear modeling has been explored to the level that most linear identification problems can be solved efficiently with fairly standard and well-known tools. Extensions to complex nonlinear models are often desirable, though in many situations Hammerstein models may result in good approximations. Hammerstein models are composed of a memoryless static nonlinearity followed by a linear dynamical system.

Many techniques have been proposed for the black-box estimation of Hammerstein systems from given input–output measurements. These techniques mainly differ in the way the static nonlinearity is represented and in the type of optimization problem that is finally obtained. In parametric approaches, the static nonlinearity is expressed in terms of a finite number of parameters. Known approaches include the expansion of the nonlinearity as a sum of (orthogonal or non-orthogonal) basis functions (Narendra & Gallman, 1966; Pawlak, 1991; McKelvey & Hanner, 2003), the use of a finite number of cubic spline functions as presented by (Dempsey & Westwick, 2004), piecewise linear functions (van Pelt & Bernstein, 2000) and neural networks (Janczak, 2003). Regardless of the parameterization scheme that is chosen, the final cost function will involve cross-products between parameters describing the static nonlinearity and those describing the linear dynamical system. Employing a maximum likelihood criterion results in a non-convex optimization problem, where global convergence is not guaranteed (Sjöberg et al., 1995). Hence, in order to find a good optimum for these techniques, a proper initialization is necessary (Crama & Schoukens, 2001).

Different approaches were proposed in the literature to overcome this difficulty. These result in convex methods which generate models of the same, or almost the same quality as their nonconvex counterparts. Unfortunately, convexity is either obtained by placing heavy restrictions on the input sequence (e.g. whiteness) and the nonlinearity under

☆ This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor M. Verhaegen under the direction of Editor T. Söderström.
* Corresponding author.
E-mail addresses: ivan.goethals@esat.kuleuven.ac.be (I. Goethals), kristiaan.pelckmans@esat.kuleuven.ac.be (K. Pelckmans), johan.suykens@esat.kuleuven.ac.be (J.A.K. Suykens), bart.demoor@esat.kuleuven.ac.be (B. De Moor).

0005-1098/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2005.02.002

consideration (Bai, 2002) or by using a technique known as overparameterization (Chang & Luus, 1971; Bai & Fu, 1998). In the latter, one replaces every cross-product of unknowns by new independent parameters resulting in a convex but overparameterized method. In a second stage the obtained solution is projected onto the Hammerstein model class using a singular value decomposition. A classical problem with the overparameterization approach is the increased variance of the estimates due to the increased number of unknowns in the first stage.

In this paper, we explore the use of LS-SVMs for Hammerstein model identification. It will be shown that the linear model parameters and the static nonlinearity can be obtained by solving a set of linear equations with size in the order of the number of observations. Given the convexity and the large number of parameters involved, the method may be regarded as an overparameterization approach. However, due to the presence of a regularization framework (Vapnik, 1998; Schölkopf & Smola, 2002; Suykens, Van Gestel, De Brabanter, De Moor, & Vandewalle, 2002), the variance of the obtained estimates is significantly lower than in classical overparameterization approaches. Due to this decrease in variance, systems with several inputs and outputs can be estimated conveniently with the presented technique.

Another advantage of the proposed derivation is the fact that additional centering-constraints and parametric components of the linear dynamical system can naturally be included in the LS-SVM framework due to the fact that it is closely related to convex optimization. Furthermore, in contrast to classical parametric approaches, no specific model structure is imposed on the nonlinearity other than a certain shape (e.g. a degree of smoothness). Hence, the presented technique combines a nonparametric approach with parametric assumptions on the dynamical system and on the noise model. The technique distinguishes itself from existing nonparametric approaches (Greblicki & Pawlak, 1986; Greblicki, 1989; Krzyżak, 1989; Greblicki & Pawlak, 1991; Pawlak, 1991; Verhaegen & Westwick, 1996; Hasiewicz, 1999) in the flexibility to incorporate prior knowledge on the shape of the nonlinearity by plugin of an appropriate kernel (e.g. linear, polynomial, RBF, spline). Furthermore, the presented method does not rely explicitly on restrictive assumptions on the inputs (as e.g. whiteness).

The outline of this paper is as follows: Some basic aspects of LS-SVMs applied to static function estimation are reviewed in Section 2. In Sections 3 and 4, a method for the identification of nonlinear SISO Hammerstein systems is proposed. In Section 5, the method is extended to MIMO Hammerstein systems. Section 6 compares the presented method to existing overparameterization techniques for the identification of Hammerstein systems. In Section 7, the method proposed in this paper is tested and compared to existing methods on a number of SISO and MIMO examples. As a general rule, lowercase symbols will be used in this paper to denote column vectors. Uppercase symbols are used for matrices. Elements of matrices and vectors are selected using Matlab-notation, e.g. $A(:,i)$ symbolizes the ith column of A. Estimates for a parameter x will be denoted by $\hat{x}$.

2. Least squares support vector machines for function approximation

2.1. Ridge regression in feature space

Let $\{(x_t, y_t)\}_{t=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ be the set of given input/output training data with input x_t and output y_t. Consider the regression model $y_t = f(x_t) + e_t$, where $x_1, \ldots, x_N$ are deterministic points, $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown real-valued smooth function and $e_1, \ldots, e_N$ are uncorrelated random errors with $E[e_t] = 0$, $E[e_t^2] = \sigma_e^2 < \infty$. In recent years, support vector machines (SVMs) and its variations have been used for the purpose of estimating the nonlinear f. The following model is assumed:

$$f(x) = w^{\mathrm{T}}\varphi(x) + b,$$

where $\varphi : \mathbb{R}^d \to \mathbb{R}^{n_{\mathrm{H}}}$ denotes a potentially infinite ($n_{\mathrm{H}} = \infty$) dimensional feature map. The regularized cost function of the least squares SVM (LS-SVM) is given as

$$\min_{w,b,e} \quad \mathcal{J}(w,e) = \frac{1}{2}w^{\mathrm{T}}w + \frac{\gamma}{2}\sum_{t=1}^{n} e_t^2$$
$$\text{s.t.} \quad y_t = w^{\mathrm{T}}\varphi(x_t) + b + e_t, \quad t = 1, \ldots, N.$$

The relative importance between the smoothness of the solution and the data fitting is governed by the scalar $\gamma \in \mathbb{R}_0^+$ referred to as the regularization constant. The optimization performed corresponds to ridge regression (Golub & Van Loan, 1989) in feature space. In order to solve the constrained optimization problem, a Lagrangian is constructed:

$$\mathcal{L}(w, b, e; \alpha) = \mathcal{J}(w,e) - \sum_{t=1}^{N} \alpha_t (w^{\mathrm{T}}\varphi(x_t) + b + e_t - y_t),$$

with $\alpha_t \in \mathbb{R}$ the Lagrange multipliers. The conditions for optimality are given as:

$$\frac{\partial \mathcal{L}}{\partial w} = 0 \to w = \sum_{t=1}^{N} \alpha_t \varphi(x_t), \tag{1}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \to \sum_{t=1}^{N} \alpha_t = 0, \tag{2}$$

$$\frac{\partial \mathcal{L}}{\partial e_t} = 0 \to \alpha_t = \gamma e_t, \quad t = 1, \ldots, N, \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_t} = 0 \to y_t = w^{\mathrm{T}}\varphi(x_t) + b + e_t, \quad t = 1, \ldots, N. \tag{4}$$

Substituting (1)–(3) into (4) yields the following set of linear equations:

$$\left[\begin{array}{c|c} 0 & 1_N{}^T \\ \hline 1_N & \Omega + \gamma^{-1} I_N \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right], \tag{5}$$

where $y = [y_1 \ \ldots \ y_N]^{\mathrm{T}} \in \mathbb{R}^N$, $1_N = [1 \ \ldots \ 1]^{\mathrm{T}} \in \mathbb{R}^N$, $\alpha = [\alpha_1 \ \ldots \ \alpha_N]^{\mathrm{T}} \in \mathbb{R}^N$, $\Omega_{ij} = K(x_i, x_j) = \varphi(x_i)^{\mathrm{T}}\varphi(x_j)$, $\forall i, j = 1, \ldots, N$ with K the positive definite kernel function. Note that in order to solve the set of Eqs. (5), the feature map φ is never to be defined explicitly. Only its inner product in the form of a positive definite kernel, is needed. This is called the kernel trick (Vapnik, 1998; Schölkopf & Smola, 2002). For the choice of the kernel $K(\cdot,\cdot)$ see e.g. (Schölkopf & Smola, 2002). Typical examples are the use of a polynomial kernel $K(x_i, x_j) = (\tau + x_i^{\mathrm{T}} x_j)^d$ of degree d or the RBF kernel $K(x_i, x_j) = \exp(-\|x_i - x_j\|_2^2/\sigma^2)$, where $\sigma \in \mathbb{R}^+$ denotes the bandwidth of the kernel. The resulting LS-SVM model for function estimation can be evaluated at a new point x_* as

$$\hat{f}(x_*) = \sum_{t=1}^{N} \alpha_t K(x_*, x_t) + b,$$

where (b, α) is the solution to (5).

2.2. Similarities and differences with other kernel based learning methods

At this point, let us motivate why the described primal–dual approach based on convex optimization is useful to model nonlinear functions. Other methods based on splines (Wahba, 1990), Gaussian processes (Williams, 1998) and results from estimation in reproducing kernel Hilbert space (RKHS) lead to somewhat similar methods in the case of non-dynamical data. These methods often approach the subject from the point of view of functional analysis (estimation in RKHS) and Bayesian inference (Gaussian processes). An advantage of the primal–dual framework over such methods is found in the ease in which one can incorporate structure (as a bias term, parametric components or additive structure) in the estimation problem itself, which will be seen to be particularly relevant in the Hammerstein case. Moreover, the optimization point of view provides a natural point of view towards approximation techniques for handling large scale datasets (Suykens et al., 2002). The primal problem is more convenient for large datasets while the dual is suitable in high-dimensional input spaces. In the case of a finite-dimensional feature map one has the choice between the primal or the dual, but in the case of $n_{\mathrm{H}} = \infty$ only the dual can be solved exactly, while fixed size LS-SVM formulations can be used to obtain approximations for the primal problem (see fixed-size LS-SVM and its application to the Silver-box benchmark (Espinoza, Pelckmans, Hoegaerts, Suykens, & De Moor, 2004)).

3. Identification of nonlinear ARX Hammerstein models

Hammerstein systems, in their most basic form, consist of a static memoryless nonlinearity, followed by a linear dynamical system (Fig. 1). The aim of Hammerstein identification is to model the nonlinearity and to estimate the model parameters of the linear system from input/output measurements. In the following derivation, we will restrict ourselves to SISO systems (single input–single output), but as will be shown in Section 5, the presented method is applicable to the MIMO case as well. For the linear dynamical part, we will assume a model structure of the ARX form (Ljung, 1999)

$$y_t = \sum_{i=1}^{n} a_i y_{t-i} + \sum_{j=0}^{m} b_j u_{t-j} + e_t, \tag{6}$$

with $u_t, y_t \in \mathbb{R}$, $t \in \mathbb{Z}$ and $\{(u_t, y_t)\}$ a set of input and output measurements. The so-called equation error e_t is assumed to be white and m and n denote the order of the numerator and denominator in the transfer function of the linear model. The model structure (6) is generally known as the "Auto-Regressive model with eXogeneous inputs" (ARX) and is one of the most studied model structures in linear identification. Adding a static nonlinearity $f : \mathbb{R} \to \mathbb{R} : x \to f(x)$ to the input in (6) leads to

$$y_t = \sum_{i=1}^{n} a_i y_{t-i} + \sum_{j=0}^{m} b_j f(u_{t-j}) + e_t, \quad \forall t, \tag{7}$$

which is the general model structure that is assumed in this paper.

In order to apply LS-SVM function estimation as outlined in the previous section, we assume the following structure for the static nonlinearity f:

$$f(u) = w^{\mathrm{T}}\varphi(u) + d_0,$$

with $\Omega_{ij} = K(u_i, u_j) = \varphi(u_i)^{\mathrm{T}}\varphi(u_j)$ a kernel of choice. Hence, Eq. (7) can be rewritten as follows:

$$y_t = \sum_{i=1}^{n} a_i y_{t-i} + \sum_{j=0}^{m} b_j (w^{\mathrm{T}}\varphi(u_{t-j}) + d_0) + e_t. \tag{8}$$

With $r = \max(m, n) + 1$, estimates for the a_i, b_j and f follow from a finite set of measurements $\{u_t, y_t\}$, $t = 1, \ldots, N$ by solving:

$$\min_{w,a,b,d_0,e} \mathcal{J}(w, e) = \frac{1}{2} w^{\mathrm{T}} w + \gamma \frac{1}{2} \sum_{t=r}^{\mathrm{T}} e_t^2,$$

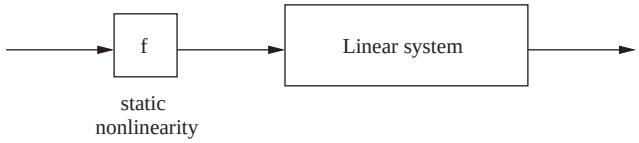


Fig. 1. A Hammerstein system consists of a memoryless static nonlinearity f followed by a linear dynamical system.

subject to (8). The Lagrangian of this constraint optimization problem is given as

$$\mathscr{L}(w, d_0, b, e, a; \alpha) = \mathscr{J}(w, e) - \sum_{t=r}^{N} \alpha_t \left(\sum_{i=1}^{n} a_i y_{t-i} + \sum_{j=0}^{m} b_j (w^{\mathrm{T}} \varphi(u_{t-j}) + d_0) + e_t - y_t \right). \quad (9)$$

The conditions for optimality are given as:

$$\frac{\partial \mathscr{L}}{\partial w} = 0 \to w = \sum_{t=r}^{N} \sum_{j=0}^{m} \alpha_t b_j \varphi(u_{t-j}), \quad (10)$$

$$\frac{\partial \mathscr{L}}{\partial d_0} = 0 \to \sum_{t=r}^{N} \sum_{j=0}^{m} \alpha_t b_j = 0,$$

$$\frac{\partial \mathscr{L}}{\partial a_i} = 0 \to \sum_{t=r}^{N} \alpha_t y_{t-i} = 0, \quad i = 1, \ldots, n,$$

$$\frac{\partial \mathscr{L}}{\partial b_j} = 0 \to \sum_{t=r}^{N} \alpha_t (w^{\mathrm{T}} \varphi(u_{t-j}) + d_0) = 0, \quad j = 0, \ldots, m,$$

$$\frac{\partial \mathscr{L}}{\partial e_t} = 0 \to \alpha_t = \gamma e_t, \quad t = r, \ldots, N, \quad (11)$$

$$\frac{\partial \mathscr{L}}{\partial \alpha_t} = 0 \to (8), \quad t = r, \ldots, N. \quad (12)$$

Substituting (10) and (11) in (12) results in the following set of nonlinear equations:

$$\sum_{j=0}^{m} \sum_{q=r}^{N} \sum_{p=0}^{m} b_j (b_p \alpha_q \varphi(u_{q-p})^{\mathrm{T}} \varphi(u_{t-j}) + d_0) + \sum_{i=1}^{n} a_i y_{t-i} + e_t - y_t = 0, \quad t = r, \ldots, N. \quad (13)$$

If the b_j values were known, the resulting problem would be linear in the unknowns and easy to solve as

$$\left[\begin{array}{c|c|c} 0 & 0 & \tilde{b} \cdot 1_{N-r+1}^{T} \\ \hline 0 & 0 & \mathcal{Y}_p \\ \hline \tilde{b} \cdot 1_{N-r+1} & \mathcal{Y}_p^{T} & \mathcal{K} + \gamma^{-1} I \end{array}\right] \left[\begin{array}{c} d_0 \\ \hline a \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline 0 \\ \hline \mathcal{Y}_f \end{array}\right], \quad (14)$$

with

$$\alpha = [\alpha_r \ \ldots \ \alpha_N]^{\mathrm{T}}, \ \tilde{b} = \sum_{j=0}^{m} b_j,$$

$$a = [a_1 \ \ldots a_n]^{\mathrm{T}}, \ \mathcal{Y}_f = [y_r \ \ldots \ y_N]^{\mathrm{T}},$$

$$\mathcal{Y}_p = \begin{bmatrix} y_{r-1} & y_r & \cdots & y_{N-1} \\ y_{r-2} & y_{r-1} & \cdots & y_{N-2} \\ \vdots & \vdots & & \vdots \\ y_{r-n} & y_{r-n+1} & \cdots & y_{N-n} \end{bmatrix},$$

$$\mathcal{K}(p, q) = \sum_{j=0}^{m} \sum_{l=0}^{m} b_j b_l \Omega_{p+r-j-1, q+r-l-1},$$

$$\Omega_{k,l} = \varphi(u_k)^{\mathrm{T}} \varphi(u_l), \quad \forall k, \ l = 1, \ldots, N.$$

Since the b_j values are in general not known and the solution to the resulting third order estimation problem (13) is by no means trivial, we will use an approximative method to obtain models of the form (7).

4. An approximative method

4.1. Optimization using collinearity constraints

In order to avoid solving problem (13), we rewrite (8) as follows:

$$y_t = \sum_{i=1}^{n} a_i y_{t-i} + \sum_{j=0}^{m} w_j^{\mathrm{T}} \varphi(u_{t-j}) + d + e_t, \quad (15)$$

which can conveniently be solved using LS-SVMs (Pelckmans, Goethals, De Brabanter, Suykens, & De Moor, 2004). Note, however, that the resulting model class is wider than (8) due to the replacement of one single w by several vectors $w_j, j = 0, \ldots, m$. The model class (15) is, therefore, not necessarily limited to the description of Hammerstein systems. A sufficient condition for the estimated model to belong to this class of systems is that the obtained w_j must be collinear in which case w_j is seen as a replacement for $b_j w$. Taking this into account during the estimation leads to extra constraints requiring the angles between any pair $\{w_j, w_k\}$, $j, k = 0, \ldots, m$ to be zero, or $(w_j^{\mathrm{T}} w_k)^2 = \sqrt{w_j^{\mathrm{T}} w_j} \sqrt{w_k^{\mathrm{T}} w_k}$. Alternatively, the collinearity constraint can be written as: $\mathrm{rank}\,[w_0 \ \ldots \ w_m] = 1$, which is equivalent to ensuring that a set of $m(m+1)n_{\mathrm{H}}(n_{\mathrm{H}}-1)/4$ 2×2 determinants are zero. As n_{H} (the dimension of w) is unknown and possibly very high, it is well-known that including such constraints in the Lagrangian would again lead to a non-convex optimization problem.

Considering the fact that ARX Hammerstein models are contained in the set of models of the form (15), we therefore propose to remove the collinearity constraints from the Lagrangian altogether, solve the more general problem (15), and project the obtained model onto the model-set (8) later. Hereby, we assume that collinearity is almost satisfied in the estimated model of the form (15) as the data originate from the Hammerstein model class. Although this approach may seem ad hoc at first, it is essentially an application of Bai's overparameterization approach (Bai & Fu, 1998)

to LS-SVMs. The key ideas behind the overparameterization approach are introduced in Section 6. Some examples of overparameterization approaches applied to the Hammerstein identification problem are found in Chang and Luus (1971), Pawlak (1991), and McKelvey and Hanner (2003).

4.2. Optimization without collinearity constraints

Disregarding the collinearity constraints, the optimization problem that is ultimately solved is the following:

$$\min_{w_j,a,d,e} \quad \mathscr{J}(w_j,e)=\frac{1}{2}\sum_{j=0}^{m} w_j^{\mathrm{T}}w_j+\gamma\frac{1}{2}\sum_{t=r}^{N} e_t^2, \tag{16}$$

$$\text{s.t.} \quad \sum_{t=1}^{N} w_j^{\mathrm{T}}\varphi(u_t)=0, \tag{17}$$

$$\sum_{j=0}^{m} w_j^{\mathrm{T}}\varphi(u_{t-j})+\sum_{i=1}^{n} a_i y_{t-i}+d+e_t-y_t=0, \tag{18}$$

with $t=r,\ldots,N$ and $j=0,\ldots,m$. Note the additional constraints (17) which center the nonlinear functions $w_j^{\mathrm{T}}\varphi(\cdot)$, $j=0,\ldots,m$ around their average over the training set. This removes the uncertainty resulting from the fact that any set of constants can be added to the terms of the additive nonlinear function (15), as long as the sum of the constants is zero (Hastie, Tibshirani, & Friedman, 2001). Removing this uncertainty will facilitate the extraction of the parameters b_j in (7) later. Furthermore, this constraint enables us to give a clear meaning for the bias parameter d, namely $d=\sum_{j=0}^{m} b_j((1/N)\sum_{k=1}^{N} f(u_k))$.

Lemma 4.1 (*Primal–dual derivation*). *Given system* (15), *the LS-SVM estimates for the nonlinear functions* $w_j^{\mathrm{T}}\varphi : \mathbb{R}\to\mathbb{R}$, $j=0,\ldots,m$ *are given as*

$$w_j^{\mathrm{T}}\varphi(u_*)=\sum_{t=r}^{N}\alpha_t K(u_{t-j},u_*)+\beta_j\sum_{t=1}^{N}K(u_t,u_*), \tag{19}$$

where the parameters $\alpha_t, t=r,\ldots,N$, $\beta_j, j=0,\ldots,m$, *as well as the linear model parameters* $a_i, i=1,\ldots,n$ *and* d *are obtained from the following set of linear equations*:

$$\left[\begin{array}{c|c|c|c} 0 & 0 & 1^T & 0 \\ \hline 0 & 0 & \mathcal{Y}_p & 0 \\ \hline 1 & \mathcal{Y}_p^T & \mathcal{K}+\gamma^{-1}I & K^0 \\ \hline 0 & 0 & K^{0T} & 1_N^T\Omega 1_N\cdot I_{m+1} \end{array}\right]\left[\begin{array}{c} d \\ \hline a \\ \hline \alpha \\ \hline \beta \end{array}\right]=\left[\begin{array}{c} 0 \\ \hline 0 \\ \hline \mathcal{Y}_f \\ \hline 0 \end{array}\right], \tag{20}$$

with $\beta=[\beta_0\ \ldots\ \beta_m]^{\mathrm{T}}$, $K^0(p,q)=\sum_{t=1}^{N}\Omega_{t,r+p-q}$, $\mathcal{K}(p,q)=\sum_{j=0}^{m}\Omega_{p+r-j-1,q+r-j-1}$, *and* 1_N *is a column vector of length* N *with elements* 1.

Proof. This directly follows from the Lagrangian:

$$\begin{aligned}\mathscr{L}(w_j,d,a,e;\alpha,\beta)&=\mathscr{J}(w_j,e)-\sum_{j=0}^{m}\beta_j\left(\sum_{t=1}^{N}w_j^{\mathrm{T}}\varphi(u_t)\right)-\sum_{t=r}^{N}\alpha_t\\&\quad\times\left(\sum_{i=1}^{n}a_i y_{t-i}+\sum_{j=0}^{m}w_j^{\mathrm{T}}\varphi(u_{t-j})+d+e_t-y_t\right),\end{aligned} \tag{21}$$

by taking the conditions for optimality: $\partial\mathscr{L}/\partial w_j=0$, $\partial\mathscr{L}/\partial a_i=0$, $\partial\mathscr{L}/\partial d=0$, $\partial\mathscr{L}/\partial e_t=0$, $\partial\mathscr{L}/\partial\alpha_t=0$, $\partial\mathscr{L}/\partial\beta_j=0$. □

4.3. Projecting the unconstrained solution onto the class of NARX Hammerstein models

The projection of the obtained model onto (7) goes as follows. Estimates for the autoregressive parameters $a_i, i=1,\ldots,n$ are directly obtained from (20). Furthermore, for the training input sequence $[u_1\ \ldots\ u_N]$, we have

$$\begin{aligned}&\begin{bmatrix} b_0 \\ \vdots \\ b_m \end{bmatrix}\begin{bmatrix} \underline{\hat{f}}(u_1) \\ \vdots \\ \underline{\hat{f}}(u_N) \end{bmatrix}^{\mathrm{T}}\\&=\begin{bmatrix} \alpha_N & \ldots & \alpha_r & & 0 \\ & \alpha_N & \ldots & \alpha_r & \\ & & \ddots & & \ddots \\ 0 & & & \alpha_N & \ldots & \alpha_r \end{bmatrix}\\&\quad\times\begin{bmatrix} \Omega_{N,1} & \Omega_{N,2} & \ldots & \Omega_{N,N} \\ \Omega_{N-1,1} & \Omega_{N-1,2} & \ldots & \Omega_{N-1,N} \\ \vdots & \vdots & & \vdots \\ \Omega_{r-m,1} & \Omega_{r-m,2} & \ldots & \Omega_{r-m,N} \end{bmatrix}\\&\quad+\begin{bmatrix} \beta_0 \\ \vdots \\ \beta_m \end{bmatrix}\sum_{t=1}^{N}\begin{bmatrix} \Omega_{t,1} \\ \vdots \\ \Omega_{t,N} \end{bmatrix}^{\mathrm{T}},\end{aligned} \tag{22}$$

with $\underline{\hat{f}}(u)$ an estimate for $\underline{f}(u)=f(u)-(1/N)\sum_{t=1}^{N}f(u_t)$. Hence, estimates for b_j and the static nonlinearity f can be obtained from a rank 1 approximation of the right-hand side of (22), for instance using a singular value decomposition. Again, this is equivalent to the SVD-step that is generally encountered in overparameterization methods (Chang & Luus, 1971; Bai & Fu, 1998). Once all the elements b_j are known, $\sum_{t=1}^{N}f(u_k)$ can be obtained as $\sum_{t=1}^{N}f(u_t)=Nd/\sum_{j=0}^{m}b_j$.

5. Extension to the MIMO case

Technically, an extension of the algorithms presented in the former section to the MIMO case is straightforward, but

the calculations involved are quite extensive. Assuming a MIMO Hammerstein system of the form:

$$y_t = \sum_{i=1}^{n} A_i y_{t-i} + \sum_{j=0}^{m} B_j f(u_{t-j}) + e_t, \quad \forall t \tag{23}$$

with $y_t, e_t \in \mathbb{R}^{n_y}, u_t \in \mathbb{R}^{n_u}, A_i \in \mathbb{R}^{n_y \times n_y}, B_j \in \mathbb{R}^{n_y \times n_u}, t = 1, \ldots, N, i = 1, \ldots, n, j = 0, \ldots, m$, and $f : \mathbb{R}^{n_u} \to \mathbb{R}^{n_u} : u \to f(u) = [f_1(u) \ \ldots \ f_{n_u}(u)]^{\mathrm{T}}$, we have for every row s in (23), that

$$y_t(s) = \sum_{i=1}^{n} A_i(s,:) y_{t-i} + \sum_{j=0}^{m} B_j(s,:) f(u_{t-j}) + e_t(s). \tag{24}$$

Substituting $f(u) = [f_1(u) \ \ldots \ f_{n_u}(u)]^{\mathrm{T}}$ in (24) leads to:

$$y_t(s) = \sum_{i=1}^{n} A_i(s,:) y_{t-i} + \sum_{j=0}^{m} \sum_{k=1}^{n_u} B_j(s,k) f_k(u_{t-j}) + e_t(s), \quad \forall t, s. \tag{25}$$

By replacing $\sum_{k=1}^{n_u} B_j(s,k) f_k(u_{t-j})$ by $w_{j,s}^{\mathrm{T}} \varphi(u_{t-j}) + d_{s,j}$ this reduces to

$$y_t(s) = \sum_{i=1}^{n} A_i(s,:) y_{t-i} + \sum_{j=0}^{m} \omega_{j,s}^{\mathrm{T}} \varphi(u_{t-j}) + d_s + e_t(s), \quad \forall t, s, \tag{26}$$

where $d_s = \sum_{j=0}^{m} d_{s,j}$. The optimization problem that is solved then is the following:

$$\mathscr{J}(\omega_{j,s}, e) = \sum_{j=0}^{m} \sum_{s=1}^{n_y} \frac{1}{2} \omega_{j,s}^{\mathrm{T}} \omega_{j,s} + \frac{\gamma_s}{2} \sum_{s=1}^{n_y} \sum_{t=r}^{N} e_t(s)^2, \tag{27}$$

subject to (26) and $\sum_{t=1}^{N} w_{j,s}^{\mathrm{T}} \varphi(u_t) = 0, j = 0, \ldots, m, s = 1, \ldots, n_y$.

Lemma 5.1 (*Primal–dual derivation of the MIMO case*). *Given system* (26), *the LS-SVM estimates for the nonlinear functions* $w_{j,s}^{\mathrm{T}} \varphi : \mathbb{R} \to \mathbb{R}, j = 0, \ldots, m, s = 1, \ldots, n_y$, *are given as*:

$$w_{j,s}^{\mathrm{T}} \varphi(u_*) = \sum_{t=r}^{N} \alpha_{t,s} K(u_{t-j}, u_*) + \beta_{j,s} \sum_{t=1}^{N} K(u_t, u_*), \tag{28}$$

where the unknowns $\alpha_{t,s}, t = r, \ldots, N, s = 1, \ldots, n_y$, $\beta_{j,s}, j = 0, \ldots, m, s = 1, \ldots, n_y$ *as well as the linear model parameters* $A_i, i = 1, \ldots, n$ *and* $d_s, s = 1, \ldots, n_y$ *are obtained from the following set of linear equations*:

$$\begin{bmatrix} L_1 & & \\ & \ddots & \\ & & L_{n_y} \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_{n_y} \end{bmatrix} = \begin{bmatrix} R_1 \\ \vdots \\ R_{n_y} \end{bmatrix}, \tag{29}$$

where

$$L_s = \left[\begin{array}{c|c|c|c} 0 & 0 & 1^T & 0 \\ \hline 0 & 0 & \mathcal{Y}_p & 0 \\ \hline 1 & \mathcal{Y}_p^T & \mathcal{K} + \gamma_s^{-1} I & \mathcal{S} \\ \hline 0 & 0 & \mathcal{S}^T & \mathcal{T} \end{array}\right], X_s = \left[\begin{array}{c} d_s \\ \hline \mathcal{A}_s \\ \hline \bar{\alpha}_s \\ \hline \bar{\beta}_s \end{array}\right],$$

$$R_s = \left[\begin{array}{c|c|c|c} 0 & 0 & \mathcal{Y}_{f,s}^T & 0 \end{array}\right]^T, \mathcal{Y}_{f,s} = \left[y_r(s)^T \ \ldots \ y_N(s)^T\right]^T,$$

$$\mathcal{A}_s = \begin{bmatrix} A_1(s,:)^T \\ \vdots \\ A_m(s,:)^T \end{bmatrix}, \bar{\alpha}_s = \begin{bmatrix} \alpha_{r,s} \\ \vdots \\ \alpha_{N,s} \end{bmatrix},$$

$$\bar{\beta}_s = \left[\beta_{0,s} \ \ldots \ \beta_{m,s}\right]^T, \Omega_{p,q} = \varphi(u_p)^T \varphi(u_q),$$

$$\mathcal{K}(p,q) = \sum_{j=0}^{m} \Omega_{p+r-j-1, q+r-j-1},$$

$$\mathcal{T} = 1_N^T \Omega 1_N \cdot I_{m+1}, \ \mathcal{S}(p,q) = \sum_{t=1}^{N} \Omega_{t,r+p-q},$$

Proof. This directly follows from the Lagrangian:

$$\begin{aligned} &\mathscr{L}(\omega_{j,s}, d_s, A, e; \alpha, \beta) \\ &= \mathscr{J}(\omega_{j,s}, e) - \sum_{t=r}^{N} \sum_{s=1}^{n_y} \alpha_{t,s} \left(\sum_{i=1}^{n} A_i(s,:) y_{t-i} \right. \\ &\quad \left. + \sum_{j=0}^{m} \omega_{j,s}^{\mathrm{T}} \varphi(u_{t-j}) + d_s + e_t(s) - y_t(s) \right) \\ &\quad - \sum_{j=0}^{m} \sum_{s=1}^{n_y} \beta_{j,s} \left(\sum_{t=1}^{N} \omega_{j,s}^{\mathrm{T}} \varphi(u_t) \right), \end{aligned} \tag{30}$$

by taking the conditions for optimality: $\partial \mathscr{L}/\partial \omega_{j,s} = 0$, $\partial \mathscr{L}/\partial A_i(s,:) = 0$, $\partial \mathscr{L}/\partial d_s = 0$, $\partial \mathscr{L}/\partial e_t(s) = 0$, $\partial \mathscr{L}/\partial \alpha_{t,s} = 0$, $\partial \mathscr{L}/\partial \beta_{j,s} = 0$. □

Note that the matrices $L_s, s = 1, \ldots, n_y$ in (29) are almost identical, except for the different regularization constants γ_s. In many practical cases, however, and if there is no reason to assume that a certain output is more important than another, it is recommended to set $\gamma_1 = \gamma_2 = \cdots = \gamma_{n_y}$. This will reduce the number of hyper-parameters to be tuned and will speed up the estimation algorithm since $L_1 = L_2 = \cdots = L_{n_y}$ needs to be calculated only once.

The projection of the obtained model onto (25) is similar as in the SISO case. Estimates for the autoregressive matrices $A_i, i = 1, \ldots, n$ are directly obtained from (29). For the training input sequence $[u_1 \ \ldots \ u_N]$ and every $k = 1, \ldots, n_u$,

we have

$$\begin{bmatrix} B_0(1,:) \\ \vdots \\ B_m(1,:) \\ \hline \vdots \\ \hline B_0(n_y,:) \\ \vdots \\ B_m(n_y,:) \end{bmatrix} \begin{bmatrix} \underline{\hat{f}}^T(u_1) \\ \vdots \\ \underline{\hat{f}}^T(u_N) \end{bmatrix}^T = \begin{bmatrix} \beta_{0,1} \\ \vdots \\ \beta_{m,1} \\ \hline \vdots \\ \hline \beta_{0,n_y} \\ \vdots \\ \beta_{m,n_y} \end{bmatrix} \sum_{t=1}^{N} \begin{bmatrix} \Omega_{t,1} \\ \vdots \\ \Omega_{t,N} \end{bmatrix}^T + \mathcal{A} \times \begin{bmatrix} \Omega_{N,1} & \Omega_{N,2} & \dots & \Omega_{N,N} \\ \Omega_{N-1,1} & \Omega_{N-1,2} & \dots & \Omega_{N-1,N} \\ \vdots & \vdots & & \vdots \\ \Omega_{r-m,1} & \Omega_{r-m,2} & \dots & \Omega_{r-m,N} \end{bmatrix} \tag{31}$$

with $\underline{\hat{f}}(u)$ an estimate for

$$\underline{f}(u) = f(u) - g, \tag{32}$$

and g a constant vector such that:

$$\sum_{j=0}^{m} B_j g = [d_1 \ \cdots \ d_{n_y}]^{\mathrm{T}}. \tag{33}$$

Estimates for $\underline{f}$ and the B_j, $j = 0, \dots, m$, can be obtained through a rank-n_u approximation of the right-hand side of (31). From $\underline{f}$ in (32) and g in (33), finally, an estimate for the nonlinear function f can be obtained. Note that if the row-rank of $\sum_{j=0}^{m} B_j$ is smaller than the column-rank, multiple choices for g are possible. This results as an inherent property of blind MIMO Hammerstein identification.

6. Comparison with existing overparameterization algorithms

As was mentioned in Section 4.1, the presented technique is closely related to the overparameterization approach (Chang & Luus, 1971; Bai & Fu, 1998). The idea of overparameterization can be summarized as writing the static nonlinearity f as a linear combination of general nonlinear basis functions f_k. In this framework, each basis function has a certain weight c_k, $f(u_t) = \sum_{k=1}^{n_f} c_k f_k(u_t)$. The functions f_1, f_2, and f_{n_f} are chosen beforehand. Starting from (7) and substituting the expansion for f leads to

$$y_t = \sum_{i=1}^{n} a_i y_{t-i} + \sum_{k=0}^{n_f} \sum_{j=0}^{m} b_j c_k f_k(u_{t-j}) + e_t \tag{34}$$

$$= \sum_{i=1}^{n} a_i y_{t-i} + \sum_{k=1}^{n_f} \sum_{j=0}^{m} \theta_{j,k} f_k(u_{t-j}) + e_t, \tag{35}$$

which can be solved for $\theta_{j,k} = b_j c_k$, $j = 0, \dots, m, k = 1, \dots, n_f$ using a least squares algorithm. Estimates for the b_j and c_k are recovered from the SVD of

$$\begin{bmatrix} \hat{\theta}_{0,1} & \theta_{0,2} & \dots & \hat{\theta}_{0,n_f} \\ \hat{\theta}_{1,1} & \theta_{1,2} & \dots & \hat{\theta}_{1,n_f} \\ \vdots & \vdots & & \vdots \\ \hat{\theta}_{m,1} & \theta_{m,2} & \dots & \hat{\theta}_{m,n_f} \end{bmatrix}. \tag{36}$$

Note that for any set of variables ε_k, $k=1, \dots, n_f$ with $\forall u \in \mathbb{R}$, $\sum_{k=1}^{n_j} \varepsilon_k f_k(u) =$ constant and any set α_j, $j = 0, \dots, m$ such that $\sum_{j=0}^{m} \alpha_j = 0$, $\theta'_{j,k} = \theta_{j,k} + \alpha_j \varepsilon_k$ is also a solution to (35). This problem is often overlooked in existing overparameterization techniques and may lead to conditioning problems and destroy the low-rank property of (36). In fact, for the examples presented in the following section, existing overparameterization approaches lead to results which are far from optimal if no measures are taken to overcome this problem. One possible solution is to calculate

$$A = \begin{bmatrix} \hat{\theta}_{0,1} & \hat{\theta}_{0,2} & \dots & \hat{\theta}_{0,n_f} \\ \hat{\theta}_{1,1} & \hat{\theta}_{1,2} & \dots & \hat{\theta}_{1,n_f} \\ \vdots & \vdots & & \vdots \\ \hat{\theta}_{m,1} & \hat{\theta}_{m,2} & \dots & \hat{\theta}_{m,n_f} \end{bmatrix} \times \begin{bmatrix} f_1(u_1) & \dots & f_1(u_N) \\ f_2(u_1) & \dots & f_2(u_N) \\ \vdots & & \vdots \\ f_{n_f}(u_1) & \dots & f_{n_f}(u_N) \end{bmatrix},$$

with u_k, $k = 1, \dots, N$ the inputs of the system, subtract the mean of every row in A and take the SVD of the remaining matrix, from which estimates for the b_j can be extracted. Estimates for the c_k can then be found in a second round by solving (34). It is this approach that will be used for the implementation of classical overparameterization approaches in the following section. Note that the approach amounts to setting the mean of $\hat{f} = \sum_{k=1}^{N} \hat{f}_k$ over the inputs $u_1, \dots, u_N$ to zero, which is similar to what was done for the LS-SVM, with the exception that in the latter case this constraint was explicitly introduced in the Lagrangian (21).

7. Illustrative examples

7.1. SISO system

The algorithm proposed in this paper was used for identification on the following SISO Hammerstein system:

$$A(z)y = B(z)f(u) + e, \tag{37}$$

with A and B polynomials in the forward shift operator z, where $B(z) = z^6 + 0.8z^5 + 0.3z^4 + 0.4z^3$, $A(z) = (z - 0.98\mathrm{e}^{\pm \mathrm{i}})(z - 0.98\mathrm{e}^{\pm 1.6\mathrm{i}})(z - 0.97\mathrm{e}^{\pm 0.4\mathrm{i}})$, and $f : \mathbb{R} \to \mathbb{R} : f(u) = \sin \mathrm{c}(u)u^2$ the static nonlinearity. A white Gaussian input sequence u with length 400, zero mean and standard deviation 2 was generated and fed into system (37). During the simulation the equation noise was

Table 1
Mean and variances of obtained distances between estimated and true nonlinearities in a SISO example

Method	mean(d)	std(d)
LS-SVM $\gamma = 500$	0.0064	0.0041
Hermite $n_f = 15$	0.2203	0.7842
Hermite $n_f = 20$	0.7241	2.3065
Hermite $n_f = 25$	1.1217	2.9660
Hermite $n_f = 30$	1.0118	2.9169
Gaussian $n_f = 18$	0.0142	0.0141
Gaussian $n_f = 24$	0.0193	0.1055
Gaussian $n_f = 30$	0.0168	0.0693
Gaussian $n_f = 36$	0.0188	0.0764

Table 2
Mean and variances of obtained distances between estimated and true nonlinearities in a SISO example

Method	mean(d)	std(d)
LS-SVM $\gamma = 500$	0.0064	0.0041
Gaussian $\gamma = 10^{13}$	0.0457	0.1028
Gaussian $\gamma = 10^{12}$	0.0089	0.0071
Gaussian $\gamma = 10^{11}$	0.0088	0.0060
Gaussian $\gamma = 10^{10}$	0.0112	0.0086

chosen white Gaussian with zero mean and as standard deviation 10% of the standard deviation of the sequence $f(u)$. The last 200 datapoints of u and the generated output y were used for identification using the following three techniques:

- *LS-SVM*: The LS-SVM estimation procedure as described in Section 4: The linear system (20) is solved for d, a, α, β. An SVD of the right-hand side of (22) is thereafter performed to obtain estimates for the linear system and the static nonlinearity. For the example, an RBF-kernel with $\sigma = 1$ was used. Different values for the regularization parameter γ were tested by applying the obtained model to an independent validation sequence. From these tests $\gamma = 500$ was selected as the best candidate.
- *Hermite*: The general algorithm described in Section 6 with $f_k(u) = \mathrm{e}^{u^2}(d^{k-1}/\mathrm{d}u^{k-1})\mathrm{e}^{-u^2}$, the Hermite polynomial of order $k-1$. This expansion was used in (Greblicki, 1989) for Hammerstein- and (Greblicki, 1994) for Wiener systems.
- *RBF network* (*Gaussian*): The general algorithm described in Section 6 with $f_k(\cdot), k = 1, \ldots, n_f$ localised Gaussian radial basis functions with location depending on the value of k. As no prior information about the nature of the static nonlinearity is assumed during the identification step, the locations of the Gaussian nonlinearities were chosen equidistantly spread between -4 and 4. The bandwidth was chosen equal to one, in line with the $\sigma = 1$, choice for LS-SVM. The main reason for considering this algorithm is that it is a parametric counterpart to the LS-SVM approach with an RBF-kernel, where the final solution is expressed as a sum of Gaussian basis functions around the training datapoints.

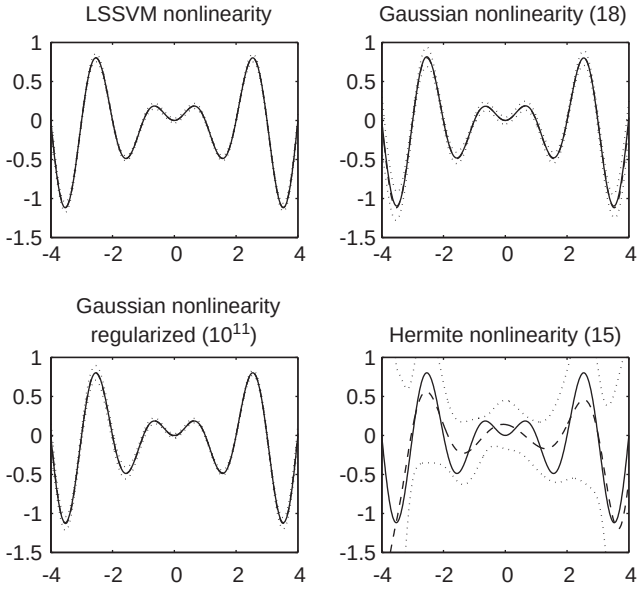


Fig. 2. True nonlinearity (solid) and mean estimated nonlinearity (dashed) for the different techniques compared in a Monte-Carlo simulation of a SISO system. Results for the LS-SVM algorithm with $\gamma = 500$ are displayed in the top-left figure, those for the Gaussian approach with $n_f = 18$ and without regularization in the top-right figure. The bottom-left figure displays the results for the Gaussian algorithm with $n_f = 46$ and constant $\gamma = 10^{11}$ tuned using validation on an independent dataset. The bottom-right figure displays the results for the Hermitian algorithm with $n_f = 15$. 90% confidence bounds on the estimated nonlinearities, following from the Monte-Carlo simulation, are included in each plot (dotted). The Hermite-approach is obviously inferior to the Gaussian and the LS-SVM technique. The best performance is obtained with the LS-SVM algorithm.

Hundred Monte-Carlo experiments were performed following the description above with $n = 6$, $m = 3$. For each experiment and each obtained estimate $\hat{f}$ for the static nonlinearity f, the distance $d = \int_{-4}^{4} \| f(x) - \hat{f}(x) \| \, \mathrm{d}x$ was calculated. The mean and variance of the distances so obtained using the LS-SVM technique are compared to those obtained from the Hermite- and Gaussian approach using different values for n_f. The results are displayed in Table 1. Note that the LS-SVM technique clearly performs better than the Hermite-approach and about three times better than the Gaussian approach. The Gaussian and the LS-SVM technique are similar in nature as in both cases the estimated nonlinearity is written as a sum of Gaussian basis functions with fixed bandwidth 1. However, it should be noted at this point that the RBF-kernel is but one possible choice in the LS-SVM algorithm, and that in principle any positive definite kernel can be chosen. A big disadvantage for the Gaussian approach is that it suffers from overfitting once the parameter n_f is chosen too high, even though with the 200 datapoints available

and $n = 6$, $m = 3$, one could easily go to $n_f = 46$ before the resulting linear equations become underdetermined. To avoid the increased variance, an extra regularization term $\gamma^{-1}\sum_{k=1}^{n_f}\sum_{j=0}^{n}\theta_{j,k}^2$ can be applied to the estimation problem (35). Results for the Gaussian approach including such a regularization term, and with $n_f = 46$, are displayed in Table 2. Note that the performance of the Gaussian estimator has drastically improved, but is still about 50% worse than the LS-SVM estimator (Fig. 2).

8. Conclusions

In this paper, we have proposed a new technique for the identification of MIMO Hammerstein ARX systems. The method is based on least squares support vector machines function approximation and allows to determine the memoryless static nonlinearity as well as the linear model parameters from a linear set of equations. The method was compared to results of two other Hammerstein identification algorithms to illustrate its performance. This combined with the straightforward derivation of the results, the availability of a strong regularization framework (Vapnik, 1998; Schölkopf & Smola, 2002; Suykens et al., 2002), and the freedom that one gets in modelling the nonlinearity by the design of an appropriate positive definite kernel makes the proposed technique an excellent candidate for Hammerstein model identification.

Acknowledgements

I. Goethals is a Research Assistant with the Fund for Scientific Research-Flanders (FWO-Vlaanderen). K. Pelckmans is a Research Assistant with KULeuven. J.A.K. Suykens is an Associate Professor with KULeuven. B. De Moor is a Full Professor with KULeuven. Our research is supported by: Research Council KUL: GOA-Mefisto 666, GOA AMBioRICS, several Ph.D./postdoc & fellow grants; Flemish Government: FWO: Ph.D./postdoc grants, projects, G.0240.99, G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0211.05, G.0080.01 research communities (ICCoS, ANMMM, MLDM); AWI: Bil. Int. Collaboration Hungary/Poland; IWT: Ph.D. Grants, GBOU (McKnow), Belgian Federal Science Policy Office: IUAP P5/22; PODO-II; EU: FP5-Quprodis; ERNSI; Eureka 2063-IMPACT; Eureka 2419-FliTE; Contract Research/agreements: ISMC/IPCOS, Data4s, TML, Elia, LMS, Mastercard.

References

Bai, E. W. (2002). A blind approach to Hammerstein model identification. *IEEE Transactions on Signal Processing*, *50*(7), 1610–1619.

Bai, E. W., & Fu, M. (1998). An optimal two-stage identification algorithm for Hammerstein–Wiener nonlinear systems. *Automatica*, *4*(3), 333–338.

Chang, F. H. I., & Luus, R. (1971). A noniterative method for identification using the Hammerstein model. *IEEE Transactions on Automatic Control*, *16*, 464–468.

Crama, P., & Schoukens, J. (2001). Initial estimates of Wiener and Hammerstein systems using multisine excitation. *IEEE Transactions on Measurement and Instrumentation*, *50*(6), 1791–1795.

Dempsey, E. J., & Westwick, D. T. (2004). Identification of Hammerstein models with cubic spline nonlinearities. *IEEE Transactions on Biomedical Engineering*, *51*, 237–245.

Espinoza, M., Pelckmans, K., Hoegaerts, L., Suykens, J., & De Moor, B. (2004). A comparative study of ls-svms applied to the silver box identification problem. *Symposium on nonlinear control systems (NOLCOS 2004)*, (pp. 513–518), Stuttgart, Germany.

Golub, G. H., & Van Loan, C. F. (1989). *Matrix computations*. Baltimore, MD: John Hopkins University Press.

Greblicki, W. (1989). Non-parametric orthogonal series identification of Hammerstein systems. *International Journal of Systems Science*, *20*(12), 2355–2367.

Greblicki, W. (1994). Nonparametric identification of Wiener systems by orthogonal series. *IEEE Transactions on Automatic Control*, *39*(10), 2077–2086.

Greblicki, W., & Pawlak, M. (1986). Identification of discrete Hammerstein systems using kernel regression estimates. *IEEE Transactions on Automatic Control*, *31*, 74–77.

Greblicki, W., & Pawlak, M. (1991). Nonparametric identification of a cascade nonlinear time series system. *Signal Processing*, *22*, 61–75.

Hasiewicz, Z. (1999). Hammerstein system identification by the Haar multiresolution approximation. *International Journal of Adaptive Control and Signal Processing*, *13*(8), 691–717.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Heidelberg: Springer.

Janczak, A. (2003). Neural network approach for identification of Hammerstein systems. *International Journal of Control*, *76*(17), 1749–1766.

Krzyżak, A. (1989). Identification of discrete Hammerstein systems by the Fourier series regression estimate. *International Journal of Systems Science*, *20*, 1729–1744.

Ljung, L. (1999). *System identification—theory for the user*. 2nd ed., Upper Saddle River, NJ: Prentice-Hall.

McKelvey, T., & Hanner, C. (2003). On identification of Hammerstein systems using excitation with a finite number of levels. *Proceedings of the 13th International Symposium on System Identification (SYSID2003)*, (pp. 57–60).

Narendra, K. S., & Gallman, P. G. (1966). An iterative method for the identification of nonlinear systems using the Hammerstein model. *IEEE Transactions on Automatic Control*, *11*, 546–550.

Pawlak, M. (1991). On the series expansion approach to the identification of Hammerstein systems. *IEEE Transactions on Automatic Control*, *36*, 736–767.

Pelckmans, K., Goethals, I., De Brabanter, J., Suykens, J. A. K., De Moor, B., & Wang, L. (Ed.), (2004). Componentwise least squares support vector machines. *Support vector machines*: *Theory and applications*, Springer, in press.

Schölkopf, B., & Smola, A. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P., Hjalmarsson, H., & Juditsky, A. (1995). Nonlinear black-box modeling in system identification: A unified overview. *Automatica*, *31*(12), 1691–1724.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. Singapore: World Scientific.

van Pelt, T. H., & Bernstein, D. S. (2000). Nonlinear system identification using Hammerstein and nonlinear feedback models with piecewise linear static maps—Part i: Theory. *Proceedings of the American control conference (ACC2000)*, (pp. 225–229).

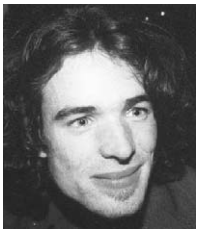
Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.

Verhaegen, M., & Westwick, D. (1996). Identifying MIMO Hammerstein systems in the context of subspace model identification methods. *International Journal of Control*, *63*, 331–349.

Wahba, G. (1990). *Spline models for observational data*. Philadelphia, PA: SIAM.

Williams, C. K. I. (1998). Prediction with Gaussian processes: from linear regression to linear prediction and beyond. in: M.I. Jordan (Ed.), *Learning and inference in graphical models* (pp. 599–621). Dordrecht: Kluwer Academic.

Ivan Goethals was born in Wilrijk, Belgium, in 1978. He obtained the M.Sc. degree in Nuclear Physics in 2000 from the "Katholieke Universiteit Leuven", Leuven, Belgium. Currently he is working towards a Ph.D. in the SCD-SISTA research group of the Department of Electrical Engineering (ESAT) from the "Katholieke Universiteit Leuven". His main research interests are in the fields of linear and nonlinear system identification.



Kristiaan Pelckmans was born on 3 November 1978 in Merksplas, Belgium. He received a M.Sc. degree in Computer Science in 2000 from the Katholieke Universiteit Leuven. After a projectwork for an implementation of kernel machines and LS-SVMs (LS-SVM1ab), he currently pursues a Ph.D. at the Kuleuven in the faculty of Applied Sciences, Department of Electrical Engineering in the SCD SISTA laboratory. His research mainly focusses on machine learning and statistical inference using primal–dual kernel machines.

Johan A.K. Suykens was born in Willebroek Belgium, May 18 1966. He received the degree in Electro-Mechanical Engineering and the Ph.D. degree in Applied Sciences from the Katholieke Universiteit Leuven, in 1989 and 1995, respectively. In 1996, he has been a Visiting Postdoctoral Researcher at the University of California, Berkeley. He has been a Postdoctoral Researcher with the Fund for Scientific Research FWO Flanders and is currently an Associate Professor with K.U.Leuven.

His research interests are mainly in the areas of the theory and application of neural networks and nonlinear systems. He is author of the books "Artificial Neural Networks for Modelling and Control of Nonlinear Systems" (Kluwer Academic Publishers) and "Least Squares Support Vector Machines" (World Scientific) and editor of the books "Nonlinear Modeling: Advanced Black-Box Techniques" (Kluwer Academic Publishers) and "Advances in Learning Theory: Methods, Models and Applications" (IOS Press). In 1998 he organized an International Workshop on Nonlinear Modelling with Time-series Prediction Competition. He is a Senior IEEE member and has served as associate editor for the IEEE Transactions on Circuits and Systems—Part I (1997–1999) and Part II (since 2004) and since 1998, he is serving as associate editor for the IEEE Transactions on Neural Networks. He received an IEEE Signal Processing Society 1999 Best Paper (Senior) Award and several Best Paper Awards at International Conferences. He is a recipient of the International Neural Networks Society (INNS) 2000, Young Investigator Award for significant contributions in the field of neural networks. He has served as Director and Organizer of a NATO Advanced Study Institute on Learning Theory and Practice (Leuven 2002) and as a program co-chair for the International Joint Conference on Neural Networks IJCNN 2004.

Bart De Moor was born Tuesday July 12, 1960 in Halle, Belgium. He is married and has three children. In 1983, he obtained his Master (Engineering) Degree in Electrical Engineering at the Katholieke Universiteit Leuven, Belgium, and a Ph.D. in Engineering at the same university in 1988. He spent 2 years as a Visiting Research Associate at Stanford University (1988–1990) at the Departments of EE (ISL, Prof. Kailath) and CS (Prof. Golub). Currently, he is a full professor at the Department of Electrical Engineering (http://www.esat.kuleuven.ac.be) of the K.U.Leuven.

His research interests are in numerical linear algebra and optimization, system theory and identification, quantum information theory, control theory, data-mining, information retrieval and bioinformatics, areas in which he has (co) authored several books and hundreds of research papers (consult the publication search engine at www.esat.kuleuven.ac.be/sista-cosic-docarch/template.php).

Currently, he is leading a research group of 39 Ph.D. students of 8 postdocs and in the recent past, 16 Ph.D.s were obtained under his guidance. He has been teaching at and been a member of Ph.D. jury's in several universities in Europe and the US. He is also a member of several scientific and professional organizations.

His work has won him several scientific awards (Leybold–Heraeus Prize (1986), Leslie Fox Prize (1989), Guillemin–Cauer Best Paper Award of the IEEE Transaction on Circuits and Systems (1990), Laureate of the Belgian Royal Academy of Sciences (1992), Bi-annual Siemens Award (1994), Best Paper award of Automatica (IFAC, 1996), IEEE Signal Processing Society Best Paper Award (1999)). Since 2004 he is a fellow of the IEEE (www.ieee.org). He is an associate editor of several scientific journals.

From 1991–1999 he was the chief advisor on Science and Technology of several ministers of the Belgian Federal Government (Demeester, Martens) and the Flanders Regional Governments (Demeester, Van den Brande).

He was and/or is in the board of 3 spin-off companies (www.ipcos.be, www.data4s.com, www.tml.be), of the Flemish Interuniversity Institute for Biotechnology (www.vib.be), the Study Center for Nuclear Energy (www.sck.be) and several other scientific and cultural organizations. He was a member of the Academic Council of the Katholieke Universiteit Leuven, and still is a member of its Research Policy Council. Since 2002 he also makes regular television appearances in the Science Show 'Hoe?Zo!' on national television in Belgium (www.tv1.be).