On Model Complexity Control in Identification of Hammerstein Systems

K. Pelckmans, I. Goethals, J.A.K. Suykens and B. De Moor
KULeuven - ESAT - SCD/SISTA
Kasteelpark Arenberg 10, B-3001 Leuven - Belgium
{Kristiaan.Pelckmans,Johan.suykens}@esat.kuleuven.ac.be

***Abstract*— Model complexity control and regularization play a crucial role in statistical learning theory and also for problems in system identification. This text discusses the potential of the issue of regularization in identification of Hammerstein systems in the context of primal-dual kernel machines and Least Squares Support Vector Machines (LS-SVMs) and proposes an extension of the Hammerstein class to finite order Volterra series and methods resulting in structure detection.**
keywords: **Identification, Hammerstein Systems, Model complexity and regularization, Kernel Methods**

I. Introduction

A. Identification of Hammerstein models

Hammerstein models consist of a sequence of a static nonlinear function and a linear dynamic submodel and provide a useful compromise between the sometimes conflicting requirements of flexibility of general nonlinear dynamic systems and the interpretability of linear dynamic systems. Most identification procedures for this model class can be classified as a statistical averaging method [10] or a method based on the over-parameterization technique [2]. Classical techniques often rely on basis-function expansions [15] mostly stressing the property of consistency [23].

Recently, applications of the primal-dual kernel machine framework were proposed towards the identification of Hammerstein models based on a linear ARX model [7] and a state-space model [8] in combination with subspace identification techniques. This paper extends those results by investigation of the role of regularization in the approach and elaboration of an alternative to the property of consistency in the form of the bias-variance trade-off for Hammerstein models. A further result is found in the definition and study of the class of generalized Hammerstein models not only extending the Hammerstein class but also containing the class of finite Volterra series. To make the approach practically workable, a relationship between model complexity control, model order and model class selection is proposed analogous to common techniques in statistical inference and machine learning, see e.g. [12].

B. Regularization in system identification

A classical qualification of an estimator $\hat{\theta} \in \Theta$ of a parameter θ contained in the set Θ based on a set of observations $\{(u_t, y_t)\}_{t=1}^T$ is the property whether it is consistent, i.e. $\lim_{T\to\infty} \hat{\theta} = \theta$. It was argued in [28] that this qualification is less relevant in predictive settings based on a finite number of observations where the only goal is to make predictions with minimal theoretical risk. Statistical learning theory studies whether this can be obtained based on an estimator minimizing the empirical risk. Especially in the context of nonlinear models, the task of recovering the (parameters of the) true model is much more involved than the subtask of prediction, see e.g. [28].

The focus of the analysis of nonlinear estimators shifts more towards the issue of appropriate model complexity control or regularization. Intuitively, one restricts here the solution-space explicitly or implicitly in order to obtain increased generalization. In order to obtain model complexity control in the context of parametric components, one may restrict the parameter-space to a ball with pre-specified radius [11], [26]. In the context of non-parametric models, one may impose smoothness constraints on the estimated output as e.g. in the case of smoothing splines [29]. A classical way to analyze the properties of estimators based on a finite number of observations and which include a form of regularization is found in the decomposition of the theoretical risk of all predictions, also referred to as the Total Mean Squared Error (TMSE). As classically, attention is restricted towards the risk of the estimator evaluated at the observed input observations [13]. Let $Y^* \in \mathbb{R}^T$ denote the true outputs at the sampling points and let $\hat{Y} \in \mathbb{R}^T$ denote the estimated (smoothed) outputs. Then the TMSE discretized to the observed observations can be decomposed as follows

$$E\left[\hat{Y} - Y^*\right]^2 = E\left[\hat{Y} - E(\hat{Y})\right]^2 + E\left[E(\hat{Y}) - Y^*\right]^2, \quad (1)$$

with the right hand-side respectively denoting as the bias and the variance of the estimator.

This paper is organized as follows. Section II studies the general issue of identification of Hammerstein models using primal-dual kernel machines and different measures of model complexity control. Section III then proceeds with broadening the class of Hammerstein models. Section IV establishes a connection between model order selection and complexity control. Section V presents an example of the methods.

II. Identification of Hammerstein models

Given a sequence of observations $\{(u_t, y_t)\}_{t=1}^T \subset \mathbb{R}^D \times \mathbb{R}$, the observations satisfy a Hammerstein model with an Auto-Regressive dynamic system with eXogenous inputs (ARX) when the following equalities hold. Let $p = \max(M, N) + 1$

and $T_p = T - p + 1$.

$$y_t = \sum_{m=1}^{M} a_m y_{t-m} + \sum_{n=1}^{N} b_n f\left(u_{t-n+1}\right) + e_t, \qquad (2)$$

for all $t = p, \ldots, T$ and assume $e = (e_p, \ldots, e_T)^T \in \mathbb{R}^{T-p}$ are i.i.d. error terms. Let in general a model complexity of a nonlinear function f be denoted as $\mathcal{C}(f)$ and of the linear dynamic model with parameters a, b as $\mathcal{C}(a, b)$. The general estimation of the Hammerstein model subject to various complexity constraints then amounts to solving

$$\begin{aligned} &(\hat{a}, \hat{b}, \hat{f}, \hat{e}) = \underset{a,b,f,e}{\arg\min}\, \mathcal{J}_{\varrho,\varsigma}(e) = \frac{1}{2}\|e\|_2^2 \\ &\text{s.t.} \quad \begin{cases} \mathcal{C}(f) \leq \varrho & (a) \\ \mathcal{C}(a,b) \leq \varsigma & (b) \\ y_t = \sum_{m=1}^{M} a_m y_{t-m} & \\ \quad + \sum_{n=1}^{N} b_n f\left(u_{t-n+1}\right) + e_t \quad \forall t, & (c) \end{cases} \end{aligned} \qquad (3)$$

where $a = (a_1, \ldots, a_M)^T \in \mathbb{R}^M$, $b = (b_1, \ldots, b_N)^T \in \mathbb{R}^N$ be vectors.

A. Controlling the model complexity of the nonlinear model

We hereby assume that the reader is somewhat familiar with the LS-SVM based approach to Hammerstein identification as reported in [7]. This framework enables the use of various regularization mechanisms. The following regularization method is prototypical. Let the nonlinear function $f : \mathbb{R}^D \to \mathbb{R}$ be approximated as $f(u) = w^T \varphi(u)$ where $\varphi : \mathbb{R}^D \to \mathbb{R}^{D_\varphi}$, the vector of parameters $w \in \mathbb{R}^{D_\varphi}$ and $D_\varphi \in \mathbb{N}_0$. Controlling the model complexity then amounts to finding a $\rho > 0$ such that

$$\mathcal{C}(f) \;\triangleq\; \frac{1}{2}\|w\|_2^2 \;\text{ s.t. }\; f(u) \;=\; w^T\varphi(u) \,+\, u, \quad (4)$$

which quantifies the distance of the fit f with the identical function $f(u) = u$. The classical over-parameterization technique as adopted in [7] then amounts to replacing the cross-products $b_n f$ by a new set of functions $\{f_n : \mathbb{R}^D \to \mathbb{R}\}_{n=1}^N$ (*the over-parameterization step*), leading to the modified model complexity definition

$$\mathcal{C}(f) \triangleq \frac{1}{2}\sum_{n=1}^{N} \|w_n\|_2^2 \;\text{ s.t. }\; f(u) = \sum_{n=1}^{N} w_n^T\varphi(u) + b_n u. \qquad (5)$$

The constrained optimization problem becomes

$$\begin{aligned} &(\hat{a}, \hat{b}, \hat{w}_n, \hat{e}) = \underset{a,b,w,e}{\arg\min}\, \mathcal{J}_\varrho(e) = \|e\|_2^2 \\ &\text{s.t.} \quad \begin{cases} \frac{1}{2}\sum_{n=1}^{N} w_n^T w_n \leq \varrho & (a) \\ y_t = \sum_{m=1}^{M} a_m y_{t-m} + \sum_{n=1}^{N} b_n u_t & \\ \quad + w_n^T \varphi\left(u_{t-n}\right) + e_t & \forall t = p, \ldots, T \quad (b) \\ b_n w^T \varphi(u_t) = w_n^T \varphi(u_t), & \forall t = p, \ldots, T. \quad (c) \end{cases} \end{aligned} \qquad (6)$$

where the constraints (6.c) are referred to as collinearity constraints which are known to be hard to be imposed. The typical procedure then is to solve a broader problem by omitting those and consequently project the estimated system onto the Hammerstein model class. We focus here on the former step. Let $Y_M \in \mathbb{R}^{T_p \times M}$ and $Y_M \in \mathbb{R}^{T_p \times N}$ be vectors defined as $Y_{M,tm} = y_{t-M+m-1}$ and $U_{N,tn} = u_{t-N+n-1}$ respectively. Let the matrix X be defined as $X = [U_M \;\; Y_N] \in \mathbb{R}^{(T_p)\times(M+N)}$, let $c = (a; b) \in \mathbb{R}^{N+M}$ and let $I_N \in \mathbb{R}^{N\times N}$ denote the identity matrix.

Lemma 1: The solution to the constrained problem (6) is given as the solution to the dual linear system

$$\begin{bmatrix} 0_{(M+N)\times(M+N)} & X^T \\ X & (\Omega_N + \gamma I_N) \end{bmatrix} \begin{bmatrix} c \\ \alpha \end{bmatrix} = \begin{bmatrix} 0_{M+N} \\ Y_p \end{bmatrix}, \quad (7)$$

where the kernel matrix Ω_N is defined as $\Omega_{N,st} = \sum_{n=1}^{N} K(u_{s-n}, u_t)$. The result can be used for prediction at timestep $t+1$ as follows

$$\hat{y}_{t+1} = \sum_{m=1}^{M} \hat{a}_m y_{t-m+1} + \sum_{n=1}^{N} \hat{b}_n u_{t-n+2} + \sum_{s=p}^{T} \hat{\alpha}_s \sum_{n=1}^{N} K(u_{s-n}, u_{t-n}), \quad (8)$$

where $\hat{c} = (\hat{a}; \hat{b})$ and $(\hat{c}, \hat{\alpha})$ are the solution to the dual problem (7).

Proof: The proof follows straightforwardly from results in convex optimization, see [4], [24]. Let $\mathcal{L}_\varrho$ be the Lagrangian of the constrained optimization problem (6)

$$\begin{aligned} \mathcal{L}_\varrho(a, b, w, e; \tilde{\alpha}, \gamma) = \frac{1}{2}e^T e + \gamma(\frac{1}{2}w^T w - \varrho) \\ + \sum_{t=p}^{T} \tilde{\alpha} \left(\sum_{m=1}^{M} a_m y_{t-m} + \sum_{n=1}^{N} (b_n u_{t-n+1} \right. \\ \left. + w_n^T \varphi\left(u_{t-n+1}\right)\right) + e_t - y_t \Big). \end{aligned} \quad (9)$$

where $\tilde{\alpha} = (\tilde{\alpha}_p, \ldots, \tilde{\alpha}_T)^T \in \mathbb{R}^{T_p+1}$ and $\gamma \in \mathbb{R}^+$ are the Lagrange multipliers. Taking the first order conditions for optimality $\frac{\partial \mathcal{L}_\varrho}{\partial w_n} = 0$, $\frac{\partial \mathcal{L}_\varrho}{\partial c} = 0$, $\frac{\partial \mathcal{L}_\varrho}{\partial e_t} = 0$ and $\frac{\partial \mathcal{L}_\varrho}{\partial \tilde{\alpha}_t} = 0$; and elimination of the primal variables w_n and e results in the dual system (7) where $\gamma\alpha = \tilde{\alpha}$. The multiplier γ and the hyper-parameter ϱ are related via a monotone secular equation as explicified in [21]. Using the conditions $\frac{\partial \mathcal{L}_\varrho}{\partial w_n} = 0$ it follows that $w_n = \sum_{t=p}^{T} \alpha_{t-n}\varphi(u_{t-n})$ for all $n = 1, \ldots, N$, one can evaluate the estimate as described in (8). ■

From this derivation it also follows that the estimated submodels $\hat{f}_n$ can be evaluated in a new point $u_* \in \mathbb{R}$ as follows

$$\hat{f}_n(u_*) = \hat{b}_n u_* + \sum_{t=p}^{N} \hat{\alpha} K(u_{t-n}, u_*), \qquad (10)$$

where $\hat{c} = (\hat{a}; \hat{b})$ solves (7).

B. Regularization of the linear sub-systems

In addition to the model complexity control of the non-linear function as described in Section II.A, complexity measures can be formulated to quantify the simplicity of the linear model

$$\mathcal{C}(a,b) = \sum_{m=1}^{M} a_m^2 + \sum_{n=1}^{N} b_m^2 \leq \varsigma. \tag{11}$$

This kind of regularization term was also employed in [18], [27] and [9] to impose stability and positive realness respectively on the identified linear system. Slightly related to those approaches, [5] studied related convex methods to impose constraints on the estimated transfer function of a linear system. In general, control of model complexity can be adopted to decrease the variance in the estimates of the parameters when the number of parameters is large in relation to the number of observations [12], [13].

In the case of the regularization mechanism as described in (11), the following convex optimization problem is obtained

$$\begin{aligned} &(\hat{a},\hat{b},\hat{w}_n,\hat{e}) = \arg\min_{a,b,w,e} \mathcal{J}_{\varrho,\varsigma}(e) = \|e\|_2^2 \\ &\text{s.t.} \quad \begin{cases} \frac{1}{2}\sum_{n=1}^{N} w_n^T w_n \leq \varrho & (a) \\ a^T a + b^T b \leq \varsigma & (b) \\ y_t = e_t + \sum_{m=1}^{M} a_m y_{t-m} \\ \quad + \sum_{n=1}^{N} b_n u_t + w_n^T \varphi(u_{t-n}) \quad \forall t. & (c) \end{cases} \end{aligned} \tag{12}$$

The primal-dual derivation is summarized as follows

Lemma 2: The dual problem to (12) becomes

$$\left(\Omega_N + \frac{\gamma}{\lambda} XX^T + \gamma I_N\right) \alpha = Y. \tag{13}$$

The parameters of the linear model can be recovered as

$$\frac{\lambda}{\gamma}\hat{c} = X^T\hat{\alpha}, \tag{14}$$

where $\hat{c} = (\hat{a};\hat{b}) \in \mathbb{R}^{N+M}$ and $\hat{\alpha} \in \mathbb{R}^{T_p}$ solve (26).

Proof: The proof follows along the same lines as in Lemma 2.1 where $\gamma > 0$ and $\lambda > 0$ are the Lagrange multipliers associated with the model complexity constraints quantified by ϱ and ς respectively. The occurrence of the extra model complexity constraints allows for eliminating the variables a and b in the dual formulation. ■

C. Smoother matrix of the Hammerstein model

It turns out that the Hammerstein identification process can be written as a linear operator as follows. The smoother matrix $S^H_{\gamma,\lambda} \in \mathbb{R}^{(T_p)\times(T_p)}$ becomes

$$\begin{aligned} &\hat{Y} = S^H_{\gamma,\lambda} Y \quad \text{s.t.} \\ &S^H_{\gamma,\lambda} = \left(\Omega_N + \frac{\gamma}{\lambda} XX^T\right)\left(\Omega + \frac{\gamma}{\lambda} XX^T + \gamma I_N\right)^{-1}. \end{aligned} \tag{15}$$

The trace of the smoother matrix was proposed in [17] as a way to quantify the degrees of freedom of a linear operator. Let USU^T be the unique SVD of the matrix $(\Omega_N + \frac{\gamma}{\lambda}XX^T)$ with $U \in \mathbb{R}^{T_p\times T_p}$ and $D = \mathrm{diag}(\sigma_1,\ldots,\sigma_{T_p})$ with $\sigma_i \geq 0$ the singular values.

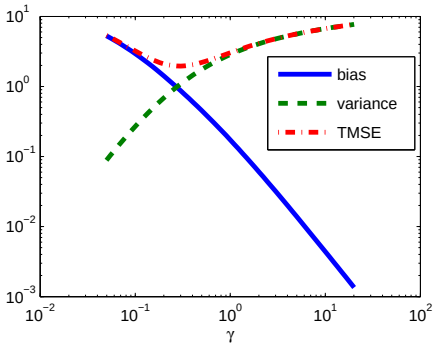

Fig. 1. *In the case $\gamma = \lambda$, the bias (solid line) and the variance (dashed line) can be plotted in terms of the sole hyper-parameter as illustrated here on the example as described in Section V. The total MSE (dashed-dotted line) typically has a global minimum at a $\gamma > 0$ implying that a little amount of regularization often increases the generalization performance of the estimates.*

Definition 1: The degrees of freedom or the effective dimension of the model are defined as

$$D_{\text{eff}}(S^H_\gamma) = \sum_{i=1}^{T_p} \frac{\sigma_i}{\sigma_i + \gamma}, \tag{16}$$

where σ_i are the singular values of the positive semi-definite matrix $(\Omega_N + \frac{\gamma}{\lambda}XX^T)$.

This formulation allows us to quantify bias and variance of the Hammerstein model as summarized as follows

Lemma 3: The total mean squared error of the estimated model can be decomposed as $\mathrm{TMSE}(\hat{Y},Y^*) = \mathrm{Bias}(\hat{Y},Y^*) + \mathrm{var}(\hat{Y})$ where the bias and variance become respectively

$$\begin{cases} \mathrm{var}(\hat{Y})^2 = \sigma_e^2 \sum_{i=1}^{T_p+1} \frac{\sigma_i^2}{(\sigma_i+\gamma^{-1})^2} \\ \mathrm{bias}(Y^*,\hat{Y})^2 = \gamma^{-2} \sum_{i=1}^{T_p+1} \frac{p_i^2}{(\sigma_i+\gamma^{-1})^2}, \end{cases} \tag{17}$$

with $p_i \in \mathbb{R}^{T_p}$ defined as $U_i^T Y^*$.

This derivation goes completely along the lines of the derivation of the bias-variance trade-off derived in the case of smoothing splines [29].

Note that only the effect of the regularization parameters γ and λ does not follow straightforwardly as their role in the matrix $(\Omega + \frac{\gamma}{\lambda}XX^T)$ is not directly reflected in the SVD matrix. In the case both hyper-parameters are taken equal, the decomposition can be used to describe the evolution of the total mean squared error as a function of this parameter as denoted in Figure 1.

III. Generalization of the Hammerstein class

This section elaborates on the measure of model complexity in terms of the model orders M and N.

A. Nonlinear over-parameterization

The classical over-parameterization approach proceeds by projecting the estimated functions f_p on the class of Hammerstein models using a rank one approximation. Let Y_n be vectors for all $n = 1,\ldots,N$ defined as $Y_n = (f_n(u_{p-t+1}),\ldots,f_n(u_{T-n+1}))^T \in \mathbb{R}^{T_p}$ and Y_f as $Y_f =$

$\sum_{n=1}^{N} Y_n \in \mathbb{R}^{T_p}$, then the following equality follows from the definition of the Hammerstein class (2):

$$\begin{bmatrix} Y_1 & Y_2 & \dots & Y_N \end{bmatrix} = Yb^T. \tag{18}$$

When estimates for Y_n are obtained via the described method, the best rank-one approximation can be found using the Singular Value Decomposition (SVD).

B. Linear over-parameterization

An alternative route towards the constraining of the estimate onto the class of the Hammerstein models can be taken. A Hammerstein model (2) can be rewritten as follows

$$H_{M,N}(u_t) = \frac{B(q)}{A(q)} f(u_t) = \frac{\sum_{n=0}^{N-1} b_n q^n}{\sum_{m=1} a_m q^m} f(u_t), \tag{19}$$

where q denotes the generalized back-shift operator $qf(u_t) = f(u_{t-1})$ and $qy_t = y_{t-1}$. It is a classical result that any fractional polynomial $\frac{A(q)}{B(q)}$ can be approximated well as $\frac{1}{\tilde{A}(q)}$ such that one can write [16]

$$H_{M,N}(u_t) \approx \tilde{H}_{P,1}(u_t) = \frac{1}{\tilde{A}(q)} f(u_t) = \frac{1}{\sum_{m=1} \tilde{a}_m q^m} f(u_t) \tag{20}$$

where P is in general much larger than N and $\{\tilde{a}_p\}_{p=1}^{P}$ is an appropriate set of coefficients. Identification of models of the form $H_{P,1}$ do apparently avoid the need for the over-parameterization of the cross-products as in the previous subsection and make the projection step based on the SVD obsolete. The over-parameterization now occurs in the linear system as described in equation (20). However, the regularization mechanism of the linear subsystem as described previously enables the successful identification of the set of linear parameters $\{\tilde{a}_p\}_{p=1}^{P}$ even if P approaches T.

A disadvantage of the linear over-parameterization technique is that the noise model is not preserved as can be seen easily

$$\begin{aligned} y_t &= \frac{B(q)}{A(q)} f(u_t) + e_t \approx \\ y_t &= \frac{\beta_0}{\tilde{A}(q)} f(u_t) + e_t \Leftrightarrow \tilde{A}(q) y_t = \beta_0 f(u_t) + \tilde{A}(q) e_t, \end{aligned} \tag{21}$$

from which it follows that this model formulation cannot handle output noise straightforwardly.

C. Generalization of the Hammerstein class

The previous elaboration motivates the following definition of the class of generalized Hammerstein models. Let for all $j = 1, \dots, N'$ the denominator $I(j)$ denote the jth set of indices $I(j) = (n_1, n_2, \dots, n_{r(j)})$ where $n_1 < n_2 < \dots < n_{r(j)}$ denote the indices of the jth function $f_j : \mathbb{R}^{r(j)} \to \mathbb{R}$ such that $f_j(u_t) = f\left(u_{t-I(j)_1}, \dots, u_{t-I(j)_r}\right)$. Let $R \in \mathbb{N}$ define the maximum length of the sets I_j. A generalization to the Hammerstein class can be defined as follows.

Definition 2: Let $M, N, R \in \mathbb{N}_0$ denote the orders of the model, $\{a_m\}_{m=1}^{M} \subset \mathbb{R}$ be a set of linear parameters and let $\{f_{I(j)}\}_{j=1}^{N'}$ be a set of given functions. The class of generalized Hammerstein models may be defined as follows

$$y_t = \sum_{m=1}^{M} a_m y_{t-m} + \sum_{j=1}^{N'} f_{I(j)}\left(u_{t-I(j)_1}, \dots, u_{t-I(j)_r}\right) \tag{22}$$

which reduces to the class of Hammerstein models if $R = 1$ and the functions $\{f_n\}_{n=1}^{N}$ are collinear (or $N = 1$).

This class of models does not only contain the class of Hammerstein models but also generalizes the class of finite Volterra series which possess a property of general approximator, see e.g. [3]. Therefor, $M = 0$ and the class of nonlinearities must take the form

$$\begin{aligned} &f_{I(j)}\left(u_{t-I(j)_1}, \dots, u_{t-I(j)_r}\right) \\ &\quad = k(I(j)_1, \dots, I(j)_r) \;\; u_{t-I(j)_1} \dots u_{t-I(j)_r}, \end{aligned} \tag{23}$$

where $k : \mathbb{R}^{r(j)} \to \mathbb{R}$ is an appropriate function. The primal-dual derivation of an appropriate kernel machine is summarized as follows.

Lemma 4: Consider the models

$$\begin{aligned} &f_{I(j)}\left(u_{t-I(j)_1}, \dots, u_{t-I(j)_r}\right) \\ &\qquad = w_{I(j)}^T \varphi_j\left(u_{t-I(j)_1}, \dots, u_{t-I(j)_r}\right) \end{aligned} \tag{24}$$

for all $j = 1, \dots, N'$. Modification of the estimation problem (12) with the model complexity of the nonlinear model replaced by the following definition

$$\frac{1}{2} \sum_{j=1}^{N'} w_{I(j)}^T w_{I(j)} \leq \varrho, \tag{25}$$

results in the dual system

$$\left(\Omega_I + \frac{\gamma}{\lambda} XX^T + \gamma I_N\right) \alpha = Y, \tag{26}$$

with γ the Lagrange multiplier proportional to ϱ and where $\Omega_I \in \mathbb{R}^{T_p \times T_p}$ is defined as

$$\Omega_{I,st} = \sum_{j=1}^{N'} K\left(u_{s-I(j)_1}, ..u_{s-I(j)_r}, u_{t-I(j)_1}, ..u_{t-I(j)_r}\right). \tag{27}$$

and the estimate can be evaluated similarly as in (8).

Note that the difference with Lemma 2 only reflects in the design of the kernel. The proof again follows along the same lines as in [19] and [7]. The main disadvantage (at least in a practice) is the presence of at least 6 hyper-parameters N, M, R, λ, γ and the kernel parameter which need to be set a priori or using a suitable model selection procedure. The following section proposes a way to circumvent this problem.

IV. Structure detection and model selection

One can relate the previous discussion with the task of model order selection as discussed next. Here we differentiate between model order selection of the linear AR part and of the nonlinear eXogenous part respectively. Structure detection (sparseness) of the parameters a_m, b_n and the functions f_n may be obtained by employing a regularization

scheme based on the L_1 norm. This approach has gained recent interest in the research on machine learning and statistical inference as in the LASSO [25] and basis pursuit [6]. The linear model can be equipped with a L_1 norm such that the model complexity of the linear system becomes

$$\sum_{m=1}^{M} |a_m| + \sum_{n=1}^{N} |b_n| \leq \varsigma. \tag{28}$$

The property that the use of L_1 norms results in sparseness on the parameters depending on the hyper-parameter ς was already exploited in the LASSO estimator [25], projection pursuit [6] and SVMs [28]. It was extended to nonlinear kernel machines in [19], [22] by using a measure of maximal variation defined as follows.

Definition 3: (Maximal Variation) The theoretical maximal variation of a function $f : \mathbb{R}^D \to \mathbb{R}$ is defined as

$$\mathcal{M}(f) = \sup_{x \in \mathbb{R}^D} |f(x)| . \tag{29}$$

Its empirical counterpart is defined as

$$\hat{\mathcal{M}}(f) = \max_{x_i \in \mathcal{D}} |f(x_i)| , \tag{30}$$

where $\mathcal{D}$ denotes the set of training data. and may be used as an approximation to (29).

It was argued in e.g. [1] that the L_1 norm is not optimal in the sense of obtaining sparseness and other measures can be employed satisfying the so-called oracle constraints. A disadvantage of those norms is that the property of convexity is lost and iterative approaches need to be adopted, see e.g. [19]. The corresponding optimization problem including the linear model complexity measure (28) and an L_1 based control of the sum of the (empirical) maximal variations become

$$\begin{aligned} &(\hat{a}, \hat{b}, \hat{w}, \hat{e}) = \underset{a,b,w,e}{\arg\min}\, \mathcal{J}_{\varrho,\varsigma}(e) = \|e\|_2^2 \\ &\text{s.t.} \quad \begin{cases} \frac{1}{2}\sum_{n=1}^{N} \hat{\mathcal{M}} \leq \varrho & (a) \\ \sum_{m=1}^{M} |a_m| + \sum_{n=1}^{N} |b_n| \leq \varsigma & (b) \\ y_t = e_t + \sum_{m=1}^{M} a_m y_{t-m} \\ \quad + \sum_{n=1}^{N} b_n u_t + w_n^T \varphi(u_{t-n}) \quad \forall t. & (c) \end{cases} \end{aligned} \tag{31}$$

Primal-dual interpretations of this kind of models were described in [22]. Those are omitted from the current paper due to space limitations. Practical re-formulations based on the additive regularization trade-off framework [20] result in the following constrained optimization problem

$$\begin{aligned} &\min_{\alpha,q,u,c} \mathcal{J}_{\gamma,\lambda}(e,q,u) = \|e\|_2^2 + \gamma\|t\|_1 + \lambda\|u\|_1 \quad \text{s.t.} \\ &\begin{cases} \begin{bmatrix} 0_{M+N\times N+M} & X^T \\ X & \Omega_N \end{bmatrix} \begin{bmatrix} c \\ \alpha \end{bmatrix} + \begin{bmatrix} 0_{N+M} \\ e \end{bmatrix} = \begin{bmatrix} 0_{M+N} \\ Y_{T_p} \end{bmatrix} \\ -q\ 1_{T_p} \leq \Omega_n \alpha \leq q\ 1_{T_p} \quad \forall n = 1, \ldots, N \\ -u\ 1_{M+N} \leq c \leq u\ 1_{N+M}, \end{cases} \end{aligned} \tag{32}$$

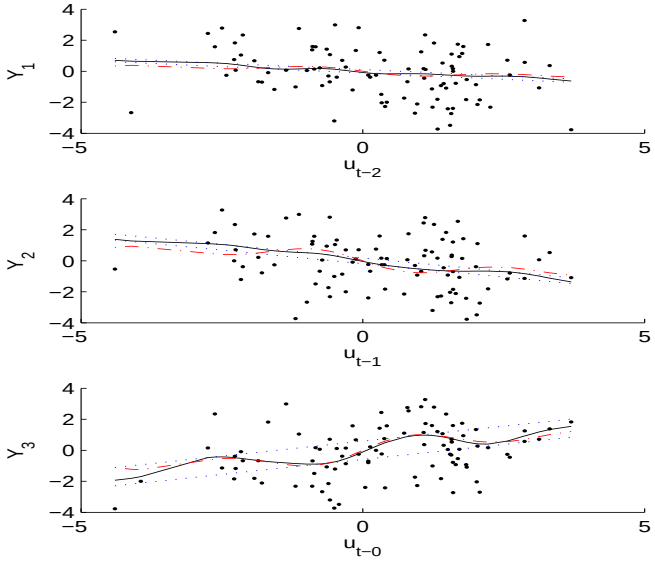

Fig. 2. *Examples of the three nonlinear functions $f_2(u_{t-2}) + f_1(u_{t-1}) + f_0(u_t)$ (dashed line) and its estimation (solid line) given a finite set of observations (dots). The dotted lines indicate the maximal variation of the nonlinearity, suggesting a second order model.*

where $q = (q_1, \ldots, q_{T_p})^T \in \mathbb{R}^{T_p}$ and $u = (u_1, \ldots, u_{N+M})^T \in \mathbb{R}^{N+M}$ are vectors of slack-variables. This convex optimization problem can be solved efficiently as a Quadratic Programming problem. Extension with the output saturation as described in the previous subsection is straightforward.

V. Illustrative Example

An artificial example illustrates the practical relevance of the discussed methods. A dataset was constructed as follows. Let $\{(u_t, y_t)\}_{t=1}^{T}$ satisfy the equality

$$\begin{aligned} y_t &= f(u_t) - 0.5 f(u_{t-1}) \\ &\quad + 0.75 y_{t-1} - 0.25 y_{t-2} + 0.2 y_{t-3} + 0.1 y_{t-3} + e_t \\ &\text{s.t.} \quad f(u_t) = (\tanh(u_t) + 0.4\sin(2u_t)) , \end{aligned} \tag{33}$$

where $\{u_t\} \sim N(0, 2)$, the innovations $\{e_t\} \sim N(0, 1.5)$ and $T = 200$. Presented model complexity control mechanisms are employed in the black-box identification of an Hammerstein model from observed data. All methods take a fixed order of the estimate of (N, M) equal to $(4, 4)$ in order to avoid the issue of model order selection in the comparison of performance and degrees of freedom. The individual hyper-parameters were tuned on a separate validation set of size $T^v = 50$.

(1) Classical approaches amount often to an expansion in basis-functions sequenced by an over-parameterization [2]. The number of effective parameters becomes MN_e where $N_e \in \mathbb{N}$ denotes the number of basis-functions. A major drawback is the amount of variance which is rapidly increasing when the order M becomes larger.

(2) The approach proposed in [7] possesses a form of complexity control on the estimated nonlinearities. The effective degrees of freedom then becomes $D_{\text{eff}}(w) + M$ where

$D_{\text{eff}}(w)$ denotes the effective dimensions of the nonlinear part [24].

(3) Including a small amount of complexity control of the linear subsystem as introduced in Subsection II.B often results in improvements of the estimate. A drawback is the occurrence of an additional hyper-parameter λ.

(4) Linear over-parameterization as presented in Subsection III.B avoids the need for nonlinear over-parameterization as argued. However, the properties of the noise model are affected by this approach resulting in a lower performance on the example.

(5) The mechanism as described in Subsection IV is illustrated in order to obtain automatic model order selection. This reduces the number of hyper-parameters to three (γ, λ and the kernel parameters). The automatic structure detection property is traded against a small decrease in performance. As a measure of the effective degrees of freedom, the dimension of the eigenspace of the Hessian is taken as in [25]. Numerical results or given in Table 1.

	Prediction	**Order**	**Eff. order**	**# Hyp.par.**
Basis func.	2.2503	$(4, 4)$	52	4
[7]	1.8295	$(4, 4)$	35.9584	4
Compl. Contr.	1.7825	$(4, 4)$	28.6609	5
Lin. overp.	2.0803	$(1, 20)$	32.6437	4
Struct. Det.	1.8006	$(10, 10)$	12.4534	3

TABLE 1.

VI. CONCLUSIONS

A study of the impact and relevance of model complexity control in the identification of Hammerstein models and its generalization is presented. Of direct practical relevance is the established relationship between model order selection and model complexity control resulting in a decreased number of hyper-parameters which need to be tuned. Future directions include the investigation of the use of the presented method towards control of Hammerstein systems, a research track which was initiated e.g. in [14]. A further important research track involves the study of the relationship between persistency of excitation and model complexity.

ACKNOWLEDGMENTS - This research work was carried out at the ESAT laboratory of the KUL. Research Council KU Leuven: Concerted Research Action GOA-Mefisto 666, GOA-Ambiorics IDO, several PhD/postdoc & fellow grants; Flemish Government: Fund for Scientific Research Flanders (several PhD/postdoc grants, projects G.0407.02, G.0256.97, G.0115.01, G.0240.99, G.0197.02, G.0499.04, G.0211.05, G.0080.01, research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT (Soft4s, STWW-Genprom, GBOU-McKnow, Eureka-Impact, Eureka-FLiTE, several PhD grants); Belgian Federal Government: DWTC IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006) (2002-2006), Program Sustainable Development PODO-II (CP/40); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS. JS is an associate professor and BDM is a full professor at K.U.Leuven Belgium, respectively.

REFERENCES

[1] A. Antoniadis and J. Fan. Regularized wavelet approximations (with discussion). *Jour. of the Am. Stat. Ass.*, 96:939–967, 2001.

[2] E.W. Bai. An optimal two-stage identification algorithm for Hammerstein-Wiener nonlinear systems. *Automatica*, 34(3):333–338, 1998.

[3] S. Boyd, L.O. Chua, and C. A. Desoer. Analytical foundations of Volterra series. *IMA Journal of Mathematical Control and Information*, 1:243–282, 1984.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] T.N. Davidson, Z.-Q. Luo, and J.F. Sturm. Linear matrix inequality formulation of spectral mask constraints with applications to FIR filter design. *IEEE Transactions on Signal Processing*, 50(11):2702–2715, 2000.

[6] J. H. Friedmann and W. Stuetzle. Projection pursuit regression. *Jour. of the Am. Stat. Assoc.*, 76:817–823, 1981.

[7] I. Goethals, K. Pelckmans, J.A.K. Suykens, and B. De Moor. Identification of MIMO Hammerstein models using least squares support vector machines. *Automatica*, In Press, 2005. available online at `www.esat.kuleuven.ac.be/scd`.

[8] I. Goethals, K. Pelckmans, J.A.K. Suykens, and B. De Moor. Subspace identification of Hammerstein systems using least squares support vector machines. Technical report, ESAT-SISTA, K.U.Leuven, Leuven, Belgium, 2004.

[9] I. Goethals, T. Van Gestel, J. Suykens, P. Van Dooren, and B. De Moor. Identification of positive real models in subspace identification by using regularization. *IEEE Transactions on Automatic Control*, 48(10):1843–1847, 2003.

[10] W. Greblicki and M. Pawlak. Identification of discrete Hammerstein systems using kernel regression estimates. *IEEE Trans. Automatic Control*, 31:74–77, 1986.

[11] P.C. Hansen. *Rank-deficient and Discrete Ill-posed Problems*. SIAM, 1998.

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, Heidelberg, 2001.

[13] A.E. Hoerl and R.W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–82, 1970.

[14] Z.H. Lang. Controller design oriented model identification method for Hammerstein systems. *Automatica*, 29:767–771, 1993.

[15] Z.H. Lang. A nonparametric polynomial identification algorithm for the Hammerstein system. *IEEE Trans. Automatic Control*, 42:1435–1441, 1997.

[16] L. Ljung. *System Identification, Theory for the User*. Prentice Hall, 1987.

[17] C.L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[18] J. Marí, P. Stoica, and T. McKelvey. Vector ARMA estimation: A reliable subspace approach. *IEEE Transactions on Signal Processing*, 48:2092–2104, 2000.

[19] K. Pelckmans, I. Goethals, J. De Brabanter, J.A.K. Suykens, and B. De Moor. Componentwise least squares support vector machines. *Support Vector Machines: Theory and Applications*, L. Wang (Ed.), Springer, 2004, *In press*.

[20] K. Pelckmans, J.A.K. Suykens, and B. De Moor. Additive regularization: Fusion of training and validation levels in kernel methods. *Internal Report 03-184, ESAT-SISTA, K.U.Leuven, Belgium, submitted*, 2003.

[21] K. Pelckmans, J.A.K. Suykens, and B. De Moor. Morozov, Ivanov and Tikhonov regularization based LS-SVMs. In *Proc. of the11th International Conference on Neural Information Processing (ICONIP 2004)*, Kolkata, India, November 2004.

[22] K. Pelckmans, J.A.K. Suykens, and B. De Moor. Building sparse representations and structure determination on LS-SVM substrates. *Neurocomputing, in press*, 2005.

[23] P. Stoica. On the convergence of an iterative algorithm used for Hammerstein system identification. *IEEE Trans. Automatic Control*, 26:967–969, 1981.

[24] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

[25] R.J. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society*, 58:267–288, 1996.

[26] A. N. Tikhonov and V. Y. Arsenin. *Solution of Ill-Posed Problems*. Winston, Washington DC, 1977.

[27] T. Van Gestel, J.A.K. Suykens, P. Van Dooren, and B. De Moor. Identification of stable models in subspace identification by using regularization. *IEEE Transactions on Automatic Control*, 46(9):1416–1420, 2001.

[28] V. Vapnik. *Statistical Learning Theory*. Wiley and Sons, 1998.

[29] G. Wahba. *Spline models for observational data*. SIAM, 1990.