Katholieke Universiteit Leuven

Departement Elektrotechniek

ESAT-SISTA/TR 04-57

Partially Linear Models and Least Squares Support Vector Machines[1]

Marcelo Espinoza [2], Johan A.K. Suykens, Bart De Moor

February 2004
Accepted for Presentation at the 43rd IEEE Conference of Decision and Control (CDC) 2004.

[1]This report is available by anonymous ftp from *ftp.esat.kuleuven.ac.be* in the directory *pub/sista/mespinoz/reports/04-57.pdf.gz*

[2]K.U.Leuven, Dept. of Electrical Engineering (ESAT), Research group SCD-SISTA, Kasteelpark Arenberg 10, 3001 Leuven, Belgium, Tel. 32/16/32 17 09, Fax 32/16/32 19 70, WWW: *http://www.esat.kuleuven.ac.be/scd*. E-mail: *marcelo.espinoza@esat.kuleuven.ac.be*.

Abstract

Within the context of nonlinear system identification, the LS-SVM formulation is extended to define a Partially Linear LS-SVM in order to identify a model containing a linear part and a nonlinear component. For a given kernel, a unique solution exists when the parametric part has full column rank, although identifiability problems can arise for certain structures. The solution has close links with traditional semiparametric techniques from the statistical literature. The properties of the model are illustrated by Monte Carlo simulations over different structures, and iterative forecasting examples for Hammerstein and other systems show a good global performance and an accurate identification of the linear part.

Partially Linear Models and Least Squares Support Vector Machines

Marcelo Espinoza, Johan A.K. Suykens, Bart De Moor
K.U. Leuven, ESAT-SCD-SISTA, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium.
{marcelo.espinoza,johan.suykens}@esat.kuleuven.ac.be

***Abstract*— Within the context of nonlinear system identification, the LS-SVM formulation is extended to define a Partially Linear LS-SVM in order to identify a model containing a linear part and a nonlinear component. For a given kernel, a unique solution exists when the parametric part has full column rank, although identifiability problems can arise for certain structures. The solution has close links with traditional semiparametric techniques from the statistical literature. The properties of the model are illustrated by Monte Carlo simulations over different structures, and iterative forecasting examples for Hammerstein and other systems show a good global performance and an accurate identification of the linear part.**

I. Introduction

Within the framework of nonlinear system identification, a common approach is to estimate a nonlinear black-box model in order to produce accurate forecasts starting from a set of observations. In this way, it is possible to define a regression vector from a set of inputs [15] and a nonlinear mapping in order to finally estimate a model that is suitable for prediction or control. Kernel based estimation techniques, such as Support Vector Machines (SVMs) and Least Squares Support Vector Machines (LS-SVMs) have shown to be powerful nonlinear black-box regression methods [12], [20]. Both techniques build a linear model in the so-called feature space where the inputs have been transformed by means of a (possibly infinite dimensional) nonlinear mapping φ. This is converted to the dual space by means of the Mercer's theorem and the use of a positive definite kernel, without computing explicitly the mapping φ. The SVM model solves a quadratic programming problem in dual space, obtaining a sparse solution [1]. The LS-SVM formulation, on the other hand, solves a linear system in dual space under a least-squares cost function [19], where the sparseness property can be obtained by e.g. sequentially pruning the support value spectrum [17] or via a fixed-size subset selection approach [18]. The LS-SVM training procedure involves the selection of a kernel parameter and the regularization parameter of the cost function, which can be done e.g. by cross-validation, Bayesian techniques [10] or others.

However, the fully nonlinear black-box model may be too general for some situations when there are reasons to include a linear part in the model. The practical rule of "do not estimate what you already know" would require to define an ad-hoc model structure if we know that the system contains a linear part, as in a Hammerstein model. Moreover, stronger motivations to explicitly include a linear part in the model can arise from practical considerations, as to check whether the system can be described within a certain accuracy by imposing a linear structure (e.g. for control purposes). Furthermore, the goal of a specific problem can be to identify a linear part that is based on first-principles, while including a nonlinear black-box part in order to keep the overall model accuracy within satisfactory levels. Or as noted in [15], the model structure is nonlinear in the inputs and contains non-white disturbances that can be described by a linear transformation of a white noise sequence; in this case, the predictor will contain a linear term containing past values of the output. In all these cases it would be desirable to have a technique that can lead to identification of a model containing both a linear and a nonlinear structure. Within the statistical literature, so-called "partially linear models" [16], [13], [7] have been developed since the mid-80s. These are models containing a linear parametric part and also a nonparametric component that is estimated using (local) smoothing techniques, usually restricted to low dimensional input vectors [6]. The concept can be extended to the LS-SVM framework by defining a model for which the LS-SVM can capture a nonlinear component while a parametric linear part can be simultaneously identified, allowing the inclusion of large dimensional input vectors for the nonlinear part. Therefore, in this paper we extend the LS-SVM formulation in order to include a linear parametric part in the primal space. The derivation and the results imply close links with statistical techniques. This paper is organized as follows. Section II describes the LS-SVM setting for nonlinear regression. In Section III the Partially Linear LS-SVM is developed and some issues about the solution and identifiability are addressed. Practical applications are reported in Section IV.

II. Least Squares Support Vector Machines for Nonlinear Regression

The standard framework for LS-SVM estimation is based on a primal-dual formulation. Given the dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$, with $\boldsymbol{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$, the goal is to estimate a model of the form

$$y_i = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i, i = 1, \ldots, N, \qquad (1)$$

where $\varphi(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^{n_h}$ is the mapping to a high dimensional (and possibly infinite dimensional) feature space, and the error terms e_i are assumed to be i.i.d. with zero mean and constant (and finite) variance. The following optimization problem is formulated:

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{e}} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\boldsymbol{e}^T\boldsymbol{e} \\ \text{s.t.} \quad & \boldsymbol{y} = \boldsymbol{\Phi}(\boldsymbol{x})\boldsymbol{w} + \mathbf{1}b + \boldsymbol{e}, \end{aligned} \tag{2}$$

where $\boldsymbol{y} = [y_1, y_2, \ldots, y_N]^T$ is the vector of dependent variables, $\boldsymbol{e}$ is the vector of error terms, $\mathbf{1}$ is a vector of ones and $\boldsymbol{\Phi}$ is the $N \times n_h$ matrix of stacked nonlinear regressors $\boldsymbol{\Phi} = [\varphi(\boldsymbol{x}_1)^T; \varphi(\boldsymbol{x}_2)^T; \ldots; \varphi(\boldsymbol{x}_N)^T]^T$ where each row i contains the feature vector $\varphi(\boldsymbol{x}_i)$. With the application of the Mercer's theorem to build the kernel matrix $\boldsymbol{\Omega}$ as $\boldsymbol{\Omega}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi(\boldsymbol{x}_i)^T\varphi(\boldsymbol{x}_j), i, j = 1, \ldots, N$ it is not required to compute explicitly the nonlinear mapping $\varphi(\cdot)$ as this is done implicitly through the use of positive definite kernel functions K. For $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ there are usually the following choices: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T\boldsymbol{x}_j$ (linear kernel); $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T\boldsymbol{x}_j + c)^d$ (polynomial of degree d, with c a tuning parameter); $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2/\sigma^2)$ (RBF kernel), where σ is a tuning parameter.

From the Lagrangian $\mathcal{L}(\boldsymbol{w}, b, e; \alpha) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\boldsymbol{e}^T\boldsymbol{e} - \boldsymbol{\alpha}^T(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w} - \mathbf{1}b - \boldsymbol{e})$, where $\boldsymbol{\alpha} \in \mathbb{R}^N$ is the vector of Lagrange multipliers, the conditions for optimality are given by:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = 0 & \rightarrow \quad \boldsymbol{w} = \boldsymbol{\Phi}^T\boldsymbol{\alpha} \\ \frac{\partial \mathcal{L}}{\partial b} = 0 & \rightarrow \quad \boldsymbol{\alpha}^T\mathbf{1} = 0 \\ \frac{\partial \mathcal{L}}{\partial \boldsymbol{e}} = 0 & \rightarrow \quad \boldsymbol{\alpha} = \gamma\boldsymbol{e} \\ \frac{\partial \mathcal{L}}{\partial \boldsymbol{\alpha}} = 0 & \rightarrow \quad \boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{w} + \mathbf{1}b + \boldsymbol{e}. \end{cases} \tag{3}$$

By elimination of $\boldsymbol{w}$ and $\boldsymbol{e}$, and using Mercer's theorem $\boldsymbol{\Omega} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T$, the following linear system is obtained:

$$\left[\begin{array}{c|c} \boldsymbol{\Omega} + \gamma^{-1}\mathbf{I} & \mathbf{1} \\ \hline \mathbf{1}^T & 0 \end{array}\right] \left[\begin{array}{c} \boldsymbol{\alpha} \\ \hline b \end{array}\right] = \left[\begin{array}{c} \boldsymbol{y} \\ \hline 0 \end{array}\right]. \tag{4}$$

The resulting LS-SVM model in dual space becomes

$$\hat{y}(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b. \tag{5}$$

Usually the training of the LS-SVM model involves an optimal selection of the kernel parameters and the regularization parameter, which can be done using e.g. cross-validation techniques or Bayesian inference [10]. For similarities and differences of this methodology with respect to Gaussian processes, regularization networks, kriging, etc., the reader is referred to [18].

III. Partially Linear Models with LS-SVM

In this paper our intention is to extend the LS-SVM formulation in order to allow a parametric part in the model structure, in such a way that the nonlinear regression and the parametric part are identified at the same time. Here we present the derivation plus some theoretical remarks.

A. Derivation of a Partially Linear LS-SVM

Let us consider the following model structure

$$y_i = \boldsymbol{\beta}^T\boldsymbol{z}_i + f(\boldsymbol{x}_i) + e_i, \qquad i = 1, \ldots, N, \tag{6}$$

where $\boldsymbol{z}_i \in \mathbb{R}^p, \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{x} \in \mathbb{R}^n$, and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is an unknown nonlinear function.The terms e_i are assumed to be i.i.d. random errors with zero mean and constant (finite) variance. To avoid identifiability problems, we assume that the variables $\boldsymbol{z}$ are not identical to $\boldsymbol{x}$, and in general, that $\boldsymbol{z}$ can not be mapped to $\boldsymbol{x}$, as it will be further explained later. For instance, within the context of system identification, the following simple examples illustrate the type of models that can be identified using the above structure (6) by obvious definitions of $\boldsymbol{z}_i$ and $f(\boldsymbol{x}_i)$:

$$\begin{aligned} y_t &= \sum_{i=1}^{p} a_i y_{t-i} + \sum_{j=1}^{q} b_j G_1(u_{t-j}, \ldots, u_{t-j-k}) + \varepsilon_t, \\ y_t &= \sum_{i=1}^{p} a_i y_{t-i} + G_2(y_{t-p-1}, y_{t-p-2}, \ldots, y_{t-p-r}) + \varepsilon_t, \end{aligned}$$

with G_1, G_2 any nonlinear functions. Using a regularized cost function as in (2), a Partially Linear model where f is estimated by LS-SVM (PL-LSSVM),

$$y_i = \boldsymbol{\beta}^T\boldsymbol{z}_i + \boldsymbol{w}^T\varphi(\boldsymbol{x}_i) + b + e_i, \qquad i = 1, \ldots, N, \tag{7}$$

can be formulated as follows:

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{e},\boldsymbol{\beta}} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\boldsymbol{e}^T\boldsymbol{e} \\ \text{s.t.} \quad & \boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\Phi}\boldsymbol{w} + \mathbf{1}b + \boldsymbol{e}, \end{aligned} \tag{8}$$

with $\boldsymbol{Z} \in \mathbb{R}^{N \times p}$ the matrix of linear regressors $\boldsymbol{z}_i$. In order to ensure the existence of a unique solution, it is assumed that $\boldsymbol{Z}$ has full column rank and that there is no constant term among the variables in $\boldsymbol{z}_i$. Again, we can build the Lagrangian $\mathcal{L}(\boldsymbol{w}, b, e, \boldsymbol{\beta}; \alpha) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\boldsymbol{e}^T\boldsymbol{e} - \boldsymbol{\alpha}^T(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta} - \boldsymbol{\Phi}\boldsymbol{w} - \mathbf{1}b - \boldsymbol{e})$, with $\boldsymbol{\alpha} \in \mathbb{R}^N$ the vector of Lagrange multipliers. The optimality conditions are obtained as follows:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = 0 & \rightarrow \quad \boldsymbol{w} = \boldsymbol{\Phi}^T\boldsymbol{\alpha} \\ \frac{\partial \mathcal{L}}{\partial b} = 0 & \rightarrow \quad \boldsymbol{\alpha}^T\mathbf{1} = 0, \\ \frac{\partial \mathcal{L}}{\partial \boldsymbol{e}} = 0 & \rightarrow \quad \boldsymbol{\alpha} = \gamma\boldsymbol{e} \\ \frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = 0 & \rightarrow \quad \boldsymbol{Z}^T\boldsymbol{\alpha} = \mathbf{0}_{p\times 1} \\ \frac{\partial \mathcal{L}}{\partial \boldsymbol{\alpha}} = 0 & \rightarrow \quad \boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\Phi}\boldsymbol{w} + \mathbf{1}b + \boldsymbol{e}, \end{cases} \tag{9}$$

where $\mathbf{0}_{p\times 1}$ is a zero-valued vector of dimension $p \times 1$. After elimination of $\boldsymbol{w}$ and $\boldsymbol{e}$, we obtain the system

$$\left[\begin{array}{c|c|c} \boldsymbol{\Omega} + \gamma^{-1}\mathbf{I} & \mathbf{1} & \boldsymbol{Z} \\ \hline \mathbf{1}^T & 0 & \mathbf{0}_{1\times p} \\ \hline \boldsymbol{Z}^T & \mathbf{0}_{p\times 1} & \mathbf{0}_{p\times p} \end{array}\right] \left[\begin{array}{c} \boldsymbol{\alpha} \\ \hline b \\ \hline \boldsymbol{\beta} \end{array}\right] = \left[\begin{array}{c} \boldsymbol{y} \\ \hline 0 \\ \hline \mathbf{0}_{1\times p} \end{array}\right], \tag{10}$$

with solution $\boldsymbol{\alpha}, b$ and[1] $\hat{\boldsymbol{\beta}}$. The final model becomes

$$\hat{y}(\boldsymbol{x}, \boldsymbol{z}) = \hat{\boldsymbol{\beta}}^T \boldsymbol{z} + \sum_{i=1}^{N} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b. \qquad (11)$$

The constant term b, as in any regression, has to do with the fact that the data may or may not have been centered around their mean. So far, no assumptions have been made in that respect. From the related kernel theoretical results [14], [18] it is known that if the variables in feature space $\boldsymbol{\varphi}(\boldsymbol{x})$ are centered, then there is no need for the b term. Therefore, if one wishes not to have the b term in the above formulation, all $\boldsymbol{y}$, $\boldsymbol{Z}$ and $\boldsymbol{\Phi}$ have to be centered about their corresponding means.

B. Links with traditional statistical techniques

Partially linear models of the form (6) have been used in many interesting applications, starting from the famous study of Engle *et al.* on the relation between electricity consumption and temperature [2]. Statistical inference on the estimated parameters has been developed based on asymptotic theory and consistency results from nonparametric estimation theory [7]. Within the statistical literature, the model (6) is estimated by approximating f by a local smoother and solving a set of normal equations [16]

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^T(I - \mathcal{S})\boldsymbol{Z})^{-1}\boldsymbol{Z}^T(I - \mathcal{S})\boldsymbol{y}, \qquad (12)$$

where $\mathcal{S}$ is a smoother matrix. Usually $\mathcal{S}$ is related to local splines, or variants of the so-called Nadaraya-Watson estimator [11]. In practice, usually $\boldsymbol{x}$ is constrained to have a very low dimensionality (typically one-dimensional).

Within our framework, by working with the equations from system (10), and assuming the data have been centered, it is possible to write

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\Omega}[(\boldsymbol{\Omega} + \gamma^{-1}\mathbf{I})^{-1}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})] + \boldsymbol{e}. \qquad (13)$$

Pre-multiplying by $\boldsymbol{Z}^T$, and noting that $\boldsymbol{Z}^T\boldsymbol{e} = \boldsymbol{Z}^T\boldsymbol{\alpha}/\gamma = \mathbf{0}$ as given by one of the optimality conditions, we obtain

$$\boldsymbol{Z}^T\boldsymbol{y} = \boldsymbol{Z}^T S\boldsymbol{y} - \boldsymbol{Z}^T S\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{\beta}, \qquad (14)$$

where

$$S = \boldsymbol{\Omega}(\boldsymbol{\Omega} + \gamma^{-1}\mathbf{I})^{-1}, \qquad (15)$$

can be interpreted as the equivalent smoothing matrix obtained under the LS-SVM estimator. After solving for $\boldsymbol{\beta}$ in (14), one obtains (12), thus showing that the PL-LSSVM estimate for $\boldsymbol{\beta}$ is linked to the traditional statistical techniques by using the smoother S defined by (15). Moreover, in our context of linear and nonlinear identification, the use of the LS-SVM improves over the traditional local techniques as it can make use of a more general set of regressors in $\boldsymbol{x}$, regardless of its dimensionality; a unique solution is obtained for the global model, and the nonlinear behavior of f can be correctly identified by using the kernel trick over the variables $\boldsymbol{x}$. Additionally, non-local basis functions can be used for the approximation of the nonlinear function f as by using e.g. a polynomial kernel.

C. Conditions and Requirements

The following condition has to do with the requirements on $\boldsymbol{Z}$ in order to ensure the existence of a unique solution for the system (10).

Condition 1: The PL-LSSVM model defined by system (10) always admits a unique solution for $(\boldsymbol{\alpha}, b, \hat{\boldsymbol{\beta}})$ if and only if both of the following conditions hold:

- $\boldsymbol{Z}$ has full column rank, and
- $\boldsymbol{Z}$ should not contain a column $c\mathbf{1}_N, c \in \mathbb{R}$.

In order to prove this condition we state the following Lemma.

Lemma 1: Let $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, a positive definite matrix; $\boldsymbol{B} \in \mathbb{R}^{N\times p}; \boldsymbol{d}_1, \boldsymbol{a}_1 \in \mathbb{R}^N$, and $\boldsymbol{d}_2, \boldsymbol{a}_2 \in \mathbb{R}^p$. Then the linear system defined by

$$\left[\begin{array}{c|c} \boldsymbol{A} & \boldsymbol{B} \\ \hline \boldsymbol{B}^T & \mathbf{0} \end{array}\right] \left[\begin{array}{c} \boldsymbol{d}_1 \\ \hline \boldsymbol{d}_2 \end{array}\right] = \left[\begin{array}{c} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{array}\right], \qquad (16)$$

has a unique solution if and only if $\boldsymbol{B}$ has full column rank.

Proof: The solutions for $\boldsymbol{d}_1, \boldsymbol{d}_2$ can be written as

$$\begin{aligned} \boldsymbol{d}_1 &= \boldsymbol{A}^{-1}\boldsymbol{a}_1 - \boldsymbol{A}^{-1}(\boldsymbol{B}^T\boldsymbol{A}\boldsymbol{B})^{-1}(\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{a}_1 - \boldsymbol{a}_2) \\ \boldsymbol{d}_2 &= (\boldsymbol{B}^T\boldsymbol{A}\boldsymbol{B})^{-1}(\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{a}_1 - \boldsymbol{a}_2). \end{aligned}$$

The unique solution exists if and only if the matrices $\boldsymbol{A}$ and $\boldsymbol{B}^T\boldsymbol{A}\boldsymbol{B}$ are invertible. As $\boldsymbol{A}$ is positive definite, it is always invertible. And it is known that if $\boldsymbol{A}$ is positive definite, then $(\boldsymbol{B}^T\boldsymbol{A}\boldsymbol{B})$ is also positive definite, and therefore invertible, if and only if $\boldsymbol{B}$ has full column rank ([8], Observation 7.1.6, p.399). ■

For the case of a pure LS-SVM (4), the matrix $\boldsymbol{A} = \boldsymbol{\Omega} + \gamma^{-1}\mathbf{I}$ is positive definite, and the matrix $\boldsymbol{B}$ corresponds to a vector of ones, which has full rank and thus a unique solution always exists. For the case of the PL-LSSVM, one has $\boldsymbol{B} = [\mathbf{1}, \boldsymbol{Z}]$. By Lemma 1 a unique solution exists only if $\boldsymbol{B}$ has full rank. We need $\boldsymbol{Z}$ to have full rank. As the first column in $\boldsymbol{B}$ is a vector of ones, it is also required that no such column (up to a constant) is found within $\boldsymbol{Z}$, otherwise there would be 2 linearly dependent columns within $\boldsymbol{B}$.

The following condition has to do with identifiability of the parameter vector of the linear part, as it was briefly mentioned above.

Condition 2: The linear parameter vector $\boldsymbol{\beta}$ in (7) is not identifiable if there exists a mapping $g : \mathbb{R}^n \to \mathbb{R}^p$ such that $g(\boldsymbol{x}) = \boldsymbol{z}$.

This problem was already noted in [13]. It implies, for instance, that if a model contains only $\boldsymbol{x}$ variables, as $y_i = \boldsymbol{\beta}^T\boldsymbol{x}_i + \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i$, then the linear parameter is unidentifiable [2]. To see this, we can write $y_i = \boldsymbol{\beta}^T\boldsymbol{x}_i + \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i = \boldsymbol{\delta}^T\boldsymbol{x}_i + \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_i) - \boldsymbol{\delta}^T\boldsymbol{x}_i + \boldsymbol{\beta}^T\boldsymbol{x}_i + b + e_i$

[1] The computed $\boldsymbol{\alpha}, b$ do not have a "hat" symbol as it is not the standard notation within the machine learning framework. Besides, $\boldsymbol{\alpha}$ is a dual variable (Lagrange multiplier) and not a parameter vector to be estimated in a parametric setting.

[2] Although by Condition 1 a solution will always exist if the matrix of linear regressors has full column rank.

$= \boldsymbol{\delta}^T \boldsymbol{x}_i + \tilde{\boldsymbol{w}}^T \tilde{\boldsymbol{\varphi}}(\boldsymbol{x}_i) + b + e_i$, $\forall \boldsymbol{\delta} \in \mathbb{R}^p$. In this case we can define equivalently a new nonlinear component $\tilde{\boldsymbol{w}}^T \tilde{\boldsymbol{\varphi}}(\boldsymbol{x}_i) = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}_i) - \boldsymbol{\delta}^T \boldsymbol{x}_i + \boldsymbol{\beta}^T \boldsymbol{x}_i$. Thus a linear part has been subsumed by the new nonlinearity defined by $\tilde{\boldsymbol{w}} = [\boldsymbol{w}; -\boldsymbol{\delta} + \boldsymbol{\beta}], \tilde{\boldsymbol{\varphi}}(\boldsymbol{x}_i) = [\boldsymbol{\varphi}(\boldsymbol{x}_i); \boldsymbol{x}_i]$. As the function f in (6) is defined in a general way, it can approximate a general class of functions, including (obviously) linear functions of the same inputs. The same reasoning can be applied for a function g such that $g(\boldsymbol{x}) = \boldsymbol{z}$, leading to the same identifiability problem. However, it is important noting that Condition 2 does not rule out a relation between $\boldsymbol{x}$ and $\boldsymbol{z}$ as, for instance, each component of $\boldsymbol{x}$ being the squared of the corresponding component of $\boldsymbol{z}$, $\boldsymbol{x}^j = (\boldsymbol{z}^j)^2$, $\forall j = 1, \ldots, p$. In this case there is no problem of identifiability as $\boldsymbol{z}$ can not be perfectly recovered from $\boldsymbol{x}$. In practice, the partially linear model is used to estimate a linear response over certain variables when it is suspected that the total response also depends nonlinearly over a different set of variables, so the identifiability problem will rarely happen in applied work.

D. Relation with Hammerstein Models

In general, a Hammerstein (SISO) model $y_t = \sum_{i=1}^{p} a_i y_{t-i} + \sum_{j=1}^{q} b_j h(u_t) + \varepsilon_t$ contains a static nonlinearity h applied over the input u_t. The Generalized Hammerstein model extends the concept to include a NFIR formulation instead of a static nonlinearity, as $y_t = \sum_{i=1}^{p} a_i y_{t-i} + \sum_{j=1}^{q} b_j h(u_t, u_{t-1}, \ldots, u_{t-k}) + \varepsilon_t$. In these formulations it is possible to apply PL-LSSVM to identify the coefficients of the linear part and the nonlinear total component by an obvious definition of f in (6) as $f(u_t) = \sum_{j=1}^{q} b_j h(u_t)$ in the first case and $f(u_t, \ldots, u_{t-k}) = \sum_{j=1}^{q} b_j h(u_t, u_{t-1}, \ldots, u_{t-k})$ in the second case. However, with the exception of simple cases ($q = 1$), the identification of f does not translate directly to an identification of h; for a detailed identification of the function h eventually an ad-hoc structure is required [4] where further restrictions are imposed to the function f.

IV. Applications

In this section, we show some examples of the PL-LSSVM performance. Its ability to identify correctly the linear and nonlinear components for some examples is assessed by Monte Carlo simulations. Its out-of-sample forecasting performance is examined for 3 model examples. On all cases, an RBF kernel is used and the parameters σ and γ are found by 10-fold cross validation over the corresponding training sample.

A. Methodology

The test cases are defined as follows:

- **Case I: Linear trend + static nonlinearity.** The model to be estimated is of the form $y_t = a_1 t + 2\text{sinc}(x_t) + \varepsilon_t$, where the true value is $a_1 = 1.5$ and x_t is drawn from a uniform distribution over [0,2.5]; ε_t is a Gaussian white noise of variance 0.02.
- **Case II: Static linearity + static nonlinearity.** The model to be estimated is of the form $y_t = a_1 z_t + 2\text{sinc}(x_t) + \varepsilon_t$, where the true value is $a_1 = 1.5$; z_t and x_t are drawn from a uniform distribution over [0,2.5] and [0,1.5], respectively; ε_t is a Gaussian white noise of variance 0.02.
- **Case III: Linear autoregression + static nonlinearity** The model to be estimated is of the form $y_t = a_1 y_{t-1} + a_2 y_{t-2} + 2\text{sinc}(x_t) + \varepsilon_t$, where the true value are $a_1 = 0.6, a_2 = 0.3$; the x_t is drawn from a normal distribution with zero mean and variance 5; ε is a Gaussian white noise of variance 0.02. This corresponds to a simple Hammerstein system.
- **Case IV: Autoregression with linear and nonlinear components.** The model to be estimated is of the form $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \text{sinc}(y_{t-3}) + \varepsilon_t$, where the true values are $a_1 = 0.6, a_2 = 0.3$; ε is a Gaussian white noise of variance 0.02.
- **Case V: Hammerstein Model.** The true model is $y_t = a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3} + b_1 \text{sinc}(u_{t-1}) + b_2 \text{sinc}(u_{t-2}) + \varepsilon_t$, with $a_1 = 0.6, a_2 = 0.2, a_3 = 0.1, b_1 = 0.4, b_2 = 0.2$. The input u_t comes from a Gaussian distribution with mean 0 and variance 2, and ε_i is Gaussian noise with variance 0.1.
- **Case VI: Generalized Hammerstein Model.** The true model is a Generalized Hammerstein model $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \text{arctan}(u_t) u_{t-1}^2 + \varepsilon_t$, with $a_1 = -0.6, a_2 = -0.1$ and the input series is generated by $u_t = b_1 u_{t-1} + \varepsilon_{t-1} + \varepsilon_{t-2}$ where ε_t is Gaussian noise with variance 1 (this example is taken from [3]).

It worth noting that although the regressors contained in the linear part might be correlated with the regressors under the nonlinear part, they are neither identical nor perfectly related to each other. Therefore, there are no identification problems under Condition 2.

Identification of the linear and nonlinear components: Cases I to IV are used for Monte Carlo simulations. In order to compare the PL-LSSVM model with traditional techniques, Ordinary Least Squares (OLS) regression using all the variables (in linear form) is implemented, as well as the partially linear model with the Nadaraya-Watson (NW) [11] smoother as in [16]. For all cases above, data is generated by sampling the respective distributions and/or using the autoregressive forms where it corresponds. For all cases the number of datapoints is $N = 200$, and the number of Monte Carlo repetitions is taken equal to 1,000.

Forecasting performance: The out-of-sample performance, on an iterative basis (simulation mode) is examined for the models defined in Cases IV, V and VI above. 1,000 datapoints are generated and the first 400 are dismissed to remove any transient effect. 500 datapoints are then used for training, and the performance is measured over the next 50 out-of-sample points running the model iteratively in simulation mode, each time using past predictions as inputs to produce the next forecasts.

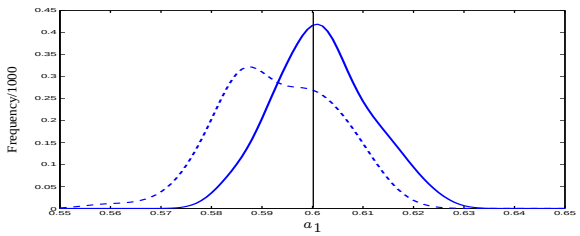


Fig. 1. Empirical Distributions for the estimated $\hat{a}_1$ under Case III using PL-LSSVM (full line) and NW (dashed line) over 1,000 repetitions. The vertical line shows the 'true' value.

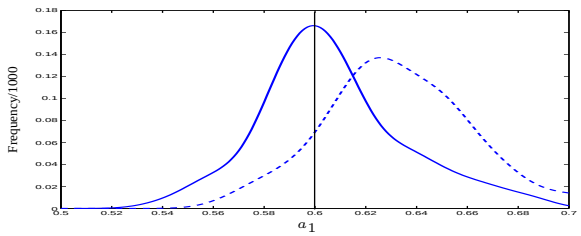


Fig. 2. Empirical Distributions for the estimated $\hat{a}_1$ under Case IV using PL-LSSVM (full line) and NW (dashed line) over 1,000 repetitions. The vertical line shows the 'true' value.

B. Results

Identification: Table 1 shows the results for the first 4 cases, as averages and standard deviations of the estimated parameters over 1,000 repetitions, together with the 10-fold crossvalidation mean squared error (CV-MSE). In the simple cases (I and II), all techniques give similar performance for the identification of the linear parameters. For the Case III a bias is present in the OLS-based estimation of the linear parameter, due to the time-series nature of the problem; and in Case IV both the NW and OLS show an important bias on each one of their estimates. The empirical distributions of the estimates obtained with this sampling can be visualized in Figures 1 and 2, for the comparison between the estimated parameter $\hat{a}_1$ using PL-LSSVM (full line) and NW (dashed line) for Case III and Case IV, respectively. Although the general conditions for asymptotic consistence for the NW partially linear model estimator have been studied extensively, in practice it is not straightforward to verify if these are being fulfilled by the problem at hand. By using Monte Carlo simulations for particular types of problems it is possible to verify the properties of each estimator, especially when temporal or serial correlation is present in the data [5]. Regarding identification of the nonlinear part, Figure 3 shows the identified nonlinear component of Case III. The 'x' show the estimated nonlinear $\hat{f} = \boldsymbol{\Omega\alpha}+b$ from the model, and the line shows the true value for the nonlinear 2sinc function, with excellent performance. For the above examples it is clear that the PL-LSSVM estimator gives a satisfactory global accuracy, and at the same time it identifies the linear part of each example with less bias problems than the other techniques.

Out-of-sample Forecasting: The results for Cases IV,V and VI regarding their estimation results and out-of-sample performance are reported on Table 2. The MSE obtained in the out-of-sample exercise (MSE simulation) is very close to the MSE level obtained within the training procedure by 10-fold cross validation (CV-MSE). At the same time, the linear parameters for each model were identified successfully. The out-of-sample iterative prediction is computed by sequentially using past predictions as new inputs for the autoregressive part of these models, in simulation mode with iterative predictions over time [9]. All models perform substantially well, as it is shown in Figure 4 for Case IV (top), Case V (middle) and Case VI (bottom), for the comparison between the predictions and the true values for the next 50 points out-of-sample.

	Estimates				**CV-MSE**	
	$\hat{a}_1$	$\sigma_{\hat{a}_1}$	$\hat{b}_1$	$\sigma_{\hat{b}_1}$	Mean	S.Dev
Case I						
PL-LSSVM	1.500	0.001	-	-	0.007	0.001
NW	1.500	0.003	-	-	0.09	0.01
OLS	1.498	0.007	-	-	0.19	0.02
Case II						
PL-LSSVM	1.50	0.01	-	-	0.008	0.001
NW	1.50	0.04	-	-	0.11	0.01
OLS	1.50	0.08	-	-	0.21	0.02
Case III						
PL-LSSVM	0.60	0.01	0.30	0.01	0.009	0.001
NW	0.59	0.01	0.30	0.01	0.17	0.01
OLS	0.57	0.01	0.32	0.01	0.25	0.01
Case IV						
PL-LSSVM	0.60	0.03	0.30	0.04	0.006	0.001
NW	0.63	0.03	0.26	0.04	0.07	0.01
OLS	1.16	0.05	-0.5	0.06	0.28	0.04

TABLE I
MEAN AND STD. DEV. FOR THE PARAMETER ESTIMATES AND THE CV-MSE OVER 1,000 REPETITIONS.

V. CONCLUSION

Starting from the definition of LS-SVMs, it is possible to define a feasible estimator for a partially linear model by extending the model formulation in order to include a parametric part. The solution is shown to be unique and to exist under the usual requirements for a set of

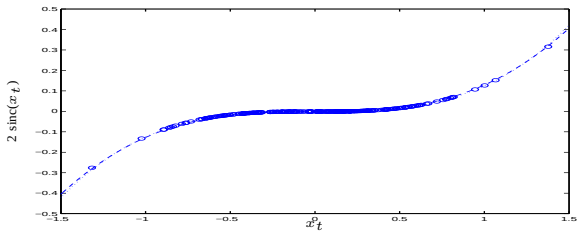


Fig. 3. The nonlinear part for the model of Case III. The 'x' shows the estimated nonlinear part as computed by the model, the line shows the 'true' nonlinear function.

	Estimates			MSE	
	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	CV (train)	Simulation
Case IV	0.598	0.302	-	0.006	0.005
Case V	0.597	0.195	0.11	0.007	0.010
Case VI	-0.592	-0.098	-	1.19	1.18

TABLE II
PARAMETER ESTIMATES, MSE (CV-TRAINING AND SIMULATION) FOR CASES IV (NAR), CASE V (HAMMERSTEIN) AND CASE VI (GENERALIZED HAMMERSTEIN).

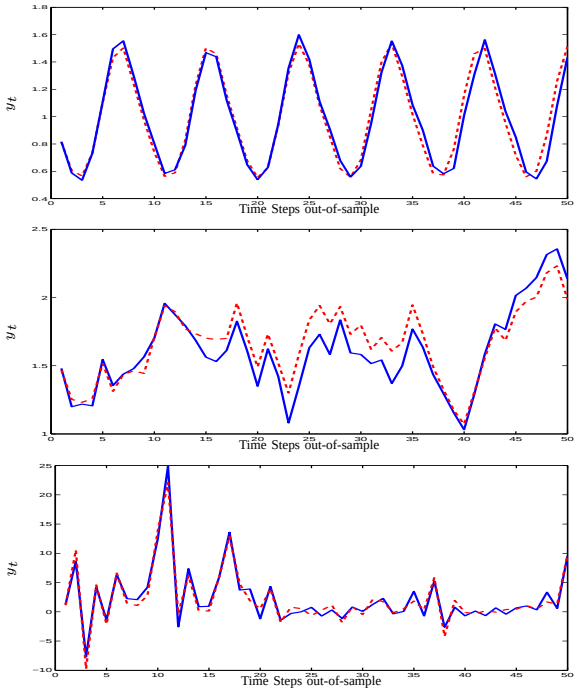


Fig. 4. Simulated (dashed) and Observed (full) values for the next 50 time steps out-of-sample for Cases IV (top), V (middle) and VI (bottom).

linear parametric regressors. This Partially Linear LS-SVM formulation is optimal in a least-squares sense, and it allows to identify a general class of model structures. Its parametric part has the same structural form as the classical statistical methods, and it extends the classical notion of semiparametric regression by allowing explicitly to include any potential nonlinear regressor as the dimensionality of the system is defined in terms of the kernel matrix under Mercer's theorem. In particular, the PL-LSSVM formulation makes it possible to tackle the identification of Hammerstein models in a simple way for forecasting purposes.

Practical examples over 4 particular types of models show the overall ability of the PL-LSSVM to identify the linear and nonlinear parts. Using Monte Carlo methods over 1,000 repetitions, it is clear that this method has a better global accuracy for the models and a better identification performance when compared with traditional techniques. In addition, good out-of-sample forecasting performance is illustrated in 3 examples. Further research is focused on mechanisms for order and input selection, asymptotic properties of the estimated parameters and further applications within the nonlinear system identification framework.

ACKNOWLEDGMENTS

This work was supported by grants and projects for the Research Council K.U.Leuven (GOA-Mefisto 666, IDO, PhD/Postdocs & fellow grants), the Flemish Government (FWO: PhD/Postdocs grants, projects G.0240.99, G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, ICCoS, ANMMM; AWI;IWT:PhD grants, Soft4s), the Belgian Federal Government (DWTC: IUAP IV-02, IUAP V-22; PODO-II CP/40), the EU(CAGE, ERNSI, Eureka 2063-Impact;Eureka 2419-FLiTE) and Contracts Research/Agreements (Data4s, Electrabel, Elia, LMS, IPCOS, VIB). J. Suykens and B. De Moor are an associated professor and a full professor at the K.U.Leuven, Belgium, respectively. The scientific responsibility is assumed by its authors.

REFERENCES

[1] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.
[2] R. Engle, C.W. Granger, J. Rice, and A. Weiss. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 81(394):310–320, 1986.
[3] M. Enqvist. Linear Models of Nonlinear FIR Systems with Gaussian Inputs. Technical Report LiTH-ISY-R-2462, Linköping Universitet, Sweden, 2002.
[4] I. Goethals, K. Pelckmans, J.A.K. Suykens, and B. De Moor. NARX Identification of Hammerstein Models Using Least Squares Support Vector Machines. Internal Report 04-40, ESAT-SISTA, K.U.Leuven, 2004.
[5] J. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.
[6] W. Härdle. *Applied Nonparametric Regression*. Econometric Society Monographs. Cambridge University Press, 1989.
[7] W. Härdle, H. Liang, and J. Gao. *Partially Linear Models*. Physica-Verlag, Heidelberg, 2000.
[8] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.
[9] L. Ljung. *System Identification: Theory for the User*. Prentice Hall, New Jersey, 1987.
[10] D.J.C. MacKay. Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11:1035–1068, 1999.
[11] E.A. Nadaraya. On estimating regression. *Theory of Probability and its Application*, 10:186–190, 1964.
[12] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78:1481–1497, 1990.
[13] P.M. Robinson. Root n-consistent Semiparametric Regression. *Econometrica*, 56(4):931–954, 1988.
[14] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
[15] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Deylon, P. Glorennec, H. Hjalmarsson, and A. Juditsky. Nonlinear Black-box Modelling in System Identification: a Unified Overview. *Automatica*, 31:1691–1724, 1995.
[16] P. Speckman. Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, 1988.
[17] J.A.K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle. Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, 48(1-4):85–105, 2002.
[18] 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.
[19] J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9:293–300, 1999.
[20] V. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, New-York, 1995.