An Instrumental Least Squares Support Vector Machine for Nonlinear System Identification ⋆

Vincent Laurain [a,b], Roland Tóth [c], Dario Piga [d], Wei Xing Zheng [e],

[a] *Université de Lorraine, CRAN, UMR 7039, 2 rue Jean Lamour, 54519 Vandoeuvre-lès-Nancy Cedex, France.*

[b] *CNRS, CRAN, UMR 7039, France.*

[c] *Control Systems Group, Department of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB, Eindhoven, The Netherlands.*

[d] *IMT Institute for Advanced Studies Lucca, Piazza San Ponziano 6, 55100 Lucca, Italy.*

[e] *School of Computing and Mathematics, University of Western Sydney, Penrith NSW 2751, Australia.*

Abstract

Least-Squares Support Vector Machines (LS-SVMs), originating from Statistical Learning and Reproducing Kernel Hilbert Space (RKHS) theories, represent a promising approach to identify nonlinear systems via nonparametric estimation of the involved nonlinearities in a computationally and stochastically attractive way. However, application of LS-SVMs and other RKHS variants in the identification context is formulated as a regularized linear regression aiming at the minimization of the ℓ_2 loss of the prediction error. This formulation corresponds to the assumption of an auto-regressive noise structure, which is often found to be too restrictive in practical applications. In this paper, Instrumental Variable (IV) based estimation is integrated into the LS-SVM approach, providing, under minor conditions, consistent identification of nonlinear systems regarding the noise modeling error. It is shown how the cost function of the LS-SVM is modified to achieve an IV-based solution. Although, a practically well applicable choice of the instrumental variable is proposed for the derived approach, optimal choice of this instrument in terms of the estimates associated variance still remains to be an open problem. The effectiveness of the proposed IV based LS-SVM scheme is also demonstrated by a Monte Carlo study based simulation example.

Key words: support vector machines; reproducing kernel Hilbert space; instrumental variables; nonlinear identification; machine learning; non-parametric estimation.

1 Introduction

Support vector machines (SVMs) have been originally developed as a class of *supervised learning* methods in stochastic learning theory. Their original purpose was to provide efficient tools for data analysis and pattern recognition in classification problems and regression analysis [1,2]. SVMs have had a paramount impact on the *machine learning* field since their extension as a theoretical framework in that setting [3]. These methods also offer an attractive, so-called non-parametric way of data-driven dynamic modeling, *i.e.*, *system identification*, especially in the nonlinear context. In that context, these approaches are part of the data-driven model learning avenue [4–6], focusing on the paradigm of estimation of the targeted system without posing prior assumptions on its dynamical nature or the nonlinearities involved. Most of the research interest regarding identification with SVMs has been dedicated to *nonlinear block models* so far, using various *least-square* SVM (LS-SVM) approaches where the original nonlinear estimation problem is posed as a linear regression [7,8]. In general, LS-SVMs are particular variations of the original support vector machine approach using a regularized ℓ_2 loss function instead of a so called ϵ-insensitive loss function on the prediction error of the

⋆ This paper was not presented at any IFAC meeting. Corresponding author Roland Tóth. Tel. +31-40-247-2655. Fax +31-40-243-4582.

⋆⋆This work was supported by the Netherlands Organization for Scientific Research (NWO, grant no.: 639.021.127) and by the French ministries of Foreign Affairs, Education and Research and the French-Dutch Academy (PHC Van Gogh project, n. 29342QL).

Email addresses: vincent.laurain@univ-lorraine.fr (Vincent Laurain), r.toth@tue.nl (Roland Tóth), dario.piga@imtlucca.it (Dario Piga), W.Zheng@uws.edu.au (Wei Xing Zheng).

model. A particular advantage of expressing both the regularization and the loss in the ℓ_2 norm is that the solution of the corresponding optimization problem is obtained by solving a system of linear equations and an attractive trade-off between regularization bias and variance of the estimates is present [8]. LS-SVMs are also related to Kriging [9] in geostatistics and *Gaussian processes* (GPs) in machine learning, *e.g.*, [10,11], which approaches can be seen as different variants of the *reproducing kernel Hilbert space* (RKHS) theory based function estimators. The relation between these methods is analyzed in [12,6].

A particular handicap of the variants of LS-SVMs (and also GPs) is that the used linear regression form under the ℓ_2 loss function corresponds to the assumption that all disturbances affecting the data-generating system can be expressed as a white noise disturbance on the equation error level, which can be seen as the assumption of a *nonlinear auto-regressive* noise structure. Such an assumption is often found to be too restrictive in practical applications. In the classical identification literature, significant research efforts have been devoted to achieve consistent estimation in case of rather general noise assumptions corresponding to the situations commonly encountered in practice [13]. To introduce the same generality of noise structures, some steps have been taken in the LS-SVM context such as the recurrent LS-SVM developed in [14] and the linear parametric noise model equipped SVM derived in [15]. However, the classical results in identification highlight that the chosen noise model, *i.e.*, the assumed noise properties, plays an important role in the consistency of the estimates. Therefore, in the light of a non-parametric prior-free modeling objective, the question rises why we should bound ourself to *a priori* specified noise assumptions, especially in the general nonlinear context. For example, in the GPs related literature for LTI models, consistency under general noise conditions is established by identifying the one-step-ahead predictor of the output, which, due to linearity, allows factorization of high order linear regression based estimates to obtain estimates of the process and the noise dynamics without posing any priors on the noise [16]. However, in the nonlinear case, the loss of linearity of this predictor in the inputs and outputs prevents applicability of this methodology, allowing consistency only under restrictive assumptions, see [5]. So, the important question that rises is how we can achieve similar generality in the nonlinear case.

By turning to the classical results, we can find that variants of linear regression based methods, *e.g.*, *instrumental variable* (IV) approaches, have been developed to cope with realistic assumptions on the noise without specifying a direct parametrization or structure [13]. The strength of IV methods in the LTI case has been found in delivering consistent estimates independently of the chosen noise model assumption in a computationally attractive way [17]. Therefore, to extend consistency of non-parametric identification with LS-SVMs in the nonlinear case, in this paper, we consider the idea of introducing the IV scheme into the LS-SVM regression structure, which was first [1] proposed in [19]. As a significant improvement of the initial scheme described in [19], in this paper, we provide a rigorous treatment of instrumental variables based LS-SVMs and showing the applicability of IV based techniques both in non-parametric identification and in regularized contexts. Furthermore, this contribution not only preserves the computationally attractive feature of the original approach by satisfying the *Mercer conditions*, but also provides unbiased estimates under general noise model structures/conditions; opening a large set of application areas for data-driven nonlinear model learning.

The paper is organized as follows: the considered problem setting and the motivation for improving the LS-SVM method are discussed in Section 2. In Section 3, the optimization problem associated with the IV-based, non-parametric model estimation is introduced together with its solution. This is followed by integrating the IV solution into the LS-SVM estimation scheme for nonlinear dynamic systems resulting in the IV-SVM method. In Section 4, the choice of the instrumental variables are discussed from the variance point of view together with the selection of kernel functions and tuning of the hyper parameters. To demonstrate the advantages of the IV-SVM, a Monte Carlo study in Section 5 is provided in which the identification of a nonlinear system under colored noise is analyzed. Finally, conclusions and some future directions of research are given in Section 6.

2 Problem description

To set the stage for the upcoming discussion, the considered identification problem is defined in this section.

2.1 The data-generating system

As an objective of the identification scenario, the data-driven estimation of a rather general class of nonlinear discrete-time systems is considered. For the sake of simplicity of the upcoming derivations, the system $\mathcal{S}_\mathrm{o}$ is assumed to be *single-input single-output* (SISO). The behavior of $\mathcal{S}_\mathrm{o}$ is described by the following difference equation

$$y(k) = f_\mathrm{o}(x(k)) + v_\mathrm{o}(k), \tag{1}$$

where $x(k) \in \mathbb{R}^{n_\mathrm{g}}$ is a vector which, in the present identification context, is composed of the delayed values of the output and input signals of $\mathcal{S}_\mathrm{o}$, y and u respectively:

$$x(k) = [y(k-1) \dots y(k-n_\mathrm{a})\ u(k) \dots\ u(k-n_\mathrm{b})]^\top,$$

with $n_\mathrm{g} = n_\mathrm{a} + n_\mathrm{b} + 1$. $f_\mathrm{o} : \mathbb{R}^{n_\mathrm{g}} \to \mathbb{R}$ is assumed to be a bounded nonlinear function belonging to the set

[1] Note that IV has also been applied to nonlinear systems in [18]. However, [18] is not related to the current work as it only applies a parametric IV method to identify local LTI models of a nonlinear system around some operating conditions.

of real square integrable functions $\mathcal{L}_2(\mathbb{R}^\bullet, \mathbb{R})$. $v_\mathrm{o}(k)$ is considered as a zero-mean, quasi-stationary stochastic noise process (not necessarily white), independent of u. Note that the general structure of the system defined by (1) can be used to describe usual block structures such as *Hammerstein* and/or *Wiener* systems by *a priori* restrictions of the structure of f_o, *e.g.*:

$$y(k) = \sum_{i=1}^{n_\mathrm{a}} f_i^\mathrm{o}\big(y(k-i)\big) + \sum_{j=0}^{n_\mathrm{b}} g_j^\mathrm{o}\big(u(k-j)\big) + v_\mathrm{o}(k). \quad (2)$$

Formulation of (1) in the *multi-input multi-output* (MIMO) case is also available as shown in [8]. Note that in case $v_\mathrm{o} = e_\mathrm{o}$, where e_o is white, (1) can be seen as a *nonlinear auto-regressive with exogenous input* (NARX) model [20].

2.2 The modeling paradigm

To briefly discuss the concept behind the LS-SVM estimator and to develop the motivations for the proposed extension of this approach, let us consider the classical parametric estimation of (1), in which the the nonlinearity is assumed to have an expansion (*e.g.*, see [21]):

$$f_\mathrm{o}\big(x(k)\big) = \sum_{i=1}^{n_\mathrm{H}} \theta_i^\mathrm{o} \phi_i\big(x(k)\big), \quad (3)$$

where $\{\phi_i : \mathbb{R}^{n_\mathrm{g}} \to \mathbb{R}\}_{i=1}^{n_\mathrm{H}}$ is a set of basis functions over a function space $\mathcal{H} \subset (\mathbb{R}^{n_\mathrm{g}})^{\mathbb{R}}$, for example $\mathcal{L}_2(\mathbb{R}^{n_\mathrm{g}}, \mathbb{R})$, and $\{\theta_i^\mathrm{o} \in \mathbb{R}\}_{i=1}^{n_\mathrm{H}}$ are the associated expansion parameters. This means that the nonlinearity is conceptually modeled as the projection $\phi^\top(\bullet)\theta$ by an *a priori* specified n_H dimensional mapping $\phi : \mathbb{R}^{n_\mathrm{g}} \to \mathbb{R}^{n_\mathrm{H}}$ from the space of input-output samples to the so called *feature space* of the output samples. This concept leads to a parametrized model $\mathcal{M}_\theta$:

$$y(k) = \varphi^\top(k)\, \theta + e(k), \quad (4)$$

where $e(k)$ qualifies as the *prediction error* and $f_\theta = \varphi^\top(k)\theta$ is the function estimate. The regressor $\varphi(k)$ and the parameter vector θ are n_H-dimensional vectors defined as

$$\varphi(k) = \Big[\ \phi_1^\top\big(x(k)\big)\ \ldots\ \phi_{n_\mathrm{H}}^\top\big(x(k)\big)\ \Big]^\top, \quad (5a)$$

$$\theta = \Big[\ \theta_1^\top\ \ldots\ \theta_{n_\mathrm{H}}^\top\ \Big]^\top. \quad (5b)$$

A well known approach to obtain an estimate of θ is to minimize the *least-squares* (LS) criterion (used to formulate the quality of the model fit) $\sum_{k=1}^N e^2(k)$, where $e(k) = y(k) - \varphi^\top(k)\theta$ is the *prediction error* w.r.t. a data set $\mathcal{D}_N = \{(y(k), u(k))\}_{k=1}^N$ generated by (1). As already motivated, the major problem with the parametric approach is the *a priori* choice of the basis functions ϕ influencing the approximation error and the variance of the estimates. Therefore, it is generally important to find $\{\phi_i\}_{i=1}^{n_\mathrm{H}}$, based on a given $\mathcal{D}_N$, which can achieve a good trade-off between the following objectives:

- to minimize n_H, *i.e.*, the number of estimated parameters (minimizing the variance of θ);
- to represent the function f_o with minimal error (minimizing the structural bias).

Instead of constraining f_o to a specific parametric structure, these objectives can be achieved by searching for the estimate f in a possibly infinite dimensional function space $\mathcal{H}$. Let $\mathcal{H}$ to be a *Hilbert space*, *i.e.*, a function space equipped with an inner product $\langle \bullet, \bullet \rangle$, like $\mathcal{L}_2$, and being complete (in terms of convergence of all Cauchy sequences) with respect to the induced norm $\|f\|_\mathcal{H} = \sqrt{\langle f, f \rangle}$, see [22]. Then, an elegant way of guaranteeing well-posedness and avoiding overfitting is to introduce $\|f\|_\mathcal{H}$ as a regularizer in the criterion:

$$\mathcal{V}(f, e) = \frac{1}{2}\|f\|_\mathcal{H}^2 + \frac{\gamma}{2N}\sum_{k=1}^N e^2(k), \quad (6)$$

where the scalar $\gamma > 0$ is the *regularization parameter* which defines the trade-off between the above listed objectives and $1/N$ is a normalization. However, since we only have a finite set of observations $\mathcal{X} = \{x(k)\}_{k=1}^N$ in $\mathcal{D}_N$, *i.e.*, our information on the system is limited, hence the minimization of (6) is only well posed if we restrict our search space to the RKHS over $\mathcal{X}$, which is a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ satisfying the following boundedness criterion: $\forall f \in \mathcal{H}$ and $\forall x \in \mathcal{X}$, there is a $0 \leq c < \infty$ such that $|f(x)| < c\|f\|_\mathcal{H}$. For every such RKHS $\mathcal{H}$, there exists a unique symmetric, so called *reproducing Kernel function* $K : \mathcal{X} \times \mathcal{X} \to \mathcal{R}$ which is positive semidefinite [2] and $f(x) = \langle f(\bullet), K(\bullet, x) \rangle$ for all $(f, x) \in (\mathcal{H}, \mathcal{X})$ [23]. This observation gives a one-to-one correspondence between RKHS of functions and positive semidefinite Kernel functions and leads us to a data-driven estimation of f_o in terms of the following theorem:

Theorem 1 (Representer Theorem, [24]) *For the RKHS $\mathcal{H}$, the minimizer of* (6) *is*

$$\hat{f}_\mathrm{LS}(\bullet) = \sum_{i=1}^N \hat{\alpha}_i K_{x_i}(\bullet), \quad (7)$$

with $K_{x_i}(\bullet) = K(\bullet, x(i))$ and $\hat{\alpha}_\mathrm{LS} = [\ \hat{\alpha}_1\ \ldots \hat{\alpha}_N\ ] \in \mathbb{R}^N$ being given by

$$\hat{\alpha}_\mathrm{LS} = \left(\frac{1}{N}K_{xx} + \gamma^{-1} I_N\right)^{-1} \frac{1}{N} Y, \quad (8)$$

where $Y = \big[\ y(1)\ \ldots\ y(N)\ \big]^\top$, $I_N \in \mathbb{R}^{N \times N}$ is an identity matrix and K_{xx} is a matrix defined as:

$$K_{xx}(i, j) = K\big(x(i), x(j)\big). \quad (9)$$

This RKHS optimal function estimator has become known as the LS-SVM approach, due to the regularized ℓ_2-loss (6) [12]. Note that the estimator (7) has an interpretation in terms of GPs, Kriging or any other

[2] $K : \mathcal{X} \times \mathcal{X} \to \mathcal{R}$ is positive semidefinite, if $\forall n \in \mathbb{N}$, $\sum_{i=1}^n \sum_{j=1}^n c_i K(x_i, x_j) c_j \geq 0$, $\forall (x_i, c_i), (x_j, c_j) \in \mathcal{X} \times \mathcal{R}$.

estimators based on the result of Theorem 1. More general version of Theorem 1 also exists where the ℓ_2 loss is replaced by a convex function of the estimation error, leading to the class of general SVMs [2].

A typical choice of the kernel, which provides uniformly effective representation of a large class of smooth functions, is the *radial basis function* (RBF) kernel:

$$K\big(x(i),x(j)\big)=\exp\left(\tfrac{-\|x(i)-x(j)\|_2^2}{\sigma^2}\right), \qquad (10)$$

with $\sigma>0$ being a tunable hyper-parameter of the induced RKHS. However, other positive definite kernels, like *linear*, *polynomial*, *rational*, *spline* or *wavelet* kernels, can also be used [2]. Choosing the most appropriate kernel highly depends on the problem at hand. Automatic kernel selection for general SVM problems is possible and is discussed in [25].

It must be noted that, similarly to the classic parametrization based approaches, the optimal function estimate (7), based on the given data set, is a linear combination of basis functions, but these basis are not fixed *a priori*, they are generated by the Kernel function centered on a set of node points $\omega\in\mathbb{R}^N$, *i.e.*: $\{\phi_i(\centerdot)=K_{\omega_i}(\centerdot)\}_{i=1}^{n_{\rm H}=N}$ where the $i^{\rm th}$ node is $\omega_i=x(i)$. This has got two important consequences:

- As shown in [22], with these basis functions, (6) can be equivalently stated as

$$\mathcal{V}(\theta,e)=\frac{1}{2}\theta^\top\theta+\frac{\gamma}{2N}\sum_{k=1}^{N}e^2(k), \qquad (11)$$

  which results in a simplified derivation of LS-SVMs where minimization of (11) over θ and e leads to a dual estimation problem of the *Lagrangian parameter* vector $\alpha\in\mathbb{R}^N$ with a solution of (7); giving an other interpretation of the RKHS estimator.
- In view of (11), the estimator (7) is actually equivalent to the regularized estimation of the following model on $\mathcal{D}_N$:

$$y(k)=\sum_{i=1}^{N}\alpha_iK_{\omega_i}(x(k))+e(k), \qquad (12)$$

  or

$$Y=K_{x\omega}\alpha+E, \qquad (13)$$

  with $E=\big[\,e(1)\;\ldots\;e(N)\,\big]^\top$.

Hence, the LS-SVM estimate (8) is a particular estimate of α in (12) when using $\{\omega_i=x(i)\}_{i=1}^N$. This particular choice is an essential and unavoidable ingredient of the RKHS and the learning theories [1,2]. However, after estimation, the resulting predictor of the output takes the form $\hat{y}(k\mid k-1)=\sum_{i=1}^N\hat{\alpha}_{{\rm LS},i}K_{\omega_i}(x(k))$. This means that ω, in terms of the training data set, remains associated with the model estimate and its structure, fixing the series expansion of (3) to $f_{\rm o}(\centerdot)\approx\sum_{i=1}^N\alpha_{{\rm o},i}K_{\omega_i}(\centerdot)$ where $\alpha_{{\rm o},i}$ are interpreted as the optimal expansion coefficients in the RKHS norm sense. This observation will play a crucial role in our analysis.

2.3 Noise induced bias of the LS-SVM estimate

The main objectives of this paper are to show how non-whiteness of $v_{\rm o}$ induces a bias for RKHS based methods and to propose a solution in order to handle such cases efficiently in the nonlinear context. Hence, the analysis will be driven by discussing how the estimated functions in LS-SVM models are affected by noise in terms of their deviation from the true nonlinear function $f_{\rm o}$. In order to simplify the discussion by parting the function approximation problem from the analysis of the stochastic properties, let us consider the ideal case when the true data-generating system has a direct explicit definition in terms of the Kernel function K such as:

$$y(k)=\underbrace{\sum_{i=1}^{N}\alpha_{{\rm o},i}K(x(k),\omega_i)}_{f_{\rm o}(x(k))}+v_{\rm o}(k), \qquad (14)$$

where all ω_i are distinct and $\alpha_{{\rm o},i}=\langle f_{\rm o}(\centerdot),K(\centerdot,\omega_i)\rangle$ in terms of the RKHS inner product. Based on the considerations given in Section 2.2, the node points $\{\omega_i\}_{i=1}^N$ are assumed to be equivalent with the data points in $\mathcal{D}_N$ used to obtain the LS-SVM estimate via (8). This means that there exists a $\Omega_{\rm o}\subseteq\{\omega_i\}_{i=1}^N$ for which $f_{\rm o}(\centerdot)=\sum_{\omega_i\in\Omega_{\rm o}}\alpha_{{\rm o},i}K(\centerdot,\omega_i)$ and, in terms of the representer theorem [22], if $N\to\infty$ and $\{w_i\}_{i=1}^N$ are all distinct, then the corresponding sequence of $\{\alpha_{{\rm o},i}\}_{i=1}^\infty$ is well-defined and absolute convergent. Hence, if we can show that the estimated function is biased w.r.t. this ideal case, then such a conclusion demonstrates that RKHS methods can fail to capture the underlaying system.

Let $\mathbb{E}$ be the expectation operator and for a random process $f(x)$ with $f:\mathbb{R}^{n_{\rm g}}\to\mathbb{R}$, let $m_{\rm f}(x)=\mathbb{E}\{f(x)\}$ denote the mean function. As a shorthand notation, we will refer to $\mathbb{E}\{f(x)\}$ as a function of $x\in\mathbb{R}^{n_{\rm g}}$ with $\mathbb{E}\{f(\centerdot)\}$. Then, in view of (14) and (8),

$$\begin{aligned}\bar{\mathbb{E}}\left\{\hat{f}_{\rm LS}(\centerdot)-f_{\rm o}(\centerdot)\right\}&=\bar{\mathbb{E}}\left\{K_{\centerdot\omega}(\hat{\alpha}_{\rm LS}-\alpha_{\rm o})\right\}\\&=\mathbb{E}\left\{\sum_{i=1}^{\infty}K(\centerdot,\omega_i)(\hat{\alpha}_{{\rm LS},i}-\alpha_{{\rm o},i})\right\}\end{aligned} \qquad (15)$$

where $\bar{\mathbb{E}}\{\centerdot\}$ is the generalized expectation operator, $K_{\centerdot\omega}\in\mathbb{R}^{1\times N}=[K(\centerdot,\omega_1)\ldots K(\centerdot,\omega_N)]$ and the data set $\mathcal{D}_N$ is assumed to be quasi stationary. By defining $R(\gamma,N)=\big(\frac{1}{N}K_{x\omega}+\gamma^{-1}I_N\big)$ with $\gamma>0$, $\hat{\alpha}_{\rm LS}$ as a random variable is equal to

$$\hat{\alpha}_{\rm LS}=R^{-1}(\gamma,N)\frac{1}{N}Y=R^{-1}(\gamma,N)\frac{1}{N}(K_{x\omega}\alpha_{\rm o}+V_{\rm o}), \qquad (16)$$

where $V_{\rm o}=\big[\,v_{\rm o}(1)\;\ldots\;v_{\rm o}(N)\,\big]^\top$. Therefore,

$$\hat{\alpha}_{\rm LS}-\alpha_{\rm o}=R(\gamma,N)^{-1}\Big(\frac{1}{N}V_{\rm o}-\frac{1}{\gamma}\alpha_{\rm o}\Big). \qquad (17)$$

As $R^{-1}(\gamma,N)=\gamma I_N-\frac{\gamma}{N}R^{-1}(\gamma,N)K_{x\omega}$, Eq. (17) leads to the following formulation of the bias

$$\bar{\mathbb{E}}\left\{\hat{f}_{\mathrm{LS}}(.)-f_{\mathrm{o}}(.)\right\}=B_{\mathrm{N}}^{\mathrm{LS}}-B_{\mathrm{R}}^{\mathrm{LS}}, \tag{18}$$

where

$$B_{\mathrm{N}}^{\mathrm{LS}}=\bar{\mathbb{E}}\left\{\frac{\gamma}{N}K_{.\omega}V_{\mathrm{o}}\right\}-\bar{\mathbb{E}}\left\{\frac{\gamma}{N^2}K_{.\omega}R^{-1}(\gamma,N)K_{x\omega}V_{\mathrm{o}}\right\},$$
$$B_{\mathrm{R}}^{\mathrm{LS}}=\bar{\mathbb{E}}\left\{\frac{1}{\gamma}K_{.\omega}R^{-1}(\gamma,N)\alpha_{\mathrm{o}}\right\}.$$

Due to quasi stationarity of (u,y), for $\gamma>0$, $R_*^{\mathrm{K}}(.):=\lim_{N\to\infty}K_{.\omega}R^{-1}(\gamma,N)$ is assumed to exist.

2.3.1 The no-correlation case

If ω_k is not correlated to $v_{\mathrm{o}}(k)$, *e.g.*, v_{o} is a white noise process, then the noise induced terms composing B_N^{LS} in (18) vanish. Indeed, if v_{o} is quasi stationary and zero mean, then $\bar{\mathbb{E}}\left\{\frac{\gamma}{N^2}K_{.\omega}R^{-1}(\gamma,N)K_{x\omega}V_{\mathrm{o}}\right\}=R_*^{\mathrm{K}}(.)\bar{\mathbb{E}}\left\{\frac{\gamma}{N^2}K_{x\omega}V_{\mathrm{o}}\right\}$ and

$$\bar{\mathbb{E}}\left\{\frac{\gamma}{N}K_{.\omega}V_{\mathrm{o}}\right\}=\bar{\mathbb{E}}\left\{\frac{\gamma}{N}K_{.\omega}\right\}\bar{\mathbb{E}}\left\{V_{\mathrm{o}}\right\}\equiv 0, \tag{19a}$$
$$\bar{\mathbb{E}}\left\{\frac{\gamma}{N^2}K_{x\omega}V_{\mathrm{o}}\right\}=\bar{\mathbb{E}}\left\{\frac{\gamma}{N^2}K_{x\omega}\right\}\bar{\mathbb{E}}\left\{V_{\mathrm{o}}\right\}\equiv 0. \tag{19b}$$

Hence, the bias expression becomes

$$\bar{\mathbb{E}}\left\{\hat{f}(.)-f_{\mathrm{o}}(.)\right\}=-\frac{1}{\gamma}R_*^{\mathrm{K}}(.)\alpha_{\mathrm{o}}=-B_{\mathrm{R}}^{\mathrm{LS}}. \tag{20}$$

The term $B_{\mathrm{R}}^{\mathrm{LS}}$ can be seen as the regularization induced bias and it can be made arbitrary small by the choice of γ, which scales the trade-off between this bias and the resulting variance of the function estimate.

2.3.2 The correlation case

In case ω_k is correlated to $v_{\mathrm{o}}(k)$, conditions (19a)-(19b) are not fulfilled and the term $B_{\mathrm{N}}^{\mathrm{LS}}$ in (18) does not vanish. This means that the estimate f, besides of the regularization bias which is controlled by γ, will be deteriorated by an additional noise induced term $B_{\mathrm{N}}^{\mathrm{LS}}$. It is also important to note that while increasing γ decreases $B_{\mathrm{R}}^{\mathrm{LS}}$, it has the opposite effect on $B_{\mathrm{N}}^{\mathrm{LS}}$. This means that tuning of γ will correspond to a trade-off between balancing of the bias terms, resulting in a decreased approximation capability of the learning process.

3 The instrumental variables SVM

Among the available identification approaches used in the regression framework, the principle idea behind the *instrumental variable* (IV) approach has been successfully applied in many contexts to elegantly resolve the inconsistency problem of LS estimation under correlated noise v_{o} [26,13]. However, it has never been applied for kernel-based methods. In the sequel, our objective is therefore to develop and analyze an IV regularized criterion and to use an extension of the LS-SVM kernel theory in order to propose a solution to the nonlinearity modeling problem, allowing a much wider applicability of this identification approach in practice.

We have seen previously that the condition required for the consistency of the LS-SVM is the absence of correlation between $x(k)$ and $v_{\mathrm{o}}(k)$. In the parametric context, an IV identification criterion has been introduced which relaxes this hypothesis to a less restrictive condition and prevents the deterioration of the estimation performance [26]. The idea is to introduce a *so-called* instrument signal $\zeta:\mathbb{Z}\to\mathbb{R}^{n_\theta}$ in order to obtain an unbiased estimate. The regularized IV estimate proposed here can be seen as the minimizer of the IV criterion:

$$\mathcal{W}(f,e)=\frac{1}{2}\|f\|_{\mathcal{H}}^2+\frac{\gamma}{2N^2}\left\|\sum_{k=1}^{N}\zeta(k)e(k)\right\|_2^2$$
$$=\frac{1}{2}\|\theta\|_{\ell_2}^2+\frac{\gamma}{2N^2}\|\Gamma^\top E\|_2^2, \tag{21}$$

based on the data set $\mathcal{D}_N$ and with Γ and E defined as

$$\Gamma=\left[\,\zeta(1)\;\ldots\;\zeta(N)\,\right]^\top, \tag{22a}$$
$$E=\left[\,e(1)\;\ldots\;e(N)\,\right]^\top. \tag{22b}$$

The motivations to pursue an IV-scheme based solution for bias reduction are the following:

- In general, recent IV approaches offer similar performance as the optimal (minimum variance and unbiased estimate) prediction error methods in case of correct assumptions on the system and noise models.
- As it will be shown later, the IV-based LS-SVM estimation problem has a similar solution to the LS-SVM estimator, implying approximately the same computational load as well as the same complexity.
- Most importantly, the IV-schemes can provide consistent estimates in case of incorrect noise assumptions.

To derive the solution of the IV-SVM, note that in (21), $\|\Gamma^\top E\|_2^2=E^\top\Gamma\Gamma^\top E$, which corresponds to a weighted ℓ_2 loss by the Grammian matrix $\Gamma\Gamma^\top$. Hence, in terms of the Theorem 1, the minimizer of (21) is

$$\hat{f}_{\mathrm{IV}}(.)=\sum_{i=1}^{N}\hat{\alpha}_iK_{x_i}(.), \tag{23}$$

with $\hat{\alpha}_{\mathrm{IV}}=[\,\hat{\alpha}_1\;\ldots\hat{\alpha}_N\,]\in\mathbb{R}^N$ being given by

$$\hat{\alpha}_{\mathrm{IV}}=\left(\frac{1}{N^2}\Gamma\Gamma^\top K_{xx}+\gamma^{-1}I_N\right)^{-1}\frac{1}{N^2}\Gamma\Gamma^\top Y. \tag{24}$$

An alternative way of obtaining this solution via the dualization of (21) and applying the KKT conditions is provided in [27].

The remaining question is how the instrument $\zeta(k)\in\mathbb{R}^{n_\theta}$ should be chosen efficiently based on data such that i) the noise induced bias can be eliminated and ii) we can still derive an RKHS estimator. Based on the interpretation of the LS-SVM in terms of $\{\phi_i(.)=K_{\omega_i}(.)\}_{i=1}^{n_{\mathrm{H}}=N}$ discussed in Section 2.2, in this paper, the following choice is proposed:

$$\zeta(k)=\left[\,\phi_1^\top\big(\xi(k)\big)\;\ldots\;\phi_{n_{\mathrm{H}}}^\top\big(\xi(k)\big)\,\right]^\top, \tag{25}$$

such that the Grammian $\Gamma\Gamma^\top$ can also be rewritten in terms of the same kernel function, *i.e.*, $\Gamma\Gamma^\top = K_{\xi\xi}$. The specific choice of ξ will be detailed in Section 4. This leads to the following solution of the IV-SVM estimator:

$$\hat{\alpha}_{\mathrm{IV}} = \underbrace{\left(\frac{1}{N^2}K_{\xi\xi}K_{xx} + \gamma^{-1}I_N\right)}_{Q(\gamma,N)}^{-1}\frac{1}{N^2}K_{\xi\xi}Y. \qquad (26)$$

It is important to note that even though (21) introduces a different sum of norms criterion than (6), it provides the same model structure as the LS-SVM: once α_{IV} is computed via (26), the estimate of the nonlinear function f_{o} is given by (23). This clearly shows that this proposed estimate is also an estimate of the model (14) using $\{\omega_i = x(i)\}_{i=1}^N$ as the node points. In other words, the consistency of α_{IV} can also be analyzed in the same way as for the LS-SVM estimate.

4 The choice of the instrument

In this section, the choice of the instrument ξ in (25), which defines $K_{\xi\xi}$, is discussed both from the bias and covariance point of view, showing how an efficient elimination of the noise bias is possible via the IV-SVM method.

4.1 Bias analysis of the IV-SVM estimator

Using the same problem setting as in Section 2.3 with $\omega = x$, the bias of the IV estimator is

$$\bar{\mathbb{E}}\left\{\hat{f}_{\mathrm{IV}}(\centerdot) - f_{\mathrm{o}}(\centerdot)\right\} = B_{\mathrm{N}}^{\mathrm{IV}} - B_{\mathrm{R}}^{\mathrm{IV}}, \qquad (27)$$

where

$$B_{\mathrm{N}}^{\mathrm{IV}} = \bar{\mathbb{E}}\Big\{\frac{\gamma}{N^2}K_{\centerdot\omega}Q^{-1}(\gamma,N)K_{\xi\xi}V_{\mathrm{o}}\Big\}, \qquad (28a)$$

$$B_{\mathrm{R}}^{\mathrm{IV}} = \bar{\mathbb{E}}\left\{\frac{1}{\gamma}K_{\centerdot\omega}Q^{-1}(\gamma,N)\alpha_{\mathrm{o}}\right\}, \qquad (28b)$$

and under the condition of quasi stationarity of (y,u,ξ), $Q_*^{\mathrm{K}}(\centerdot) := \lim_{N\to\infty} K_{\centerdot\omega}Q^{-1}(\gamma,N)$, for $\gamma > 0$, is assumed to exist.

Note that $B_{\mathrm{R}}^{\mathrm{IV}}$ can be again seen as the regularization bias and it can be made arbitrary small based on the user given choice of γ. Our focus is to eliminate $B_{\mathrm{N}}^{\mathrm{IV}}$ and hence achieve the ideal setting where choosing γ corresponds to a trade-off between bias and variance not as a balancing term between two sources of bias.

Analyzing (28a) shows that if the instrument ξ fulfills the following condition:

X1 $\quad \mathbb{E}\{\xi(k)v_{\mathrm{o}}(k)\} = 0, \quad \forall k \in \mathbb{Z},$

and if v_{o} is quasi stationary and zero mean, then $\bar{\mathbb{E}}\{\frac{\gamma}{N^2}K_{\centerdot\omega}Q^{-1}(\gamma,N)K_{\xi\xi}V_{\mathrm{o}}\} = Q_*^{\mathrm{K}}(\centerdot)\bar{\mathbb{E}}\{\frac{\gamma}{N^2}K_{\xi\xi}V_{\mathrm{o}}\}$ and

$$\bar{\mathbb{E}}\left\{\frac{\gamma}{N^2}K_{\xi\xi}V_{\mathrm{o}}\right\} = \bar{\mathbb{E}}\left\{\frac{\gamma}{N^2}K_{\xi\xi}\right\}\bar{\mathbb{E}}\left\{V_{\mathrm{o}}\right\} \equiv 0. \qquad (29)$$

Eq. (29) implies that choosing the instrument ξ such that condition X1 is satisfied leads to $B_{\mathrm{N}}^{\mathrm{IV}} \equiv 0$, which can be considered as an unbiased estimate with respect to the noise. Similar to the LTI case, projection of the data on a space orthogonal to the noise eliminates the noise induced bias. This idea grants a wide range of possible choices for achieving consistency by picking instruments uncorrelated with the noise. However, these choices can have various undesired effects on other moments of the function estimate. The next section discusses the choice of an instrument ξ which is also attractive from the covariance viewpoint.

4.2 Covariance analysis of the IV-SVM estimator

Analyzing the full expression of $\mathrm{cov}\{f_{\mathrm{IV}}(\centerdot) - f_{\mathrm{o}}(\centerdot)\}$ and getting meaningful conclusions on the choice of ξ are difficult tasks in case ω_k is correlated to $v_{\mathrm{o}}(k)$ due to the effect of regularization. Hence, consider the ideal case when ω_k and $v_{\mathrm{o}}(k)$ are independent, *i.e.*, what would be the best choice of the instrument which would render the covariance to the level of the LS-SVM estimate, when no-noise bias would be present. In this case, if $\gamma \to 0$, then, under the assumption that $K_{\xi\xi}K_{x\omega}$ is positive definite,

$$\mathrm{cov}\left\{\hat{\alpha}_{\mathrm{IV}} - \alpha_{\mathrm{o}}\right\} = \bar{\mathbb{E}}\big\{(K_{\xi\xi}K_{x\omega})^{-1}\big\}\cdot \bar{\mathbb{E}}\big\{(K_{\xi\xi}V_{\mathrm{o}}V_{\mathrm{o}}^\top K_{\xi\xi}\big\}\bar{\mathbb{E}}\big\{(K_{\xi\xi}K_{x\omega})^{-1}\big\}. \qquad (30)$$

In this much simpler situation, the results of [28] can be directly used such that the optimal instrument must be chosen that it satisfies

X2 $\quad \Gamma\Gamma^\top = K_{\xi\xi} = \bar{\mathbb{E}}\{K_{xx}\}.$

Even if the mathematical justification is only intuitive, condition X2 suggests that i) $\zeta(k)$ should be chosen as in (25) and ii) the choice of ξ is strongly related to the kernel K used and to the deterministic dynamics of the system. To approximately fulfill this condition under an unknown correlation structure of $x(k)$, $\xi(k)$ is chosen as

$$[\breve{y}(k-1)\ldots\breve{y}(k-n_{\mathrm{a}})\ u(k)\ldots\ u(k-n_{\mathrm{b}})]^\top \qquad (31)$$

where $\breve{y}$ is the simulated output of an estimated model of the system, *e.g.*, a model obtained via the LS-SVM approach. This choice of the instrument resembles to the widely used IV solution for linear regression [26,13].

To refine such a choice, an iterative IV-SVM scheme described by Algorithm 1 can be implemented in order to iteratively refine ξ. Note that due to the analytic solution of the IV-SVM estimator (24), computational complexity of Algorithm 1 in each step is the same as for the standard LS-SVM, *i.e.*, $\mathcal{O}(N^3)$ (main difference is due to model simulation which is $\mathcal{O}(N^2)$). Furthermore, in case the resulting estimate $\mathcal{M}^{(\tau-1)}$ is unstable, Step 6 needs to applied as a forward-backward nonlinear filtering to avoid divergence.

4.3 The choice of γ and the kernels

As it has been briefly explained in Section 2.2, the choice of the most appropriate kernel for the modeling problem at hand highly depends on the structure of the system to be identified and on the available data. However,

Algorithm 1 Refined IV-SVM

Require: model structure (4) in terms of model orders n_a and n_b with $n_\mathrm{g} := n_\mathrm{a} + n_\mathrm{b} + 1$, data set $\mathcal{D}_N = \{y(k), u(k)\}_{k=1}^N$, regularization parameter γ, kernel function K.

1: set $\tau \leftarrow 0$.
2: compute the matrix K_{xx} based on $\mathcal{D}_N$.
3: estimate $\alpha^{(0)} = \left(\frac{1}{N}K_{xx} + \gamma^{-1}I_N\right)^{-1}\frac{1}{N}Y$ via the LS-SVM resulting in the model estimate $\mathcal{M}^{(0)}$.
4: **repeat**
5: set $\tau \leftarrow \tau + 1$
6: use $\mathcal{M}^{(\tau-1)}$ to generate, by simulation, $\{\breve{y}^{(\tau)}(k)\}_{k=1}^N$.
7: calculate $\{\xi(k)\}_{k=1}^N$ via (31) using $\{\breve{y}^{(\tau)}(k), u(k)\}_{k=1}^N$ and compute $K_{\xi\xi}$.
8: estimate $\alpha^{(\tau)}$ via (24) resulting in the model estimate $\mathcal{M}^{\tau}$.
9: **until** $\alpha^{(\tau)}$ has converged.
10: **return** $\mathcal{M}^{(\tau)}$ with the estimate of the nonlinear function $\hat{f}$ obtained via (23).

these choices have an impact on the function class, *i.e.*, the RKHS $\mathcal{H}$ in which the expansion (23) is made rather than the actual decay rate of the expansion error. So it becomes a question, how the particular parameters of these kernel functions, like σ in (10), should be chosen to maximize the decay rate of the expansion w.r.t. the estimated unknown functional terms and hence the accuracy of the obtained model. Furthermore, the optimal choice of the regularization parameter γ is dependent on the choice of kernel functions, hence all such hyper-parameters must be simultaneously optimized.

If we restrict our attention to RBF kernels, a simple methodology can be used to optimize the kernel functions K and γ for the system to be estimated. The parameters σ and γ are tuned via cross-validation based optimization. For instance, the values of σ and γ providing the most accurate model w.r.t. an independent "validation" data set can be computed through a two-dimensional grid-search procedure over the space of hyper-parameters. Other numerically efficient techniques for the computation of the optimal hyper-parameters by means of genetic algorithms, particle swarm optimization and marginalization of the likelihood under a Bayesian setting are discussed in [29–31,5].

5 Simulation example

To demonstrate the results of this paper, the performance of the IV-SVM and LS-SVM approaches are compared using a Monte-Carlo study based on a simulation example.

5.1 The data-generating system

The noise-free data-generating system $\mathcal{S}_\mathrm{o}$ considered in this study is described by the difference equation

$$\breve{y}(k) \;=\; f_\mathrm{o}\big(\breve{y}(k-1), \breve{y}(k-2), u(k), u(k-1)\big), \quad (32)$$

where

$$f_\mathrm{o}\big(x_1, x_2, x_3, x_4\big) = f_1^\mathrm{o}\big(x_1\big) + f_2^\mathrm{o}\big(x_2\big) + g_0^\mathrm{o}\big(x_3\big) + g_1^\mathrm{o}\big(x_4\big),$$

specified by:

$$f_1^\mathrm{o}(x) \;=\; -0.7x, \qquad f_2^\mathrm{o}(x) \;=\; \tfrac{1}{8}x^2$$

$$g_0^\mathrm{o}(x) = \begin{cases} 0.5 & \text{if} \quad x \geq 0.5, \\ x & \text{if} \quad -0.5 < x < 0.5, \\ -0.5 & \text{if} \quad x \leq -0.5. \end{cases} \qquad g_1^\mathrm{o}(x) \;=\; -0.4x.$$

In order to study the effect of the noise on the non-parametric identification of this system, four different, realistic noise scenarios, corresponding to four data generating systems, are considered:

$\mathcal{S}_1$**:** $y(k) = f_\mathrm{o}\big(y(k-1), y(k-2), u(k), u(k-1)\big) + e_\mathrm{o}(k)$, which corresponds to a NARX structure.
$\mathcal{S}_2$**:** $y(k) = \breve{y}(k) + e_\mathrm{o}(k)$, corresponding to a NOE structure.
$\mathcal{S}_3$**:** $y(k) = f_\mathrm{o}\big(y(k-1), y(k-2), u(k), u(k-1)\big) + v_\mathrm{o}(k)$, corresponding an NARMA structure.
$\mathcal{S}_4$**:** $y(k) = \breve{y}(k) + v_\mathrm{o}(k)$ which is a NBJ structure.

In $\mathcal{S}_1$ and $\mathcal{S}_2$, e_o is a white noise with $e_\mathrm{o}(k) \sim \mathcal{N}(0, \sigma_\mathrm{e}^2)$. In $\mathcal{S}_3$ and $\mathcal{S}_4$, $v_\mathrm{o}(k)$ is a zero-mean colored noise generated by filtering $e_\mathrm{o}(k) \sim \mathcal{N}(0, \sigma_\mathrm{e}^2)$ as

$$v_\mathrm{o}(k) = a_1 v_\mathrm{o}(k) + b_0 e_\mathrm{o}(k) + b_1 e_\mathrm{o}(k-1), \quad (33)$$

where $a_1 = 0.95$, $b_0 = 1.5$ and $b_1 = -0.3$. To generate data sets $\mathcal{D}_N = \{u(k), y(k)\}_{k=1}^N$, the system is excited with a $N = 1000$ long white input sequence with uniform distribution $\mathcal{U}(-1, 1)$ starting with zero initial conditions. In order to provide representative results, 100 of such data sets are generated with independent realizations of the noise, setting up a Monte Carlo study with $N_\mathrm{MC} = 100$ runs. The variance σ_e^2 has been calculated separately for each structure in order to provide a *signal-to-noise ratio*, $\mathrm{SNR} = 10\log(\sum_{k=1}^N \breve{y}^2(k) / \sum_{k=1}^N v_\mathrm{o}^2(k))$, equal to 11dB.

5.2 The model structure

In all estimation scenarios using both the the LS-SVM and IV-SVM, the model structure is assumed to be:

$$y(k) = f(x(k)) + e(k), \quad (34a)$$

$$x(k) = [\, y(k-1)\ y(k-2)\ u(k)\ u(k-1)\,]^\top, \quad (34b)$$

where $e(k)$ is the residual error. A four dimensional RBF kernel (10) is used to characterize this nonlinear function. The hyper-parameters are optimized using cross-validation by maximizing the *best fit rate* (BFR) on the noisy validation dataset. The BFR is defined as:

$$\mathrm{BFR} = \max\left\{1 - \frac{\|y(k) - \hat{y}(k)\|_{\ell_2}}{\|y(k) - \overline{y}\|_{\ell_2}}, 0\right\} \cdot 100\%, \quad (35)$$

Table 1
Hyper-parameter values

		$\mathcal{S}_1$		$\mathcal{S}_2$		$\mathcal{S}_3$		$\mathcal{S}_4$	
σ_{LS}	$\frac{\gamma_{\mathrm{LS}}}{N}$	5.6	800	4.4	800	3.8	600	4.4	600
σ_{IV}	$\frac{\gamma_{\mathrm{IV}}}{N^2}$	5.3	800	4.9	800	4.1	800	4.9	800

Table 2
BFR values computed on noise-free validation data sets (mean±std [%])

	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$
LS-SVM	88±1.3	86.5±1.2	84.5±2	77±2.5
IV-SVM	87±2.5	88±1.8	89.5±2.7	86.5±2.6

with $\overline{y}$ denoting the mean value of the output sequence $y(k)$, while $\hat{y}(k)$ is the simulated output of the estimated model. Based on a coarse grid search, the optimized hyper-parameter values are displayed in Table 1.

5.3 Estimation results

Usually, quality of non-parametric model identification is solely assessed by comparing the achieved fit w.r.t. validation data using the same excitation conditions. Therefore, bias with respect to a true nonlinear function is never really assessed. In order to cope with usual assessment scores, the achieved mean and standard deviation of the BFR of both methods on a noise-free validation data set is displayed in Table 2 which is in accordance with the statistical properties of the estimates. In other words, it can be seen in Table 2 that, similarly to the linear regression case, the variance of the IV method is always higher than for LS-based methods. Nevertheless, it can also be seen that the mean of the BFR is much less affected by the noise structure in the IV-SVM case than in the LS-SVM case. This emphasizes the unbiased properties of the IV-based methods. Finally, these results also show that, in the realistic case where the noise generating system is unknown, the IV-SVM offers better prediction capabilities than the LS-SVM estimate.

To visualize the function estimates on the assumed 4-dimensional domain, a set of grid points is used, displayed in Figure 1, which covers a subset of $\mathbb{R}^4$ excited during identification. The mean value and standard deviation of the estimated nonlinear functions computed using the LS-SVM and IV-SVM estimates are displayed on these grid points and compared to the true nonlinear function f_{o} in Figure 2. It is interesting to notice that the LS-SVM is strongly affected by the structure of the noise (the SNR is constant in all examples) while the unbiased IV-SVM estimate produces a nonlinear function centered on the true one.

6 Conclusions

In this paper, an instrumental variable based formulation of the LS-SVM approach has been introduced in order to cope with the limitations of the assumed noise

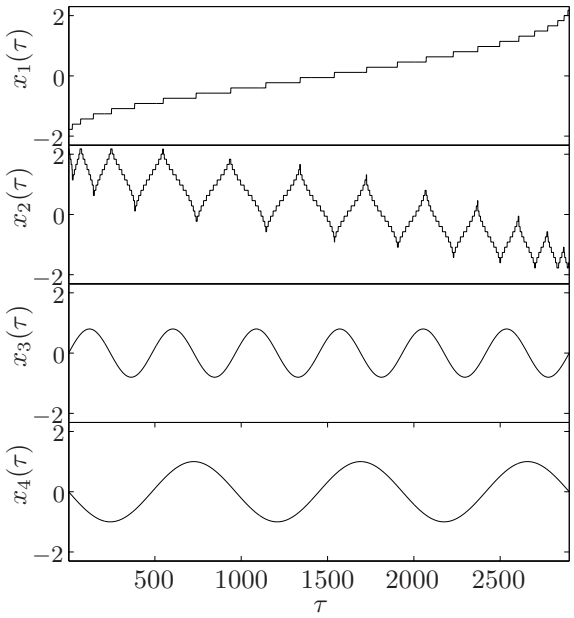

Fig. 1. Grid points for comparing the estimates.

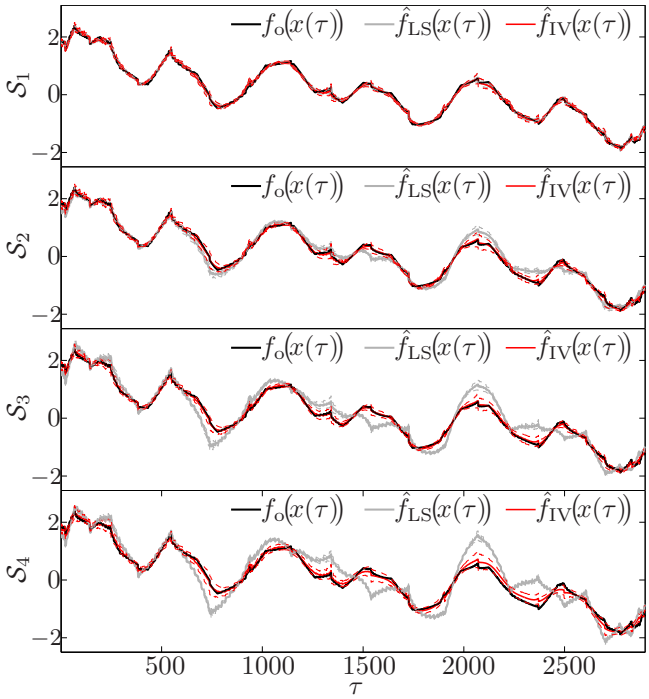

Fig. 2. The true nonlinear function and the estimated functions (mean: solid line, std: dashed line) displayed at the grid points given in Figure 1 over the Monte Carlo study.

structure, but, at the same time, preserving its attractive computational properties. It has been shown that the proposed IV-SVM scheme allows the elimination of the noise induced bias in case the noise process is additive and zero-mean. Hence, the proposed scheme considerably widens the applicability of LS-SVM (and the re-

lated GP and Kriging) based methods. A suitable choice for the required instrument has been discussed and an iterative instrument refining scheme, inspired by the LTI IV methods, has been proposed. The performance of the resulting IV-SVM algorithm with respect to the regular LS-SVM method has been demonstrated in a Monte Carlo study. Generalization of the approach for non-zero mean and/or nonlinearly distorted noise is technically more demanding and remains the objective of future research in the LS-SVM context.

References

[1] V. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.

[2] B. Schölkopf and A. Smola, *Learning with kernels*. Cambridge MA: MIT Press, 2002.

[3] N. Cristianini and J. Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press, 2000.

[4] L. Ljung, "Perspectives on system identification," *Annual Reviews in Control*, vol. 34, pp. 1–12, 2010.

[5] G. Pillonetto, M. H. Quang, and A. Chiuso, "A new kernel-based approach for nonlinearsystem identification," *IEEE Trans. on Automatic Control*, vol. 56, no. 12, pp. 2825–2840, 2011.

[6] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, vol. 50, no. 3, pp. 657–682, 2014.

[7] T. Falck, K. Pelckmans, J. Suykens, and B. De Moor, "Identification of Wiener-Hammerstein systems using LS-SVMs," in *15th IFAC symposium on System Identification*, Saint Malo, France, July 2009.

[8] I. Goethals, K. Pelckmans, J. Suykens, and B. De Moor, "Identification of MIMO Hammerstein models using least squares support vector machines," *Automatica*, vol. 41, no. 7, pp. 1263–1272, 2005.

[9] D. G. Krige, "A study of gold and uranium distribution patterns in the klerksdorp gold field," *Geoexploration*, vol. 4, no. 1, pp. 43 – 53, 1966.

[10] R. Frigola and C. E. Rasmussen, "Integrated pre-processing for bayesian nonlinear system identification with Gaussian processes," in *Proc. of the 52nd IEEE Conference on Decision and Control*, Florence, Italy, Dec. 2013.

[11] J. Kocijan, A. Girard, B. Banko, and R. Murray-Smith, "Dynamic systems identification with gaussian processes," *Mathematical and Computer Modelling of Dynamical Systems*, vol. 11, no. 4, pp. 411–424, 2005.

[12] T. Van Gestel, J. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. Vandewalle, "Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis," *Neural Computation*, vol. 14, no. 5, pp. 1115–1147, 2002.

[13] L. Ljung, *System Identification, theory for the user*, 2nd ed. Prentice-Hall, 1999.

[14] J. Suykens and J. Vandewalle, "Recurrent least squares support vector machines," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 47, no. 7, pp. 1109–1114, 2000.

[15] T. Falck, J. Suykens, and B. De Moor, "Linear parametric noise models for least squares support vector machines," in *Proc. of the 49th IEEE Conf. on Decision and Control*, Atlanta, USA, Dec. 2010, pp. 6389–6394.

[16] G. Pillonetto, A. Chiuso, and G. De Nicolao, "Prediction error identification of linear systems: a nonparametric Gaussian regression approach," *Automatica*, vol. 47, no. 2, pp. 291–305, 2011.

[17] P. C. Young, *Recursive estimation and time-series analysis. An introduction to the student and the practionner*. Berlin: Springer-Verlag, 2011.

[18] A. Lundgren and J. Sjöberg, "Nonlinear instrument variable methods based on local linear models," in *Proc. of the 6th IFAC Symposium on Nonlinear Control Systems*, 2004, pp. 501–506.

[19] V. Laurain, W. Zheng, and R. Tóth, "Introducing instrumental variables in the LS-SVM based identification framework," in *Proc. of the 50th IEEE Conf. on Decision and Control*, Orlando, Florida, USA, Dec. 2011, pp. 3198–3203.

[20] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P. Glorennec, H. Hjalmarsson, and A. Juditsky, "Nonlinear black-box modeling in system identification: a unified overview," *Automatica*, vol. 31, no. 12, pp. 1691–1724, 1995.

[21] E.-W. Bai, "An optimal two-stage identification algorithm for Hammerstein-Wiener nonlinear systems," *Automatica*, vol. 34, no. 3, pp. 333–338, 1998.

[22] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 39, no. 1, pp. 1–49, 2001.

[23] N. Aronszajn, "Theroy of reproducing kernels," *Transactions of the American Mathematical Society*, no. 68, pp. 337–404, 1950.

[24] G. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions," *Journal of Mathematical Analysis and Applications*, pp. 82–95, 1971.

[25] T. Howley and M. G. Madden, "The genetic kernel support vector machine: Description and evaluation," *Artificial Intelligence Review*, vol. 24, no. 3-4, pp. 379–395, 2005.

[26] T. Söderström and P. Stoica, *Instrumental Variable Methods for System Identification*. Springer-Verlag, New York, 1983.

[27] V. Laurain, R. Tóth, D. Piga, and W. X. Zheng, "Instrumental varaibles based least squares support vector machine for identification of nonlinear systems," Eindhoven University of Technology, Tech. Rep. TUE-CS-2013-005, 2013.

[28] P. Stoica and T. Söderström, "Optimal instrumental variable estimation and approximate implementations," *IEEE Trans. on Automatic Control*, vol. 28, no. 7, pp. 757 – 772, 1983.

[29] M. W. Chang and C. J. Lin, "Leave-one-out bounds for support vector regression model selection," *Neural Computation*, vol. 17, no. 5, pp. 1188–1222, 2005.

[30] S. An, W. Liu, and S. Venkatesh, "Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression," *Pattern Recognition*, vol. 40, no. 8, pp. 2154–2162, 2007.

[31] X. C. Guo, J. H. Yang, C. G. Wu, C. Y. Wang, and Y. C. Liang, "A novel LS-SVMs hyper-parameter selection based on particle swarm optimization," *Neurocomputing*, vol. 71, no. 16, pp. 3211–3215, 2008.