Introducing instrumental variables in the LS-SVM based identification framework

Vincent Laurain, Wei Xing Zheng, Roland Toth

Introducing Instrumental Variables in the LS-SVM Based Identification Framework

Vincent Laurain, Wei Xing Zheng and Roland Tóth

***Abstract*— Least-Squares Support Vector Machines (LS-SVM) represent a promising approach to identify nonlinear systems via nonparametric estimation of the nonlinearities in a computationally and stochastically attractive way. All the methods dedicated to the solution of this problem rely on the minimization of a squared-error criterion. In the identification literature, an instrumental variable based optimization criterion was introduced in order to cope with estimation bias in case of a noise modeling error. This principle has never been used in the LS-SVM context so far. Consequently, an instrumental variable scheme is introduced into the LS-SVM regression structure, which not only preserves the computationally attractive feature of the original approach, but also provides unbiased estimates under general noise model structures. The effectiveness of the proposed scheme is demonstrated by a representative example.**

I. Introduction

Support Vector Machines (SVMs) have been originally developed as a class of *supervised learning* methods aiming at 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 approach to *system identification*, especially in the nonlinear context. In nonlinear system identification, most of the research interest has been dedicated to *nonlinear block models* using various *Least Square-SVM* (LS-SVM) approaches [4]–[6]. In general, LS-SVMs are particular variations of the original support vector machine approach using an ℓ_2 loss function. Their main advantage is the uniqueness of the solution, which is obtained by solving a set of linear equations.

Given the convexity of the estimation problem and the large number of parameters typically involved in LS-SVMs, these approaches can be regarded as so-called *over-parametrization approaches* in the nonlinear framework [7], [8]. However, due to the existence of powerful regularization methods for SVMs [1], [2], the variance of the estimated nonlinear functions is significantly lower than in the classical over-parametrization methods. On the other hand, SVMs also offer the possibility of incorporating a model structure and prior knowledge on the nonlinearities unlike other nonparametric methods (*e.g.*, [9]).

Variants of linear regression based methods in identification have been developed in order to cope with realistic assumptions on the noise [10]–[12]. To introduce the same generality of noise structures, some recurrent LS-SVM have been developed in [13], while in [14], a particular linear parametric noise model has been introduced in the LS-SVM framework. However, the chosen noise model plays an important role in the consistency of the estimates. In the parametric identification framework, the strength of IV methods is to deliver consistent estimates independently on the chosen noise model assumption in a computationally attractive way. Consequently, the use of an instrumental variable based criterion in the LS-SVM framework can lead to a performance improvement of the current LS-SVM approaches. Nonetheless, such a method would require the dual solution of the IV optimization problem [10], [15], which has not been developed so far. To overcome this gap, this paper aims to derive a dual solution to the regularized IV optimization problem and to introduce the use of the *Instrumental Variable* (IV) scheme into the LS-SVM regression structure. This contribution not only preserves the computationally attractive feature of the original approach, but also provides unbiased estimates for general noise model structures/conditions.

The rest of the paper is organized as follows: after defining the problem setting considered in Section II, both the primal and the dual solution of the usual optimization problem used in LS-SVM methods are presented in Section III. In Section IV, the IV optimization problem is introduced both in the primal form and in the newly introduced dual form. In Section V, the use of the dual IV solution to the LS-SVM framework is developed, resulting in an IV-LS-SVM method. The statistical performance of the proposed IV-LS-SVM method is compared in Section VI to the traditional LS-SVM approach via a Monte Carlo study of the identification of a nonlinear system with an *Output Error* (OE) noise structure. Finally, conclusions and some future directions of the research are given in Section VII.

II. Problem description

Consider the general description of an affine, *Single-Input Single-Output* (SISO), nonlinear, discrete-time and *AutoRegressive with eXogeneous input* (ARX) system $\mathcal{S}_{\rm o}$ given by

$$y(k) = \sum_{i=1}^{n_{\rm a}} f_i^{\rm o}(y(k-i)) + \sum_{j=0}^{n_{\rm b}} g_j^{\rm o}(u(k-j)) + e_{\rm o}(k), \quad (1)$$

This work was supported by a research grant of the Australian Research Council and by the Netherlands Organization for Scientific Research (Grant No. 680-50-0927).

V. Laurain and W. X. Zheng are with the School of Computing and Mathematics, University of Western Sydney, Penrith NSW 2751, Australia. vlaurain@scm.uws.edu.au; w.zheng@uws.edu.au

R. Tóth is with Delft Center for Systems and Control, Delft University of Technology, Mekelweg 2, 2628 CD, Delft, The Netherlands. r.toth@tudelft.nl

where u and y are the input and output signals respectively, k denotes the discrete time, $f_i^{\mathrm{o}}, g_j^{\mathrm{o}} : \mathbb{R} \to \mathbb{R}$ are a set of possibly nonlinear functions and $e_{\mathrm{o}}(k)$ is a zero-mean white noise sequence with $e_{\mathrm{o}}(k) \in \mathcal{N}(0, \sigma^2_{e_{\mathrm{o}}})$. Note that representation (1) is general enough to describe usual block structures such as *Hammerstein* or *Wiener* systems. Formulation of (1) in the *Multi-Input Multi-Output* (MIMO) case is also available as shown in [5]. It is important to note that the considered system class is more restrictive than the nonlinear NARX class presented in [16]. This simplification is used to present the underlying idea behind this contribution in a clear fashion.

The nonlinearities involved in (1) are supposed to be *a priori* unknown. In the LS-SVM context, the assumption is made that each nonlinearity f_i can be modeled using an n_{H} dimensional feature map $\phi_i : \mathbb{R} \to \mathbb{R}^{n_{\mathrm{H}}}$ (where n_{H} is potentially infinite). A feature map in this setting represents nonlinear mappings from the extended input-output space to the output space (feature space). Nevertheless, before properly addressing the LS-SVM problem and in order to clearly develop the motivations for the proposed approach, it is is assumed that each nonlinearity has an explicit description:

$$f_i(y(k-i)) = \sum_{j=0}^{n_{\mathrm{H}}} \rho_{i,j} \phi_{i,j}(y(k-i)). \tag{2}$$

$$g_j(u(k-i)) = \sum_{l=0}^{n_{\mathrm{H}}} \rho_{\tilde{j},l} \phi_{\tilde{j},l}(u(k-i)). \tag{3}$$

with $\tilde{j} = n_a + 1 + j$. This assumption leads to the parametrized model $\mathcal{M}_\rho$

$$y(k) = \varphi(k)^\top \rho + e(k), \tag{4}$$

where e(k) is the equation error and the regressor φ is defined as

$$\begin{aligned}\varphi(k) = \big[\ & \phi_1^\top(y(k-1)) \quad \dots \quad \phi_{n_{\mathrm{a}}}^\top(y(k-n_a)) \\ & \phi_{n_{\mathrm{a}}+1}^\top(u(k)) \quad \dots \quad \phi_{n_{\mathrm{a}}+n_{\mathrm{b}}+1}^\top(u(k-n_{\mathrm{b}}))\ \big]^\top\end{aligned} \tag{5}$$

with $\phi_i : \mathbb{R} \to \mathbb{R}^{n_{\mathrm{H}}}$ being n_{H} dimensional basis functions, $\rho = [\ \rho_1^\top \quad \dots \quad \rho_{n_{\mathrm{a}}+n_{\mathrm{b}}+1}^\top\ ]^\top \in \mathbb{R}^{n_\rho}$ is the parameter vector, $\rho_i \in \mathbb{R}^{n_{\mathrm{H}}}$ and $n_\rho = (n_{\mathrm{a}} + n_{\mathrm{b}} + 1) n_{\mathrm{H}}$.

Let $\mathcal{M} = \{\mathcal{M}_\rho \mid \rho \in \mathbb{R}^{n_\rho}\}$ be the collection of all models in the form of (4). $\mathcal{M}$ represents the set of models in which we are searching for the "best" $\mathcal{M}_\rho$ that describes $\mathcal{S}_{\mathrm{o}}$ given a data set $\mathcal{D}_N = \{y(k), u(k)\}_{k=1}^N$ generated by $\mathcal{S}_{\mathrm{o}}$.

In the considered problem setting it is assumed that the system belongs to the model set defined and therefore there exists a $\rho_{\mathrm{o}} \in \mathbb{R}^{n_\rho}$ such that

$$y(k) = \varphi(k)^\top \rho_{\mathrm{o}} + e_{\mathrm{o}}(k). \tag{6}$$

III. Optimization criterion

The quality of the model fit is formulated in terms of a cost function $\mathcal{J}(\rho, e)$, where e is given by (4). Minimization of $\mathcal{J}(\rho, e)$ corresponds to the estimation of the parameter vector ρ. In the LS-SVM framework, the used minimization criterion is the LS error criterion on e. However, the dimension n_{H} of the regressor ϕ involved is usually large (and potentially infinite). Hence, a regularization term on ρ is applied, leading to the minimization of the cost function

$$\mathcal{J}(\rho, e) = \frac{1}{2}\rho^\top \rho + \frac{\gamma}{2}\sum_{k=1}^N e^2(k) = \frac{1}{2}\|\rho\|_{\ell_2}^2 + \frac{\gamma}{2}\|e(k)\|_{\ell_2}^2, \tag{7}$$

where the scalar $\gamma \in \mathbb{R}_0^+$ is the *regularization parameter*. Note that (7) is a so-called *sum-of-norms* criterion as it contains both the equation error term $e(k)$ and a regularization term: the ℓ_2 cost of ρ scaled by γ.

The solution of this optimization problem both in the primal and dual forms are presented in the next subsections.

A. Solution in primal form

The primal solution to minimize the criterion (7) is obtained by simply deriving the analytical solution of

$$\frac{\partial \mathcal{J}(\rho, e)}{\partial \rho} = 0. \tag{8}$$

This leads to the minimum at:

$$\hat{\rho}_{\mathrm{P}} = \left[\gamma^{-1} I_{n_\rho} + \sum_{k=1}^N \varphi(k)\varphi(k)^\top\right]^{-1} \cdot \left[\sum_{k=1}^N \varphi(k) y(k)\right]. \tag{9}$$

It can be further noticed that by using the notation

$$Y = [\ y(1) \quad \dots \quad y(N)\ ]^\top \in \mathbb{R}^N, \tag{10a}$$

$$\Phi = [\ \varphi(1) \quad \dots \quad \varphi(N)\ ]^\top \in \mathbb{R}^{N \times n_\rho}, \tag{10b}$$

the primal solution can be written as:

$$\hat{\rho}_{\mathrm{P}} = \underbrace{\left[\Phi^\top \Phi + \gamma^{-1} I_{n_\rho}\right]^{-1}}_{R_{\mathrm{P}}(\gamma, N)} \Phi^\top Y. \tag{11}$$

B. Solution in the dual form

The optimization problem (7) w.r.t. the constraints (4) can also be solved by constructing the *Lagrangian*:

$$\mathcal{L}(\rho, e, \alpha) = \mathcal{J}(\rho, e) - \sum_{k=1}^N \alpha_k \left(\varphi(k)^\top \rho + e(k) - y(k)\right) \tag{12}$$

with $\alpha_k \in \mathbb{R}$ being the Lagrangian multipliers. The global optimum is obtained when

$$\frac{\partial \mathcal{L}}{\partial e} = 0 \quad \to \quad \alpha_k = \gamma e(k), \tag{13a}$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_k} = 0 \quad \to \quad y(k) = \rho^\top \varphi(k) + e(k), \tag{13b}$$

$$\frac{\partial \mathcal{L}}{\partial \rho} = 0 \quad \to \quad \rho = \sum_{k=1}^N \alpha_k \varphi(k). \tag{13c}$$

Substituting (13a) and (13c) into (13b) leads to

$$y(k) = \varphi(k)^\top \underbrace{\left(\sum_{k=1}^N \alpha_k \varphi(k)\right)}_{\rho} + \underbrace{\gamma^{-1}\alpha_k}_{e(k)} \tag{14}$$

for $k \in \{1, \dots, N\}$. This set of equations is equivalent to:

$$Y = \left[\Phi\Phi^\top + \gamma^{-1} I_N\right] \alpha, \tag{15}$$

where $\alpha = [\alpha_1 \ \dots \ \alpha_N]^\top \in \mathbb{R}^N$. This linear problem admits the solution:

$$\alpha = \left[\Phi\Phi^\top + \gamma^{-1} I_N\right]^{-1} Y. \tag{16}$$

According to (13c), $\rho = \Phi^\top \alpha$ and therefore

$$\hat{\rho}_{\mathrm{D}} = \Phi^\top \underbrace{\left[\Phi\Phi^\top + \gamma^{-1} I_N\right]}_{R_{\mathrm{D}}(\gamma,N)}^{-1} Y. \tag{17}$$

C. Equivalence and bias of the solutions

It is important to notice that, under the condition that both $R_{\mathrm{D}}(\gamma, N) \in \mathbb{R}^{N\times N}$ in (17) and $R_{\mathrm{P}}(\gamma, N) \in \mathbb{R}^{n_\rho \times n_\rho}$ in (11) are non-singular, then the dual and primal solutions are equivalent. Assuming that both $R_{\mathrm{D}}(\gamma, N)$ and $R_{\mathrm{P}}(\gamma, N)$ are non-singular, then it can be proven using the well-known properties of the primal solution that the estimate is consistent ($\mathbb{E}\{\rho\} = \rho_{\mathrm{o}}$) under the conditions:

C1 $\gamma \to \infty$.
C2 $\mathbb{E}\{\varphi(k)e_{\mathrm{o}}(k)\} = 0, \quad \forall k \in \mathbb{Z}$.

This implies that both C1 and C2 must also hold for the dual estimate to be consistent. For the system class considered, C2 only holds if e_{o} is white as $\varphi(k)$ is constructed using past signals values of y and the input signal u which is uncorrelated to the noise. Nonetheless the ARX structure as described in Section II is unrealistic in most practical applications as it implies that the noise on the output has the same dynamics and nonlinearities as the system itself. Consequently, in most practical applications, the minimization of criterion (7) will lead to a biased estimate. The next section introduces an IV method in order to cope with this issue.

IV. Instrumental Variable Approach

Among the available identification approaches used in the regression framework, the *Instrumental Variable* (IV) approach has been successfully applied to resolve in a simple and highly efficient fashion the inconsistency problem of LS regression under a noise-modeling [10], [11], [15], [17]. The most restrictive condition guaranteeing consistency is condition C2. In most problems, including the LS-SVM case, the regressor is correlated (implicitly or explicitly) to the noise and C2 does not hold. Thus, in the parametric context, a IV identification criterion has been introduced which relaxes C2 to a less restrictive condition and prevents the deterioration of the estimation performance [15]. The idea is to introduce a *so-called* instrument $\zeta(k) \in n_\rho$ such that the consistency conditions become:

X1 $\gamma \to \infty$.
X2 $\mathbb{E}\{\zeta(k)e_{\mathrm{o}}(k)\} = 0, \quad \forall k \in \mathbb{Z}$.

While the condition C2 depends on $\varphi(k)$ and therefore on the model assumed, X2 depends on $\zeta(k)$ which can be chosen by the user. There is a wide range of possible solutions to pick an instrument uncorrelated to the noise. To respect the consistency conditions, the IV estimate corresponds to the solution of the criterion

$$\hat{\rho}^{\mathrm{IV}} = \mathrm{sol}\left\{ \frac{1}{N}\sum_{k=1}^{N} \rho + \gamma\zeta(k)\left[y(k) - \varphi^\top \rho\right] = 0 \right\}. \tag{18}$$

Similarly to (7), a regularization term on ρ weighted by γ is also involved in this estimation scheme.

The motivation to pursue an IV-scheme based solution for bias elimination are the following:

- In general, the recent IV approaches offer a similar performance as the optimal (minimum variance and unbiased estimates) 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 problem can be solved in a very similar way to the LS-SVM problem, implying approximately the same computational load as well as the same complexity.
- Most importantly, the IV-schemes provide consistent estimates in case of incorrect noise assumptions. This feature is really important in practical situations as usually no physical models of the noise are available.

Nonetheless, while the IV methods are now widely used under the primal form of the optimization problem, they have never been introduced in a dual setting to the best of the authors' knowledge. Thus, the question arises: Can the parallelism between the primal and dual solutions, explored in Section II, be used to introduce an IV scheme for the dual form without any performance degradation?

A. IV in the primal form

The primal solution of (18) is straightforwardly given as

$$\hat{\rho}_{\mathrm{P}}^{\mathrm{IV}} = \left[\gamma^{-1} I_{n_\rho} + \sum_{k=1}^{N} \zeta(k)\varphi(k)^\top\right]^{-1} \cdot \left[\sum_{k=1}^{N} \zeta(k)y(k)\right]. \tag{19}$$

By using the notation (10b) and by declaring

$$Z = \left[\ \zeta(1) \quad \dots \quad \zeta(N)\ \right]^\top \in \mathbb{R}^{N\times n_\rho}, \tag{20}$$

the primal IV estimate can be expressed as:

$$\hat{\rho}_{\mathrm{P}}^{\mathrm{IV}} = \underbrace{\left[Z^\top\Phi + \gamma^{-1} I_{n_\rho}\right]}_{R_{\mathrm{P}}^{\mathrm{IV}}(\gamma,N)}^{-1} Z^\top Y. \tag{21}$$

Many instruments can be chosen in order to fulfill X2. Nonetheless, the existence of the estimate is now constrained by the non-singularity of $R_{\mathrm{P}}^{\mathrm{IV}}(\gamma, N)$ in (21). The discussion about the choice of a suitable instrument guaranteeing this property is too technical. Hence, due to the space restriction, the authors refer to [15] for a discussion about this issue.

B. IV in the dual form

The main contribution of this paper is to introduce the solution of the instrumental variable optimization (21) in the dual form. Introduce α_k and $\zeta(k)$ satisfying:

$$\alpha_k = \gamma e(k), \tag{22a}$$
$$y(k) = \varphi(k)^\top \rho + e(k), \tag{22b}$$
$$\rho = \sum_{k=1}^{N} \alpha_k \zeta(k). \tag{22c}$$

We will prove that the choice of (22c) is necessary to obtain the dual solution of the optimization criterion (18). Substituting (22a) and (22c) into (22b) yields the following set of linear equations:

$$y(k) = \varphi(k)^\top \underbrace{\left(\sum_{k=1}^{N} \alpha_k \zeta(k)\right)}_{\rho} + \underbrace{\gamma^{-1}\alpha_k}_{e(k)}, \tag{23}$$

for $k \in \{1, \ldots, N\}$, which leads to the solution

$$\alpha = \left[\Phi Z^\top + \gamma^{-1} I_N\right]^{-1} Y, \tag{24}$$

where $\alpha = [\alpha_1 \ \ldots \ \alpha_N]^\top \in \mathbb{R}^N$. According to (22c), $\rho = Z^\top \alpha$ and therefore

$$\hat{\rho}_{\mathrm{a}} = Z^\top \underbrace{\left[\Phi Z^\top + \gamma^{-1} I_N\right]}_{R_{\mathrm{D}}^{\mathrm{IV}}(\gamma, N)}^{-1} Y, \tag{25}$$

which is equivalent to $\hat{\rho}_{\mathrm{P}}^{\mathrm{IV}}$ (see (21)) if both $R_{\mathrm{D}}^{\mathrm{IV}}(\gamma, N)$ and $R_{\mathrm{P}}^{\mathrm{IV}}(\gamma, N)$ are non-singular. Consequently, $\hat{\rho}_{\mathrm{a}}$ is the dual solution of the IV optimization problem (18), $\hat{\rho}_{\mathrm{a}} = \hat{\rho}_{\mathrm{D}}^{\mathrm{IV}}$ and this estimate is consistent, independently of the noise model assumed under the conditions X1 and X2. In conclusion the IV optimization solution has been introduced in the dual representation and the next section describes its application to the LS-SVM framework.

V. Instrumental variable in the LS-SVM context

So far in this paper, the studied system was considered to lie in the model set defined by $\mathcal{M}$ and could be described using a finite dimensional parameter vector. It allowed to derive the statistical properties for both the primal and dual solutions of different optimization criteria (LS based and IV based). Nonetheless, in a nonlinear context, finding an appropriate model set can be a tedious task. In most linear regression methods, explicit feature maps are defined (for example polynomial) along with their dimension. Nonetheless, this implies the quality of the model will highly depend on the structure chosen and in most cases, will lead to a structural bias. A possible way to avoid this structural bias is to increase the dimension of the feature maps : $n_{\mathrm{H}} \to \infty$ and therefore $n_\rho \to \infty$. In this case, $n_\rho \gg N$, and the use of the dual solution becomes necessary. It must be pointed out that defining explicitly an infinite dimensional feature map and therefore an infinite dimensional regressor is not feasible in practice. Hence, the main advantage of the LS-SVM method is to be able to handle infinite dimensional feature maps with a low computational load via a dual solution.

A. LS-SVM method

In the LS-SVM context, $\varphi(k)$ is composed of possibly infinite dimensional feature maps $n_{\mathrm{H}} \to \infty$: therefore $n_\rho \to \infty$ and ρ cannot be explicitly computed. The main feature of the LS-SVM method is that the vector α can be explicitly computed without the proper knowledge of the feature maps Φ. Introduce the so-called *Grammian matrix* as $G = \Phi\Phi^\top$ in (16), which can be defined without the explicit knowledge of Φ. Notice that

$$[G]_{j,k} = \sum_{i=1}^{n_\rho} [G^i]_{j,k} \tag{26}$$

with

$$[G^i]_{j,k} = \langle \phi_i(x_i(j)), \phi_i(x_i(k)) \rangle = K^i(x_i(j), x_i(k)), \tag{27}$$

where K^i is a positive definite *kernel function* and

$$x_i(k) = y(k-i), \quad i = 1, \ldots, n_{\mathrm{a}}, \tag{28a}$$

$$x_{n_{\mathrm{a}}+1+j}(k) = u(k-j), \quad j = 0, \ldots, n_{\mathrm{b}}. \tag{28b}$$

Consequently, given a set of kernel functions K^i defines G and hence characterizes Φ. This is called the *kernel trick* [1], [2], which allows the identification of the nonlinear functions f_i, g_j without explicitly defining the feature maps involved. A typical type of kernel is, for example, the *Radial Basis Function* (RBF) kernel:

$$K^i_{j,k} = K^i(x_i(j), x_i(k)) = \exp\left(\frac{-\|x_i(j) - x_i(k)\|^2_{\ell_2}}{\sigma_i^2}\right), \tag{29}$$

but other kernels, like *polynomial* kernels, can also be used. Another remark is that the parameter vector $\hat{\rho}_{\mathrm{D}}$ is never accessible in the LS-SVM framework, and only the combined estimation ${\rho_i}^\top \phi_i(\centerdot) = f_i(\centerdot)$ is computable using the kernel functions defined. Nonetheless, even if the estimate of ρ is not accessible, the consistency properties C1 and C2 hold.

B. Instrumental variable for the LS-SVM framework

The final aim of this paper is to introduce the IV solution in the the LS-SVM framework. The conditions on the instrument in order to obtain a consistent estimate have been derived in the previous section. It must be emphasized that in a nonlinear context, the choice of an optimal instrument depends highly on the system structure and the noise model assumed, and is mostly an open problem. Consequently, the instrument chosen to address the IV-LS-SVM solution is inspired by the instrument proposed in [10] which leads to the IV4 solution in the primal form:

$$\begin{aligned}\zeta(k) = \big[\ & \phi_1^\top(y_{\mathrm{LS}}(k-1)) \quad \ldots \quad \phi_{n_{\mathrm{a}}}^\top y_{\mathrm{LS}}(k-n_{\mathrm{a}}) \\ & \phi_{n_{\mathrm{a}}+1}^\top(u(k)) \quad \ldots \quad \phi_{n_{\mathrm{a}}+n_{\mathrm{b}}+1}^\top(u(k-n_{\mathrm{b}})) \ \big]^\top,\end{aligned} \tag{30}$$

where y_{LS} is the simulated output of the model given by the LS-SVM method and ϕ_i are the same as in (5). This instrument always guarantees X2 in the considered case and it has been successfully used in the primal context.

In the same fashion as in (26), the IV *Grammian matrix* $J = \Phi Z^\top$ is defined as

$$[J]_{j,k} = \sum_{i=1}^{n_\rho} [J^i]_{j,k} \tag{31}$$

with

$$[J^i]_{j,k} = \langle \phi_i(x_i(j)), \phi_i(\xi_i(k)) \rangle = K^i(x_i(j), \xi_i(k)), \tag{32}$$

$$\xi_i(k) = y_{\mathrm{LS}}(k-i), \quad i = 1, \ldots, n_{\mathrm{a}}, \tag{33a}$$

$$\xi_{n_{\mathrm{a}}+1+j}(k) = u(k-j), \quad j = 0, \ldots, n_{\mathrm{b}}. \tag{33b}$$

It is possible to derive the conditions on the instrument for applying the kernel trick. Nonetheless, this issue is not discussed here due to space restrictions. The definition of the kernel functions K^i allows an explicit expression of α. Consequently, it can be concluded from (22c) that the resulting IV4-LS-SVM estimate is given by

$$f_i(\centerdot) = \phi_i^\top(\centerdot)\rho_i = \sum_{k=1}^{N} \alpha_k K^i(\xi_i(k), \centerdot), \tag{34a}$$

$$g_j(\centerdot) = \phi_{\tilde{j}}^\top(\centerdot)\rho_{\tilde{j}} = \sum_{k=1}^{N} \alpha_k K^{\tilde{j}}(\xi_{\tilde{j}}(k), \centerdot), \tag{34b}$$

where $\tilde{j} = n_{\mathrm{a}} + 1 + j$. The IV4-LS-SVM algorithm w.r.t. the instrument (30) is summarized as Algorithm 1.

Algorithm 1 IV4-LS-SVM

1: use the LS-SVM method to obtain a model $\mathcal{M}_{\mathrm{LS-SVM}}$
2: use $\mathcal{M}_{\mathrm{LS-SVM}}$ to generate y_{LS} by simulation
3: compute ξ and J via (33a-b) and (32)
4: compute α by solving (24)

VI. Simulation Example

A. Data generating system

The main advantage of the IV methods is their robustness when facing modeling errors in the noise structure. Consequently, in order to compare the LS-SVM and the IV4-LS-SVM methods under realistic noise conditions, a nonlinear *Output Error* (OE) system $\mathcal{S}_{\mathrm{o}}$ is considered :

$$\chi(k) = -0.7\chi(k-1) + f(\chi(k-2)) + g(u(k)), \quad (35a)$$

$$y(k) = \chi(k) + e_{\mathrm{o}}(k), \quad (35b)$$

where

$$f(x) = \begin{cases} -0.2x^2 & \text{if} \quad x > 0 \\ 0 & \text{else} \end{cases} \quad (35c)$$

$$g(x) = -x + 2x^2 + \cos(10x). \quad (35d)$$

In the sequel, the input $u(k)$ is taken as a zero-mean white noise process with a uniform distribution $\mathcal{U}(-0.5, 0.5)$ and with length $N = 1200$ to generate data sets $\mathcal{D}_N$ of $\mathcal{S}_{\mathrm{o}}$. $e_{\mathrm{o}}(k)$ is taken as a zero-mean white noise sequence with $e_{\mathrm{o}}(k) \in \mathcal{N}(0, \sigma^2_{e_{\mathrm{o}}})$.

B. Model structures

Both the conventional LS-SVM approach and the proposed IV4-LS-SVM approach use the same ARX model structure $\mathcal{M}_\rho$ given as:

$$y(k) = \rho_1 y(k-1) + \phi_2^\top(y(k-2))\rho_2 + \phi_3^\top(u(k))\rho_3 + e(k).$$

Note that the equation error $e(k)$ is not white. The robustness of the proposed IV4-LS-SVM and the existing LS-SVM algorithms are analyzed under a *signal-to-noise ratio* $\mathrm{SNR} = 10\log\frac{P_{\chi_{\mathrm{o}}}}{P_{e_{\mathrm{o}}}} = 7\mathrm{dB}$, where $P_{\chi_{\mathrm{o}}}$ and $P_{e_{\mathrm{o}}}$ are the average power of the signals χ_{o} and e_{o} respectively. To provide representative results, a Monte Carlo simulation of $N_{\mathrm{MC}} = 100$ runs with new noise realization in each run is applied.

One of the advantage of the LS-SVM algorithm is to be able to use some *a priori* knowledge which, in this case, means the explicit definition of $\phi^1(y(k-1)) = y(k-1)$. To characterize the nonlinearities, RBF kernels are used for K^2 and K^3. It is important to note that the main contribution of this paper is the introduction of the IV optimization criterion (18) and its solution in the LS-SVM framework (24). Therefore, in order to evaluate the impact of this criterion only, it is important that the model structure is the same for both the LS-SVM method and the IV4-LS-SVM methods. In the present context, where the feature maps are implicitly defined, the model structure is defined by the kernels used and therefore by the σ parameters. The model structure is chosen such that it maximizes the *Best Fit Rate* (BFR) on the estimation data set for the LS-SVM method (using an exhaustive search) where:

TABLE I

Mean and standard deviation of the estimated parameter ρ_1 and the BFR computed on validation data.

	Mean $\hat{\rho}_1$	std $\hat{\rho}_1$	Mean BFR	std BFR
True value	-0.7	–	–	–
LS-SVM	-0.528	0.0142	82.61	1.29
IV4-LS-SVM	-0.699	0.0202	91.99	1.86

$$\mathrm{BFR} = 100\% \cdot \max\left(1 - \frac{\|\chi(k) - \hat{\chi}(k)\|_{\ell_2}}{\|\chi(k) - \bar{\chi}\|_{\ell_2}}, 0\right), \quad (36)$$

with $\bar{\chi}$ being the mean of χ. This search has resulted in $\sigma_2 = 3$, $\sigma_3 = 0.5$. The γ parameters have been however optimized separately (by exhaustive search too) as they are directly linked to the different optimization problems considered. This leads to $\gamma_{\mathrm{LS}} = 3500$ and $\gamma_{\mathrm{IV}} = 500$.

C. Simulation results

Table I displays the mean and standard deviation of the estimated parameter ρ_1. It can be seen that, in line with the theory, the LS-SVM algorithm is biased while the proposed IV4-LS-SVM method is unbiased. Like in the linear regression framework, the IV based method displays a slightly larger variance than the LS method. Note that ϕ_1 in this case is explicitly defined so ρ_1 can be directly accessed, while this is not the case for the other parameters ρ_2 and ρ_3.

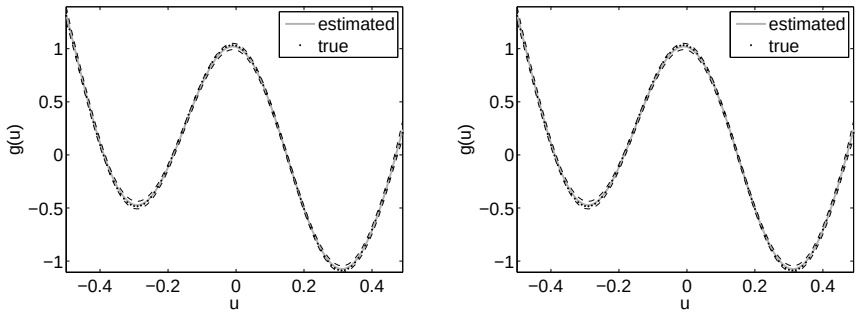
Figure 1 shows the estimation results of $g(u)$ by the IV4-LS-SVM and the LS-SVM algorithms and exposes the mean estimated function together with the standard deviation interval. As expected, both algorithms perform similarly in estimating g as $u(k)$ is uncorrelated with $e(k)$ and therefore $\varphi_2(k) = \zeta_2(k)$. Figure 2 shows the estimation results of $f(y)$ by the IV4-LS-SVM and the LS-SVM algorithms in terms of the mean and standard deviation of the estimates. The bias of the LS-SVM method algorithm clearly appears in this figure. In contrast, the mean estimate of f by the IV4-LS-SVM algorithm is centered on the original one. Note that a more advanced instrument might lead to even better results.

Table I also displays the mean and standard deviation of the BFR for both algorithms on a validation set. This clearly shows that on the validation set, the proposed IV4-LS-SVM method achieves, even for this simple model, significantly better performance than the usual LS-SVM algorithm. As the computation time of $\hat{\chi}$ is negligible, this implies that the execution time of the IV4-LS-SVM is only approximately two times of the LS-SVM method, where the latter is known to be computationally efficient.

Finally, it needs to be pointed out that w.r.t. (35a-b), condition X2 holds only if $\chi(k-2) < 0$. Even though, the achieved estimation performance by the proposed approach has considerably increased on the whole feature space (even for $\chi(k-2) \geq 0$). This highlights that condition X2 cannot be asserted for any nonlinear structures, but it holds in general for structures which are linear in the output (Hammerstein, Linear parameter varying, etc.).

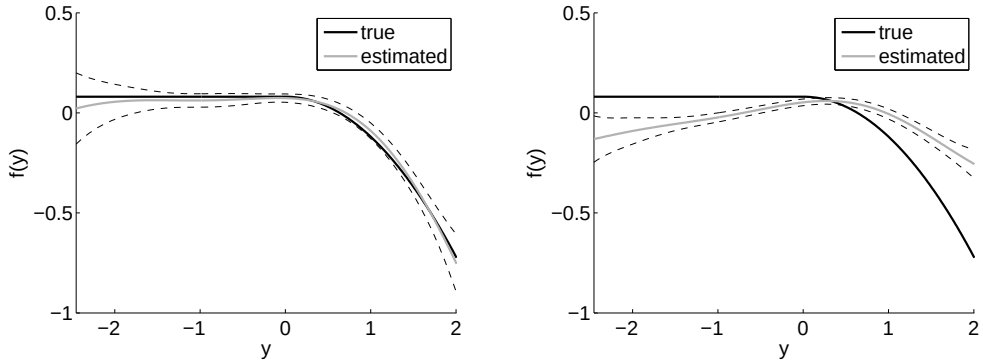
VII. Conclusion

In this paper, an instrumental variable estimation scheme has been proposed for the SVM framework, which signif-



(a) IV4-LS-SVM *(b)* LS-SVM

Fig. 1. True nonlinearity $g(u)$ (dotted black) together with the mean estimate (solid grey) $+/-$ standard deviation (dashed black) over the Monte-Carlo simulation.



(a) IV4-LS-SVM *(b)* LS-SVM

Fig. 2. True nonlinearity $f(y)$ (solid black) together with the mean estimate (solid grey) $+/-$ standard deviation (dashed black) over the Monte-Carlo simulation.

icantly extends the applicability of the LS-SVM algorithm to general noise cases while maintaining its computational efficiency. To the authors' knowledge, this method is among the first of the LS-SVM approaches designed to be consistent under modeling error of the noise. Via a simulation example, it has been demonstrated that the proposed IV4-LS-SVM method performs better than the LS-SVM algorithm w.r.t. data generated by a non-ARX system. It has also been observed that the computational load of the IV4-LS-SVM scheme is at the same magnitude as the LS-SVM method. Future research concerns the introduction of optimal instruments for different system classes and the refinement of the proposed IV4-LS-SVM scheme for extended noise models directly which is hoped to decrease variance of the estimates.

REFERENCES

[1] V. Vapnik, *Statistical Learning Theory*. New York: 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] 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.

[5] 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.

[6] F. Giri and E.-W. Bai, *Lecture Notes in Control and Information Sciences*, ser. Block-oriented Nonlinear System Identification. Berlin: Springer-Germany, 2010.

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

[8] F. H. I. Chang and R. Luus, "A noniterative method for identification using the Hammerstein model," *IEEE Trans. Automatic Control*, vol. 16, no. 5, pp. 464–468, 1971.

[9] W. Greblicki and M. Pawlak, *Non-Parametric System Identification*. Cambridge, UK: Cambridge University Press, 2008.

[10] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[11] H. Garnier and L. Wang, editors, *Identification of Continuous-time Models from Sampled Data*. Berlin: Springer-Verlag, 2008.

[12] V. Laurain, M. Gilson, R. Tóth, and H. Garnier, "Refined instrumental variable methods for identification of LPV Box-Jenkins models," *Automatica*, vol. 46, no. 6, pp. 959–967, 2010.

[13] J. Suykens and J. Vandewalle, "Recurrent least squares support vector machines," *IEEE Trans. Circuits Syst. I Fundam. Theory Appl.*, vol. 47, no. 7, pp. 1109–1114, 2000.

[14] 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.

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

[16] 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.

[17] P. C. Young, *Recursive Estimation and Time-Series Analysis*. Berlin: Springer-Verlag, 1984.