Sparse RKHS Estimation via Globaly Convex Optimization and its Application in LPV-IO Identification ⋆

V. Laurain [a,b], R.Tóth [c], D. Piga [d], M.A.H. Darwish [e],

[a] *Université de Lorraine, CRAN, UMR 7039, 2 rue Jean Lamour, F-54519, Vandoeuvre-lès-Nancy, 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] *IDSIA Dalle Molle Institute for Artificial Intelligence, SUPSI-USI, Via Cantonale 2C, 6928 Manno, Switzerland.*

[e] *Electrical Engineering Department, Faculty of Engineering, Assiut University, 71515 Assiut, Egypt.*

Abstract

Function estimation using the Reproducing Kernel Hilbert Space (RKHS) framework is a powerful tool for identification of a general class of nonlinear dynamical systems without requiring much *a priori* information on model orders and nonlinearities involved. However, the high degrees-of-freedom (DOFs) of RKHS estimators has its price, as in case of large scale function estimation problems, they often require a serious amount of data samples to explore the search space adequately for providing high-performance model estimates. In cases where nonlinear dynamic relations can be expressed as a sum of functions, the literature proposes solutions to this issue by enforcing sparsity for adequate restriction of the DOFs of the estimator, resulting in parsimonious model estimates. Unfortunately, all existing solutions are based on greedy approaches, leading to optimization schemes which cannot guarantee convergence to the global optimum. In this paper, we propose an an ℓ_1-regularized non-parametric RKHS estimator which is the solution of a quadratic optimization problem. Effectiveness of the scheme is demonstrated on the non-parametric identification problem of LPV-IO models where the method solves simultaneously (i) the model order selection problem (in terms of number of input-output lags and input delay in the model structure) and (ii) determining the unknown functional dependency of the model coefficients on the scheduling variable directly from data. The paper also provides an extensive simulation study to illustrate effectiveness of the proposed scheme.

Key words: Reproducing kernel Hilbert spaces; elastic net; support vector machines; Gaussian processes; non-parametric estimation; Linear parameter-varying systems; model order selection.

1 Introduction

For decades, estimation of parsimonious models of functional relations between signal variables using data has been a central problem in many scientific fields like statistics, systems and control engineering, but also in newly developing fields like machine learning which has its roots in *statistical learning theory*. While the latter has mainly focused on the estimation of static functional relations with possibly a large-number of candidate signals, in systems and control engineering, the focus has been on the estimation, or so called *system identification*, of possibly nonlinear dynamic relationships between signals. In both cases, selection of a suitable model structure to capture the unknown functional relations is characterized by two **main challenges**:

(i) determining which variables contribute to the relationship (e.g., in identification, this corresponds to the selection of the "suitable" dynamic order, input delay and noise structure);
(ii) determining/parametrizing a class of functional relations s.t. they have the least possible complexity for adequately representing the signal relations.

Optimal choice in these questions is rarely known *a priori* (especially for (ii)). A possible solution leads through the parametrization of functional dependencies in terms of an extensive set of basis functions such that the resulting model structure is capable of explaining a rich set of possible relations, and let the data decide which substructure is appropriate to use. This can be achieved by the use of classical model structure selection tools from statistics like *Akaike's information criterion* (AIC), etc.,

⋆ This paper was not presented at any IFAC meeting. Corresponding author V. Laurain.

Email addresses: vincent.laurain@univ-lorraine.fr (V. Laurain), r.toth@tue.nl (R.Tóth), dario.piga@supsi.ch (D. Piga), mohamed.darwish@eng.au.edu.eg (M.A.H. Darwish).

or ℓ_1 regularization based sparse estimators and shrinkage methods: *non-negative garrote* (NNG) [1], *least absolute shrinkage and selection operator* (LASSO) [2] and SPARSEVA [3]. Although, these methods are capable of achieving model structure selection in terms of (i) and (ii), their efficiency strongly depends on adequate *a priori* selection of the basis functions which is left to rest on the shoulders of the user.

Alternatively, selection problem (ii) can be equivalently seen as an unknown function "reconstruction" problem based on measured data for which an important set of the methods have emerged from the theory of *reproducing kernel Hilbert space* (RKHS) estimators [4] in various forms, such as *least-squares support vector machines* (LS-SVM) [5], *Gaussian processes* (GP) [6] and *Kriging* [7]. The core idea is that instead of using *a priori* set of basis to parametrize the to-be-estimated functional relation, a *kernel function* is introduced that acts as a basis generator driven by the observed data. In this sense, the function class is defined only *a priori* for which the basis are restricted to a subspace that can be distinguished based upon the given measurements. These methods have been successfully applied in system identification both in the linear and nonlinear cases [8–10].

To jointly address sub-problems (i) and (ii), in the machine learning community, various concepts of RKHS estimators have been developed where sparsity is mostly enforced using ℓ_1 norm regularization and by assuming that the nonlinear function relation at hand can be decomposed as a *"sum of nonlinear functions"* [11–13]. The results are mostly theoretical and are derived under restrictive statistical assumptions (whiteness, joint independence of signals). Hence, for estimation problems encountered in practice, such as in system identification, the resulting optimization problem has only been solved *via* various relaxations (e.g. greedy method [14]) without any guarantees on the convergence of the resulting approach. Hence in this paper, the following *contributions* are provided

(1) Proposing an ℓ_1-regularized sparse estimator based on the RKHS framework that corresponds to a directly solvable global quadratic optimization problem with *linear matrix inequality* (LMI) constraints;
(2) Extension of the *representer theorem* for the proposed estimator, allowing joint selection of functional terms and their reconstruction from data, without prior parametrization (data-driven model structure selection with (i) and (ii) in one step).

The proposed approach is demonstrated on the identification problem of *linear parameter-varying* (LPV) systems (see [15]), which corresponds to a generalization of the *"sum of nonlinear functions"* problem. It is shown that the proposed method allows joint reconstruction of the scheduling-variable dependencies and the model coefficient structure from data, extending the capabilities of previous non-parametric estimators developed for this model class [10, 16, 17].

The paper is organized as follows. In Section 2, fundamental concepts of the RKHS theory needed to derive the results of the paper are provided. Then, a novel sparsity enforcement ℓ_1 regularization term for RKHS estimation is introduced in Section 3. In Section 4, the considered LPV identification setting is introduced. This is followed by detailing how the LPV sparse estimation problem is solved from the RKHS point of view in Section 5. In Section 6, the proposed method is compared to existing solutions via a simulation study. Finally, the conclusions are presented in Section 7.

Notation

$\mathbb{R}$ and $\mathbb{Z}$ are the sets of the real and integer numbers, respectively, while $\mathbb{N}$ is the set of all positive integers. $\|x\|_1$, $\|x\|_2$ and $\|x\|_\infty$ represent the ℓ_1, ℓ_2 and ℓ_∞ norms of a possibly infinite dimensional vector x. I_n is the n-dimensional identity matrix and δ_{ij} denotes the Kronecker delta. $\mathbb{I}_{n_1}^{n_2} = \{n_1, n_1+1, \ldots, n_2\} \subset \mathbb{Z}$ is an index set. $\mathcal{N}$ indicates a Gaussian distribution, while $\mathcal{U}$ denotes a uniform distribution.

2 RKHS theory

To introduce the preliminaries for the main results of the paper, a brief introduction to nonlinear function estimation via the RKHS theory is provided in this section.

2.1 Data-generating system

In standard regression problems, a set of observations, $\{(z_k, w_k)\}_{t=1}^N$ is assumed to be available, generated by

$$z_k = f(w_k) + \epsilon_k, \tag{1}$$

where $w_k \in \mathscr{W}$ is the input sequence, $z_k \in \mathbb{R}$ is the output, $f : \mathscr{W} \to \mathbb{R}$ is an unknown nonlinear function and $\epsilon_k \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right)$ is an independent Gaussian additive noise with σ_ϵ^2 being its variance. Our goal is to provide an estimate of f that describes the observed data, but also, for an arbitrary new pair (w, z), the predicted value of $f(w)$ is close to z in the *mean squared error* (MSE) sense.

2.2 Kernel functions and RKHSs

Recall the following definitions:

Definition 1 (Positive definite kernel, [5]) *Let $\mathscr{W}$ be a metric space. A real-valued function $K : \mathscr{W} \times \mathscr{W} \to \mathbb{R}$ is called a positive definite kernel if it is continuous, symmetric and $\sum_{k=1,l=1}^{n,n} a_k a_l K(w_k, w_l) \geq 0$ for any finite set of points $\{w_1, \ldots, w_n\} \subset \mathscr{W}$ and $\{a_1, \ldots, a_n\} \subset \mathbb{R}$.*

Definition 2 (Reproducing kernel) *Let $\mathscr{H}$ be a Hilbert space of real-valued functions on $\mathscr{W}$ with inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}}$. A positive definite kernel function $K : \mathscr{W} \times \mathscr{W} \to \mathbb{R}$ is a reproducing kernel for $\mathscr{H}$ if*

(1) $\forall w \in \mathscr{W}, K_w(\cdot) = K(w, \cdot) \in \mathscr{H}$, where K_w is the so-called kernel section centered at w;
(2) The reproducing property holds, meaning that

$$f(w) = \langle f(\cdot), K(w, \cdot) \rangle_{\mathscr{H}}, \ \forall w \in \mathscr{W}, \forall f \in \mathscr{H}.$$

A Hilbert space of real-valued functions which possesses a reproducing kernel is called an RKHS [18]. Due to the Moore-Aronszajn theorem [19], for every positive definite kernel K, there is a unique RKHS $\mathscr{H}$ with K as its reproducing kernel and vice versa. In the sequel. we denote that RKHS as $\mathscr{H}_K$ and its inner product as $\langle \cdot, \cdot \rangle_K$ with the associated norm $\| \cdot \|_K$. Additionally, due to the symmetric and reproducing property: $K(w, w') = K(w', w) = \langle K(w, \cdot), K(w', \cdot) \rangle_{\mathscr{H}} = \langle K(w', \cdot), K(w, \cdot) \rangle_{\mathscr{H}}$. Due to identification setting and without any loss of generality, we restrict the scope to countable metric spaces $\mathscr{W}$.

Definition 3 (RKHS) *Let K be a positive definite kernel function and $\mathscr{H}_K$ is the associated RKHS. Then,*

$$\mathscr{H}_K = \Bigg\{ f : \mathscr{W} \to \mathbb{R} \mid f(\cdot) = \sum_{k=1}^{\infty} a_k K_{w_k}(\cdot), w_k \in \mathscr{W}, \\ a_k \in \mathbb{R},\ \|f\|_{\mathscr{H}} < +\infty \Bigg\}, \quad (2)$$

where $\|f\|_K = \sqrt{\langle f, f \rangle_K}$ is the norm in $\mathscr{H}_K$ induced by the inner product $\langle f, f' \rangle_K = \sum_{k=1,l=1}^{\infty,\infty} a_k b_l K(w_k, w_l)$, for $f = \sum_{k=1}^{\infty} a_k K_{w_k}$ and $f' = \sum_{l=1}^{\infty} b_l K_{w_l}$.

Definition 3 implies that all $f \in \mathscr{H}_K$ inherit their properties from the kernel, e.g., the continuity of K implies the continuity of all $f \in \mathscr{H}_K$ [4]. Hence, the main advantage of RKHS-based estimators is that expected properties, e.g., smoothness, integrability, etc., can be easily encoded in $\mathscr{H}_K$ via the associated kernel function K.

2.3 Regularization in RKHSs

The main idea for solving the estimation of f in (1) based on $\mathcal{D}_N = \{(z_k, w_k)\}_{k=1}^N$ is to have a real-valued loss function $\mathcal{V}$ that consists of two terms, i.e., a *"data-fit"* term denoted by $\mathcal{C}$ and a *"regularizer"* term denoted by $\mathcal{R}$ forming

$$\mathcal{V}(f) = \mathcal{C}\big(\{w_k, z_k, f(w_k)\}_{k=1}^N\big) + \gamma \mathcal{R}\big(\|f\|_K\big), \quad (3)$$

where $\gamma > 0$ is a parameter which defines the trade-off between these contradicting terms. Common choices are

$$\mathcal{C}\big(\{w_k, z_k, f(w_k)\}_{k=1}^N\big) = \sum_{k=1}^{N} \big(z_k - f(w_k)\big)^2, \quad (4a)$$

$$\mathcal{R}(\|f\|_K) = \|f\|_K^2, \quad (4b)$$

i.e., a quadratic loss function for the "data-fit" term (similar to the *prediction error minimization* (PEM) setting commonly considered in LTI and LPV system identification) and the squared norm of $\mathscr{H}_K$ used for the "regularizer". Once both $\mathcal{C}$ and $\mathcal{R}$ are chosen, the unknown function f is estimated by solving

$$\hat{f} = \underset{f \in \mathscr{H}_K}{\operatorname{argmin}}\ \mathcal{V}(f). \quad (5)$$

The RKHS framework allows to obtain a closed-form and unique solution of (5), even if the employed RKHS is an infinite-dimensional space:

Theorem 1 (Generalized representer, [20–23]) *For a given RKHS $\mathscr{H}_K$ with reproducing kernel $K : \mathscr{W} \times \mathscr{W} \to \mathbb{R}$, the minimizer of (5) for any positive $\mathcal{C}$ and any strictly monotonically increasing $\mathcal{R}$ on $[0, \infty)$ can be represented as*

$$\hat{f}(\cdot) = \sum_{k=1}^{N} c_k K(w_k, \cdot), \quad \{c_k\}_{k=1}^N \subset \mathbb{R}. \quad (6)$$

Theorem 1 indicates that using criterion (3), the estimate of f can be expressed as a finite sum of kernel slices/sections centered on the available observations. In case $\mathcal{C}$ and $\mathcal{R}$ are chosen as in (4), the parameters $c = [c_1 \ \cdots \ c_N]^\top \in \mathbb{R}^N$ defining the estimated function $\hat{f}$ in (6) (minimizer of $\mathcal{V}(f)$) can be computed as

$$c = (\mathcal{K} + \gamma I_N)^{-1} Z_N, \quad (7)$$

where the (k, l)-th entry of $\mathcal{K} \in \mathbb{R}^{N \times N}$ is $K(w_k, w_l)$.

3 Sparsity in RKHS

In the specific case where $f(w_k)$ can be written as a sum of nonlinear functions

$$f(w_k) = \sum_{i=1}^{n_g} f_i(w_{k,i}), \quad (8)$$

where $w_{k,i}$ is the i-th component of w_k, it is well-known that the Kernel function K can be also expressed as a sum of kernels $K_i : \mathscr{W} \times \mathscr{W} \to \mathbb{R}$ (see section 8.1.1): $K(w_k, \cdot) = \sum_{i=1}^{n_g} K_i(w_{k,i}, \cdot)$. In theses cases, an important issue is to be able to enforce sparsity in order to keep the number of nonzero functions $f_i(\cdot)$ as low as possible in order to be able to identify the most parsimonious structure given the data.

To this end, the literature proposes an additive regularization term that complements (3). More specifically, the new cost function consists of three terms [11–13]:

(1) $\mathcal{C}$: "data-fit" term $\sum_{k=1}^{N} \big(z_k - f(w_k)\big)^2$ that quantifies the fit (empirical loss) w.r.t. the measured data;
(2) $\mathcal{R}$: "regularizer" term $\|f\|_K^2$ to prevent overfitting;
(3) $\mathcal{S}$: "sparsity" term aiming at shrinking the functions f_i to zero to minimize the number of non-zero coefficient functions f_i characterizing the chosen model structure.

The "sparsity" term proposed in the literature is

$$\big\| \big[\ \|f_1\|_2 \ \cdots \ \|f_{n_g}\|_2 \ \big] \big\|_1 \quad (9)$$

where $\| \cdot \|_1$ is a convex approximation of the ℓ_0-pseudo norm. [1] Unfortunately, the resulting optimization problem must be solved via some relaxation [14] which cannot guarantee the convergence of the resulting estima-

[1] The ℓ_0-pseudo norm of a vector x equals to the number of nonzero elements of x. Minimization under an ℓ_0 objective is a non-convex NP-hard problem.

tion approach. In order to tackle this problem, we propose, to use as a sparsity enforcing term,

$$\mathcal{S}(g) = \left\| \begin{bmatrix} \|f_1\|_\infty & \cdots & \|f_{n_\mathrm{g}}\|_\infty \end{bmatrix} \right\|_1 \tag{10}$$

where $\|f_i\|_\infty$ is the ℓ_∞-norm (maximum absolute value) of the function f_i over $\mathscr{W}$. It will be shown that using this term leads to a quadratic optimization problem by simply approximating the incomputable $\|f_i\|_\infty$ by

$$S_i = \max_{j \in \mathbb{I}_1^M} |f_i(m_j)|,$$

where $\mathbb{M} = \{m_j\}_{j=1}^M \subset \mathscr{W}$ are a set of node points. The set $\mathbb{M}$ can be chosen by gridding of $\mathscr{W}$ or by random selection using a prior distribution. Nevertheless, the difference between the infinity norm and the proposed approximation scheme is expected to be small if the kernels K_i enforce a sufficient smoothness on $f_i(w_i)$. As a result of such an approximation, the "sparsity" term can be expressed as $\left\| \begin{bmatrix} S_1 & \cdots & S_{n_\mathrm{g}} \end{bmatrix} \right\|_1$. Thus, the estimation of the model $f(w_k)$ in (1) can be formulated as

$$\begin{aligned} \min_f \quad & \sum_{k=1}^N \left(z_k - g(w_k)\right)^2 + \gamma_\mathrm{s} \left\| \begin{bmatrix} S_1 \cdots S_{n_\mathrm{g}} \end{bmatrix} \right\|_1 + \gamma \|f\|_K^2 \\ \text{s.t.} \quad & S_i = \max_{j \in \mathbb{I}_1^M} |f_i(m_j)|, \end{aligned} \tag{11}$$

where $\gamma_\mathrm{s} > 0$ is the hyper-parameter scaling the influence of the sparsity term. As a contribution of the paper, next we show that the solution of (11) admits a representer and is the solution of a quadratic optimization scheme, which hence guarantees the convergence of the resutling estimation approach

Theorem 2 (Representer under sparsity) *Let $\mathscr{H}_K$ be an RKHS over an n_g-dimensional $\mathscr{W}$ and with reproducing kernel K such that $K(w, w') = \sum_{i=1}^{n_\mathrm{g}} K_i(w_i, w_i')$. Then, the minimizer of* (11) *can be expressed as a representer in the form*

$$\hat{f}(\cdot) = \sum_{k=1}^N c_k K_{w_k}(\cdot) + \sum_{i=1}^{n_\mathrm{g}} \left(\sum_{j=1}^M \overline{c}_{i,j} K_{m_j}(\cdot) \right). \tag{12}$$

with $\{c_k\}_{k=1}^N \subset \mathbb{R}$ and $\{\overline{c}_{i,j}\}_{i=1,j=1}^{n_\mathrm{g},M} \subset \mathbb{R}$.

PROOF. Our goal is to express (11) in the form of (3) suited for applying Theorem 1. Introduce

$$\mathcal{C}_{\gamma_\mathrm{s}}(*) = \sum_{k=1}^N \left(z_k - f(w_k)\right)^2 + \gamma_\mathrm{s} \left\| \begin{bmatrix} S_1 \cdots S_{n_\mathrm{g}} \end{bmatrix} \right\|_1 \tag{13}$$

which depends both on the observation points $\{w_k\}_{k=1}$ and the grid points $\{m_j\}_{j=1}^M$. Append observations in $\mathscr{W}$ as $\{w_k\}_{k=1}^N \cup \mathbb{M}$. Hence, the cost function $\mathcal{V}(f)$ in (11) is now the sum of $\mathcal{C}_{\gamma_\mathrm{s}}$ and $\mathcal{R}$. As $\mathcal{C}_{\gamma_\mathrm{s}}$ is a positive function, (11) is expressed as an optimization criterion in the form of (3) with observation points $\{w_k\}_{k=1}^N \cup \mathbb{M}$, which allows direct application of Theorem 1, implying (12). ■

Note that in (12), the coefficients $\overline{c}_{i,j}$ appear due to the added sparsity constraint. Problem (11) can be reformulated as a convex quadratic *optimization problem* (QP) that can be efficiently solved by standard solvers. Due to space restrictions, the required steps are only given when this sparse estimator concept is applied in LPV system identification, which is treated in the next section.

4 The LPV identification problem

In order to illustrate the benefits of the sparse estimator concept introduced in Section 3, we consider the problem of joint model order selection and non-parametric identification of LPV-IO models, which corresponds to a sparse sum-of-nonlinear-functions problem.

4.1 Data-generating system

The simplest form of discrete-time LPV systems considered in identification is the *autoregressive with exogenous input* (ARX) structure, which is defined, in the *single-input single-output* (SISO) case, as

$$y_k = \sum_{i=1}^{n_\mathrm{a}^\mathrm{o}} a_i^\mathrm{o}(p_k) y_{k-i} + \sum_{j=d^\mathrm{o}}^{n_\mathrm{b}^\mathrm{o}} b_j^\mathrm{o}(p_k) u_{k-j} + e_k^\mathrm{o}, \tag{14}$$

where $u_k \in \mathbb{R}$, $y_k \in \mathbb{R}$, $p_k \in \mathbb{P}$ are values of the input, output and scheduling signals at discrete time instant $k \in \mathbb{Z}$, respectively, $\mathbb{P} \subseteq \mathbb{R}^{n_\mathrm{p}}$ is a compact set, $d^\mathrm{o} \geq 0$ is the input delay, $n_\mathrm{a}^\mathrm{o}, n_\mathrm{b}^\mathrm{o} \geq 0$, while $e_k^\mathrm{o} \in \mathbb{R}$ is a zero-mean white noise. The coefficients a_i^o and b_j^o are static [2] functions of the measurable scheduling signal p. By introducing $n_\mathrm{g}^\mathrm{o} = n_\mathrm{a}^\mathrm{o} + n_\mathrm{b}^\mathrm{o} - d^\mathrm{o} + 1$ and $x_{k,i}^\mathrm{o}$, where the latter is the i-th component of $x_k^\mathrm{o} = [y_{k-1} \ \cdots \ y_{k-n_\mathrm{a}^\mathrm{o}} \ u_{k-d^\mathrm{o}} \ \cdots \ u_{k-n_\mathrm{b}^\mathrm{o}}]^\top$, (14) can be written in the compact form :

$$y(k) = f^\mathrm{o}(x_k^\mathrm{o}, p_k) + e_k^\mathrm{o} = \sum_{i=1}^{n_\mathrm{g}^\mathrm{o}} g_i^\mathrm{o}(p_k) x_{k,i}^\mathrm{o} + e_k^\mathrm{o}. \tag{15}$$

The model structure to estimate (15) is considered as

$$y_k = \underbrace{\sum_{i=1}^{n_\mathrm{a}} a_i(p_k) q^{-i} y_k + \sum_{j=d}^{n_\mathrm{b}} b_j(p_k) q^{-j} u_k}_{f(x_k, p_k) = \sum_{i=1}^{n_\mathrm{g}} g_i(p_k) x_{k,i}} + e_k, \tag{16}$$

where e_k denotes the residual term, $n_\mathrm{a}, n_\mathrm{b}, d$ are nonnegative and not necessary equal with n_a^o, n_b^o, d^o. Furthermore, $n_\mathrm{g} = n_\mathrm{a} + n_\mathrm{b} - d + 1$ and $x_{k,i}$ is the i-th component of $x_k = [y_{k-1} \ \cdots \ y_{k-n_\mathrm{a}} \ u_{k-d} \ \cdots \ u_{k-n_\mathrm{b}}]^\top$. Please note that LPV models can also be seen as a special case of (8) with $w_k = [x_k \ p_k]$ and $f_i(x_{k,i}, p_k) = g_i(p_k) x_{k,i}$.

[2] For clarity of exposition, we assume that a_i^o and b_j^o have static dependence. Extension of the results of this paper to the dynamic dependency case, i.e., dependence on $p_k, p_{k-1}, p_{k-2}, \ldots$ follows straightforwardly.

4.2 Problem statement

Our goal is to jointly reconstruct the model structure, i.e., model order, number of effective coefficient functions, delay, etc. (Challenge (i)), and the scheduling variable dependencies (Challenge (ii)) directly from data. Specifically, based on a finite record of input, output and scheduling variable measurements, i.e., $\mathcal{D}_N = \{y_k, u_k, p_k\}_{k=1}^N$, we want to

- A1 Enforce sparsity over estimation of the functions $a_i(p_k)$ (with $i \in \mathbb{I}_1^{n_\mathrm{a}}$) and $b_j(p_k)$ (with $j \in \mathbb{I}_0^{n_\mathrm{b}}$).
- A2 Estimate the possibly nonlinear functions $a_i(p_k)$ and $b_j(p_k)$, characterizing the estimated relationship, directly from data;

In the next section, we solve the joint problem of A2 and (A2) by applying Theorem 2 via suitable kernels for LPV-IO models.

5 Sparse non-parametric LPV-IO identification

5.1 Kernel choice for LPV-IO models

In the considered LPV identification problem, the model (16) corresponds to $w_k = [\ x_k^\top \quad p_k^\top\ ]^\top$ and $z_k = y_k$, and the aim is to find a kernel K that is a representer for the function $f(x,p)$ with a specific structure $f(x,p) = \sum_{i=1}^{n_\mathrm{g}} g_i(p)x_i$. Naturally, such a Hilbert space is not unique[3], hence it is important to choose $\mathscr{H}_K$ to represent f with the least possible degrees of freedom, i.e., taking into account the linear dependency of LPV models on x_i, to reduce variance of the estimates.

Lemma 1 (Reproducing kernel for LPV-IO models) *Given the LPV-IO structure $f(x,p)$ in (16), the function $f(x,p)$ is embedded in the RKHS $\mathscr{H}_K$, whose reproducing kernel $K : \mathbb{R}^{n_\mathrm{g}+n_\mathbb{P}} \times \mathbb{R}^{n_\mathrm{g}+n_\mathbb{P}} \to \mathbb{R}$ is defined as:*

$$K\big((x,p),(x',p')\big) = \sum_{i=1}^{n_\mathrm{g}} x_i K_i(p,p')x_i', \qquad (17)$$

where each sub-kernel $K_i(p,p') : \mathbb{R}^{n_\mathbb{P}} \times \mathbb{R}^{n_\mathbb{P}} \to \mathbb{R}$ defines an RKHS $\mathscr{H}_{K_i}$ embedding $g_i(p) : \mathbb{R}^{n_\mathbb{P}} \to \mathbb{R}$.

PROOF. It is well-known that the RKHS $\mathscr{L}_i$, embedding static linear functions for $x_i \in \mathbb{R}$, is defined by the 1-dimensional kernel $L(x_i, x_i') : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ that is equal to $x_i x_i'$. Let $K_i(p,p') : \mathbb{R}^{n_\mathbb{P}} \times \mathbb{R}^{n_\mathbb{P}} \to \mathbb{R}$ define an RKHS $\mathscr{H}_{K_i}$ embedding of $g_i(p)$. By the *Aronszajn Theorem* on RKHS products [19] (see Appendix 8.1.2), the function $g_i(p)x_i$ is embedded in the RKHS product $\mathscr{H}_{K_i} \otimes \mathscr{L}_i$, where $\otimes$ denotes the direct product. Then, by the *Aronszajn Theorem* [19] on RKHS sums (see Appendix 8.1.1), $f(x,p)$ is embedded in the RKHS given by:

$$\mathscr{H}_K = \sum_{i=1}^{n_\mathrm{g}} \mathscr{H}_{K_i} \otimes \mathscr{L}_i, \qquad (18)$$

and the associated kernel defined in (17) reproduces $\mathscr{H}_K$, which ends the proof. ■

For K_i, any positive definite kernel, e.g., linear, polynomial, rational, spline or wavelet kernels, can be used. Choosing an appropriate K_i highly depends on the problem at hand. More details on this topic can be found in [5]. For the LPV case, *radial basis functions* (RBF) are typically chosen as kernels to describe the expected structural dependency on p. Such a choice is motivated by the fact that, in typical applications, $a_i(p_k)$ and $b_i(p_k)$ are nonlinear and smooth functions and RBF kernels are proven to perform well in capturing such functions.

5.2 Estimation of the coefficient functions from data

In general, operation $\mathscr{H}_{K_1} \otimes \mathscr{H}_{K_2}$ embeds all products between functions g_1 of $\mathscr{H}_{K_1}$ and g_2 of $\mathscr{H}_{K_2}$. After estimating such a product, it is not trivial to deduce g_1 and g_2 separately. In the LPV context, due to the linear nature of $\mathscr{L}_i$, it is possible to get a direct estimation of each of the functions $g_i(p)$, as described in Lemma 2.

Lemma 2 (Estimating the coefficient functions) *Let $f(x',p')$ be embedded in an RKHS $\mathscr{H}_K$ with reproducing kernel K as in (17). If f has a representer $f(x',p') = \sum_{k=1}^N c_k K\big((x_k,p_k),(x',p')\big)$, then each subfunction $g_i(p')$ of $f(x',p')$ is represented as*

$$g_i(p') = \sum_{k=1}^{N} c_k x_{k,i} K_i(p_k, p'). \qquad (19)$$

PROOF. For each $i \in \mathbb{I}_1^{n_\mathrm{g}}$, consider f at $(x^{\varnothing_i}, p')$ with

$$x_j^{\varnothing_i} = \delta_{i,j}, \quad \forall j \in \mathbb{I}_1^{n_\mathrm{g}}. \qquad (20)$$

In other words, $x^{\varnothing_i}$ is 0 except at its i-th element which is equal to 1. Then, $f(x^{\varnothing_i}, p') = \sum_{j=1}^{n_\mathrm{g}} g_j(p')x_j^{\varnothing_i} = g_i(p')$. In other words, $g_i(p')$ can be expressed as f evaluated at $(x^{\varnothing_i}, p')$. Hence, given the representer $f(x',p') = \sum_{k=1}^N c_k K\big((x_k,p_k),(x',p')\big)$ it can be easily seen that $K\big((x_k,p_k),(x^{\varnothing_i},p')\big) = x_{k,i}K_i(p_k,p')$. ■

Lemma 2 formalizes the estimation of the coefficient functions g_i in the general RKHS framework. It is interesting to note that, for the estimation criterion defined by (4), Lemma 2 confirms the results obtained in [24] from the LS-SVM and in [25] from the GP viewpoints.

5.3 Sparse RKHS estimator for structure selection

In this section, we propose to solve the joint problem of model structure selection A2 in terms of model order and delay and non-parametric estimation of the coefficient dependencies (A2), by applying the sparsity enforcing solution (11) which can be formulated in this LPV problem as:

[3] Any Hilbert space of functions $\sum_{i=1}^{n_\mathrm{g}} g_i(p)\pi_i(x_i)$ with π_i a polynomial would also embed $f(x,p)$.

$$\mathcal{V}(f) = \mathcal{C}_{\gamma_s}\big(\mathcal{D}_N, \{f(x_k,p_k)\}_{k=1}^N, \{g_i(m_j)\}_{i,j=1}^{n_g,M}\big) + \gamma\mathcal{R}\big(\|f\|_K\big) \quad (21)$$

where

$$\mathcal{C}_{\gamma_s}(*) = \sum_{k=1}^{N}\big(y_k - f(x_k,p_k)\big)^2 + \gamma_s\big\|\begin{bmatrix} S_1 & \cdots & S_{n_g}\end{bmatrix}\big\|_1,$$

$$S_i = \max_{j\in\mathbb{I}_1^M}|g_i(m_j)| \quad \text{and} \quad f(x_k,p_k) = \sum_{i=1}^{n_g} g_i(p_k)x_{k,i},$$

$$\mathcal{R}(f) = \|f\|_K^2.$$

Note that this problem is equivalent to (11), since $g_i(m_j) = f(x^{\varnothing_i}, m_j) = f_i(1, m_j)$. To estimate the function f minimizing (11), the representer of f based on Theorem 2 readily follows:

Corollary 1 (Representer under sparsity for LPV-IO models) *Let $\mathscr{H}_K$ be an RKHS embedding LPV-IO model* (16) *with reproducing kernel K as in* (17)*. Then the minimizer of* (11) *can be expressed as a representer in the form:*

$$\hat{f}(\cdot) = \sum_{k=1}^{N} c_k K_{(x_k,p_k)}(\cdot) + \sum_{i=1}^{n_g}\left(\sum_{j=1}^{M}\overline{c}_{i,j}K_{(x^{\varnothing_i},m_j)}(\cdot)\right). \quad (22)$$

PROOF. Theorem 2 directly applies as, due to the linear structure, $\{g_i(m_j)\}_{i,j=1}^{n_g,M} = \{f(x^{\varnothing_i}, m_j)\}_{i,j=1}^{n_g,M}$ and $K_{(x^{\varnothing_i},m_j)}(x',p') = K\big((x^{\varnothing_i},m_j),(x',p')\big) = K_i(m_j,p')\,x'$. ■

Having defined an optimization criterion, an LPV kernel structure as well as a representer for the problem at hand, the estimation problem can be defined as follows:

Problem 1 (Joint non-parametric LPV-IO estimation and structural selection) *Consider a data set $\mathcal{D}_N = \{y_k, u_k, p_k\}_{k=1}^N$ measured from a data-generating system* (15) *and a set of nodes $\{m_j\}_{j=1}^M$. Using the representer* (22) *with K defined in* (17)*, estimate the coefficients $\{c_k\}_{k=1}^N$ and $\{\overline{c}_j^i\}_{i=1,j=1}^{n_g,M}$ which minimize* (11)*.*

The solution of Problem 1 in terms of minimization of (11) can be reformulated as a quadratic *optimization problem* (QP) that can be efficiently solved by standard solvers. All the steps necessary for this reformulation are detailed in Appendix 8.2. Based on Lemma 2, each coefficient function $g_i(\cdot), i \in \mathbb{I}_1^{n_g}$ is obtained by computing $f(x^{\varnothing_i}, \cdot) = g_i(\cdot)$, which reads as:

$$g_i(\cdot) = \sum_{k=1}^{N} c_k x_{k,i} K_i(p_k,\cdot) + \sum_{j=1}^{M}\overline{c}_j^i K_i(m_j,\cdot). \quad (23)$$

Remark 1 *Due to the ℓ_1-penalty term introduced in* (21)*, i.e., $\gamma_s\big\|\begin{bmatrix} S_1 & \cdots & S_{n_g}\end{bmatrix}\big\|_1$, to shrink the coefficient functions a_i and b_j to zero, the resulting estimates of a_i and b_j will be biased. To reach an unbiased estimate, a two step procedure is applied, where* (21) *is used to determine indices $\mathcal{I}_y$ and $\mathcal{I}_u$ which correspond to significant functions a_i ($i \in \mathcal{I}_y$) and b_j ($j \in \mathcal{I}_u$). In the second step, by restricting the estimator corresponding to* (3) *to only these functions, the following lower-complexity LPV model is re-estimated:*

$$y_k = \sum_{i\in\mathcal{I}_y} a_i(p_k)q^{-i}y_k + \sum_{j\in\mathcal{I}_u} b_j(p_k)q^{-j}u_k + e_k. \quad (24)$$

Remark 2 *In order to obtain the estimate* (22)*, the following hyper-parameters are required to be chosen:*

- *β_w: hyper-parameters of the kernels;*
- *γ: ℓ_2 regularization parameter;*
- *γ_s : sparsity regularization parameter.*

For RKHS estimators, a large variety of hyper-parameter estimation methods such as empirical Bayes (EB), C_p statistics, Stein's unbiased risk estimator (SURE) and various forms of cross-validation (CV) have been proposed (see [8] for an overview). As the estimation of the proposed sparse representer is solved via a QP, most of these methods are not directly applicable. In this paper, we apply CV based on a validation data set which can be implemented by nonlinear optimization, gridding or Bayesian optimization.

6 Simulation example

The effectiveness of the developed RKHS approach is shown in this section on a Monte-Carlo study based on a simulation example. The identification of an LPV system with a sparse dynamic relation using an over-parameterized LPV-IO model is considered.

6.1 Data-generating system

The LPV data-generating system is a *Multi-Input Single-Output* (MISO) system described by

$$y_k = a_1^o(p_k)y_{k-1} + b_{15,1}^o(p_k)u_{k-15,1} + b_{4,2}^o(p_k)u_{k-4,2} + b_{5,2}^o(p_k)u_{k-5,2} + e_k^o, \quad (25)$$

where e_k^o is a white noise process with Gaussian distribution $\mathcal{N}\left(0,\sigma_e^2\right)$ and standard deviation $\sigma_e = 0.3$. The coefficient functions are described by the nonlinear maps:

$$a_1^o(p_k) = 0.9p_k^3, \quad (26a)$$

$$b_{15,1}^o(p_k) = 2\frac{\sin(2\pi p_k)}{2\pi p_k}, \quad b_{5,2}^o(p_k) = 2p_k^2. \quad (26b)$$

$$b_{4,2}^o(p_k) = \begin{cases} -1 & \text{if } p_k > 0.5; \\ -2p_k & \text{if } -0.5 \le p_k \le 0.5; \\ 1 & \text{if } p_k < -0.5; \end{cases} \quad (26c)$$

The system is estimated from a data set $\mathcal{D}_N = \{y_k, u_k, p_k\}_{k=1}^N$ with $N = 600$. To gather data, the input $u_{k,1}$ and the scheduling signal p_k are chosen to be mutually independent white-noise sequences with uniform distribution $\mathcal{U}(-1,1)$. The second input $u_{k,2}$ is a white noise process with Gaussian distribution $\mathcal{N}(0,1)$. For a

validation data set, independent realizations of u_k and p_k with the same distributions is used to generate $\mathcal{D}_{N_\mathrm{v}}^{\mathrm{val}}$ with $N_\mathrm{v} = 200$ samples.

6.2 LPV model structure

The identification problem is formulated in the considered RKHS setting by using an over-parameterized LPV model structure:

$$y_k = \sum_{i=1}^{n_\mathrm{a}} a_i(p_k) y_{k-i} + \sum_{j=1}^{n_{\mathrm{b},1}} b_{j,1}(p_k) u_{k-j,1} + \sum_{j=1}^{n_{\mathrm{b},2}} b_{j,2}(p_k) u_{k-j,2} + e_k. \quad (27)$$

with $n_\mathrm{a} = 20, n_{\mathrm{b},1} = 20$ and $n_{\mathrm{b},2} = 20$. According to the RKHS identification setting considered in this paper, the dependence of the functions $a_i(p_k)$, $b_{j,1}(p_k)$ and $b_{j,2}(p_k)$ on the scheduling signal p is not specified a priori.

6.3 Kernel structure

An RBF kernel is used for each K_i, i.e.,

$$K_i(p, p') = \exp\left(-\frac{(p-p')^2}{{\beta_{\mathrm{w}i}}^2}\right).$$

This kernel defines an RKHS encompassing a wide variety of nonlinear functions. In this example, all $\beta_{\mathrm{w}i}$ parameters are enforced to be the same value β_w in order to simplify the hyper-parameter optimization. In order to minimize the multi-objective function $\mathcal{V}(f)$ in (21), for $\gamma_\mathrm{s} > 0$, the interval $\mathbb{P} = [-1, 1]$ is gridded into $M = 11$ equidistant nodes m_j.

6.4 Methodology

To demonstrate the efficiency of the proposed approach, the methodology outlined in Algorithm 1 is applied with $T = 10^{-2}$ and with a CV score defined in terms of

$$\mathrm{BFR} = \max\left\{0, 1 - \sqrt{\frac{\sum_{k=1}^{N_\mathrm{V}} (y_k - \hat{y}_k)^2}{\sum_{k=1}^{N_\mathrm{V}} (y_k - \overline{y})^2}}\right\} \cdot 100\%,$$

with $\hat{y}_k$ denoting the simulated model output and $\overline{y}$ the sample mean of the measured output over the validation set. For Algorithm 1, the grid Γ is chosen such that $\max(\gamma^{(\tau)})$ and $\max(\gamma_\mathrm{s}^{(\tau)})$ produce $\hat{f} = 0$, while $\max(\beta_\mathrm{w}^{(\tau)})$ is set to be three times the length of the interval $\mathbb{P}$. Furthermore, for each parameter, the minimum is taken to be zero and the resulting region is covered by 100 grid-points.

6.5 Coefficient selection results

In order to provide representative results, a Monte-Carlo simulation (MCs) of $N_\mathrm{MC} = 50$ runs is performed. At each run, new realizations of the data sets $\mathcal{D}_N$, $\mathcal{D}_{N_\mathrm{v}}^{\mathrm{val}}$ are considered. The average of the *Signal-to-Noise Ratio* (SNR) over the MCs is equal to 13dB. Tuning of the hyper-parameters in terms of Algorithm 1 has been realized on the first dataset and $\mathcal{D}_{N_\mathrm{v}}^{\mathrm{val}}$. The results are: $\gamma = 0.01$, $\gamma_\mathrm{s} = 0.3$ and $\beta_{\mathrm{w},i} = 0.7$. These are then used in the 50 runs based MCs, which corresponds to Steps 4-9 in Algorithm 1 with these parameters fixed. The results are compared against a regular RKHS method using model (27) without sparse regularization and with optimized hyper-parameters $\gamma = 1$ and $\beta_\mathrm{w} = 0.7$. Furthermore, in order to evaluate the performance of the proposed approach with respect to other sparse estimation methods, a LASSO estimator is also applied using a 6^{th}-order monomial basis based parametrization of each nonlinearity. The regularization parameter of the LASSO has been optimized using the grid search process. For $a_i(p_k)$, $b_{j,1}(p_k)$ and $b_{j,2}(p_k)$ estimated via the RKHS estimator (i.e., with $\gamma_\mathrm{s} = 0$), its sparse version (S-RKHS) and the LASSO, the maximum absolute values $\overline{a}_i$, $\overline{b}_{j,1}$ and $\overline{b}_{j,2}$ of these functions are computed over $\mathbb{P}$. The average and standard deviation of these maximums over the 50 MC runs is reported in Figure 1.

Algorithm 1 Sparse estimation & hyper-para. tuning

Require: model structure (27), data sets $\mathcal{D}_N$, $\mathcal{D}_{N_\mathrm{v}}^{\mathrm{val}}$, node points $\mathbb{M} = \{m_j\}_{j=1}^{M} \subset \mathbb{P}$, kernel function K, grid $\Gamma = \{\gamma^{(\tau)}, \gamma_\mathrm{s}^{(\tau)}, \beta_\mathrm{w}^{(\tau)}\}_{\tau=1}^{N_\mathrm{g}}$ and threshold $T > 0$.

1: set $\tau \leftarrow 0$.
2: **repeat**
3: Set $\tau \leftarrow \tau + 1$.
4: Set hyper-parameters to $\gamma^{(\tau)}, \gamma_\mathrm{s}^{(\tau)}, \beta_\mathrm{w}^{(\tau)}$.
5: Minimize $\mathcal{V}(f)$ in (21) to estimate (22). The results of this sparse estimator is referred to as S-RKHS.
6: For each $i \in \mathbb{I}_1^{n_\mathrm{g}}$, test if $\max_{j \in \mathbb{I}_1^M} |g_i(m_j)| > T$.
7: Collect the index of the significant a_i, $b_{j,1}$ and $b_{j,2}$ in Step 6 into the sets $\mathcal{I}_\mathrm{y}$, $\mathcal{I}_{\mathrm{u}_1}$ and $\mathcal{I}_{\mathrm{u}_2}$.
8: With the same γ, β_w, estimate the low-complexity model (24) via the standard RKHS method (6) and (7). The result $\hat{f}^{(\tau)}$ is denoted as LC-RKHS.
9: Compute the BFR score of the simulated output of $\hat{f}^{(\tau)}$ on $\mathcal{D}_{N_\mathrm{v}}^{\mathrm{val}}$.
10: **until** $\tau = N_\mathrm{g}$.
11: **return** $\hat{f}^{(\tau)}$ with the lowest BFR score.

Figure 1 reveals that the S-RKHS approach correctly detects the nonzero coefficient functions as a_1, $b_{15,1}$, $b_{4,2}$ and $b_{5,2}$ (see (25)). These are the only functions with maximum absolute value greater than the threshold 10^{-2}. It is also worth remarking that the true coefficient structure of the system is detected in 47 out of 50 Monte-Carlo runs, while in the other 3 runs, 5 nonzero functions were detected instead of 4. As can be seen in Figure 1, the estimated maximum values of $|a_1(\centerdot)|$, $|b_{15,1}(\centerdot)|$, $|b_{4,2}(\centerdot)|$ and $|b_{5,2}(\centerdot)|$ over the interval $\mathbb{P}$ are 0.27, 0.12, 0.67 and 0.75, respectively, while the corresponding true values are 0.9, 2, 1 and 2. This is a common phenomenon with ℓ_1 regularization indicating the necessity of the second re-estimation step. When comparing the selection quality with the LASSO approach, the perfor-

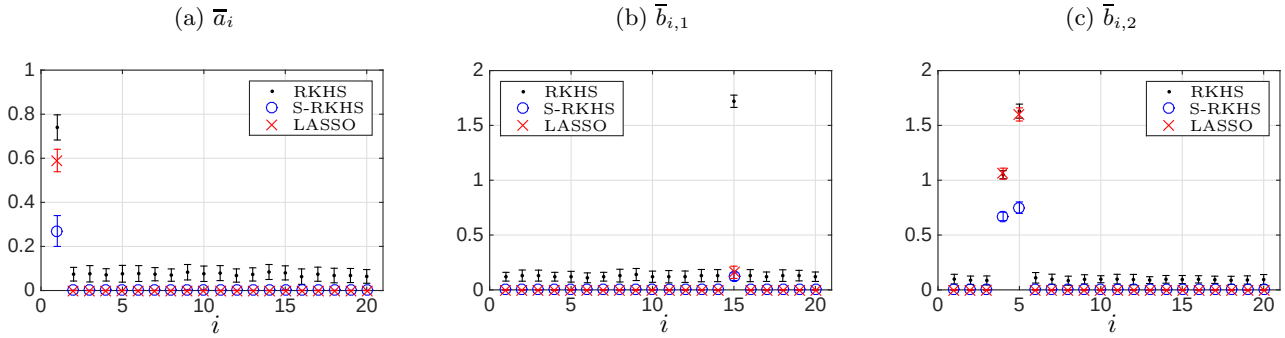

Fig. 1. Maximum of the coefficient function estimates over 50 Monte Carlo runs.

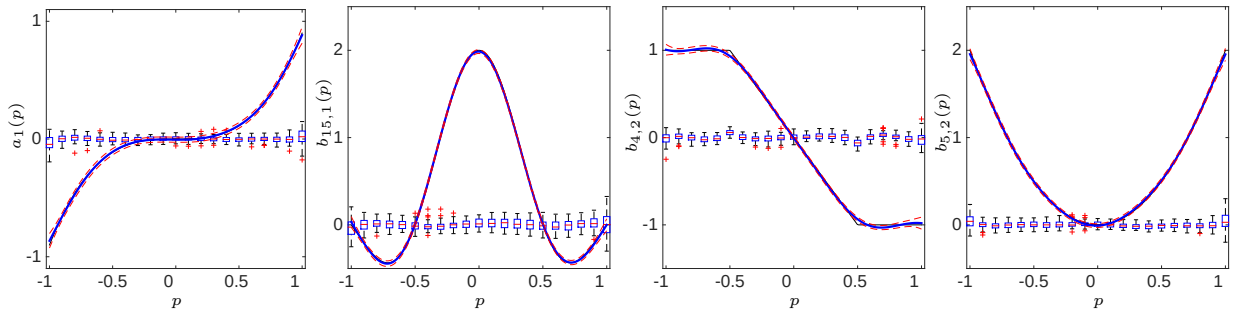

Fig. 2. Estimated coefficient functions $a_1(p_k)$, $b_{15,1}(p_k)$, $b_{4,2}(p_k)$, $b_{5,2}(p_k)$ after model structure selection (LC-RKHS). True function (solid black line), mean estimate (solid blue line), the standard deviation intervals (dashed red line), and the box plot of the error over the 50 Monte Carlo runs.

mance of both approaches are equivalent. Naturally the average values of significant functions are different since the LASSO is based on an explicit and a priori fixed polynomial parametrization. This shows that on an example where relatively low order polynomial parametrization is suitable, the proposed approach is competitive with LASSO. However, the RKHS guarantees a more accurate modeling capability in a larger variety of nonlinear structures. In cases when the explicit parametrization is not adequate (*e.g.*, non-symmetric functions), LASSO selection process can potentially run into difficulties.

6.6 Final estimation results

Here, the results of the final estimation step of the LC-RKHS after complexity shrinking is presented. The estimates of the nonzero coefficient functions a_1, $b_{15,1}$, $b_{4,2}$ and $b_{5,2}$ are plotted in Figure 2. It can be seen that the true nature of the nonlinear scheduling functions are well captured and accurately estimated by the RKHS model. The box-plots of the BFR on the validation dataset (used neither for training nor to tuning the hyper-parameters γ, γ_s and β_{w_i}) obtained with the RKHS and S-RKHS approaches are computed and reported in Figure 3. The obtained results clearly indicate that, by exploiting the sparsity structure, the LC-RKHS dramatically improves the model quality with respect to overparameterized one.

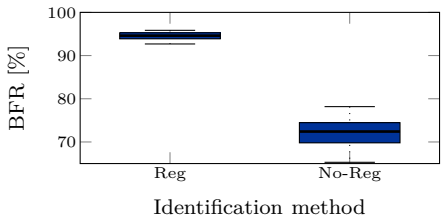

Fig. 3. Box-plot of the Monte-Carlo results for the BFR obtained with the LC-RKHS (left) and standard RKHS (right) estimators.

7 Conclusions

In this paper, we have presented a sparse Reproducing Kernel Hilbert Space (RKHS) estimator for a general class of nonlinear dynamical problems where the underlaying nonlinearity can be expressed in a sum-of-functions form. The main strength of this estimator is that its solution can be computed by convex (quadratic) optimization and therefore it avoids possible convergence problems of perviously proposed iterative greedy solutions. Applicability of the estimator is demonstrated on non-parametric estimation of LPV input-output models with an ARX noise structure. The resulting estimator avoids parameterization of the dependencies of the model coefficients on the scheduling variable and capable of automatic selection of model structure based on data, leading to a truly black-box identification method in the LPV setting without the need of user interaction. Such properties are achieved due to the combination of RKHS estimator and ℓ_1 regularization, which, next to the formulation of the resulting estimator is the main contribution of the paper. Behavior of the proposed estimator is empirically analyzed showing consistency of the identification method in recovering the functional dependency and dynamic order of the system together

with possible sparsity pattern of the model coefficients. The method is further extendable for more general noise conditions of the Box-Jenkins type using an instrumental variable based modification which is the target of future research.

8 Appendix

8.1 Aronszajn's Theorems [19]

8.1.1 Sum of kernels

If $K_i(x,x')$ is the reproducing kernel of the RKHS $\mathscr{H}_{K_i}$ with the norm $\|\cdot\|_{K_i}$, then $K(x,x') = \sum_{i=1}^n K_i(x,x')$ is the reproducing kernel of the RKHS $\mathscr{H}_K$ containing all functions $f = \sum_{i=1}^n f_i$ with $f_i \in \mathscr{H}_{K_i}$ and with norm $\|f\|_K^2 = \min\{\sum_{i=1}^n \|f_i\|_{K_i}^2\}$, where the minimum is taken for all the decompositions $f = \sum_{i=1}^n f_i$ with $f_i \in \mathscr{H}_{K_i}$. If all $\mathscr{H}_{K_i}$ are disjoint and therefore do not include any common functions beside 0, then the norm in $\mathscr{H}_K$ is simply given by $\sum_{i=1}^n \|f_i\|_{K_i}^2$.

8.1.2 Product of kernels

Let $\mathscr{H}_{K_1}$ and $\mathscr{H}_{K_2}$ be RKHS's defined by the reproducing kernels $K_1(x_1,x_1')$ and $K_2(x_2,x_2')$. The direct product of $\mathscr{H}_{K_1} \otimes \mathscr{H}_{K_2}$ is an RKHS $\mathscr{H}_K$ defined by the reproducing kernel $K\big((x_1,x_1'),(x_2,x_2')\big) = K_1(x_1,x_1')K_2(x_2,x_2')$ and has the norm $\|(f_1,f_2)\|_K = \|f_1\|_{K_1}^2\|f_2\|_{K_2}^2$. $\mathscr{H}_K$ embeds all functions of type $f(x_1,x_2) = f_1(x_1)f_2(x_2)$ with $f_1 \in \mathscr{H}_{K_1}$ and $f_2 \in \mathscr{H}_{K_2}$.

8.2 Quadratic optimization problem

8.2.1 Formulation

Under the considered kernel structure (17) and sparsity objectives of the LPV case expressed in terms of $\mathbb{M}$, the optimization problem (11) can be written by introducing the slack variables $r = \{r_i\}_{i=1}^{n_g}$ as:

$$\begin{aligned} &\min_{f,r_i} \sum_{k=1}^N (y_k - f(x_k,p_k))^2 + \gamma||f||_K^2 + \gamma_s \sum_{i=1}^{n_g} r_i \\ &\text{s.t.} -r_i \le f(x^{\varnothing_i}, m_j) \le r_i, \quad i \in \mathbb{I}_1^{n_g}, \quad j \in \mathbb{I}_1^M. \end{aligned} \tag{28}$$

Notice that the specific structure of the LPV problem simplifies the general expression of K given in (17), when it is computed at the specific points $x^{\varnothing_i}$ in (20) due to the reproducing property:

$$\begin{aligned} \langle K_{(x_k,p_k)}(\cdot), K_{(x^{\varnothing_i},m_j)}(\cdot)\rangle_K &= K((x_k,p_k),(x^{\varnothing_i},m_j)) \\ &= x_{k,i}K_i(p_k,m_j) \\ \langle K_{(x^{\varnothing_i},m_j)}(\cdot), K_{(x^{\varnothing_j},m_s)}(\cdot)\rangle_K &= K((x^{\varnothing_i},m_j),(x^{\varnothing_j},m_s)) \\ &= K_i(m_j,m_s) \quad \text{if} \quad i=j \\ &= 0 \quad \forall i \ne j \end{aligned} \tag{29}$$

By noticing these equalities, it becomes obvious that (21) is equivalent with (28) and that the sparse representer (12), i.e., (22), can be used to express $||y_k - f(x_k,p_k)||_2^2$, $||f||_K^2$ and $f(x^{\varnothing_i},m_j)$ in terms of the coefficients c_k and $\overline{c}_j^i$. Their matrix expression, leading to the QP are detailed in the subsequent subsections.

8.2.2 Expressing the error term

In order to compute $\sum_{t=1}^N (y_k - f(x_k,p_k))^2$, the value of f must be computed for all measurements points (x_k,p_k) by using directly the sparse representer (22).

Stack column-wise the parameters as

$$\breve{c}^\top = [c^\top \overline{c}^\top] = [c_1 \ \dots \ c_N \ \overline{c}_{1,1} \ \dots \overline{c}_{1,M} \ \overline{c}_{2,1} \ \dots \ \overline{c}_{n_g,M}],$$

the output $Y_N = [y_1 \ \dots \ y_N]^\top$ and introduce the following notation $\forall i \in \mathbb{I}_1^{n_g}$:

$$\begin{aligned} X_i &= \operatorname{diag}(x_{1,i},\dots,x_{N,i}), \\ \mathcal{K}_i^{NM}(k,j) &= K_i(p_k,m_j), \qquad \forall k,j \in \mathbb{I}_1^M, \\ \mathcal{K}_i^{MN} &= \mathcal{K}_i^{NM\top}, \\ \mathcal{K}_i^{MM}(j,s) &= K_i(m_j,m_s), \qquad j,s \in \mathbb{I}_1^M, \\ \mathcal{K}_i^{NN}(k,l) &= K_i(p_k,p_l), \qquad k,l \in \mathbb{I}_1^N. \end{aligned}$$

Then, the following matrix formulation is obtained:

$$\sum_{k=1}^N (y_k - f(x_k,p_k))^2 = \| Y_N - \breve{\mathcal{K}}\breve{c} \|_2^2, \tag{30}$$

where $\breve{\mathcal{K}} \in \mathbb{R}^{N\times(N+Mn_g)}$ is given as

$$\breve{\mathcal{K}} = \left[\begin{array}{c|c|c|c} \sum_{i=1}^{n_g} X_i\mathcal{K}_i^{NN}X_i & X_1\mathcal{K}_1^{NM} & \dots & X_{n_g}\mathcal{K}_{n_g}^{NM} \end{array}\right]. \tag{31}$$

8.2.3 Expressing the regularizer

Using the expression $||f||_K^2 = \langle f,f\rangle_K$ under the sparse representer (22):

$$\begin{aligned} ||f||_K^2 = &\sum_{i=1}^{n_g}\Big(\sum_{k=1}^N\sum_{l=1}^N c_k x_{k,i} K_i(p_k,p_l) x_{l,i} c_l\Big) \\ &+ \sum_{i=1}^{n_g}\Big(\sum_{j=1}^M\sum_{l=1}^N \overline{c}_{i,j} K_i(m_j,p_l) x_{l,i} c_l\Big) \\ &+ \sum_{i=1}^{n_g}\Big(\sum_{k=1}^N\sum_{s=1}^M c_k x_{k,i} K_i(p_k,m_s) \overline{c}_{i,s}\Big) \\ &+ \sum_{i=1}^{n_g}\Big(\sum_{j=1}^M\sum_{s=1}^M \overline{c}_{i,j} K_i(m_j,m_s) \overline{c}_{i,s}\Big), \end{aligned} \tag{32}$$

which corresponds in a matrix form with $\breve{c}^\top = [c^\top \overline{c}^\top]$ to

$$||f||_K^2 = \breve{c}^\top \Omega \breve{c}, \tag{33}$$

where $\Omega \in \mathbb{R}^{(N+Mn_g)\times(N+Mn_g)}$ is given as

$$\Omega = \left[\begin{array}{c|c|c|c|c} \sum_{i=1}^{n_g} X_i\mathcal{K}_i^{NN}X_i & X_1\mathcal{K}_1^{NM} & X_2\mathcal{K}_2^{NM} & \dots & X_{n_g}\mathcal{K}_{n_g}^{NM} \\ \hline \mathcal{K}_1^{MN}X_1 & \mathcal{K}_1^{MM} & 0 & \dots & 0 \\ \hline \mathcal{K}_2^{MN}X_2 & 0 & \mathcal{K}_2^{MM} & \ddots & \vdots \\ \hline \vdots & \vdots & \ddots & \ddots & 0 \\ \hline \mathcal{K}_{n_g}^{MN}X_{n_g} & 0 & \dots & 0 & \mathcal{K}_{n_g}^{MM} \end{array}\right].$$

8.2.4 Expressing the sparsity term

Finally by using that $g_i(m_j) = f(x^{\emptyset_i}, m_j)$ and the sparse representer (22):

$$g_i(m_j) = f(x^{\emptyset_i}, m_j) = \big(\sum_{k=1}^{N} c_k x_{k,i} K_i(p_k, m_j)\big) + \big(\sum_{s=1}^{M} \overline{c}_{i,s} K_i(m_s, m_j)\big), \quad \forall i \in \mathbb{I}_1^{n_g}, \forall j \in \mathbb{I}_1^M, \quad (34)$$

or equivalently

$$g_i(m_j) = [c^\top \overline{c}_i^\top] \begin{bmatrix} \mathcal{K}_i^{Nj} \\ \mathcal{K}_i^{Mj}, \end{bmatrix} \quad (35)$$

using the additional notation

$$\begin{aligned} \mathcal{K}_i^{Nj}(k) &= x_{k,i} K_i(p_k, m_j), \qquad k \in \mathbb{I}_1^N, \\ \mathcal{K}_i^{Mj}(s) &= K_i(m_s, m_j), \qquad s \in \mathbb{I}_1^M. \end{aligned} \quad (36)$$

8.2.5 Matrix expression of the optimization criterion

Under the previous derivation, the optimization problem (21) can be equivalently considered as:

$$\min_{c,\overline{c},r} \parallel Y_N - \breve{\mathcal{K}}\breve{c} \parallel_2^2 + \gamma_s \sum_{i=1}^{n_g} r_i + \gamma \breve{c}^\top \Omega \breve{c}$$

$$\text{s.t.} \begin{cases} -r_i \le [c^\top \overline{c}_i^\top] \begin{bmatrix} \mathcal{K}_i^{Nj} \\ \mathcal{K}_i^{Mj} \end{bmatrix} \le r_i, \quad i \in \mathbb{I}_1^{n_g}, \quad j \in \mathbb{I}_1^M \\ r_i > 0 \end{cases}$$

which is a quadratic programming problem.

References

[1] L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, pp. 373–384, 1995.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. of the Royal Statistical Society. Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[3] C. R. Rojas, R. Tóth, and H. Hjalmarsson, "Sparse estimation of polynomial and rational dynamic models," *Special issue, IEEE Transactions on Automatic Control*, vol. 59, pp. 2962–2977, 2014.

[4] 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.

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

[6] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. The MIT Press, 2006.

[7] 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.

[8] 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.

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

[10] M. A. H. Darwish, P. B. Cox, I. Proimadis, G. Pillonetto, and R. Tóth, "Prediction-error identification of LPV systems: A nonparametric gaussian regression approach," *Automatica*, vol. 97, pp. 92–103, 2018.

[11] V. Koltchinskii and M. Yuan, "Sparsity in multiple kernel learning," *The Annals of Statistics*, vol. 38, no. 6, pp. 3660–3695, 2010.

[12] C. Micchelli and M. Pontil, "Learning the kernel function via regularization," *Journal of Machine Learning Research*, vol. 6, pp. 1099–1125, 2005.

[13] F. R. Bach, "Consistency of the group lasso and multiple kernel learning," *Journal of Machine Learning Research*, vol. 9, pp. 1179–1225, 2008.

[14] H. Liu, L. Wasserman, J. D. Lafferty, and P. K. Ravikumar, "Spam: Sparse additive models," in *Advances in Neural Information Processing Systems 20* (J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, eds.), pp. 1201–1208, Curran Associates, Inc., 2008.

[15] R. Tóth, *Modeling and identification of linear parameter-varying systems*. Lecture Notes in Control and Information Sciences, Vol. 403, Springer, Heidelberg, 2010.

[16] V. Laurain, R. Tóth, W. Zheng, and M. Gilson, "Nonparametric identification of LPV models under general noise conditions, an LS-SVM based approach," in *Proc. of the the 16th IFAC Symposium on System Identification*, (Brussels, Belgium), pp. 1761–1766, July 2012.

[17] P. Lopes dos Santos, T. P. Azevedo-Perdicoúlis, J. A. Ramos, S. Deshpande, D. E. Rivera, and J. L. Martins de Carvalho, "LPV system identification using a separable least squares support vector machines approach," in *Proc. of the 53rd IEEE Conference on Decision and Control*, (Los Angeles, CA, USA), pp. 2548–2554, Dec. 2014.

[18] G. Wahba, *Spline models for observational data*. Philadelphia, PA, USA: Siam, 1990.

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

[20] G. S. Kimeldorf and G. Wahba, "A correspondence between Bayesian estimation on stochastic processes and smoothing by splines," *The Annals of Mathematical Statistics*, vol. 41, no. 2, pp. 495–502, 1970.

[21] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Computational Learning Theory* (D. Helmbold and B. Williamson, eds.), pp. 416–426, Springer Berlin Heidelberg, 2001.

[22] A. Argyriou and F. Dinuzzo, "A unifying view of representer theorems.," in *Proc. of the 31th Internatinal Conference on Machine Learning (ICML)*, (Beijing, China), pp. 748–756, June 2014.

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

[24] R. Tóth, V. Laurain, W. Zheng, and K. Poolla, "Model structure learning: A support vector machine approach for LPV linear-regression models," in *Proc. of the 50th IEEE Conf. on Decision and Control*, (Orlando, Florida, USA), pp. 3192–3197, Dec. 2011.

[25] A. Golabi, N. Meskin, R. Toth, and J. Mohammadpour, "A bayesian approach for lpv model identification and its application to complex processes," *IEEE Transactions on Control Systems Technology*, vol. 25, pp. 2160–2167, 11 2017.