Prediction-Error Identification of LPV Systems: A Nonparametric Gaussian Regression Approach ⋆

Mohamed A. H. Darwish [a], Pepijn B. Cox [b], Ioannis Proimadis [b], Gianluigi Pillonetto [c], Roland Tóth [b].

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

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

[c] *Information Engineering Department, University of Padova, Padova 35131, Italy.*

Abstract

In this paper, a Bayesian nonparametric approach is introduced to estimate multi-input multi-output (MIMO) linear parameter-varying (LPV) models under the general noise model structure of Box-Jenkins (BJ) type. The approach is based on the estimation of the one-step-ahead predictor of general LPV-BJ structures. Parts of the predictors associated with the input and output signals are modeled as asymptotically stable infinite impulse response (IIR) models. Then, these IIR models are identified in a completely nonparametric sense: not only the coefficients are estimated as functions, but also the whole time evolution of the impulse response w.r.t. the scheduling signal of the LPV system. In this Bayesian setting, the estimate of the one-step-ahead predictor is a realization from a zero-mean Gaussian random field, where the covariance function is a multidimensional Gaussian kernel that encodes both the possible structural dependencies and the stability of the predictor. Two different kernel formulations are presented for the LPV setting, namely a diagonal (DI) like and tuned/correlated (TC) like kernels, where the TC-like kernel is able to describe the correlation between coefficient functions associated with different time indices. The unknown hyperparameters that parameterize the DI or TC kernel are tuned by maximizing the marginal likelihood w.r.t. the observed data. Moreover, we provide a nonparametric realization scheme to recover the original process and noise IIRs from the identified one-step-ahead predictor. The performance of the presented identification approach is tested on a MIMO LPV-BJ simulation example by means of an extensive Monte-Carlo study.

Key words: Bayesian identification; System identification; Reproducing kernel Hilbert space; Linear parameter-varying systems; Machine learning; Regularization; Prediction-error identification; Gaussian processes; Box-Jenkins models.

1 Introduction

Linear parameter-varying (LPV) systems, introduced in [1], have received considerable attention [2,3], as they offer an attractive modeling framework to capture nonlinear and/or non-stationary behavior of physical and chemical processes [4,5]. Most of the existing LPV *identification* (ID) approaches are formulated in *discrete-time* (DT) [2] to identify state-space or linear-fractional representation forms, e.g., [6,7,8,9]; series-expansion based models, e.g., [10]; and various *input-output* (IO) model structures, e.g., [4,11,12].

Identification of LPV-IO models gained popularity, as *prediction-error minimization* (PEM) methods have been successfully extended to LPV models, providing a well-understood framework for consistency and stochastic interpretation of the estimates together with low computational complexity of the resulting identification procedures [11]. Moreover, the PEM framework is well suited to identify a large variety of noise and plant models, see [2] for an overview. Although LPV-IO models cover a variety of process and noise representations,

⋆ This research has benefited from the financial support of the Student Mission, Ministry of Higher Education, Government of Egypt. The first three authors have contributed equally to the paper. Corresponding author Mohamed A. H. Darwish. Tel. +31-(0)65-939-7952.

Email addresses: mohamed.darwish@eng.au.edu.eg (Mohamed A. H. Darwish), p.b.cox@tue.nl (Pepijn B. Cox), i.proimadis@tue.nl (Ioannis Proimadis), giapi@dei.unipd.it (Gianluigi Pillonetto), r.toth@tue.nl (Roland Tóth).

where the Box-Jenkins (BJ) model is the most general form, PEM identification of BJ models leads to a nonlinear optimization problem [11], which is sensitive to local minima. Alternatively, the *instrumental variable* (IV) method provides an attractive approach that deals with the general noise scenario and avoids the nonlinear optimization [13]. Another important issue in the identification of LPV-IO models is capturing the structural dependency on the scheduling signal. In the parametric case, the structural dependency is generally characterized by using a pre-specified set of basis functions, which either require significant prior knowledge of the underlying system or tedious repetitive execution of methods to synthesize an acceptable basis [11]. In addition, the choice of the number of these bases is challenging as it induces a bias/variance trade-off, i.e., by using fewer basis functions, the under-modeling (bias) error increases while increasing their number results in an increase of the variance of the parameters of the estimated models.

The so-called nonparametric methods offer an attractive alternative approach to capture the underlying dependencies directly from data without specifying any parameterization in terms of fixed basis functions. The main approaches of LPV nonparametric identification in the literature are: i) the dispersion function method [14], ii) the *least squares-support vector machine* (LS-SVM) methods, e.g., [15,16], and iii) the Bayesian setting based approaches [17,18]. However, in i)-iii) the considered noise models are restricted to output error type (LPV *finite impulse response* (FIR) model) and autoregressive type (LPV *autoregressive with exogenous input* (ARX) model). Additionally, both LS-SVM and Bayesian approaches have roots in the *reproducing kernel Hilbert space* (RKHS) theory [19] and admit an ℓ_2-regularization interpretation [20], such that consistency and convergence notions of the resulting estimator can be formulated.

This work is inspired by recent advances in nonparametric identification of *linear time-invariant* (LTI) models in the PEM setting [21] and novel results for optimal kernel design [22]. Here, we aim at formulating a nonparametric estimator of the one-step-ahead predictor for an LPV-BJ model, preserving the generality of the noise class and the asymptotic optimality of PEM. More specifically, we consider the one-step-ahead predictor as the summation of two sub-predictors associated with the input and output signals, where these sub-predictors are modeled as asymptotically stable LPV *infinite impulse response* (IIR) models. These LPV-IIR sub-predictors are identified in a nonparametric sense, where not only the coefficients are estimated as functions, but also the whole time evolution of the impulse response.

We follow a Bayesian approach for the nonparametric estimation by modeling the sub-predictors as realizations of zero-mean Gaussian random fields, which can be completely characterized by covariance (kernel) functions that implicitly act as a basis generator to describe both the functional dependencies and the time evolution of the impulse response of the sub-predictors. To this end, inspired by [23], we introduce a multidimensional Gaussian kernel which encodes: i) the possible structural dependencies on the scheduling signal by using *radial basis functions* (RBF) and ii) the stability of the predictor by including a decay term, which models the vanishing influence of the past input-scheduling-output pairs on the predicted output. The hyperparameters that parameterize the kernel can be efficiently estimated from data by maximizing the marginal likelihood w.r.t. the observations [24]. A preliminary work in this direction can be found in [25], however, here we provide the following extensions:

(1) Kernel formulation for the *multi-input multi-output* (MIMO) case;
(2) Enriching the kernel to take into account (nominal) LTI dynamics of the model, independent of the scheduling variables;
(3) Introduce a *tuned/correlated* (TC)-like kernel for the LPV setting to encode correlation between coefficient functions associated with different time indices;
(4) Introducing a nonparametric realization scheme to recover the original process and noise IIRs from the identified one-step-ahead sub-predictors.

The paper is organized as follows. In Section 2, the considered model structure is defined and the corresponding optimal one-step-ahead predictor is derived. The considered Gaussian regression framework is reviewed in Section 3. In Section 4, Bayesian identification of LPV-IO models and the problem of estimating the predictor and the hyperparameters of the kernel from data are presented. This is followed by a realization approach to get a nonparametric estimate for the process and noise dynamics from the identified predictor in Section 5. In Section 6, the effectiveness of the introduced approach is demonstrated by means of extensive Monte Carlo study. Finally, the paper is ended with conclusions in Section 7.

2 Problem Statement

2.1 LPV-BJ model

Consider a MIMO data-generating LPV system described in DT [1] by the following difference equations:

$$(A_0(q^{-1}) \diamond p)_k \breve{y}(k) = (B_0(q^{-1}) \diamond p)_k u(k), \tag{1a}$$
$$(D_0(q^{-1}) \diamond p)_k v(k) = (C_0(q^{-1}) \diamond p)_k e(k), \tag{1b}$$
$$y(k) = \breve{y}(k) + v(k), \tag{1c}$$

[1] Equivalence between DT and *continuous-time* (CT) LPV systems can be understood in terms of considering all free signals of the CT system (e.g. the input and the scheduling signals) to be generated by an ideal *zero order hold* (ZOH), i.e., they are piecewise constant and the output is sampled in a perfectly synchronized manner (for details see [26]).

where $k \in \mathbb{Z}$ is the discrete time, q is the forward time-shift operator, i.e., $qx(k) = x(k+1)$, $u : \mathbb{Z} \to \mathbb{U} = \mathbb{R}^{n_\mathrm{u}}$ is the input, $\breve{y}, y : \mathbb{Z} \to \mathbb{Y} = \mathbb{R}^{n_\mathrm{y}}$ are the noiseless and noisy outputs, respectively, $p : \mathbb{Z} \to \mathbb{P}$ is the so-called scheduling variable with compact range $\mathbb{P} \subset \mathbb{R}^{n_\mathrm{p}}$, and $e : \mathbb{Z} \to \mathbb{Y}$ is a white noise process with normal (Gaussian) distribution, i.e., $e(k) \sim \mathcal{N}(0, \Sigma_\mathrm{e})$ with covariance $\Sigma_\mathrm{e} \in \mathbb{R}^{n_\mathrm{y} \times n_\mathrm{y}}$, generating the colored noise signal $v : \mathbb{Z} \to \mathbb{Y}$. The p-dependent operators $A_0(q^{-1})$ and $B_0(q^{-1})$ that define the process model (1a), are matrix polynomials in q^{-1} of degree n_a and n_b respectively:

$$(A_0(q^{-1}) \diamond p)_k = I + \sum_{i=1}^{n_\mathrm{a}} (a_i \diamond p)_k \, q^{-i}, \tag{2a}$$

$$(B_0(q^{-1}) \diamond p)_k = \sum_{j=0}^{n_\mathrm{b}} (b_j \diamond p)_k \, q^{-j}, \tag{2b}$$

where I is the identity matrix with the appropriate dimension and $a_i : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}^{n_\mathrm{y} \times n_\mathrm{y}}$ and $b_j : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}^{n_\mathrm{y} \times n_\mathrm{u}}$ are matrix functions while $(a_i \diamond p)_k$ and $(b_j \diamond p)_k$ are shorthand notations for $(a_i \diamond p)_k = a_i(p(k), \ldots, p(k-i))$ and $(b_j \diamond p)_k = b_j(p(k), \ldots, p(k-j))$. These functions are assumed to be smooth and bounded on $\mathbb{P}$. In a similar fashion, the noise model relation (1b) is characterized by $C_0(q^{-1})$ and $D_0(q^{-1})$ corresponding to

$$(C_0(q^{-1}) \diamond p)_k = I + \sum_{i=1}^{n_\mathrm{c}} (c_i \diamond p)_k \, q^{-i}, \tag{3a}$$

$$(D_0(q^{-1}) \diamond p)_k = I + \sum_{j=1}^{n_\mathrm{d}} (d_j \diamond p)_k \, q^{-j}, \tag{3b}$$

where $c_i : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}^{n_\mathrm{y} \times n_\mathrm{y}}$ and $d_j : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}^{n_\mathrm{y} \times n_\mathrm{y}}$ are the coefficient function matrices of the monic polynomials (3) in q^{-1} of degree n_c and n_d, respectively.

2.2 The IIR form

As a first step to identify (1), the system representation is reformulated in an equivalent IIR formulation. For the reformulation to be well-posed, both (1a), (1b), and the inverse noise dynamics of (1b) [2] need to correspond to asymptotically stable LPV filters:

Definition 1 *An LPV filter* $(A_(q^{-1}) \diamond p)_k \, y(k) = (B(q^{-1}) \diamond p)_k \, u(k)$ *is called* globally asymptotically stable*, if, for all trajectories* $\{u(k), p(k), y(k)\}$ *satisfying the filter equation with* $u(k) = 0$ *for* $k \geq 0$ *and* $p(k) \in \mathbb{P}$*, it holds that* $\lim_{k \to \infty} y(k) = 0$.

A computational approach to verify asymptotic stability of (1) in terms of Definition 1 can be found in [27].

[2] Hence, we assume the existence of a stable left inverse of the corresponding IIR.

In classical identification approaches, one is interested in finding the process G_0 and noise H_0 dynamics, see [28], of (1) in terms of the following equivalent representation

$$y(k) = (G_0(q^{-1}) \diamond p)_k \, u(k) + (H_0(q^{-1}) \diamond p)_k \, e(k), \tag{4}$$

where the process and noise models, given that the underlying system has the BJ form (1), are defined as

$$G_0(q^{-1}) = A_0^\dagger(q^{-1}) B_0(q^{-1}) = \sum_{i=0}^{\infty} g_{0,i} \, q^{-i}, \tag{5a}$$

$$H_0(q^{-1}) = D_0^\dagger(q^{-1}) C_0(q^{-1}) = I + \sum_{i=1}^{\infty} h_{0,i} \, q^{-i}, \tag{5b}$$

where $A^\dagger$ denotes the left inverse of the polynomial A, see Lemma 2. Note that (5) represent functions that are dependent on the scheduling signal and they can be evaluated for a given scheduling trajectory, i.e., $(g_{0,i} \diamond p)_k$ and $(G_0(q^{-1}) \diamond p)_k$, similar to (2) and (3). Furthermore, (4) defines the infinite impulse response representation of the underlying system. The IIRs are key in formulating the identification problem, because the notion of transfer functions, that are applied in the LTI case, do not exit in the LPV setting.

Lemma 2 *Given a monic parameter-varying polynomial filter* $A(q^{-1})$ *with finite order* n_a*. If* $(A(q^{-1}) \diamond p)_k \, y(k) = u(k)$ *is asymptotically stable in the sense of Definition 1, then the left inverse* $A^\dagger(q^{-1})$ *of* $A(q^{-1})$ *is given by*

$$A^\dagger(q^{-1}) = \sum_{i=0}^{\infty} \left(I - A(q^{-1})\right)^i, \tag{6}$$

such that $A^\dagger(q^{-1}) A(q^{-1}) = I$*. In addition, if* (6) *is approximated by a finite truncation order* n*:*

$$\tilde{A}^\dagger(q^{-1}) = \sum_{i=0}^{n} \left(I - A(q^{-1})\right)^i, \tag{7}$$

then the normed truncation-error is given by

$$\begin{aligned} \epsilon_n &= \sup_{\substack{\|u\|_{\ell_2} = 1 \\ p \in \mathbb{P}^{\mathbb{Z}}}} \left\| \left((A^\dagger(q^{-1}) - \tilde{A}^\dagger(q^{-1})) \diamond p \right) u \right\|_{\ell_2} \\ &= \sup_{\substack{\|u\|_{\ell_2} = 1 \\ p \in \mathbb{P}^{\mathbb{Z}}}} \left\| \left((I - A(q^{-1})) \diamond p \right)^{n+1} u \right\|_{\ell_2}. \end{aligned} \tag{8}$$

where $\|\cdot\|_{\ell_2}$ *denotes the* ℓ_2 *norm of a discrete signal.*

Proof: See Appendix A. ■

Note that the inverse $(A^\dagger(q^{-1}) \diamond p)_k$ is also well-defined in terms of Lemma 2, i.e., $(A^\dagger(q^{-1}) \diamond p)_k (A(q^{-1}) \diamond p)_k = I$.

By straightforward application of Lemma 2 and substituting (2)-(3) into (5), the individual coefficient functions $g_{0,i}$ and $h_{0,i}$ are given by the following recursions:

$$(g_{0,i}\diamond p)_k = (b_i\diamond p)_k - \sum_{j=1}^{\min(n_\mathrm{a},i)} (a_j\diamond p)_k (g_{0,i-j}\diamond p)_{k-j}, \quad (9a)$$

$$(h_{0,i}\diamond p)_k = (c_i\diamond p)_k - \sum_{j=1}^{\min(n_\mathrm{d},i)} (d_j\diamond p)_k (h_{0,i-j}\diamond p)_{k-j}, \quad (9b)$$

for $i\geq 1$ with $g_{0,0}=b_0$ and $h_{0,0}=I$. Note that the coefficient functions $a_i,\ldots,h_i$ are non-commutative, i.e., $q^{-1}(a_i\diamond p)_k = (a_i\diamond p)_{k-1}q^{-1}$, therefore, the time-index in (9) is of the essence.

Similar to the LTI case [28], the representation (4) can be reformulated based on the trajectory of u,p,y, and the current value of e as, see [25, Eq. (23)]:

$$y(k) = \left((I - H_0^\dagger(q^{-1}))\diamond p\right)_k y(k) + \left(H_0^\dagger(q^{-1})G_0(q^{-1})\diamond p\right)_k u(k) + e(k), \quad (10)$$

if the noise filter (1b) and its left inverse are globally asymptotically stable in terms of Definition 1. In this case, the noise v has bounded spectral density. Moreover, given the system (1), the form (10) is also an IIR:

Theorem 3 ([25]) *If process dynamics* (1a) *and noise dynamics* (1b) *are asymptotically stable according to Definition 1, and, in addition, the inverse noise process is monic and asymptotically stable, then* (1) *can be equivalently represented by the following IIR*

$$y(k) = \sum_{i=1}^{\infty} (h_{y_i}\diamond p)_k q^{-i}\, y(k) + \sum_{j=0}^{\infty} (h_{u_j}\diamond p)_k q^{-j}\, u(k) + e(k), \quad (11)$$

where, $h_{y_i}:\mathbb{P}\times\ldots\times\mathbb{P}\to\mathbb{R}^{n_\mathrm{y}\times n_\mathrm{y}}$ *and* $h_{u_j}:\mathbb{P}\times\ldots\times\mathbb{P}\to\mathbb{R}^{n_\mathrm{y}\times n_\mathrm{u}}$ *are real meromorphic*[3] *matrix coefficient functions in the scheduling signal* p.

Hence, utilizing Lemma 2 and substituting (2)-(3) into (10), the coefficient functions h_{y_i} and h_{u_j} in (11) are given by

$$(h_{y_i}\diamond p)_k = -(d_i\diamond p)_k - \sum_{j=1}^{\min(n_\mathrm{c},i)} (c_j\diamond p)_k (h_{y_{i-j}}\diamond p)_{k-j}, \quad (12a)$$

$$(h_{u_i}\diamond p)_k = -\sum_{j=0}^{i} (h_{y_{i-j}}\diamond p)_k (g_{0,j}\diamond p)_{k-i+j}, \quad (12b)$$

for $i\geq 0$ with $h_{y_0}=-I$. It is worth to mention that the IIR representation (11) also exists for LPV state-space models represented with a specific innovations noise structure, making representation (11) also attractive for identifying such LPV state-space models, e.g., see [7,29].

2.3 One-step-ahead predictor

In the *prediction-error* setting, identification of (1) is formulated by using (11) to define a one-step-ahead predictor of $y(k)$ based on only the observations of the input $u(\tau)$ for $\tau\leq k$, the scheduling $p(\tau)$ for $\tau\leq k$, and the output signal $y(\tau)$ for $\tau\leq k-1$. The basic idea is to consider the mean of y conditioned on the past data:

$$\hat{y}(k|k-1) = \operatorname*{argmin}_{\delta\in\mathbb{R}} \ \mathbb{E}\left\{ \|y(k)-\delta\|_2^2 \mid x^{(k)} \right\}, \quad (13)$$

where $\|\cdot\|_2$ is the Euclidean norm, $\mathbb{E}\{\cdot\}$ is the expectation operator and $x^{(k)} = \left\{u^{(k)}, p^{(k)}, y^{(k-1)}\right\}$ is the shorthand notation of the past measurements, e.g., $u^{(k)} = \{u(\tau)\}_{\tau\leq k}$. Based on (11) and the fact that $e(k)$ is considered to be white noise, straightforward application of the expectation operator on (11) gives

$$\hat{y}(k|k-1) = \sum_{i=1}^{\infty} (h_{y_i}\diamond p)_k q^{-i} y(k) + \sum_{j=0}^{\infty} (h_{u_j}\diamond p)_k q^{-j} u(k). \quad (14)$$

The resulting predictor is similar to the predictor in the LTI case as it is a summation of two IIRs. Analogous to the LTI case, parametric identification employing (14) can be performed by either: 1) parameterizing the underlying BJ model ($a_i,\ldots,d_i$ in (2)-(3)) leading to a nonlinear estimation problem, prone to local minima, or 2) parameterizing the coefficient functions h_{y_i} and h_{u_j} in (14) directly as a linear combination of a set of basis functions leading to a linear-in-the-parameter problem [28]. In the LPV case, the latter choice requires a *priori* specified set of basis functions dependent on the scheduling signal. Incorrect selection of this set leads to structural bias of the estimate, where over-parameterization results in a variance increase of the estimated parameters. Due to the larger degree of freedom of LPV models, this bias/variance trade-off plays a more prominent role compared to the LTI case. The question is how to utilize the simplicity of the IIR form (11), but to overcome the large parametrization and the high parameter variance associated with its identification. A solution can be found in the regularization framework, which can be understood from two equivalent point of views: Bayesian identification and the RKHS context [30]. More specifically, the functional dependencies h_{y_i} and h_{u_j} are estimated nonparametrically, where a regularization is introduced to keep the variance of the estimates low by allowing a small amount of estimation bias. In the next section, the Bayesian identification within the Gaussian framework and the connection to function estimation in the RKHS will be reviewed.

[3] h is a real meromorphic function, if $h(\cdot)=g(\cdot)/f(\cdot)$ with g,f analytic functions and $f(\cdot)\neq 0$.

3 Gaussian Regression Framework

In this section, the *Gaussian process* (GP) regression framework and the connection to function estimation in RKHS are reviewed.

3.1 Nonparametric Gaussian regression

Definition 4 ([31]) *A GP is a collection of random variables $f(x)$, indexed by $x \in \mathbb{R}$, any finite number of which have a joint Gaussian distribution.*

This means that a GP is completely characterized by its mean and covariance function $m(x) = \mathbb{E}\{f(x)\}$ and $\mathcal{K}(x, x') = \operatorname{cov}(f(x), f(x'))$, respectively, where $f(x)$ is a real function and it is denoted by

$$f(x) \sim \mathcal{GP}(m(x), \mathcal{K}(x, x')). \tag{15}$$

Note that this definition basically extends the notion of random variables with normal distribution to normally distributed functions with domain $\mathbb{R}$.

For the sake of simplicity, the GP framework is first introduced for the *multi-input single-output* (MISO) case. Given a data set $\mathcal{D}_N = \{x(k), \omega(k)\}_{k=1}^N$ that is generated according to:

$$\omega(k) = f(x(k)) + \epsilon(k), \tag{16}$$

where $x : \mathbb{Z} \to \mathbb{R}^d$ is considered as an input variable, $f : \mathbb{R}^d \to \mathbb{R}$ is a (non)linear function, and $\epsilon(k) \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is a zero-mean normally distributed white noise with variance σ_ϵ^2. Following the Bayesian inference techniques within the GP framework, e.g., see [31], GP regression considers the unknown function f as a random function with a GP distribution postulating the prior (15). This prior also represents our beliefs and the high-level assumptions about f, e.g., smoothness and stability. Often, a zero-mean GP prior on f is assumed [4], i.e., $f(x) \sim \mathcal{GP}(0, \mathcal{K}_\theta(x, x'))$ where the covariance is defined via a parameterized kernel function $\mathcal{K}_\theta$ in terms of the unknown hyperparameter vector θ.

The posterior distribution $\mathcal{P}(f(x_*) \mid X, W, \theta)$ is Gaussian and it can be used to make predictions about f at an arbitrary point $x_* \in \mathbb{R}^d$. Let $X = [x(1) \cdots x(N)]^\top \in \mathbb{R}^{N \times d}$ be the regression matrix and $W = [\omega(1) \cdots \omega(N)]^\top \in \mathbb{R}^{N \times 1}$ be the observed output of (16). The minimum variance estimator of $f(x_*)$ for X, W, with a given θ, i.e., $\hat{f}(x_*) = \mathbb{E}\{f(x_*) \,|\, X, W, \theta\}$, is [31]

$$\hat{f}(x_*) = \sum_{i=1}^N c_i \mathcal{K}_\theta(x(i), x_*), \tag{17}$$

where c_i denotes the (i)-th element of $c = \left(K_\theta + \sigma_\epsilon^2 I\right)^{-1} W$ and K_θ is the so-called "kernel matrix" and its (i,j)-th element is defined as $[K_\theta]_{i,j} = \mathcal{K}_\theta(x(i), x(j))$. Furthermore, the covariance of the posterior distribution, i.e., $\operatorname{cov}(f(x_*) \,|\, X, W, \theta)$, that defines the uncertainty of the estimation (17) under the prior (15), is given as

$$\operatorname{cov}(f(x_*) \,|\, X, W, \theta) = \mathcal{K}_\theta(x_*, x_*) - K_* \left[K_\theta + \sigma_\epsilon^2 I\right]^{-1} {K_*}^\top, \tag{18}$$

where $K_* = [\mathcal{K}_\theta(x_*, x(1)) \cdots \mathcal{K}_\theta(x_*, x(N))]$ encodes the covariance relation between the test point x_* and the training points x_i.

In a function estimation problem, the true underlying covariance function $\mathcal{K}$ is not known a priori. Hence, a crucial step in GP regression is to design a kernel function $\mathcal{K}_\theta$, parameterized in terms of θ, which can express a wide variety of expected properties. At the same time, θ must be low dimensional such that the estimation of θ based on data can be efficiently performed. A well-known choice for the kernel function to encode both smoothness and stationarity of the unknown function is the *radial basis function* (RBF) [31]

$$\mathcal{K}_\theta(x, x') = \alpha^2 \exp\left(-\frac{1}{2}(x - x')^\top \Gamma^{-1} (x - x')\right), \tag{19}$$

where $x, x' \in \mathbb{R}^d$, α^2 is a scaling parameter that represents the signal variance and $\Gamma = \operatorname{diag}\left(\left[\varsigma_1^2 \cdots \varsigma_d^2\right]\right)$ is a diagonal matrix of the squared kernel width parameters $\{\varsigma_i\}_{i=1}^d$. The complete hyperparameter vector θ of (19) is $\theta = [\, \alpha \;\; \varsigma_1 \;\; \cdots \;\; \varsigma_d \,]^\top$.

A popular approach to obtain the hyperparameters θ that parameterize the kernel $\mathcal{K}_\theta$ is to estimate them from data by maximizing the log-*marginal likelihood* of the output w.r.t. θ [32]:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \;\; \log \mathcal{P}(W \mid X, \theta), \tag{20}$$

where $\mathcal{P}$ denotes a *probability density function* (PDF) and the log-marginal likelihood function is

$$\log \mathcal{P}(W \mid X, \theta) = -\frac{N}{2} \log(2\pi) - \underbrace{\frac{1}{2} W^\top \left(K_\theta + \sigma_\epsilon^2 I\right)^{-1} W}_{\text{"data-fit" term}} - \underbrace{\frac{1}{2} \log\det(K_\theta + \sigma_\epsilon^2 I)}_{\text{"complexity" term}}. \tag{21}$$

[4] In the Bayesian setting, unless other prior information available, the mean of the Gaussian process is considered to be zero. Such an assumption does not penalize the sign or variation of the function, fulfilling the so-called maximal entropy principle. Introducing a parameterized mean function often only increases the variance of the estimator (and also the computational complexity). As pointed out in [31], in general the kernel is sufficiently rich and parameterizing the mean is not needed.

Such an optimization problem, i.e., maximizing the marginal likelihood (20) is a nonlinear optimization problem prone to local minima [31,24]. However, the superiority of maximizing the marginal likelihood, i.e., (20) over other classical tuning methods, e.g., C_p statistics [33], *cross-validation* [28], *predicted residual sums of squares* [34], *generalized cross-validation* [35], *Stein's unbiased risk estimator* [36], has been investigated in [37], showing that it can better balance data fit and model complexity. In the sequel, for the sake of simplicity, the dependency of the kernel function on θ will be dropped.

It is worth to mention that in the above discussion, we have considered a univariate prediction, i.e., the MISO setting with $x_* \in \mathbb{R}^d$, $\omega_* \in \mathbb{R}$. However, for the MIMO setting with $\omega_* \in \mathbb{R}^{n_\omega}$, the correlation between different function values associated with different output channels should be considered. To do so, the covariance function $\mathcal{K}(x, x')$ is replaced by a *covariance matrix function*

$$\mathcal{K}(x, x') = \begin{bmatrix} \mathcal{K}_{11}(x, x') & \cdots & \mathcal{K}_{1n_\omega}(x, x') \\ \vdots & \ddots & \vdots \\ \mathcal{K}_{n_\omega 1}(x, x') & \cdots & \mathcal{K}_{n_\omega n_\omega}(x, x') \end{bmatrix}, \quad (22)$$

where $\mathcal{K}_{ij}$ denotes the covariance between the (i)-th and (j)-th output channels. However, by assuming that the function values $f_1(x_*), \ldots, f_{n_\omega}(x_*)$ are conditionally independent given an input x_*, then the off-diagonal entries in (22) become 0 [5]. So, in case of multivariate prediction with a deterministic test input x_*, n_ω independent GP models can be trained with the same training inputs X, but with different training outputs $W_i = [\omega_i(1) \cdots \omega_i(N)]^\top$, $i = 1, \ldots, n_\omega$.

3.2 Connection to functional estimate in RKHS

The connection between Gaussian processes and RKHS is given in [30], where it has been shown that the minimum variance estimate $\hat{f}$ in (17) can be obtained as the solution of the following Tikhonov-type variational problem

$$\hat{f} = \underset{f \in \mathcal{H}_\mathcal{K}}{\operatorname{argmin}} \sum_{k=1}^{N} \|\omega(k) - f(x(k))\|_2^2 + \sigma_\epsilon^2 \|f\|_\mathcal{K}^2, \quad (23)$$

where $\mathcal{H}_\mathcal{K}$ is the RKHS associated with $\mathcal{K}$ and $\|\cdot\|_\mathcal{K}$ is the norm defined in $\mathcal{H}_\mathcal{K}$. Note that the solution of (23) can be easily obtained using the representer theorem [30]. Resorting to the RKHS point of view is useful to provide information about the properties of the estimator, e.g., consistency, by characterizing the RKHS space $\mathcal{H}_\mathcal{K}$, which is considered to be the hypothesis space or "model set" of the functions to be estimated.

[5] The covariance matrix function in (22) is assumed to be diagonal for the sake of simplicity, but such an assumption can be removed as it is not an implicitly needed assumption of the method. Assuming a full covariance matrix would significantly increase the complexity of the hyperparameters optimization problem as this results in the necessity of the simultaneous estimation of all channels.

4 Bayesian Identification of LPV-IO models

In this section, the Gaussian regression framework of Section 3 will be applied to the estimation of the LPV-IIR representation (11). First, the identification of (11) will be formulated as shown in Section 3. Second, an appropriate kernel $\mathcal{K}$ will be designed. Finally, the estimation of the unknown structural dependencies will be introduced.

The covariance on the noise $e(k)$ is assumed to be diagonal, i.e., $\Sigma_\mathrm{e} = \operatorname{diag}([\,\sigma_1^2 \cdots \sigma_{n_y}^2\,])$. Hence, the $(\lambda)-$th output channel of (11) can be written as:

$$\begin{aligned} [y(k)]_\lambda &= f_\lambda\big(x^{(k)}\big) + [e(k)]_\lambda \\ &= \underbrace{\sum_{\gamma=1}^{n_y} \underbrace{\sum_{i=1}^{\infty} [(h_{y_i} \diamond p)_k]_{\lambda,\gamma}\, q^{-i}\, [y(k)]_\gamma}_{f^\mathrm{y}_{\lambda,\gamma}\left(x^{(k)}\right)}}_{f^\mathrm{y}_\lambda\left(x^{(k)}\right)} \\ &\quad + \underbrace{\sum_{\gamma=1}^{n_u} \underbrace{\sum_{i=0}^{\infty} [(h_{u_i} \diamond p)_k]_{\lambda,\gamma}\, q^{-i}\, [u(k)]_\gamma}_{f^\mathrm{u}_{\lambda,\gamma}\left(x^{(k)}\right)}}_{f^\mathrm{u}_\lambda\left(x^{(k)}\right)} + [e(k)]_\lambda\,, \end{aligned} \quad (24)$$

where $[\bullet]_{\lambda,\gamma}$ denotes the (λ,γ)-th element of a matrix and $[\bullet]_\lambda$ is the (λ)-th element of a vector. Moreover, f^y_λ, f^u_λ represent the sub-predictors whose sum forms the one-step-ahead predictor, denoted by f_λ. Finally, $f^\mathrm{y}_{\lambda,\gamma}$, $f^\mathrm{u}_{\lambda,\gamma}$, under the stability assumption of the data-generating system, represent convergent IIRs. It is worth to remind the reader that $x^{(k)} = \big\{u^{(k)}, p^{(k)}, y^{(k-1)}\big\}$ is the shorthand notation of the past measurements, e.g., $u^{(k)} = \{u(\tau)\}_{\tau \le k}$.

From (24), the identification of the one-step-ahead predictor f_λ can be considered as a standard GP regression problem. More specifically, by following the Bayesian setting within the GP framework detailed in Section 3, the IIRs $f^\mathrm{y}_{\lambda,\gamma}$, $f^\mathrm{u}_{\lambda,\gamma}$ are assumed to be realizations of zero-mean Gaussian random fields, i.e.,

$$f^\mathrm{y}_{\lambda,\gamma} \sim \mathcal{GP}\big(0, \mathcal{K}^\mathrm{y}_{\lambda,\gamma}\big), \qquad f^\mathrm{u}_{\lambda,\gamma} \sim \mathcal{GP}\big(0, \mathcal{K}^\mathrm{u}_{\lambda,\gamma}\big), \quad (25)$$

characterized by the covariance functions $\mathcal{K}^{\mathrm{y}}_{\lambda,\gamma}$, $\mathcal{K}^{\mathrm{u}}_{\lambda,\gamma}$. In the Bayesian setting, these covariance functions encode the prior knowledge and assumptions about the to-be-estimated functional dependency. Hence, in order to successfully identify the data-generating system, the kernel function needs to be appropriately designed for the problem at hand. The design of the kernel function is concerned with choosing a parameterized form of it in terms of the hyperparameters, see Section 3 for more details.

4.1 Kernel design for LPV-IO models

First of all, within the LPV framework, the relation between the input and output is assumed to be linear, but with coefficients ($a_i, \dots, d_i$ in (2)-(3)) that are dependent on p. In many situations, the functional dependencies consist of a p-independent (LTI) part and a p-dependent part, which should be represented in the kernel. In addition, the kernel should guarantee the stability of the one-step-ahead predictors. To conclude, the kernel functions $\mathcal{K}^{\mathrm{y}}_{\lambda,\gamma}$, $\mathcal{K}^{\mathrm{u}}_{\lambda,\gamma}$ should be designed to

K1 Describe possible structural dependencies on p.
K2 Encode asymptotic stability of the predictor.
K3 Take the LTI part into account.

Next, we show how to design a kernel satisfying K1-K3 to identify the MIMO LPV-BJ system by employing the one-step-ahead predictor (14). From (24) and under the GP prior (25) of the IIRs for the output channel λ, we collect the data in the vector $Y_\lambda = \left[[y(1)]_\lambda \cdots [y(N)]_\lambda\right]^\top$. Under the assumption that $\mathbb{E}\{f^{\mathrm{y}}_{\lambda,\gamma} f^{\mathrm{u}}_{\lambda,\gamma'}\} = 0$ for all $\gamma \in \{1,\dots,n_{\mathrm{y}}\}$, and $\gamma' \in \{1,\dots,n_{\mathrm{u}}\}$, the covariance of the output channel λ is given by $\mathbb{E}\{Y_\lambda Y_\lambda^\top\}$ and its $(\xi,\eta)-$th entry is described as follows [6]

$$\mathbb{E}\left\{[y(\xi)]_\lambda [y(\eta)]_\lambda\right\} = \sum_{\gamma=1}^{n_{\mathrm{y}}} \mathcal{K}^{\mathrm{y}}_{\lambda,\gamma}\big(x^{(\xi)}, x^{(\eta)}\big) + \sum_{\gamma=1}^{n_{\mathrm{u}}} \mathcal{K}^{\mathrm{u}}_{\lambda,\gamma}\big(x^{(\xi)}, x^{(\eta)}\big) + \sigma_\lambda^2, \quad (26)$$

where $\mathcal{K}^{\mathrm{y}}_{\lambda,\gamma}$ is defined as ($\mathcal{K}^{\mathrm{u}}_{\lambda,\gamma}$ is defined in a similar fashion)

$$\mathcal{K}^{\mathrm{y}}_{\lambda,\gamma}\big(x^{(\xi)}, x^{(\eta)}\big) = \mathbb{E}\left\{f^{\mathrm{y}}_{\lambda,\gamma}\big(x^{(\xi)}\big) f^{\mathrm{y}}_{\lambda,\gamma}\big(x^{(\eta)}\big)\right\} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\Big([y(\xi-i)]_\gamma \mathcal{Q}^{\mathrm{y}}_{\lambda,\gamma}\big(p^{(\xi,i)}, p^{(\eta,j)}\big)[y(\eta-j)]_\gamma\Big), \quad (27)$$

[6] Note that, such an assumption implies that the IIRs $f^{\mathrm{y}}_{\lambda,\gamma}$ and $f^{\mathrm{u}}_{\lambda,\gamma'}$ are independent. Hence, the coefficient sequence h_{y_i} and h_{u_i} in (24) are independent, which simplifies the underlying problem by removing the cross-correlation terms.

with $p^{(\xi,i)}$ being the vector of past scheduling values $p^{(\xi,i)} = \left[p^\top(\xi) \cdots p^\top(\xi-i)\right]^\top$ and

$$\mathcal{Q}^{\mathrm{y}}_{\lambda,\gamma}\big(p^{(\xi,i)}, p^{(\eta,j)}\big) = \mathbb{E}\Big\{\big[(h_{y_i} \diamond p)_\xi\big]_{\lambda,\gamma}\big[(h_{y_j} \diamond p)_\eta\big]_{\lambda,\gamma}\Big\}.$$

We need to parameterize the kernel $\mathcal{Q}^{\mathrm{y}}_{\lambda,\gamma}$ to encode the prior knowledge, i.e., K1-K3. Interestingly, due to the linearity of the addressed system class, ideas of kernel design for LTI systems, e.g., *diagonal* (DI) kernel [20], *tuned/correlated* (TC) kernel [38], and *orthonormal basis function* (OBFs) based kernels [39,40,41], can be extended to the considered LPV setting in this paper. More specifically, in the following discussion, we show how to design DI-like and TC-like kernels for LPV systems.

To describe the underlying structural dependency on p represented in terms of the matrix coefficient functions h_{y_i}, h_{u_i}, i.e., Item K1, any positive semidefinite kernel, e.g., polynomial, spline, etc., can be used, depending on the problem at hand. In our case, h_{y_i}, h_{u_i} are assumed to be smooth matrix coefficient functions and, hence, an RBF kernel can be used to describe such dependency. Secondly, to encode asymptotic stability of the predictor or equivalently to guarantee the convergence of the estimated IIR, i.e., Item K2, a decay term needs to be included that models the vanishing influence of the past input-scheduling-output pairs on the predicted output, i.e., the effect of $(y_{k-t}, u_{k-t}, p_{k-t})$ over y_k decreases as $t \to \infty$. Thirdly, to take the LTI part into account, i.e., Item K3, the kernel function is composed of two parts, namely a part to describe the LTI dynamics and a part to describe the $p-$dependent dynamics. In view of the above discussion, a general formulation of a kernel function that encodes the prior knowledge about the underlying IIR $f^{\mathrm{y}}_{\lambda,\gamma}$, i.e., K1-K3 is

$$\mathcal{Q}^{\mathrm{y}}_{\lambda,\gamma}\big(p^{(\xi,i)}, p^{(\eta,j)}\big) = \underbrace{\mathcal{Q}^{\mathrm{y,lin}}_{\lambda,\gamma}(i,j)}_{\text{linear part}} + \underbrace{\mathcal{Q}^{\mathrm{y,p}}_{\lambda,\gamma}\big(p^{(\xi,i)}, p^{(\eta,j)}\big)}_{p\text{-dependent part}}, \quad (28)$$

with

$$\mathcal{Q}^{\mathrm{y,lin}}_{\lambda,\gamma}(i,j) = \alpha_1^2 r_1(\alpha_2), \quad (29a)$$

$$\mathcal{Q}^{\mathrm{y,p}}_{\lambda,\gamma}\big(p^{(\xi,i)}, p^{(\eta,j)}\big) = \alpha_3^2 r_2(\alpha_4) \exp\left(-\frac{\|p^{(\xi,i)} - p^{(\eta,j)}\|_2^2}{[\varsigma_{\mathrm{y}}(i,j)]^2_{\lambda,\gamma}}\right), \quad (29b)$$

where α_1, α_3 are scaling parameters and $r_1(\alpha_2), r_2(\alpha_4) \to 0$ as $i, j \to \infty$ to describe the decay rate, i.e., to ensure that the IIR is convergent. The RBF kernel in (29b) describes the possible structural dependency on p, where $[\varsigma_{\mathrm{y}}(i,j)]_{\lambda,\gamma}$ is the width of the RBF. The kernel $\mathcal{Q}^{\mathrm{u}}_{\lambda,\gamma}$ for $f^{\mathrm{u}}_{\lambda,\gamma}$ is similarly defined.

Due to the assumed existence of the one-step-ahead predictor (14) for (1), i.e., the assumed convergence of the involved IIRs, the sub-predictors $f^{\mathrm{y}}_{\lambda,\gamma}$, $f^{\mathrm{u}}_{\lambda,\gamma}$ in

(24) asymptotically decay to the zero function with the higher order terms of the expansion becoming insignificant. Hence, (24) can be arbitrarily well approximated by truncating the corresponding infinite sum:

$$\begin{aligned}\bar{f}_\lambda\big(x^{(k)}\big) &= [\bar{y}(k \mid k-1)]_\lambda \\ &= \underbrace{\sum_{\gamma=1}^{n_\mathrm{y}} \bar{f}^\mathrm{y}_{\lambda,\gamma}\big(x^{(k)}\big)}_{\bar{f}^\mathrm{y}_\lambda\left(x^{(k)}\right)} + \underbrace{\sum_{\gamma=1}^{n_\mathrm{u}} \bar{f}^\mathrm{u}_{\lambda,\gamma}\big(x^{(k)}\big)}_{\bar{f}^\mathrm{u}_\lambda\left(x^{(k)}\right)}, \end{aligned} \tag{30}$$

with

$$\bar{f}^\mathrm{y}_{\lambda,\gamma}\big(x^{(k)}\big) = \sum_{i=1}^{n_{\mathrm{f_y}}} \left[(h_{y_i} \diamond p)_k\right]_{\lambda,\gamma} q^{-i} \left[y(k)\right]_\gamma, \tag{31a}$$

$$\bar{f}^\mathrm{u}_{\lambda,\gamma}\big(x^{(k)}\big) = \sum_{i=0}^{n_{\mathrm{f_u}}} \left[(h_{u_i} \diamond p)_k\right]_{\lambda,\gamma} q^{-i} \left[u(k)\right]_\gamma, \tag{31b}$$

where $n_{\mathrm{f_y}}$ and $n_{\mathrm{f_u}}$ are sufficiently large to capture the dominant dynamic behavior of the system. As a result, the covariance function (27) is truncated as

$$\bar{\mathcal{K}}^\mathrm{y}_{\lambda,\gamma}\big(\bar{x}^{(\xi)}, \bar{x}^{(\eta)}\big) = \mathbb{E}\left\{\bar{f}^\mathrm{y}_{\lambda,\gamma}\big(\bar{x}^{(\xi)}\big)\bar{f}^\mathrm{y}_{\lambda,\gamma}\big(\bar{x}^{(\eta)}\big)\right\} = \sum_{i=1}^{n_{\mathrm{f_y}}}\sum_{j=1}^{n_{\mathrm{f_y}}}\Big(\left[y(\xi-i)\right]_\gamma \mathcal{Q}^\mathrm{y}_{\lambda,\gamma}\big(p^{(\xi,i)}, p^{(\eta,j)}\big)\left[y(\eta-j)\right]_\gamma\Big), \tag{32}$$

where $\bar{x}^{(\xi)}$ is the set of truncated past measurements, i.e., $\bar{x}^{(\xi)} = \left\{u^{(\xi,n_{\mathrm{f_u}})}, p^{(\xi,n_\mathrm{f})}, y^{(\xi,n_{\mathrm{f_y}})}\right\}$ and $n_\mathrm{f} = \max(n_{\mathrm{f_y}}, n_{\mathrm{f_u}})$ is the maximum truncation order. The truncated covariance for the input IIR $\bar{\mathcal{K}}^\mathrm{u}_{\lambda,\gamma}\left(\bar{x}^{(\xi)}, \bar{x}^{(\eta)}\right)$ is defined similar to (32), but with truncation order $n_{\mathrm{f_u}}$.

For the truncated kernel representation (32) the number of hyperparameters is (including noise variance [7])

$$\big(n_\mathrm{y}(n_{\mathrm{f_y}}+4) + n_\mathrm{u}(n_{\mathrm{f_u}}+4)+1\big)\, n_\mathrm{y}, \tag{33}$$

which grows rapidly in n_y, n_u, $n_{\mathrm{f_y}}$, and $n_{\mathrm{f_u}}$, potentially leading to computational problems. However, further assumptions could be made if necessary to reduce the number of hyperparameters:

Assumption 5 *For the output channel λ: all the IIRs associated with the sub-predictor f^y_λ, i.e., $f^\mathrm{y}_{\lambda,\gamma}$ for $\gamma = 1,\ldots,n_\mathrm{y}$, share the same decay rate, i.e., they share the same parameterization for r_1, r_2, with different scaling parameters α_1, α_3. The same assumption holds true for the IIRs associated with the sub-predictor f^u_λ, i.e., $f^\mathrm{u}_{\lambda,\gamma}$ for $\gamma = 1,\ldots,n_\mathrm{u}$.*

[7] Note that the noise variance σ^2_λ is not known a priori. One possible way to identify σ^2_λ is to regard it as an additional hyperparameter and estimate it together with the other hyperparameters by maximizing the marginal likelihood.

Assumption 6 *For every IIR $f^\mathrm{y}_{\lambda,\gamma}$ (24), the kernel width is assumed to be the same for all coefficient functions within this IIR, i.e., $[\varsigma_\mathrm{y}(i,j)]^2_{\lambda,\gamma}$ in (29) are equal for all i,j. The same holds true for $f^\mathrm{u}_{\lambda,\gamma}$.*

Under Assumptions 5 and 6, the total number of unknown hyperparameters is reduced to

$$\left(3\left(n_\mathrm{y}+n_\mathrm{u}\right)+5\right) n_\mathrm{y}. \tag{34}$$

For example, in case $n_\mathrm{y}=2$, $n_\mathrm{u}=2$, $n_{\mathrm{f_y}}=n_{\mathrm{f_u}}=10$ the original number of the hyperparameters (33) that are needed to be estimated is 114. However, by following Assumptions 5 and 6, this number is reduced to 34.

By appropriately choosing the functions $r_1(\cdot)$ and $r_2(\cdot)$, different relations between the coefficient functions associated with different time instants can be represented. In this paper, we consider two cases: i) a non-correlated, DI-like representation of the resulting kernel in (32):

$$\mathcal{Q}^\mathrm{y}_{\lambda,\gamma}\big(p^{(\xi,i)}, p^{(\eta,j)}\big) = \left[\alpha_1^2\right]_{\lambda,\gamma}\left([\alpha_2]_\lambda\right)^i \delta_{i,j} + \left[\alpha_3^2\right]_{\lambda,\gamma}\left([\alpha_4]_\lambda\right)^i \exp\left(-\frac{\|p^{(\xi,i)}-p^{(\eta,j)}\|_2^2}{[\varsigma_\mathrm{y}(i,j)]^2_{\lambda,\gamma}}\right)\delta_{i,j}, \tag{35}$$

where $\delta_{i,j}$ is the Kronecker delta function w.r.t. (i,j); ii) a correlated, TC-like formulation can be given by taking into account the correlation between the coefficient functions associated with different time instants:

$$\mathcal{Q}^\mathrm{y}_{\lambda,\gamma}\big(p^{(\xi,i)}, p^{(\eta,j)}\big) = \left[\alpha_1^2\right]_{\lambda,\gamma}\left([\alpha_2]_\lambda\right)^{\max(i,j)} + \left[\alpha_3^2\right]_{\lambda,\gamma}\left([\alpha_4]_\lambda\right)^{\max(i,j)} \exp\left(-\frac{\|p^{(\xi,i)}-p^{(\eta,j)}\|_2^2}{[\varsigma_\mathrm{y}(i,j)]^2_{\lambda,\gamma}}\right), \tag{36}$$

where $\left[\alpha_1^2\right]_{\lambda,\gamma}$, $\left[\alpha_3^2\right]_{\lambda,\gamma}$ are scaling parameters of the LTI and the $p-$dependent part of the $(\lambda,\gamma)-$th IIR, respectively, and $[\alpha_2]_\lambda$, $[\alpha_4]_\lambda$ are the parameters that determine the decay rate of the IIRs associated with the (λ)-th output channel.

Remark 7 *Regarding Assumptions 5 and 6, they can be seen as selections that can be made by the user to alleviate the computational burden and decrease the chances of ending in an unwanted local minimum. These two assumptions are a valid starting point. Nonetheless, the proposed kernel structure facilitates the incorporation of additional prior knowledge, which might be available to the user.*

4.2 Estimation of the predictor from data

With the kernel designed, the estimation of the predictor f_λ in (24) by the truncated model (30) from a

given data set $\mathcal{D}_N = \{y(k), u(k), p(k)\}_{k=1}^N$ can now be discussed. This is accomplished by minimizing the ℓ_2 norm of the prediction-error $\varepsilon(k) = y(k) - \bar{y}(k \mid k-1)$. Let θ_λ denote the vector of unknown hyperparameters related to the output channel λ. Let $Y = \big[y^\top(1) \cdots y^\top(N)\big]^\top$, $Y' = \big[y^\top(n_\mathrm{f}+1) \cdots y^\top(N)\big]^\top$, $Y'_\lambda = [[y(n_\mathrm{f}+1)]_\lambda \cdots [y(N)]_\lambda]^\top$, $U = \big[u^\top(1) \cdots u^\top(N)\big]^\top$ and $P = \big[p^\top(1) \cdots p^\top(N)\big]^\top$. Under the Gaussian regression framework, briefly introduced in Section 3, the posterior distribution of f_λ is also Gaussian. Hence, the minimum variance estimate $\hat{f}_\lambda$ of the predictor for output channel λ, i.e., $\bar{f}_\lambda$ in (30), conditioned on a fixed θ_λ can be written as [8]

$$\hat{f}_\lambda(\bullet) = \mathbb{E}\big\{\bar{f}_\lambda(\bullet) \,|\, Y, U, P, \theta_\lambda\big\} = \sum_{k=n_\mathrm{f}+1}^{N} c_{k-n_\mathrm{f}} \bar{\mathcal{K}}_\lambda\big(\bullet, \bar{x}^{(k)}\big), \tag{37}$$

where

$$\bar{\mathcal{K}}_\lambda\big(\bullet, \bar{x}^{(k)}\big) = \sum_{\gamma=1}^{n_\mathrm{y}} \bar{\mathcal{K}}^\mathrm{y}_{\lambda,\gamma}\big(\bullet, \bar{x}^{(k)}\big) + \sum_{\gamma=1}^{n_\mathrm{u}} \bar{\mathcal{K}}^\mathrm{u}_{\lambda,\gamma}\big(\bullet, \bar{x}^{(k)}\big),$$

and c_{k-n_f} is the $(k-n_\mathrm{f})$-th component of the vector

$$c = \big(\Sigma_y(\theta_\lambda)\big)^{-1} Y'_\lambda,$$

with $\Sigma_y(\theta_\lambda) \in \mathbb{R}^{N-n_\mathrm{f} \times N-n_\mathrm{f}}$ being invertible and its entries are given by

$$\big[\Sigma_y(\theta_\lambda)\big]_{i,j} = \bar{\mathcal{K}}_\lambda\big(\bar{x}^{(n_\mathrm{f}+i)}, \bar{x}^{(n_\mathrm{f}+j)}\big) + \sigma_\lambda^2 \delta_{i,j}.$$

In this work, we follow the approach of maximizing the marginal likelihood of the output w.r.t. θ_λ under $\mathcal{D}_N$ [24]. More specifically, the log-marginal likelihood of the observations Y'_λ given U, P, θ_λ:

$$\log \mathscr{P}(Y'_\lambda \mid U, P, \theta_\lambda) = -\frac{N}{2}\log(2\pi) - \frac{1}{2} Y'^\top_\lambda \big(\Sigma_y(\theta_\lambda)\big)^{-1} Y'_\lambda - \frac{1}{2}\log|\Sigma_y(\theta_\lambda)|. \tag{38}$$

Then, an estimate for θ_λ is obtained by maximizing this log-marginal likelihood or equivalently:

$$\hat{\theta}_\lambda = \underset{\theta_\lambda}{\operatorname{argmin}} \; -\log \mathscr{P}(Y'_\lambda \mid U, P, \theta_\lambda). \tag{39}$$

According to the empirical Bayes approach [42], the minimum variance estimate of the predictor, i.e., $\hat{f}_\lambda$ in (37), is obtained by substituting θ_λ by its estimate $\hat{\theta}_\lambda$ from (39). Moreover, the estimate $\hat{f}_\lambda$ is normally distributed and the estimate can be associated with a probability level or confidence region to provide a quantification for the quality of the estimate. This can be achieved by computing the variance of the posterior distribution of the prediction based on (18):

$$\operatorname{var}\big(\hat{f}_\lambda(\bullet) \mid Y, U, P, \theta_\lambda\big) = \bar{\mathcal{K}}_\lambda(\bullet, \bullet) - \psi_\lambda^\top \big(\Sigma_y(\hat{\theta}_\lambda)\big)^{-1} \psi_\lambda, \tag{40}$$

where $\psi_\lambda = \big[\bar{\mathcal{K}}_\lambda(\bullet, \bar{x}^{(n_\mathrm{f}+1)}) \cdots \bar{\mathcal{K}}_\lambda(\bullet, \bar{x}^{(N)})\big]^\top$.

4.3 The individual coefficient function estimates

Next, it is explained how the individual coefficient functions h_{y_i}, h_{u_i} can be calculated from the estimated one-step-ahead predictor.

From [23, Theorem 4], it can be seen that the kernels $\mathcal{Q}^\mathrm{y}_{\lambda,\gamma}$, $\mathcal{Q}^\mathrm{u}_{\lambda,\gamma}$ induce mutually orthogonal subspaces $\mathcal{H}_{\mathcal{Q}^\mathrm{y}_{\lambda,\gamma}}$, $\mathcal{H}_{\mathcal{Q}^\mathrm{u}_{\lambda,\gamma}}$, the associated RKHS with $\mathcal{Q}^\mathrm{y}_{\lambda,\gamma}$, $\mathcal{Q}^\mathrm{u}_{\lambda,\gamma}$, respectively. As a result, the minimum variance estimate of the individual coefficient functions, i.e., $[\hat{h}_{y_i}]_{\lambda,\gamma}$, $[\hat{h}_{u_j}]_{\lambda,\gamma}$, can be obtained as the orthogonal projection of $\hat{f}_\lambda \in \mathcal{H}_{\bar{\mathcal{K}}_\lambda}$, where $\mathcal{H}_{\bar{\mathcal{K}}_\lambda}$ is the RKHS associated with $\bar{\mathcal{K}}_\lambda$, onto $\mathcal{H}_{\mathcal{Q}^\mathrm{y}_{\lambda,\gamma}}$, $\mathcal{H}_{\mathcal{Q}^\mathrm{u}_{\lambda,\gamma}}$, respectively, as follows:

$$\begin{aligned}\big[(\hat{h}_{y_i} \diamond \bullet)\big]_{\lambda,\gamma} &= \mathbb{E}\Big\{\big[(h_{y_i} \diamond \bullet)\big]_{\lambda,\gamma} \mid Y, U, P, \theta\Big\} \\ &= \sum_{k=n_\mathrm{f}+1}^{N} c_{k-n_\mathrm{f}} [y(k-i)]_\gamma \mathcal{Q}^\mathrm{y}_{\lambda,\gamma}\big(\bullet, p^{(k,i)}\big),\end{aligned} \tag{41}$$

and the estimated variance of the corresponding posterior distribution is given by

$$\begin{aligned}\operatorname{var}\Big(\big[(h_{y_i} \diamond \bullet)\big]_{\lambda,\gamma} \mid Y, U, P, \theta\Big) = \\ \mathcal{Q}^\mathrm{y}_{\lambda,\gamma}(\bullet, \bullet) - \big(\psi^\mathrm{y}_{\lambda,\gamma}\big)^\top \big(\Sigma_y(\theta)\big)^{-1} \psi^\mathrm{y}_{\lambda,\gamma},\end{aligned} \tag{42}$$

where

$$\begin{aligned}\psi^\mathrm{y}_{\lambda,\gamma} = \Big[\mathcal{Q}^\mathrm{y}_{\lambda,\gamma}\big(\bullet, p^{(n_\mathrm{f}+1,i)}\big)\,[y(n_\mathrm{f}-i+1)]_\gamma \quad \cdots \\ \mathcal{Q}^\mathrm{y}_{\lambda,\gamma}\big(\bullet, p^{(N,i)}\big)\,[y(N-i)]_\gamma\Big]^\top.\end{aligned} \tag{43}$$

Eq. (42) provides a quantification of the uncertainty of estimated coefficient functions by highlighting the regions that suffer from poor excitation. Hence, such information can be used to further improve the estimate. The minimum variance estimate of $(h_{u_j} \diamond \bullet)$ and its associated covariance can be formulated in a similar fashion.

[8] The summation in (37) starts from $n_\mathrm{f}+1$ to avoid the estimation of the unknown initial conditions. In case n_f is not much smaller than N, it is recommended to collect more data.

5 Estimate of the Process and Noise Models

In this section, we present a novel method to construct a nonparametric estimate of the process (9a) and noise (9b) model based on the estimated one-step-ahead predictor in of Section 4.3. Such a representation of the model estimates is more useful for control synthesis [43] or analysing the dynamics of the deterministic part (1a).

To this end, we use the truncated nonparametric estimates $\hat{h}_{y_i}$, $\hat{h}_{u_j}$ (41) of h_{y_i}, h_{u_j} of order $n_{\mathrm{f_y}}$, $n_{\mathrm{f_u}}$ (described in Section 4.3) to calculate estimates of $g_{0,i}$ and $h_{0,i}$ in (5). To start, note that $\hat{h}_{y_i}$ in (41) for a finite order estimate corresponds to

$$(\hat{H}^{\dagger}(q^{-1}) \diamond \bullet)_k = I - \sum_{i=1}^{n_{\mathrm{f_y}}} (\hat{h}_{y_i} \diamond \bullet)_k q^{-i}. \tag{44}$$

To recover an estimate of $H_0(q^{-1})$, the left inverse of $\hat{H}^{\dagger}(q^{-1})$ in (44) needs to be calculated. By applying (6) on (44), after some algebra, the inverse relation boils down to

$$(\hat{h}_j \diamond \bullet)_k = \sum_{i=1}^{\min(n_{\mathrm{f_y}},j)} (\hat{h}_{y_i} \diamond \bullet)_k (\hat{h}_{j-i} \diamond \bullet)_{k-i}, \tag{45}$$

for $i \geq 1$ where $\hat{h}_{y_0} = I$. Recursion (45), in practice, is computed up to the truncation order $n_{\mathrm{f_y}}$, with a residual truncation error $\epsilon_{n_{\mathrm{f_y}}}$ expressed by (8). Key of the recursion (45) is that it is solely dependent upon $\hat{h}_{y_i}$. Therefore, the noise filter can be constructed from the estimated one-step-ahead predictor. In a similar fashion, a nonparametric estimate of the process coefficient filters $\hat{g}_i$ is found by left multiplying $\sum \hat{h}_{u_i} q^{-i}$ with $\sum \hat{h}_j q^{-j}$ of (45), i.e., left multiplying $\hat{H}^{\dagger}(q^{-1})\hat{G}(q^{-1})$ with $\hat{H}(q^{-1})$, which, after some algebra, gives

$$(\hat{g}_j \diamond \bullet)_k = (\hat{h}_{u_j} \diamond \bullet)_k + \sum_{i=1}^{\min(n_{\mathrm{f_u}},j)} (\hat{h}_i \diamond \bullet)_k (\hat{g}_{j-i} \diamond \bullet)_{k-i}, \tag{46}$$

with $j \geq 1$ and $(\hat{g}_0 \diamond \bullet)_k = (\hat{h}_{u_0} \diamond \bullet)_k$. In case the functional estimates $\hat{h}_{y_i}$ and $\hat{h}_{u_i}$ are replaced by the true functions h_{y_i} and h_{u_j}, the original coefficient functions $g_{0,i}$ and $h_{0,i}$ of $G_0(q^{-1})$ and $H_0(q^{-1})$ can be recovered with (45) and (46). The estimated individual coefficient functions $\hat{h}_{u_j}$ and $\hat{h}_{y_i}$ are assumed to be independent and normally distributed, however, due to the multiplication of these functions, the estimates $\hat{g}_i$ and $\hat{h}_i$ are not normally distributed any more. The associated variance and PDF of $\hat{g}_i$ and $\hat{h}_i$ becomes rather complex for high i due to the high order of multiplications and, therefore, calculating a confidence bound by numerical integration is required. Yet, combining the nonparametric identification approach of Section 4.2 with the nonparametric realization will lead to a relatively simple identification approach to find a nonparametric estimate of $G_0(q^{-1})$ and $H_0(q^{-1})$.

6 Numerical Simulation

In this section, the performance of the presented nonparametric approach for the identification of LPV-BJ models based on their one-step-ahead predictor is shown by means of an extensive Monte-Carlo study.

6.1 Data-generating system

The considered data-generating system is a MIMO system with $n_{\mathrm{u}} = 2$, $n_{\mathrm{y}} = 2$ and $n_{\mathrm{p}} = 2$ in the form of (1). The LPV-BJ data-generating system [9] has a plant model order of $n_{\mathrm{a}} = n_{\mathrm{b}} = 2$ and a noise model order of $n_{\mathrm{c}} = n_{\mathrm{d}} = 2$.

6.2 Identification setting

The one-step-ahead predictor is estimated using an identification data set with three different sizes $N = \{200, 500, 1000\}$ and the prediction performance of the estimated model is examined on a validation data set that contains $N_{\mathrm{val}} = 200$ samples. The identification and validation data sets are generated with independent realizations of a white noise input signal u with uniform distribution, i.e., $[u(k)]_\lambda \sim \mathcal{U}(-1, 1),\ \lambda = 1, 2$. The scheduling signals are given by

$$[p(k)]_\lambda = 0.4 \sin(0.035k + \frac{\lambda\pi}{5}) + 0.25\lambda + \mathcal{U}(-0.15, 0.15), \quad \text{for } \lambda = 1, 2.$$

The variance of the white noise e driving the noise process is chosen such that the signal-to-noise (SNR) ratio

$$\mathrm{SNR}_{[y]_\lambda} = 10 \log \frac{\sum_{k=1}^{N} [\breve{y}(k)]_\lambda^2}{\sum_{k=1}^{N} [v(k)]_\lambda^2},$$

is 20dB. To analyze the statistical properties of the presented identification approach, a Monte-Carlo study with $N_{\mathrm{MC}} = 100$ runs is carried out. At each run, a new realization of the input u, the scheduling signal p and the noise e are taken.

The predicted output $\hat{\breve{y}}$ from the estimated one-step-ahead predictor model is compared to the true output of

[9] Due to lack of space, the matrix polynomials associated with the plant and noise models are not provided. Description of the LPV-BJ system can be found in [44].

the data-generating system by the *best fit ratio* (BFR)

$$\text{BFR}=\max\left(1-\frac{\frac{1}{N}\sum_{k=1}^{N}\|y(k)-\hat{\bar{y}}(k\mid k-1)\|_2}{\frac{1}{N}\sum_{k=1}^{N}\|y(k)-\bar{\mathrm{y}}\|_2},0\right)\cdot 100\%, \tag{47}$$

where $\bar{\mathrm{y}}$ defines the mean of the true output y. Note that the definition in (47) characterizes the average performance over all output channels.

6.3 Identification results

In this section, the results of the identification of the one-step-ahead predictor of the data-generating system given in Section 6.1 with the identification setting given in Section 6.2 are discussed. The results have been obtained with a truncation order $n_{\mathrm{f_y}} = n_{\mathrm{f_u}} = n_{\mathrm{f}} = 10$. The considered estimators are

(1) Bayesian estimator with DI-like kernel in (35).
(2) Bayesian estimator with TC-like kernel in (36).
(3) Oracle estimator that knows the true underlying nonlinear functional dependencies of h_{y_i}, h_{u_j} in (12a), (12b), respectively. With such knowledge, the Oracle estimator performs an LS estimate of a high-order ARX model with a truncation order of $n = 15$, which is chosen large enough to capture the dynamics of the system. It is worth to mention that the number of parameters to be estimated via LS is $(2n+1)n_{\mathrm{y}} = 62$. In this setting, the truncation order introduces a bias/variance trade-off.

In case of the Bayesian estimator, the hyperparameters are estimated via solving (39). Figure 1 displays the first 50 samples of one realization of the true and the predicted output response on the validation data set by the one-step-ahead predictor estimated with TC-like kernel, truncation order $n_{\mathrm{f}} = 10$ and $N = 1000$ samples based identification data set. In Figure 1, 95% confidence region of the predictor is also displayed to quantify the expected variance of the predictor. The figure shows that the presented Bayesian approach is able to identify an LPV model under general BJ noise conditions. Table 1 gives the sampled mean and the standard deviation (std) of the BFR of the identified predictor over $N_{\mathrm{MC}} = 100$ runs tested on the validation data set. To gain more insights into the performance of the considered estimators, Figure 2 gives a box plot based visualization of the model fit on the validation data set for various sizes of the identification data set. It can be seen from Table 1 and Figure 2 that all the predictors benefit from increase amount of samples in the identification data set, seemingly converging in performance to the Oracle.

6.4 Realization of the process and noise models

In this section, the performance of the realization of the IIRs of the process and noise models is assessed in terms of the coefficient functions $\hat{g}_i$ and $\hat{h}_i$, respectively. We use the scheduling trajectory of the validation data set to compute the BFR of a function by

$$[\text{BFR}]_{l,m}= 100\%\cdot\max\Big(1- \frac{\frac{1}{N}\sum_{k=1}^{N}\|[(g_{0,i}\diamond p)_k]_{l,m}-[(\hat{g}_i\diamond p)_k]_{l,m}\|_2}{\frac{1}{N}\sum_{k=1}^{N}\|[(g_{0,i}\diamond p)_k]_{l,m}-[\bar{\mathbf{g}}_{0,i}]_{l,m}\|_2},0\Big), \tag{48}$$

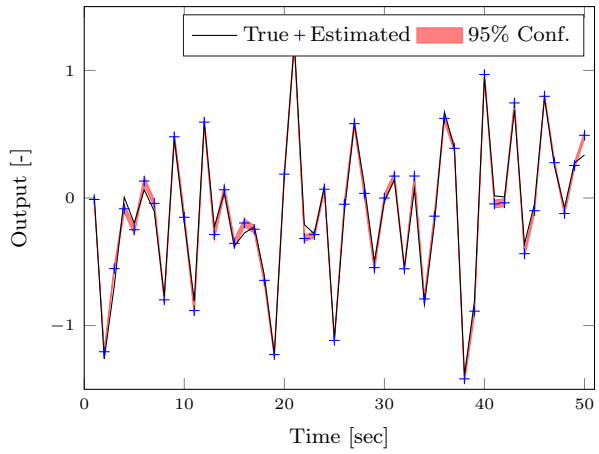
where $[\cdot]_{l,m}$ denotes the (l,m)-th element and $\bar{\mathbf{g}}_{0,i}$ is the average of the coefficient variation along the given scheduling trajectory.

Table 2 shows the mean and standard deviation (std) of the average BFR over all i, j elements in (48) for $N_{\mathrm{MC}} = 100$ runs. The table shows that the BFR decreases for a higher index number, i.e., higher i in $\hat{g}_i$. As the system is asymptotically stable, the coefficient functions decay to the zero function. Hence, for a higher index number, the contribution of the associated coefficient function is lower in the measured output signals. The relative magnitude difference also explains that a higher index number has an increased variance. Also, increasing the number of samples in the identification data set increases the prediction performance and lowers the variance of the estimate in almost all cases, as expected. In line with Table 1, no significant difference is noticed between the DI and TC kernels.

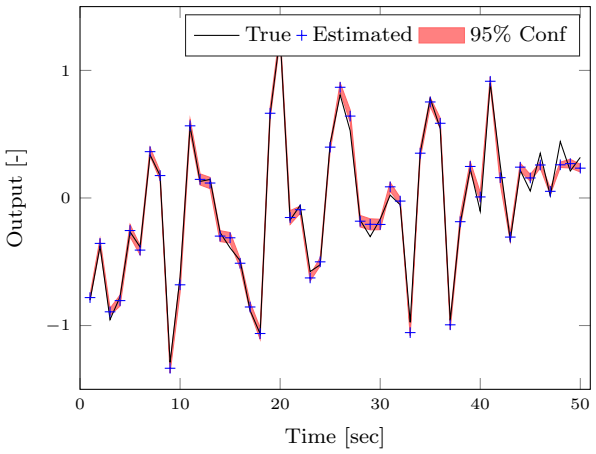
In Figure 3, the true coefficient function $g_{0,1}$ is compared to the sampled mean and sampled standard deviation (std) of the estimated coefficient function $\hat{g}_1$ for $N_{\mathrm{MC}} = 100$ runs. The scheduling trajectory applied has 190 increasing and equally distant samples between the minimum and maximum value of the scheduling signal used in the identification data set. The figure shows that the function estimate is close to the original function shape. In overall, by Table 2 and Figure 3, it is evident that the underlying data-generating parameter-varying matrix functions $g_{0,i}$ and $h_{0,i}$ can be consistently recovered.

7 Conclusion

In this paper, we have presented a nonparametric identification approach for MIMO LPV-BJ models. Similar to the LTI case, it has been shown that the one-step-ahead predictor of such models is a summation of two sub-predictors associated with the input and output signals, where under mild assumptions, these sub-predictors are shown to be convergent IIRs. To cope with issues associated with identifying such models, e.g., parameterization of parameter-varying matrix coefficient functions, a Bayesian nonparametric approach within the GP framework has been adopted. More specifically, the IIRs associated with the predictor are assumed to be realizations



(a) Output #1



(b) Output #2

Fig. 1. The first 50 samples of the true and the predicted output response on the validation data set by the one-step-ahead predictor estimated with TC-like kernel, truncation order $n_{\mathrm{f}} = 10$ and $N = 1000$ samples based identification data set. 95% confidence region of the predictor is also displayed.

Table 1
The sampled mean and std of the BFR of the identified predictor on the validation data set over $N_{\mathrm{MC}} = 100$ Monte-Carol runs.

		$N = 200, n_{\mathrm{f}} = 10$			$N = 500, n_{\mathrm{f}} = 10$			$N = 1000, n_{\mathrm{f}} = 10$		
		DI	**TC**	**Oracle**	**DI**	**TC**	**Oracle**	**DI**	**TC**	**Oracle**
BFR[%]	mean	87.18	87.20	88.11	88.36	88.39	89.01	88.80	88.80	89.32
	std	0.8222	0.8170	0.8150	0.6769	0.6804	0.7054	0.7173	0.7227	0.6828

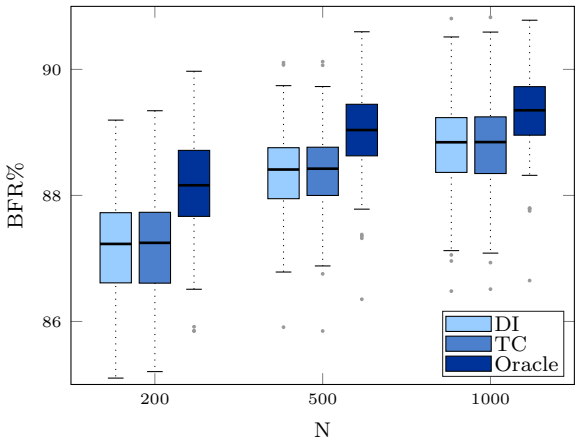


Fig. 2. BFR of the predicted response on the validation data sets using the estimated models with DI, TC kernels and the Oracle estimate under various sizes of the identification data set $N = \{200, 500, 1000\}$.

of zero-mean Gaussian random fields with suitable kernels. One of the main important contributions of this work is to show how to design such kernels in the LPV setting that encode the expected behavior of the predictor:

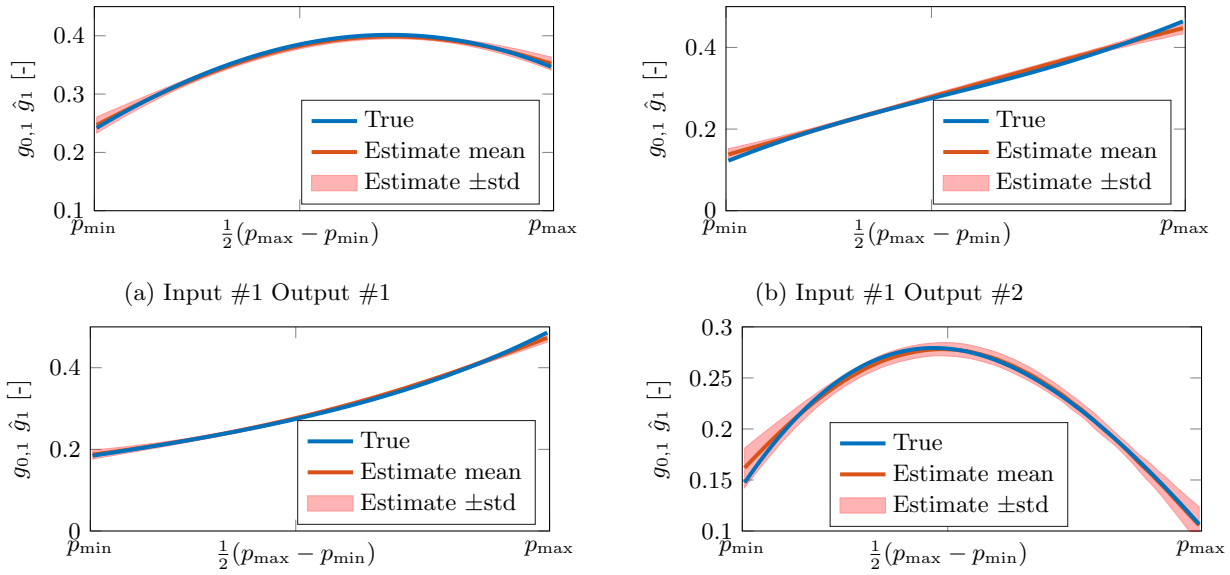
- Ensure stability of the identified predictor;
- Encode possible structural dependencies;

Table 2
The sampled mean and std of the average BFR of the realized process IIR coefficients for $N_{\mathrm{MC}} = 100$ runs. The performance criterion is based on the value of the realized functions on the scheduling trajectory in the validation data set.

BFR[%]		$N = 200$, $n = 10$		$N = 500$, $n = 10$		$N = 1000$, $n = 10$	
		DI	**TC**	**DI**	**TC**	**DI**	**TC**
$\hat{g}_0$	mean	88.90	88.85	93.15	93.18	95.27	95.28
	std	4.613	4.760	2.501	2.567	1.662	1.684
$\hat{g}_1$	mean	74.88	74.97	84.10	84.20	88.82	88.86
	std	6.787	6.928	4.344	4.269	2.883	2.867
$\hat{g}_2$	mean	64.04	65.25	78.06	78.78	84.11	84.25
	std	16.10	14.77	10.66	9.798	7.355	7.189
$\hat{g}_3$	mean	51.92	53.74	65.62	66.48	73.04	73.81
	std	24.50	25.00	17.81	18.41	14.16	14.20
$\hat{g}_4$	mean	17.81	20.27	29.21	30.43	41.74	44.00
	std	17.19	18.20	18.37	19.21	17.79	18.11

- Take into account the LTI part as well as the $p-$dependent part of the model.

Two kernel formulations have been presented: a DI-like and a TC-like kernel. The hyperparameters of the kernels are tuned by maximizing the marginal likelihood



(a) Input #1 Output #1

(b) Input #1 Output #2

(c) Input #2 Output #1

(d) Input #2 Output #2

Fig. 3. The true parameter-varying matrix function $g_{0,1}$ compared with the sample based estimate of the mean and std of the matrix function estimate $\hat{g}_1$ for $N = 1000$, $n_{\mathrm{f}} = 10$, and TC kernel on $N_{\mathrm{MC}} = 100$ runs. The functions are displayed between the minimum and maximum value of the scheduling signal used in the identification data set, i.e., $p_{\min} = [-0.2955 \;\; -0.0248]^\top$ to $p_{\max} = [0.7588 \;\; 1.0296]^\top$.

of the predictor over the observed data. Additionally, a nonparametric realization scheme has been developed to recover the estimates of the process and noise models from the identified predictor.

References

[1] J. S. Shamma and M. Athans, "Analysis of gain scheduled control for nonlinear plants," *IEEE Trans. on Automatic Control*, vol. 35, no. 8, pp. 898–907, 1990.

[2] R. Tóth, *Modeling and Identification of Linear Parameter-Varying Systems*. Springer, 2010.

[3] J. Mohammadpour and C. W. Scherer, *Control of Linear Parameter Varying Systems with Applications*. Springer-Verlag, 2011.

[4] A. A. Bachnas, R. Tóth, A. Mesbah, and J. Ludlage, "A review on data-driven linear parameter-varying modeling approaches: A high-purity distillation column case study," *Journal of Process Control*, vol. 24, pp. 272–285, 2014.

[5] M. G. Wassink, M. van de Wal, C. W. Scherer, and O. Bosgra, "LPV control for a wafer stage: Beyond the theoretical solution," *Control Engineering Practice*, vol. 13, no. 2, pp. 231–245, 2004.

[6] A. Wills and B. Ninness, "System identification of linear parameter varying state-space models," in *Linear Parameter-Varying System Identification: New Developments and Trends*, P. L. dos Santos, C. Novara, D. Rivera, J. Ramos, and T. Perdicoúlis, Eds. World Scientific Publishing, 2011, pp. 295–313.

[7] J. W. van Wingerden and M. Verhaegen, "Subspace identification of bilinear and LPV systems for open- and closed-loop data," *Automatica*, vol. 45, no. 2, pp. 372–381, 2009.

[8] P. L. dos Santos, J. A. Ramos, and J. L. M. de Carvalho, "Identification of LPV systems using successive approximations," in *Proc. of the 47th IEEE Conf. on Decision and Control*, Cancun, Mexico, Dec. 2008, pp. 4509–4515.

[9] M. Sznaier and M. C. Mazzaro, "An LMI approach to control-oriented identification and model (in) validation of LPV systems," *IEEE Trans. on Automatic Control*, vol. 48, pp. 1619–1624, 2003.

[10] R. Tóth, P. S. C. Heuberger, and P. M. J. Van den Hof, "Asymptotically optimal orthonormal basis functions for LPV system identification," *Automatica*, vol. 45, no. 6, pp. 1359–1370, 2009.

[11] ——, "Prediction-error identification of LPV systems: present and beyond," in *Control of Linear Parameter Varying Systems with Applications*, J. Mohammadpour and C. W. Scherer, Eds. Springer, 2012, pp. 27–60.

[12] B. Bamieh and L. Giarré, "Identification of linear parameter varying models," *Int. Journal of Robust and Nonlinear Control*, vol. 12, pp. 841–853, 2002.

[13] 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.

[14] K. Hsu, T. L. Vincent, and K. Poolla, "Nonparametric methods for the identification of linear parameter varying systems," in *Proc. of the Int. Symposium on Computer-Aided Control System Design*, San Antonio, Texas, USA, Sept. 2008, pp. 846–851.

[15] 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, Dec. 2011, pp. 3192–3197.

[16] D. Piga and R. Tóth, "LPV model order selection in an LS-

SVM setting," in *Proc. of the 52nd IEEE Conf. on Decision and Control*, Florence, Italy, Dec. 2013, pp. 4128–4133.

[17] A. Golabi, N. Meskin, R. Tóth, and J. Mohammadpour, "A Bayesian approach for estimation of LPV linear-regression models," in *Proc. of the 53rd IEEE Conf. on Decision and Control*, Los Angeles, CA, USA, Dec. 2014, pp. 2555–2560.

[18] A. Golabi, N. Meskin, R. Tóth, and J. Mohammadpour, "A Bayesian approach for LPV model identification and its application to complex processes," *IEEE Trans. on Control Systems Technology*, no. 99, pp. 1–8, 2017.

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

[20] T. Chen, H. Ohlsson, and L. Ljung, "On the estimation of transfer functions, regularizations and Gaussian processes–revisited," *Automatica*, vol. 48, no. 8, pp. 1525–1535, 2012.

[21] 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.

[22] 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.

[23] 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.

[24] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[25] M. A. H. Darwish, P. B. Cox, G. Pillonetto, and R. Tóth, "Bayesian identification of LPV Box-Jenkins models," in *Proc. of the 54th IEEE Conf. on Decision and Control*, Osaka, Japan, Dec. 2015, pp. 66–71.

[26] R. Tóth, P. S. C. Heuberger, and P. M. J. Van den Hof, "On the discretization of LPV state-space representations," *IET Control Theory & Applications*, vol. 4, pp. 2082–2096, 2010.

[27] S. Wollnack, H. S. Abbas, R. Tóth, and H. Werner, "Fixed-structure LPV-IO controllers: An implicit representation based approach," *Automatica*, vol. 83, pp. 282–289, 2017.

[28] L. Ljung, *System Identification, Theory for the User*, 2nd ed. Prentice-Hall, 1999.

[29] I. Proimadis, H. Bijl, and J. W. van Wingerden, "A kernel based approach for LPV subspace identification," in *Proc. of the 1st IFAC Workshop on Linear Parameter-Varying Systems*, Grenoble, France, Oct. 2015, pp. 97–102.

[30] G. Wahba, *Spline Models for Observational Data*. Siam, 1990, vol. 59.

[31] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[32] D. J. C. MacKay, "Comparison of approximate methods for handling hyperparameters," *Neural computation*, vol. 11, no. 5, pp. 1035–1068, 1999.

[33] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, 2009.

[34] L. Wang and W. R. Cluett, "Use of PRESS residuals in dynamic system identification," *Automatica*, vol. 32, no. 5, pp. 781–784, 1996.

[35] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.

[36] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The annals of Statistics*, vol. 9, no. 6, pp. 1135–1151, 1981.

[37] G. Pillonetto and A. Chiuso, "Tuning complexity in regularized kernel-based regression and linear system identification: The robustness of the marginal likelihood estimator," *Automatica*, vol. 58, pp. 106–117, 2015.

[38] G. Pillonetto and G. De Nicolao, "A new kernel-based approach for linear system identification," *Automatica*, vol. 46, no. 1, pp. 81–93, 2010.

[39] M. A. H. Darwish, G. Pillonetto, and R. Tóth, "Perspectives of orthonormal basis functions based kernels in Bayesian system identification," in *Proc. of the 54th IEEE Conf. on Decision and Control*, Osaka, Japan, Dec. 2015, pp. 2713–2718.

[40] M. A. H. Darwish, J. Lataire, and R. Tóth, "Bayesian frequency domain identification of LTI systems with OBFs kernels," in *Proc. of the 20th IFAC World Congress*, Toulouse, France, July 2017, pp. 6412–6417.

[41] M. A. H. Darwish, G. Pillonetto, and R. Toth, "The quest for the right kernel in bayesian impulse response identification: The use of obfs," *Automatica*, vol. 87, pp. 318 – 329, 2018.

[42] B. P. Carlin and T. A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*. CRC Press, 2000.

[43] S. Formentin, D. Piga, R. Tóth, and S. M. Savaresi, "Direct learning of LPV controllers from data," *Automatica*, vol. 65, pp. 98–110, 2016.

[44] M. A. H. Darwish, P. B. Cox, I. Proimadis, G. Pillonetto, and R. Tóth, "Description of the data generating system utilized in prediction-error identification of LPV systems: A nonparametric Gaussian regression approach," Eindhoven University of Technology, Tech. Rep. TUE-CS-2017-001, 2017.

A Proof of Lemma 2

Based on the notation of Lemma 2, $\dim(y) = \dim(u) = n_{\mathrm{y}}$. For notational ease, define $A_q \coloneqq A(q^{-1})$ and $R_q \coloneqq (I - A(q^{-1}))$. First, we will prove that $A_q^\dagger$ in (6) is a left inverse of A_q, i.e., $A_q^\dagger A_q = I$. So, let us expand the infinite summation $A_q^\dagger A_q$ for a finite order n, as

$$S_q = \sum_{i=0}^{n} R_q^i A_q = A_q + R_q A_q + R_q^2 A_q + \ldots + R_q^n A_q,$$

and multiply from the left with R_q

$$R_q S_q = R_q A_q + R_q^2 A_q + \ldots + R_q^{n+1} A_q.$$

Next, subtract the two previous expansions giving

$$\begin{aligned}(I - R_q) S_q &= A_q - R_q^{n+1} A_q \\ A_q S_q &= A_q - R_q^{n+1} A_q = A_q - A_q R_q^{n+1} \\ S_q &= I - R_q^{n+1}. \end{aligned} \tag{A.1}$$

Note that $R_q^i A_q = A_q R_q^i$, which trivially follows by expanding the terms and factorising them again. As the filter A_q is asymptotically stable, the filter $R_q^n \to 0$ in

terms of convergence to a zero function as $n \to \infty$. Hence, using (A.1)

$$A_q^\dagger A_q = \lim_{n\to\infty} \sum_{i=0}^{n} R_q^i A_q = I - \lim_{n\to\infty} R_q^{n+1} = I. \quad \text{(A.2)}$$

So, taking a finite order n, the approximation error of the inverse is R_q^{n+1}.

Taking a finite order n, the approximation error of the inverse is R_q^{n+1}. It remains to be proven that the ℓ_2 signal norm of the approximation error in (8) exists, i.e., that $\big(I-(A(q^{-1})\diamond p)\big)^{n+1}u$ in (8) qualifies as an ℓ_2 signal. Note that the expression is a polynomial in A with $(A(q^{-1}))^n \to 0$ for $n \to \infty$ as the filter A is asymptotically stable and that u is an ℓ_2 signal. The individual coefficients can be written as

$$\forall k \in \mathbb{Z} : \big((A(q^{-1})\diamond p)_k\big)^n u_k = u_k + \sum_{i=1}^{n_\mathrm{a}} (a_i \diamond p)_k u_{k-i} + \sum_{i=1}^{n_\mathrm{a}} (a_i \diamond p)_k \sum_{j=1}^{n_\mathrm{a}} (a_j \diamond p)_{k-i} u_{k-i-j} + \ldots,$$

where each product is a ℓ_2 signal and, hence, their finite summation is also an ℓ_2 signal. Therefore, the ℓ_2 norm in (8) exists.

Note that A being monic is a crucial property for this derivation to be valid and for nonmonic A the existence of an inverse is not guaranteed.