2015 IEEE 54th Annual Conference on Decision and Control (CDC)
December 15-18, 2015. Osaka, Japan

Bayesian Identification of LPV Box-Jenkins Models*

Mohamed Darwish†, Pepijn Cox†, Gianluigi Pillonetto‡ and Roland Tóth†

***Abstract*— In this paper, we introduce a nonparametric approach in a Bayesian setting to efficiently estimate, both in the stochastic and computational sense, linear parameter-varying (LPV) input-output models under general noise conditions of Box-Jenkins (BJ) type. The approach is based on the estimation of the one-step-ahead predictor model of general LPV-BJ structures, where the sub-predictors associated with the input and output signals are captured as asymptotically stable infinite impulse response models (IIRs). These IIR sub-predictors are identified in a completely nonparametric sense, where not only the coefficients are estimated as functions, but also the whole time evolution of the impulse response is estimated as a function. In this Bayesian setting, the one-step-ahead predictor is modelled as 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. The unknown hyperparameters that parameterize the kernel are tuned using the empirical Bayes approach, i.e., optimization of the marginal likelihood with respect to available data. It is also shown that, in case the predictor has a finite order, i.e., the true system has an ARX noise structure, our approach is able to recover the underlying structural dependencies. The performance of the identification method is demonstrated on LPV-ARX and LPV-BJ simulation examples by means of a Monte Carlo study.**

I. INTRODUCTION

Identification of *linear parameter-varying input-output* (LPV-IO) models has recently received significant attention, e.g., [1], [2], [3], as the *prediction error minimization* (PEM) methods have successfully been extended to the LPV case, providing a well-understood framework for the stochastic interpretation of the estimates [4]. Moreover, the PEM framework is well suited to a large class of noise and plant models, see [5] for an overview. Consider a *single-input single-output* (SISO) data generating LPV system described in discrete-time by the following equations:

$$A_0(p,k,q^{-1})\breve{y}(k) = B_0(p,k,q^{-1})u(k), \tag{1a}$$

$$D_0(p,k,q^{-1})v(k) = C_0(p,k,q^{-1})e(k), \tag{1b}$$

$$y(k) = \breve{y}(k) + v(k), \tag{1c}$$

where $k \in \mathbb{Z}$ is the sample time, q is the time-shift operator, i.e., $q^{-i}u(k) = u(k-i)$, $u : \mathbb{Z} \to \mathbb{R}$, $\breve{y} : \mathbb{Z} \to \mathbb{R}$ and $y : \mathbb{Z} \to \mathbb{R}$ denote the input, noise free output and noisy output signals respectively, $p : \mathbb{Z} \to \mathbb{P}$ is the so-called scheduling variable with compact range $\mathbb{P} \subseteq \mathbb{R}^{n_\mathrm{p}}$, $v : \mathbb{Z} \to \mathbb{R}$ is the coloured noise signal, and $e \in \mathbb{R}$ is a white noise process with $e(k) \sim \mathcal{N}(0, \eta^2)$. The p-dependent operators $A_0(p,k,q^{-1})$ and $B_0(p,k,q^{-1})$ that describe the process part (1a) of the model are polynomials in q^{-1} of degree n_a and n_b respectively:

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

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

where the coefficient functions $a_i(p,k,i) : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}$ and $b_j(p,k,j) : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}$ are shorthand notations for $a_i(p,k,i) = a_i(p(k), \ldots, p(k-i))$ and $b_j(p,k,j) = b_j(p(k), \ldots, p(k-j))$. These functions are assumed to be smooth and bounded functions on $\mathbb{P}$. In a similar fashion, the noise model (1b) relations $C_0(p,k,q^{-1})$ and $D_0(p,k,q^{-1})$ are defined as

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

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

where $c_i(p,k,i) : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}$ and $d_j(p,k,j) : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}$ are the coefficient functions of the monic polynomials in q^{-1} of degree n_c and n_d, respectively.

For such a general noise scenario represented by (1), classical PEM methods lead to a nonlinear optimization problem [4], which is sensitive to local minima and its performance is, for example, dependent on the parametrization of (2)-(3) and also on the initial estimate. A linear parametrization of the coefficient functions $\{a_1, \ldots, d_{n_\mathrm{d}}\}$ in terms of a user specified set of basis functions is often required [5], which either needs a significant a priori knowledge of the underlying system or tedious repetitive execution of the methods to synthesise an acceptable basis.

Alternatively, the so-called nonparametric setting offers an attractive approach to capture the underlying dependencies directly from data without specifying any parametrization in terms of fixed basis functions. The two main streams of LPV nonparametric identification presented in literature are: the *least squares-support vector machine* (LS-SVM) methods, e.g., [6], [7], and the Bayesian setting based approaches, e.g., [8]. The LS-SVM and Bayesian approaches are methods that have roots in *reproducing kernel Hilbert space* (RKHS) theory and admit a ℓ_2-regularization based interpretation [9].

In this work, we chose a Bayesian setting to formulate a nonparmaetric estimator of the one-step-ahead predictor of (1), preserving the generality of the hypothesized noise

* This research has benefited from the financial support of the Student Mission, Ministry of Higher Education, Government of Egypt and the Netherlands Organization for Scientific Research (NWO, grant no.: 639.021.127).

† Mohamed Darwish, Pepijn Cox and Roland Tóth are with the Control Systems Group, Department of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, {m.a.h.darwish,p.b.cox,r.toth}@tue.nl

‡ Gianluigi Pillonetto is with the Information Engineering Department, University of Padova, Padova 35131, Italy, giapi@dei.unipd.it

class and achieving the maximum likelihood property of the PEM without the heavy computational complexity. Formulating the LPV-BJ identification problem in a Bayesian framework is an extension of the nonparametric identification setting introduced in [8], where only LPV-FIR models are considered. For LPV-BJ setting, we define two LPV sub-predictors associated to the input and output signals by asymptotically stable *infinite impulse response* (IIR) models. The associated coefficient functions of these impulse responses are captured in a jointly nonparameteric way, for which the computational complexity remains equivalent to a linear regression problem. The nonparametric framework uses a kernel function that acts as a basis function generator to describe both the functional dependencies and the time-evolution of the impulse responses of the sub-predictors. However, an important issue is how to design the kernel structure parametrized in terms of a few hyperparameters, such that a large generality of dynamics can be captured. To this aim, inspired by the approach presented in [10], we hereby introduce a kernel function which encodes: i) the possible structural dependencies and ii) stability of the predictor by including a decaying term, which models a vanishing influence of the past input-scheduling-output pairs on the predicted output. The small set of hyperparameters related to this kernel can be efficiently estimated by the empirical Bayes approach, e.g., [11], via maximizing the *marginal likelihood* (ML).

This paper is organized as follows: in Section II, the IIR models of the one-step-ahead predictor of the LPV-BJ model (1) is presented. Then, a Bayesian approach to identify the one-step-ahead predictor is formulated in Section III. Section IV shows the performance of the proposed nonparametric identification scheme on two illustrative examples by means of Monte Carlo simulation studies. Finally, the paper is concluded in Section V.

II. General LPV-IO Model

In this section, we develop an equivalent one-step-ahead predictor representation of (1). This form allows the identification of these models under general noise scenarios in the linear regression framework.

A. The IIR form

For well-possessedness of formulating a predictor for (1) both (1a), (1b), and the inverse noise dynamics of (1b)[1] need to correspond to asymptotically stable LPV filters as it is given by Definition 1.

Definition 1: An LPV filter $A_0(p,k,q^{-1})\breve{y}(k) = B_0(p,k,q^{-1})u(k)$, is called globally asymptotically stable, if for all trajectories of $\{u(k), p(k), \breve{y}(k)\}$ satisfying (1), with $u(k) = 0$ for $k \geq 0$ and $p(k) \in \mathbb{P}$, it holds that $\lim_{k\to\infty}|\breve{y}(k)| = 0$. □

If the noise filter (1b) is global asymptotically stable in terms of Definition 1, then the noise v has bounded spectral density, which is a natural assumption in data-driven modeling. Based on this stability notion the system given in (1) is written in terms of IIRs as shown in Theorem 2.

Theorem 2: Let the process dynamics (1a) and noise dynamics (1b) be asymptotically stable according to Definition 1. If, in addition, the inverse noise process is also monic and asymptotically stable, then (1) can be equivalently represented by the following IIR form

$$y(k) = \sum_{i=1}^{\infty} h_{y_i}(p,k,i)q^{-i}y(k) + \sum_{j=0}^{\infty} h_{u_j}(p,k,j)q^{-j}u(k) + e(k), \quad (4)$$

$h_{y_i}(p,k,i) : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}$ and $h_{u_j}(p,k,j) : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}$ are real meromorphic[2] coefficient functions in the scheduling signal p. □

For a proof see the Appendix. The IIR representation makes it possible to formulate a predictor which is only given by the past inputs-scheduling-output signals, as shown next.

B. One-step-ahead predictor

In the *prediction error* (PE) setting, identification of (1) is formulated by using (4) 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, i.e.,

$$y(k|k-1) = \underset{\delta\in\mathbb{R}}{\operatorname{argmin}}\ \mathbb{E}\Big\{\|y(k)-\delta\|_2^2 \mid x^{(k)}\Big\}, \quad (5)$$

where $\mathbb{E}\{\cdot\}$ is the expectation operator and $x^{(k)} = \{u^{(k)}, p^{(k)}, y^{(k-1)}\}$ is the shorthand notation of the past measurements till time k, e.g., $u^{(k)} = \{u(\tau)\}_{\tau\leq k}$. As $e(k)$ is a white noise, straightforward application of the expectation operator gives

$$y(k|k-1) = \sum_{i=1}^{\infty} h_{y_i}(p,k,i)q^{-i}y(k) + \sum_{j=0}^{\infty} h_{u_j}(p,k,j)q^{-j}u(k). \quad (6)$$

The prediction error model (6) is a mixture of the functional dependencies (2) and (3) as given in the Appendix. Hence, direct minimization of the ℓ_2 norm of the PE, i.e., $\varepsilon(k) = y(k) - y(k|k-1)$, leads to a nonlinear optimization problem, prone to local minima. To avoid the nonlinear optimization, it is possible to create a linear-in-the-parameter problem by parametrizing the coefficient functions h_{y_i} and h_{u_j} as a linear combination of a set of basis functions but for real-world problems this often results in large-scale parametrizations [5]. Generally, such approach results in a high estimation variance with no overall gain compared to attempting to solve the nonlinear optimization problem.

III. Bayesian Identification for LPV-IO models

The question is how to utilize the simplicity of the IIR form (4) but overcome the large parametrization and the high parameter variance associated with its identification. A solution can be found in the Bayesian framework, where the

[1]Hence, we assume the existence of a stable left inverse of the corresponding IRR as detailed in the Appendix.

[2]$h : \mathbb{R}^n \to \mathbb{R}$ is a real meromorphic function, if $h = g/f$ with g, f analytic functions and $f \neq 0$.

functional dependencies h_{y_i} and h_{u_j} are estimated nonparametrically and, in addition, a regularization is introduced to keep the variance of the estimates low by a sacrifice of a small amount of estimation bias.

A. Mixture of Gaussian kernels for LPV-IO models

Inspired by the approach introduced in [10], IIRs based model (4) can be considered as a standard Gaussian process regression model, i.e., $y(k) = f(x^{(k)}) + e(k)$, where f is a particular realization from a zero-mean Gaussian random field, i.e., $f \sim \mathcal{GP}(0, \mathcal{K})$, which is the priori assumption in our Bayesian setting. The Gaussianity assumption implies that, in order to define f, the covariance $\mathcal{K}$ of the random field is required to be specified. The latter express the time and scheduling evolution of the functional form of the IIRs which is estimated. As the IIRs in (4) correspond to the sum of two "sub-models", hence, f can be viewed as a sum of two zero mean Gaussian random fields f^u, f^y, i.e., $f = \sum_{i=1}^{\infty} f_i^y + \sum_{j=0}^{\infty} f_j^u$. Due to our orthogonality assumption, it follows that $\mathbb{E}\{f^u f^y\} = 0$ and, hence, the covariance function for f can be expressed as

$$\begin{aligned}\mathcal{K}(x^{(k)}, x^{(k')}) &= \mathbb{E}\Big\{ f(x^{(k)}) f(x^{(k')}) \Big\} \\ &= \sum_{i=1}^{\infty} \mathcal{K}_i^y(x^{(k)}, x^{(k')}) + \sum_{j=0}^{\infty} \mathcal{K}_j^u(x^{(k)}, x^{(k')}), \quad (7)\end{aligned}$$

where $\mathcal{K}$ is a mixture of Gaussian kernels $\mathcal{K}_i^y, \mathcal{K}_j^u$, which represent the covariances of f_i^y, f_j^u between time k and k' respectively. In the Bayesian identification framework, the covariance (7) should be parameterized with a priori information to include possible structural dependencies and asymptotic stability of the predictor. A popular choice in system identification to capture functional dependencies is to use Gaussian kernels, also known as *radial basis functions* (RBFs) (e.g., see [11]), which only encode information on the smoothness of the underlying function to be estimated. Therefore, a decaying term, introduced in [10], is included in the kernel to express the decay of the impulse response over time. Hence, the covariance $\mathcal{K}_j^u$ of the function f_j^u takes the following form

$$\begin{aligned}\mathcal{K}_j^u(x^{(k)}, x^{(k')}) &= \mathbb{E}\Big\{ f_j^u(x^{(k)}) f_j^u(x^{(k')}) \Big\} = \\ & u(k-j) \underbrace{\beta_j^u \exp\left(-\frac{\|p^{(k,j)} - p^{(k',j)}\|_2^2}{\sigma_u^2} \right)}_{R_j^u(p^{(k,j)}, p^{(k',j)})} u(k'-j), \quad (8)\end{aligned}$$

where $p^{(k,j)}$ is a vector of the past measurements starting from time k till $k-j$, i.e., $p^{(k,j)} = \left[p^\top(k), \ldots, p^\top(k-j) \right]^\top$, $\beta_j^u = \lambda_1 e^{-j\lambda_2}$ denotes the exponential decaying term, and $\sigma_u, \lambda_1, \lambda_2 \in \mathbb{R}^+$ are the unknown hyperparameters parametrizing the kernel. Similarly, the covariance $\mathcal{K}_i^y$ is

$$\begin{aligned}\mathcal{K}_i^y(x^{(k)}, x^{(k')}) &= \mathbb{E}\Big\{ f_i^y(x^{(k)}) f_i^y(x^{(k')}) \Big\} = \\ & y(k-i) \underbrace{\beta_i^y \exp\left(-\frac{\|p^{(k,i)} - p^{(k',i)}\|_2^2}{\sigma_y^2} \right)}_{R_i^y(p^{(k,i)}, p^{(k',i)})} y(k'-i), \quad (9)\end{aligned}$$

where $\beta_i^y = \lambda_3 e^{-i\lambda_4}$ is the decaying term and $\sigma_y, \lambda_3, \lambda_4 \in \mathbb{R}^+$ are the hyperparameters.

In addition, due to the asymptotic decay of the coefficient functions, the IIR sub-models in (6) can be arbitrary well approximated by truncating the infinite sum as

$$\begin{aligned}\hat{y}(k|k-1) = &\sum_{i=1}^{n_{f_\mathrm{y}}} h_{y_i}(p,k,i) q^{-i} y(k) + \\ &\sum_{j=0}^{n_{f_\mathrm{u}}} h_{u_j}(p,k,j) q^{-j} u(k), \quad (10)\end{aligned}$$

where n_{f_y} and n_{f_u} are sufficiently large. In the same line of reasoning, the covariance functions (7) can be truncated to the same finite orders n_{f_y} and n_{f_u}.

B. Estimation from data

The objective is to estimate the predictor (6) by the truncated model (10) from a given data set $\mathscr{D}_N = \{y(k), u(k), p(k)\}_{k=1}^N$ by minimizing the ℓ_2 norm of the predictor error $\varepsilon(k) = y(k) - \hat{y}(k|k-1)$. Let θ denote the vector of the unknown hyperparameters, $\theta = \begin{bmatrix} \eta & \sigma_y & \sigma_u & \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \end{bmatrix}^\top$, with η the standard deviation of the noise e in (1b). Let $n = \max(n_{f_\mathrm{u}}, n_{f_\mathrm{y}})$ denote the maximum truncation order and let y^N denote the vector with the observations, i.e., $y^N = \begin{bmatrix} y_{n+1} & \ldots & y_N \end{bmatrix}^\top$. In the Gaussian regression framework, two estimates minimizing this ℓ_2 norm can be formulated, the first one is the minimum variance estimator of f conditioned on a fixed θ [11]:

$$\hat{f}(\cdot) = \mathbb{E}\big\{ f(\cdot) |\ y^N, \theta \big\} = \sum_{k=n+1}^{N} c_{k-n} \mathcal{K}(\cdot, x^{(k)}), \quad (11)$$

where $x^{(k)}$ is the set of truncated past measurements, i.e., $x^{(k)} = \{u^{(k,n_{f_\mathrm{u}})}, p^{(k,n)}, y^{(k,n_{f_\mathrm{y}})}\}$ and c_{k-n} is the k-th component of the vector

$$c = (\Sigma_y(\theta))^{-1} y^N,$$

and $\Sigma_y \in \mathbb{R}^{N-n \times N-n}$ is invertible with an (i,j)-entry given by

$$[\Sigma_y(\theta)]_{i,j} = \mathcal{K}(x^{(n+i)}, x^{(n+j)}) + \eta^2 \delta_{i,j},$$

where $\delta_{i,j}$ is the Kronecker delta function w.r.t. (i,j). The second estimate is the marginal likelihood of the data y^N given θ:

$$p(y^N|\theta) = \frac{\exp\left(-\frac{1}{2} (y^N)^\top (\Sigma_y(\theta))^{-1} y^N \right)}{\sqrt{\det(2\pi \Sigma_y(\theta))}}. \quad (12)$$

The latter expression is used in the empirical Bayes approach, e.g., see [12], to estimate the unknown hyperparameters vector θ parametrizing the kernel $\mathcal{K}$. This is achieved by maximizing the ML or equivalently

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \ -\log p(y^N|\theta). \quad (13)$$

This nonlinear program can be solved by a standard line-search algorithm, e.g., `fmincon` in Matlab due to the low number of hyperparameters and the structure of (12). According to the empirical Bayes approach, the estimate $\hat{f}$ in (11) can be obtained by substituting the optimized $\hat{\theta}$ of (13). Hence, the minimum variance estimate (11) of individual coefficient functions are

$$\hat{h}_{y_i}(\bullet) = \sum_{k=n+1}^{N} c_{k\text{-}n} y(k-i) R_i^y(p^{(k,i)}, \bullet), \tag{14a}$$

$$\hat{h}_{u_j}(\bullet) = \sum_{k=n+1}^{N} c_{k\text{-}n} u(k-j) R_j^u(p^{(k,j)}, \bullet). \tag{14b}$$

The corresponding covariance estimate of the individual coefficient functions is given by

$$\mathrm{cov}\hat{h}_{y_i}(\bullet) = R_i^y(\bullet,\bullet) - (\psi_i^y)^\top (\Sigma_y(\theta))^{-1} \psi_i^y, \tag{15a}$$

$$\mathrm{cov}\hat{h}_{u_j}(\bullet) = R_j^u(\bullet,\bullet) - (\psi_j^u)^\top (\Sigma_y(\theta))^{-1} \psi_j^u, \tag{15b}$$

where

$$\psi_i^y = [R_i^y(\bullet, p^{(n+1,i)}) y(n\text{-}i{+}1), \ldots, R_i^y(\bullet, p^{(N,i)}) y(N\text{-}i)]^\top,$$
$$\psi_j^u = [R_j^u(\bullet, p^{(n+1,j)}) u(n\text{-}j{+}1), \ldots, R_j^u(\bullet, p^{(N,j)}) u(N\text{-}j)]^\top.$$

In case the data-generating system has an LPV-ARX representation, then the individual coefficient functions a_i and b_j correspond to h_{y_i} and h_{u_j}, respectively; hence, the coefficient functions are identified by the quantities (14b) and (14a) directly. The decaying terms keep the estimation variance low even if a high truncation order n_{f_y}, n_{f_u} is selected, hence, the selected "truncation order" plays an insignificant role in the bias/variance trade-off.

Assume that the one-step-ahead predictor (6) belongs to the RKHS $\mathcal{H}_K$ associated to the kernel K (7) with parametrization (8)-(9). In addition, if $\lambda_i \to 0$ as $N \to \infty$ for each i, and the same technical assumptions stated in [13] hold, then the estimator functions $\hat{h}$ (14) are consistent, i.e., $\hat{h} \to h$ with probability one in the topology of $\mathcal{H}_K$.

IV. Illustrative Examples

In this section, the performance of the proposed nonparametric approach to identify the one-step-ahead predictor is demonstrated by two Monte Carlo studies. The identification scheme is applied to two different examples: 1) an LPV-ARX model and 2) an LPV-BJ model.

A. Identification setting

The one-step-ahead predictor is estimated using identification datasets with various sizes $N = \{200, 500, 1000\}$ and the prediction performance of the estimated model is examined on an independently generated validation dataset that contains $N_{\mathrm{val}} = 500$ samples. The identification and validation datasets are generated with independent realizations of a white input signal u with Gaussian distribution, i.e., $u(k) \sim \mathcal{N}(0,1)$, and a scheduling signal given by

$$p(k) = 0.4\sin(0.1k) + 0.1 + \delta_p(k),$$

where $\delta_p(k)$ is i.d.d. with $\delta_p(k) \sim \mathcal{U}(0,1)$. The variance of the white noise e driving the noise process is chosen such that the signal-to-noise (SNR) ratio

$$\mathrm{SNR}_y = 10\log\frac{\sum_{k=1}^{N} y(k)^2}{\sum_{k=1}^{N} v(k)^2},$$

is 20dB. For each of the datasets and each of the $N_{\mathrm{MC}} = 100$ Monte Carlo simulations, different realization of the noises u, δ_p, and e are taken. The LPV data-generating system has

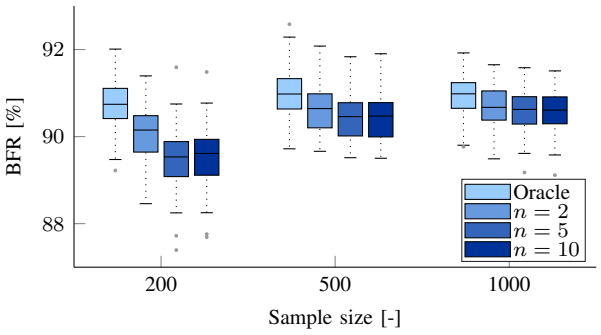

Fig. 1: The BFR of the predicted response with respect to the validation data sets using the estimated models based on different truncation orders $n = \{2, 5, 10\}$ and different sizes of the identification data set $N = \{200, 500, 1000\}$.

in both examples a plant model order of $n_{\mathrm{a}} = n_{\mathrm{b}} = 2$, and coefficient functions

$$a_1(\cdot) = 0.1p^2(k-1), \tag{16a}$$
$$a_2(\cdot) = \tan^{-1}(p(k-1))\cos(p(k-2)), \tag{16b}$$
$$b_0(\cdot) = -\exp(-p(k)), \tag{16c}$$
$$b_1(\cdot) = 1 - 0.5p^2(k) + p(k-1), \tag{16d}$$
$$b_2(\cdot) = \tan^{-1}(p(k-2)). \tag{16e}$$

Note, the nonlinear functions (16) have dynamic dependencies on the scheduling variable, i.e., the functions depend on p with different time shifts. The noise process is given in the corresponding sections, Sections IV-B and IV-C. In the given examples $\lambda_1 = \lambda_3$, $\lambda_2 = \lambda_4$, and $\sigma_u = \sigma_y$, to reduce computational complexity of the ML optimization (13). The performance is measured by means of the best-fit ratio (BFR)

$$\mathrm{BFR} = \max\left\{1 - \frac{\frac{1}{N}\sum_{k=1}^{N}\|y(k) - \hat{y}(k)\|_2}{\frac{1}{N}\sum_{k=1}^{N}\|y(k) - \bar{y}\|_2}, 0\right\} 100\%, \tag{17}$$

where $\hat{y}(k)$ is the predicted output by the estimated one-step-ahead predictor on the validation dataset and $\bar{y}$ defines the mean of the true output $y(k)$.

B. Example 1: identifying an LPV-ARX model

For the LPV-ARX model, the noise process (1b) is driven by $C_0(p,k,q^{-1}) = 1$ and $D_0(p,k,q^{-1}) = A_0(p,k,q^{-1})$. Both, plant and noise processes lie within the model set (4) for finite model order, i.e., the one-step-ahead predictor (10) has a one-to-one relation with the underlying data-generating system. Hence, the proposed nonparametric identification scheme should fully recover the structural dependency of the original coefficients functions.

In this example, the performance of an 'oracle' estimator is also displayed, which is a least squares (LS) based parameter estimate of an ARX model with the true model order and true underlying nonlinear functional dependencies, e.g., $h_{y_i} = a_i$, $h_{u_j} = b_j$. The oracle represents the maximum achievable performance given the dataset.

In Fig. 1, the BFR is shown for the oracle and the identified one-step-ahead predictor for different truncation orders $n = \{2, 5, 10\}$ and different sample sizes $N = \{200, 500, 1000\}$. Fig. 1 shows that as the sample size increases, the average BFR increases and the variance

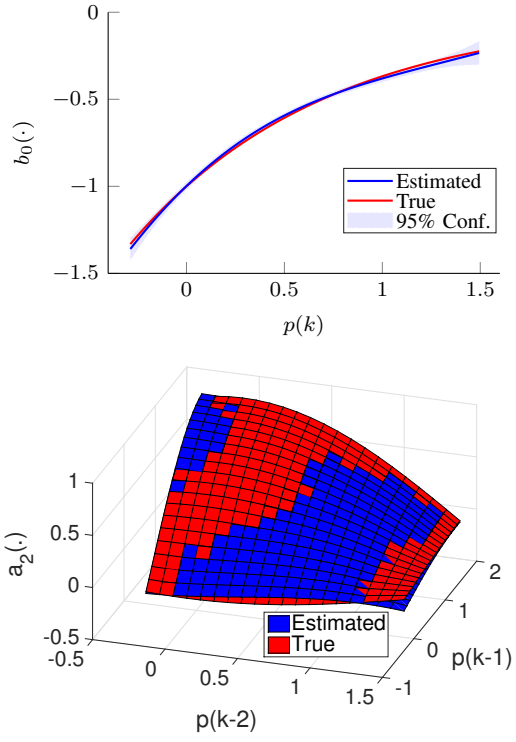

Fig. 2: Results of the function estimation with $n = 10$, $N = 1000$.

decreases, as expected, as the kernels come closer to the true covariance of (16). The nonparametric identification approach is very efficient, as it comes close to the oracle and, in contrary to the oracle, the functional dependencies in the scheduling variable are also estimated. This high performance is also evident in Fig. 2, which shows that the true and estimated functions b_0 and a_2.

Furthermore, the hyperparameters for the truncation order selection are given in Fig. 3. The ML optimization also successfully recovers the model order from the data, i.e., the function estimates $\hat{h}_{y_i}$, $\hat{h}_{u_i}$ for $i > 2$ are estimated with an insignificant magnitude. For this reason, no significant increase of variance of the BFR or decrease in mean of the BFR is noticed in Fig. 1, compared to the case where this exponential decay would not be included, e.g., see [14].

C. Example 2: identifying an LPV-BJ model

In this second example, the capabilities of the approach under general noise conditions are demonstrated. The noise process v in this case is a colored noise generated by a filtered white noise, where the coefficient functions (3) are

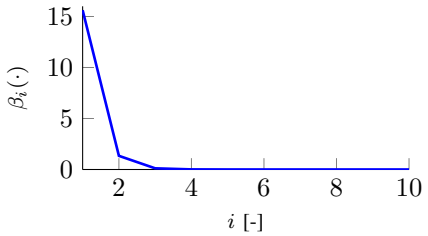

Fig. 3: The $\beta_i = \lambda_1 \exp(-i\lambda_2)$ function optimized based on $N = 1000$ samples with truncation order $n = 10$. The estimated hyperparameters are $\lambda_1 = 182$ and $\lambda_2 = 2.45$.

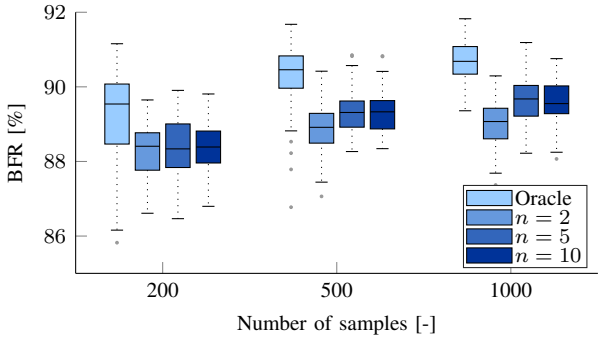

Fig. 4: The BFR of the predicted response with respect to the validation data sets using the estimation models for different truncation orders $n = \{2, 5, 10\}$, different sizes of the identification data set $N = \{200, 500, 1000\}$.

given as

$$\begin{aligned} c_1(\cdot) &= 0.8p^2(k-1), \quad c_2(\cdot) = 0.5\tan^{-1}(p(k-2)), \\ d_1(\cdot) &= 0.2p^3(k-1), \quad d_2(\cdot) = 0.5\sin(p(k-2)). \end{aligned} \tag{18}$$

Note, under this general noise assumption, the coefficient functions in (6) are summations of different products of the functions $a_i, \ldots, d_i$, i.e., see (24a) and (24b), contrary to the LPV-ARX case. For this case, the 'oracle' is an LS estimate of a high-order ARX model parametrized in terms of the true underlying nonlinear functional dependencies, i.e., these summation of the different products of the functional dependencies $a_i, \ldots, d_i$ in (24). The chosen truncation order is $n = 12$, which has been tuned to have the highest average BFR and lowest variance.

Fig. 4 shows the BFR for the identified one-step-ahead predictor with the proposed Bayesian approach under different truncation orders and sample sizes. In this figure, it can been seen that increasing the truncation order slightly increases the average BFR. According to our expectation, all predictors benefit from increasing the amount of samples in the identification dataset, which is evident in an increased BFR and a decreased BFR variance.

The figures show that the Bayesian approach estimated one-step-ahead predictor is capable of capturing the plant and noise dynamics, concluding that the proposed nonparametric identification scheme is able to identify the LPV model under general nonlinear noise conditions.

V. Conclusions

We reformulated the BJ problem as an impulse response identification problem based upon the one-step-ahead predictor form of these models. Under the assumption that the resulting sub-predictors are asymptotically stable, a nonparametric identification approach has been introduced based on the Bayesian setting with a Gaussian random field as a priori. As an important contribution, a suitable choice of a kernel to jointly express the time convolution and the scheduling dependency in terms of a few hyperparameters has been proposed. These hyperparameters are tuned by an empirical Bayes approach, i.e., maximizing the marginal likelihood based on data. It is shown that the proposed nonparametric approach is capable of achieving consistent estimation of

LPV-BJ models and it is capable of recovering the true structural dependencies.

APPENDIX
PROOF OF THEOREM 2

Due to the sake of space, only a brief overview of the proof is given. For all details see [4]. For notational ease, define $A_k(q) \triangleq A_0(p,k,q^{-1})$, ..., $D_k(q) \triangleq D_0(p,k,q^{-1})$ and let $u_k \triangleq u(k)$, ..., $e_k \triangleq e(k)$. By using recursive substituion, it is possible to rewrite the noise process (1b) as[3]

$$\begin{aligned} v_k &= C_k(q)e_k + (1-D_k(q))v_k, \\ &= C_k(q)e_k + (1-D_k(q))C_k(q)e_k \\ &\quad + (1-D_k(q))(1-D_k(q))v_k, \\ &\vdots \\ &= \underbrace{\sum_{i=0}^{\infty} (1-D_k(q))^i C_k(q)}_{H_k(q)} e_k. \end{aligned} \tag{19}$$

As asymptotic stability of (1b) in terms of Definition 1 is assumed, it is possible to show that (19) is convergent, such that

$$H_k(q) = 1 + \sum_{i=1}^{\infty} h_i(p,k,i)q^{-i}, \tag{20}$$

where h_i converges to the zero function as $i \to \infty$. Note that $C_k(q)$ and $D_k(q)$ being monic, implies that the IRR (20) is monic as well. Similarly, the plant process (1a) can be rewritten as

$$\breve{y}_k = \underbrace{\sum_{i=0}^{\infty} (1-A_k(q))^i B_k(q)}_{G_k(q)} u_k, \tag{21}$$

which, due the assumed asymptotic stability of (1a) in terms of Definition 1, is convergent, given

$$G_k(q) = \sum_{i=0}^{\infty} g_i(p,k,i)q^{-i}, \tag{22}$$

where g_i converges to the zero function as $i \to \infty$. Assume that the monic $H_k(q)$ has a stable left inverse $H_k^{\dagger}(q)$, i.e, a convergent monic IRR exists such that $H_k^{\dagger}(q)H_k(q) = 1$. As $H_k(q)$ is monic, if such a $H_k^{\dagger}(q)$ exists, then it is a bi-lateral inverse of $H_k(q)$, i.e., $H_k^{\dagger}(q)H_k(q) = H_k(q)H_k^{\dagger}(q) = 1$. Now, if we substitute (20) and (22) into (1c), then it follows

$$\begin{aligned} y_k &= G_k(q)u_k + H_k(q)e_k \\ &= \underbrace{\left(1 - H_k^{\dagger}(q)\right)}_{F_y(p,k,q^{-1})} y_k + \underbrace{H_k^{\dagger}(q)G_k(q)}_{F_u(p,k,q^{-1})} u_k + e_k. \end{aligned} \tag{23}$$

As $H_k^{\dagger}(q)$ and $G_k(q)$ are convergent, asymptotically stable IIRs, then it follows that F_y and F_u in (23) are also convergent and represent asymptotically stable LPV filters, given by

$$F_y(p,k,q^{-1}) = 1 - \sum_{i=0}^{\infty} (1-C_k(q))^i D_k(q), \tag{24a}$$

and

$$F_u(p,k,q^{-1}) = \sum_{i=0}^{\infty} (1-C_k(q))^i D_k(q) \cdot \sum_{j=0}^{\infty} (1-A_k(q))^j B_k(q). \tag{24b}$$

To conclude the proof, the maximum backward time-shifts for p in the functional dependencies of (24) does not exceed the maximum backward time-shifts for u and y in (23). Therefore, (23) can be rewritten as (4).

[3]Note, multiplication between the shift operator and a coefficient function is non-communicative, hence, $q^{-1}C_k = C_{k-1}q^{-1}$ (see [15] for details).

REFERENCES

[1] B. Bamieh and L. Giarré, "Identification of linear parameter varying models," *Int. J. of Robust and Nonlinear Control*, vol. 12, no. 9, pp. 841–853, 2002.

[2] M. Butcher, A. Karimi, and R. Longchamp, "On the consistency of certain identification methods for linear parameter varying systems," in *Proc. of the 17th IFAC World Congress*, Seoul, Korea, July 2008, pp. 4018–4023.

[3] 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.

[4] R. Tóth, P. S. C. Heuberger, and P. M. J. Van den Hof, "Pediction 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.

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

[6] 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.

[7] 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.

[8] A. Golabi, N. Meskin, R. Tóth, and M. 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.

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

[10] 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.

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

[12] B. P. Carlin and T. A. Louis, *Bayes and empirical Bayes methods for data analysis*. CRC Press, 2000.

[13] G. Pillonetto, "Consistent identification of Wiener systems: A machine learning viewpoint," *Automatica*, vol. 49, no. 9, pp. 2704–2712, 2013.

[14] 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.

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