Towards Efficient Maximum Likelihood Estimation of LPV-SS Models ⋆

Pepijn B. Cox [a], Roland Tóth [a], Mihály Petreczky [b]

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

[b] *Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL). Lille 1 University, M3, Avenue Carl Gauss, 59650 Villeneuve-d'Ascq, France.*

Abstract

To efficiently identify multiple-input multiple-output (MIMO) linear parameter-varying (LPV) discrete-time state-space (SS) models with affine dependence on the scheduling variable remains still an open question, as identification methods proposed in the literature suffer heavily from the curse of dimensionality and/or depend on over-restrictive approximation of the measured signal behaviours. However, obtaining an SS model of the targeted system is crucial for many LPV control synthesis methods, as these synthesis tools are almost exclusively formulated for the aforementioned representation of the system dynamics. Therefore, in this paper, we tackle the problem by combining state-of-the-art LPV input-output (IO) identification methods with an LPV-IO to LPV-SS realization scheme and a maximum likelihood refinement step. The resulting modular LPV-SS identification approach achieves statical efficiency with a relative low computational load. The method contains the following three steps: 1) estimation of the Markov coefficient sequence of the underlying system using correlation analysis or Bayesian impulse response estimation, then 2) LPV-SS realization of the estimated coefficients by using a basis reduced Ho-Kalman method, and 3) refinement of the LPV-SS model estimate from a maximum-likelihood point of view by a gradient-based or an expectation-maximization optimization methodology. The effectiveness of the full identification scheme is demonstrated by a Monte Carlo study where our proposed method is compared to existing schemes for identifying a MIMO LPV system.

Key words: Identification; linear parameter-varying system; state-space representation; LPV realization; correlation analysis; impulse response estimation; expectation-maximization; gradient-based search; parameter estimation.

1 Introduction

The *linear parameter-varying* (LPV) modelling paradigm offers an attractive model class to capture nonlinear and/or time-varying systems with a parsimonious parameterization. The LPV model class preserves the linear signal relation between the inputs and outputs of the system, however, these linear relations are functions of a measurable, time-varying signal, the scheduling variable, denoted as p. This scheduling signal can be any combination of inputs, measurable process states, outputs, or measurable exogenous variables and, in addition, these signals can be filtered by any arbitrary functional relation. Hence, the LPV modelling paradigm can represent both non-stationary and nonlinear behaviour of a wide variety of physical or chemical process, e.g., see [1–4].

The majority of LPV control synthesis methods are based upon the assumption that an LPV *state-space* (SS) model of the system is available, especially with static and affine dependence of the involved matrix coefficients on the scheduling variable p, e.g., [5]. Hence, to perform efficient identification of LPV-SS models has been in the focus of intensive research, but it remains an open question. Conceptually, LPV identification can be performed as: i) the interpolation of local LTI models estimated from multiple experiments around fixed operating points, i.e., with constant p, often referred to as the *local identification* setting; or ii) a direct model estimation problem, i.e., the *global identification* setting, which requires the experimental data with a varying p which is informative to uniquely identify the considered model parameters. Accordingly, global identification approaches include scheduling dynamics, see [4] for detailed comparison between the two settings. In this paper, we will focus on the global setting and the identification of discrete-time models.

In the global identification setting, a class of identification methods aims at *prediction error minimization* (PEM), where either full state measurements are available [6, 7] or the PEM is minimzed dircetly using *gradient-based* (GB)

⋆ This paper was not presented at any IFAC meeting. Corresponding author P. B. Cox. Tel. +31-40-2478188.

Email addresses: p.b.cox@tue.nl (Pepijn B. Cox), r.toth@tue.nl (Roland Tóth), mihaly.petreczky@ec-lille.fr (Mihály Petreczky).

methodologies, e.g., see [8–11]. The full knowledge of the state is in many practical cases unrealistic. On the other hand, solving the direct PEM problem is nonlinear in the parameters [8–10]. Recently, the *expectation-maximization* (EM) method is developed for LPV-SS models [11], as an addition to the GB methods. The EM algorithm is more robust to an inaccurate initial estimate compared to the GB PEM; however, its convergence rate is much lower near the optimum [12]. Problem 1 (P1): both, EM and GB methods can be computationally demanding and their convergence depence heavily on a proper initial guess. Besides PEM identification methods, LPV grey-box [13, 14] and LPV *set-membership* (SM) [15–17] identification approaches are known. Grey-box schemes need a detailed model with only few unknown parameters, which are estimated by a Kalman like filtering strategy. The SM methods characterize the measurement errors as a bounded-error or SM, opposed to the stochastic description in PEM. (P2): Unfortunately, SM approaches have, in general, a significant higher computational load compared to direct PEM and rely on convex outer-approximation.

The LPV-SS identification problem can be tackled differently by: first identifing an LPV-IO model, which is well established in literature (e.g., see [5, 18, 19]) and, secondly, to execute an exact realization on the identified LPV-IO form to get an LPV-SS model. (P3): However, such an exact realization will, in general, result in functional relations with rational, dynamic dependence on the scheduling variable or lead to a non-minimal state realization if the static, affine dependence should be preserved [20]. More relevant, these schemes have a high computational demand. However, recently introduced LPV realization theory based schemes, so called *subspace identification* (SID) methods, have greatly decreased the computational complexity. SID schemes can apply *direct* LPV Ho-Kalman like realization [21] on specific LPV-IO models that are identified by a least-squares method; or have an intermediate *projection* step, i.e., 1) identify an IO structure using convex optimization, 2) find a projection to estimate the unknown state sequence via matrix decomposition methods, then 3) estimate the SS matrices in a least-squares fashion, e.g., see [22–25]. (P4): However, to attain the convex optimization, the latter class of SID methods usually depend on over-restrictive approximations on the signal behaviours and/or the number of observed variables grows exponentially. As a consequence, the estimation problem is ill-conditioned with high parameter variance and/or has still a high computational demand, making it inapplicable for real-world systems. (P5): As it is well known, the aforementioned realization based schemes provide an LPV-SS models estimate which is not minimized w.r.t. any criterion and, therefore, it is not "optimal" in a *maximum likelihood* (ML) sense. To have ML, this LPV-SS model estimate is often used to initialize gradient-based PEM schemes.

From P1-P5 it becomes clear, how to perform efficient identification of LPV-SS models on real-world sized problems remains still an open question. Hence, the goal of this paper is to provide a maximum likelihood identification method for LPV-SS models in the global, open-loop identification setting, which can estimate moderate sized problems in a computational efficient way. To reach this goal, it is key to have a subspace method with low computational complexity that can be used as the initial starting point of a GB or EM method to reach the ML estimate. In light of recent developments in LPV realization theory, we will not extend projection type of SID methods, but we will tackle the problem at the core, i.e., at the realization side, by applying the basis reduced LPV Ho-Kalman scheme [26].

The papers starts by providing some formal definitions; an analysis between the LPV-SS model with general noise structure and the innovation form, to highlight modelling limitations of the innovation form considered in many LPV SID methods; and the formal problem statement (Sec. 2). Continuing, we propose our modular identification method, defined in three steps: 1) estimate the Markov coefficient sequence of the underlying system using *correlation analysis* (CRA), or MIMO Bayesian *finite impulse response* (FIR) estimation (Sec. 3), then 2) create an LPV-SS realization from the estimated coefficients by using an Ho-Kalman like method (Sec. 4), and 3) to have an ML estimate, refine the LPV-SS model by GB and/or EM optimization (Sec. 5). Hence, in this paper state-of-the-art LPV-IO identification, LPV-IO to LPV-SS realization, and ML estimation techniques are combined to give a unified method to achieve an efficient ML estimate. The efficiency of the combined approach is demonstrated by a Monte Carlo study and it is compared to existing LPV-SS identification schemes [15, 25, 27] by identifying a MIMO LPV-SS model (Sec. 6).

2 The LPV identification problem

Technical preliminaries

We denote a probability space as $(\Omega, \mathcal{F}_\Omega, \mathbf{P})$ where $\mathcal{F}_\Omega$ is the σ-algebra, defined over the sample space Ω; and $\mathbf{P} : \mathcal{F}_\Omega \to [0, 1]$ is the probability measure defined over the measurable space $(\Omega, \mathcal{F}_\Omega)$. Within this work, we consider random variables that take values on the Euclidean space $\mathbb{R}^n$ with associated Borel measurable space $\mathscr{B}(\mathbb{R}^n)$ [28]. More precisely, for the given probability space $(\Omega, \mathcal{F}_\Omega, \mathbf{P})$ we define a *random variable* $\mathbf{f}$ as a measurable function $\mathbf{f} : \Omega \to \mathbb{R}^n$, which induces a probability measure on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$. As such, a realization $\nu \in \Omega$ of $\mathbf{P}$, denoted $\nu \sim \mathbf{P}$, defines a realization f of $\mathbf{f}$, i.e., $f := \mathbf{f}(\nu)$. Furthermore, a *stochastic process* $\mathbf{x}$ is a collection of random variables $\mathbf{x}_t : \Omega \to \mathbb{R}^n$ indexed by the set $t \in \mathbb{Z}$ (discrete time), given as $\mathbf{x} = \{\mathbf{x}_t : t \in \mathbb{Z}\}$. A realization $\nu \in \Omega$ of the stochastic process defines a signal trajectory $x := \{\mathbf{x}_t(\nu) : t \in \mathbb{Z}\}$. We call a stochastic process $\mathbf{x}$ *stationary* if $\mathbf{x}_t$ has the same probability distribution on each time index as $\mathbf{x}_{t+\tau}$ for all $\tau \in \mathbb{N}$. In addition, a stationary process consisting of uncorrelated random variables with zero mean and finite variance is called a *white noise process*.

2.1 The data-generating system

Consider a *multiple-input multiple-output* (MIMO), discrete-time linear parameter-varying data-generating system, defined by the following first-order difference equation, i.e.,

LPV-SS representation with general noise model:

$$x_{t+1} = \mathcal{A}(p_t)x_t + \mathcal{B}(p_t)\,u_t + \mathcal{G}(p_t)w_t, \tag{1a}$$
$$y_t = \mathcal{C}(p_t)\,x_t + \mathcal{D}(p_t)u_t + \mathcal{H}(p_t)v_t, \tag{1b}$$

where $x : \mathbb{Z} \to \mathbb{X} = \mathbb{R}^{n_\mathrm{x}}$ is the state variable, $y : \mathbb{Z} \to \mathbb{Y} = \mathbb{R}^{n_\mathrm{y}}$ is the measured output signal, $u : \mathbb{Z} \to \mathbb{U} = \mathbb{R}^{n_\mathrm{u}}$ denotes the input signal, $p : \mathbb{Z} \to \mathbb{P} \subseteq \mathbb{R}^{n_\mathrm{p}}$ is the scheduling variable, subscript $t \in \mathbb{Z}$ is the discrete time, $w : \mathbb{Z} \to \mathbb{R}^{n_\mathrm{x}}$, $v : \mathbb{Z} \to \mathbb{R}^{n_\mathrm{y}}$ are the sample path realizations of the zero-mean stationary processes:

$$\begin{bmatrix} \mathbf{w}_t \\ \mathbf{v}_t \end{bmatrix} \sim \mathcal{N}(0, \Sigma), \qquad \Sigma = \begin{bmatrix} \mathcal{Q} & \mathcal{S} \\ \mathcal{S}^\top & \mathcal{R} \end{bmatrix}, \tag{2}$$

where $\mathbf{w}_t : \Omega \to \mathbb{R}^{n_\mathrm{x}}$, $\mathbf{v}_t : \Omega \to \mathbb{R}^{n_\mathrm{y}}$ are random variables, $\mathcal{Q} \in \mathbb{R}^{n_\mathrm{x} \times n_\mathrm{x}}$, $\mathcal{S} \in \mathbb{R}^{n_\mathrm{x} \times n_\mathrm{y}}$, and $\mathcal{R} \in \mathbb{R}^{n_\mathrm{y} \times n_\mathrm{y}}$ are covariance matrices, such that Σ is positive definite. Furthermore, we will assume u, p, w, v, y to have left compact support to avoid technicalities with initial conditions. As often considered in LPV control theory, the matrix functions $\mathcal{A}(\cdot), ..., \mathcal{H}(\cdot)$, defining the SS representation (1) are defined as affine combinations:

$$\begin{aligned}
\mathcal{A}(p_t) &= A_0 + \sum_{i=1}^{n_\psi} A_i \psi^{[i]}(p_t), \quad \mathcal{B}(p_t) = B_0 + \sum_{i=1}^{n_\psi} B_i \psi^{[i]}(p_t), \\
\mathcal{C}(p_t) &= C_0 + \sum_{i=1}^{n_\psi} C_i \psi^{[i]}(p_t), \quad \mathcal{D}(p_t) = D_0 + \sum_{i=1}^{n_\psi} D_i \psi^{[i]}(p_t), \\
\mathcal{G}(p_t) &= G_0 + \sum_{i=1}^{n_\psi} G_i \psi^{[i]}(p_t), \quad \mathcal{H}(p_t) = H_0 + \sum_{i=1}^{n_\psi} H_i \psi^{[i]}(p_t),
\end{aligned} \tag{3}$$

where $\psi^{[i]}(\cdot) : \mathbb{P} \to \mathbb{R}$ are bounded scalar functions on $\mathbb{P}$ and $\{A_i, B_i, C_i, D_i, G_i, H_i\}_{i=0}^{n_\psi}$ are constant matrices with appropriate dimensions. Additionally, for well-posedness, it is assumed that $\{\psi^{[i]}\}_{i=1}^{n_\psi}$ are linearly independent over an appropriate function space and are normalized w.r.t. an appropriate norm or inner product [21]. Due to the freedom to consider arbitrary functions $\psi^{[i]}$, (3) can capture a wide class of static nonlinearities and time-varying behaviour.

2.2 Properties of LPV-SS representations

In this section, we present some formal definitions needed for the analysis of the different noise structures, together with the problem statement and the description of the LPV-IO to LPV-SS realization theory. Note that, the deterministic part of (1) is governed by

$$x^\mathrm{d}_{t+1} = \mathcal{A}(p_t)x^\mathrm{d}_t + \mathcal{B}(p_t)\,u_t, \tag{4a}$$
$$y^\mathrm{d}_t = \mathcal{C}(p_t)\,x^\mathrm{d}_t + \mathcal{D}(p_t)u_t. \tag{4b}$$

The corresponding IO solution set, i.e., the manifest behaviour, of (4) is

$$\mathfrak{B}_\mathrm{d} = \big\{ (y^\mathrm{d}, u, p) \in (\mathbb{Y} \times \mathbb{U} \times \mathbb{P})^\mathbb{Z} \;\big|\; \exists x^\mathrm{d} \in (\mathbb{X})^\mathbb{Z} \text{ s.t. (4) holds} \big\}. \tag{5}$$

On the other hand, the stochastic part of (1) is

$$\mathbf{x}^\mathrm{s}_{t+1} = \mathcal{A}(p_t)\mathbf{x}^\mathrm{s}_t + \mathcal{G}(p_t)\,\mathbf{w}_t, \tag{6a}$$
$$\mathbf{y}^\mathrm{s}_t = \mathcal{C}(p_t)\,\mathbf{x}^\mathrm{s}_t + \mathcal{H}(p_t)\mathbf{v}_t. \tag{6b}$$

The corresponding manifest behaviour of (1) is

$$\mathfrak{B}_\mathrm{SS} = \big\{ (y, u, p) \in (\mathbb{Y} \times \mathbb{U} \times \mathbb{P})^\mathbb{Z} \;\big|\; \exists \nu \in \Omega \text{ and } \exists y^\mathrm{d} \in (\mathbb{Y})^\mathbb{Z} \text{ s.t. } (y^\mathrm{d}, u, p) \in \mathfrak{B}_\mathrm{d} \text{ and (6) hold } \forall t \;\big\}. \tag{7}$$

To introduce the essential details of the deterministic realization step we extend to exploit in Sec. 4, we momentary neglect the stochastic process (6). To this end, we take the expectation $\mathbb{E}\{\mathbf{y}_t\} = y^\mathrm{d}_t$, which is equivalent as taking $v_t = w_t = 0$ in (1). The manifest behaviour w.r.t. an LPV-SS representation $\mathcal{S}$ is denoted $\mathfrak{B}_\mathrm{d}(\mathcal{S})$.

In this paper, we are interested in finding the LPV-SS representation with minimal state dimension:

Definition 1 (Minimal LPV-SS representations) *The LPV-SS representation $\mathcal{S}$* (1) *is called minimal, if there exists no other LPV-SS representation $\mathcal{S}'$ with ${n_\mathrm{x}}' < n_\mathrm{x}$ and equivalent IO behavior $\mathfrak{B}_\mathrm{d}(\mathcal{S}) = \mathfrak{B}_\mathrm{d}(\mathcal{S}')$.* □

For specific subclasses of LPV-SS representations in which the functional dependency structure of $\mathcal{A}(\cdot), \ldots, \mathcal{D}(\cdot)$ differs, the minimal state dimension might differ [20]. Hence, for the remainder of the paper, state minimality is considered w.r.t. the parametrization (3). The minimality condition by Def. 1 for (1a)-(1b) is a necessary condition for complete state-observability.

Lemma 2 (Isomorphic LPV-SS representations [29,30]) *Two LPV-SS representations* (4) *S and S' with static, affine dependency structure* (3) *and equivalent state dimensions $n_\mathrm{x} = {n_\mathrm{x}}'$ are isomorphic, i.e., their input-output maps are equal $\mathfrak{B}_\mathrm{d}(\mathcal{S}) = \mathfrak{B}_\mathrm{d}(\mathcal{S}')$, if and only if there exists a non-singular isomorphism matrix $T \in \mathbb{R}^{n_\mathrm{x} \times n_\mathrm{x}}$, such that*

$$A_i' T = T A_i, \qquad B_i' = T B_i, \qquad C_i' T = C_i, \qquad D_i' = D_i,$$

for all $i \in \mathbb{I}_0^{n_\psi}$. □

Lem. 2 is a special case of equivalence relations in the LPV case as the transformation matrix T is independent of the scheduling signal, e.g., see [20, Def. 3.29] for the general case. Under dependency structure (3) and assumption of state minimality, the equivalence class of LPV-SS representations is completely characterized by the non-singular transformation matrix T, as given in Lem. 2.

We are interested in identification under open-loop conditions, hence, the underlying data generating system is considered to be asymptotically stable:

Definition 3 (Asymptotic stability) *An LPV system, represented in terms of* (4)*, is called asymptotically stable, if for all trajectories of $\{u_t, p_t, y_t\}$ satisfying* (4)*, with $u_t \equiv 0$ for $t \geq 0$, and $p_t \in \mathbb{P}$, it holds that $\lim_{t\to\infty} |y_t| = 0$.* □

2.3 LPV-SS noise models and the innovation form

A popular model for many subspace identification schemes is the innovation form, e.g., see [31]. Under some mild conditions, the LPV-SS representation (1) has the following equivalent innovation form:

Lemma 4 ([32]) *For each given trajectory of the input u and scheduling p, the LPV data-generating system* (1) *can be equivalently represented by a p-dependent innovation form*

$$\check{x}_{t+1} = \mathcal{A}(p_t)\check{x}_t + \mathcal{B}(p_t)\,u_t + \mathcal{K}_t\xi_t, \tag{8a}$$
$$y_t = \mathcal{C}(p_t)\,\check{x}_t + \mathcal{D}(p_t)u_t + \xi_t, \tag{8b}$$

where ξ_t is the sample path of $\boldsymbol{\xi}_t \sim \mathcal{N}(0, \Xi_t)$ and $\mathcal{K}_t$ can be uniquely determined by

$$\mathcal{K}_t = \left[\mathcal{A}(p_t)\mathcal{P}_{t|t-1}\mathcal{C}^\top(p_t) + \mathcal{G}(p_t)\mathcal{S}\mathcal{H}^\top(p_t)\right]\Xi_t^{-1}, \tag{8c}$$
$$\mathcal{P}_{t+1|t} = \mathcal{A}(p_t)\mathcal{P}_{t|t-1}\mathcal{A}^\top(p_t) - \mathcal{K}_t\Xi_t\mathcal{K}_t^\top + \mathcal{G}(p_t)\mathcal{Q}\mathcal{G}^\top(p_t), \tag{8d}$$
$$\Xi_t = \mathcal{C}(p_t)\mathcal{P}_{t|t-1}\mathcal{C}^\top(p_t) + \mathcal{H}(p_t)\mathcal{R}\mathcal{H}^\top(p_t), \tag{8e}$$

under the assumption that $\exists t_0 \in \mathbb{Z}$ such that $x_{t_0} = 0$ and Ξ_t is non-singular for all $t \in [t_0, \infty)$. □

In (8c)-(8e), the notation of $\mathcal{K}_t$, $\mathcal{P}_{t+1|t}$, and Ξ_t is a shorthand for $\mathcal{K}_t := (\mathcal{K} \diamond p_t) \in \mathscr{R}^{n_\mathrm{x}\times n_\mathrm{y}}$, $\mathcal{P}_{t+1|t} := (\mathcal{P}_{t+1|t} \diamond p_t) \in \mathscr{R}^{n_\mathrm{x}\times n_\mathrm{x}}$, and $\Xi_t := (\Xi \diamond p_t) \in \mathscr{R}^{n_\mathrm{y}\times n_\mathrm{y}}$, where $\mathscr{R}$ defines the ring of all real meromorphic functions with finite dimensional domain, and the operator $\diamond : (\mathscr{R}, \mathbb{P}^{\mathbb{Z}}) \to \mathbb{R}^{\mathbb{Z}}$ denotes $(\mathcal{K}_t \diamond p_t) = \mathcal{K}_t(p_{t+\tau_1}, \ldots, p_t, \ldots, p_{t-\tau_2})$ with $\tau_1, \tau_2 \in \mathbb{Z}$. The subscript notation $_{t+1|t}$ denotes that the matrix function at time $t+1$ depends only on p_i for $i = t_0, \ldots, t$.

In [32] it is shown that the setting of (1) is not equivalent to the innovation form with only static, affine matrix function $\mathcal{K}(p)$, similarly parametrized as (3). This static, affine structure is commonly used in many SID methods [22, 25, 33]. However, [32] also shows that a static, affine $\mathcal{K}(p)$ can approximate the general setting (8) if the state dimension is increased. It follows that, to guarantee state minimality of an innovation form based realization of (1), the Kalman gain $\mathcal{K}_t$ (8c) should have rational and dynamic dependency on p. However, in practice, we need to restrict overparameterization to reduce complexity of the estimation method and variance of the model estimates. Hence, despite to the possible increase of state order of the equivalent innovation form, the underlying complexity trade-off might be acceptable from a practical point of view.

2.4 Problem statement

In this paper, we are interested in identifying LPV-SS models (1) with dependency structures (3) to capture the process dynamics (1) of the underlying data-generating system. Hence, our focus is not on identifying the noise structure $(\mathcal{H}, \mathcal{G})$, but to derive a methodology which can provide consistent estimates of (4) under the general noise structure of (1). We will also assume that the scalar functions $\{\psi^{[i]}\}_{i=1}^{n_\psi}$ are known a priori. As a consequence, we are interested in estimating the parameters of (3), i.e.

$$\Lambda_0 = \begin{bmatrix} A_0 \ \ldots \ A_{n_\psi} & B_0 \ \ldots \ B_{n_\psi} \\ C_0 \ \ldots \ C_{n_\psi} & D_0 \ \ldots \ D_{n_\psi} \end{bmatrix}, \tag{9}$$

with $\Lambda_0 \in \mathbb{R}^{n_\mathrm{x}+n_\mathrm{y}\times(n_\mathrm{x}+n_\mathrm{u})(1+n_\psi)}$. Based on these, we denote by $\mathcal{S}(\Lambda_0)$ the original SS representation of the data generating system $\mathcal{S}$ with parameters Λ_0. According to Def. 2, we aim at identifying an isomorphic $\mathcal{S}(\Lambda)$ w.r.t. $\mathcal{S}(\Lambda_0)$, due to the non-uniqueness of the SS representation of the underlying input-scheduling-output behaviour $\mathfrak{B}_\mathrm{d}(\mathcal{S})$. Hence, the corresponding parameter Λ lies within the following set

$$\mathcal{Q} = \Bigg\{ \Lambda \;\Big|\; \exists T \in \mathbb{R}^{n_\mathrm{x}\times n_\mathrm{x}} \text{ sat. } \mathrm{rank}(T) = n_\mathrm{x} \text{ and } \Lambda = \begin{bmatrix} T^{-1} & 0 \\ 0 & I_{n_\mathrm{y}} \end{bmatrix} \Lambda_0 \begin{bmatrix} I_{1+n_\psi} \otimes T & 0 \\ 0 & I_{n_\mathrm{u}(1+n_\psi)} \end{bmatrix} \Bigg\}, \tag{10}$$

where $\otimes$ is the Kronecker product.

Given a dataset $\mathcal{D}_N = \{u_t, p_t, y_t\}_{t=1}^N$ and the bases $\{\psi^{[i]}\}_{i=1}^{n_\psi}$, our objective is to efficiently find, in a stochastic and computational sense, an estimate $\hat{\Lambda}$ of Λ and, accordingly, the state dimension n_x of the underlying system (1). In addition, the proposed scheme should be consistent, i.e., $\hat{\Lambda} \to \Lambda \in \mathcal{Q}$ with probability one as $N \to \infty$. We will discuss these properties per individual identification step later on. In the remaining part of this paper it is assumed that the data generating LPV-SS system (1) with dependency structure (3) is structural observable and structural reachable, i.e., the system is jointly minimal, and that the input-scheduling signals are persistently exciting, such that the parameters are uniquely identifyable. We will not address the identifiability problem nor we provide persistency of excitation conditions for the input and scheduling signals. These problems are addressed for the LPV-ARX[1] case in [34], but they are not yet fully understood w.r.t. (1).

3 Identification of LPV impulse response models

3.1 LPV Impulse response representation

In order to realize our objective defined in Sec. 2.4, an important ingredient is to rewrite (1) or (8) into its so called *infinite impulse response* (IIR) representation:

Lemma 5 (Infinite impulse response [20]) *Any asymptotically stable LPV system according to Def. 3 has a convergent series expansion in terms of the pulse-basis $\{q^{-i}\}_{i=0}^\infty$ given by*

$$y_t = \sum_{i=0}^{\infty} (h_i \diamond p_t) q^{-i} u_t + y_t^\mathrm{s}, \tag{11}$$

where $h_i \in \mathscr{R}^{n_\mathrm{y}\times n_\mathrm{u}}$ are the expansion coefficient functions, i.e., Markov coefficients*, and y_t^s is a sample path of* (6). □

The corresponding IIR of an asymptomatic stable LPV-SS representation (1) is

$$y_t = \underbrace{\mathcal{D}(p_t)}_{h_0\diamond p_t} u_t + \underbrace{\mathcal{C}(p_t)\mathcal{B}(p_{t-1})}_{h_1\diamond p_t} u_{t-1} + \underbrace{\mathcal{C}(p_t)\mathcal{A}(p_{t-1})\mathcal{B}(p_{t-2})}_{h_2\diamond p_t} u_{t-2} + \ldots + \underbrace{\mathcal{G}(p_t)v_t + \mathcal{C}(p_t)\mathcal{H}(p_{t-1})w_{t-1} + \ldots}_{y_t^\mathrm{s}}, \tag{12}$$

where h_i converges to the zero function as $i \to \infty$. The noise y_t^s in (11)-(12) is colored, as it is a combination of the IIR filtered innovation noise w and the additive output noise v of (1) and this IIR filter is asymptotically stable, as it is filtered by the asymptotically stable process dynamics. For

[1] Autoregressive model with exogenous input (ARX).

notional ease, define $\psi^{[i]}(p_t) = \psi_t^{[i]}$ and the signal vector $\psi_t = [\,1\ \psi_t^{[1]}\ \cdots\ \psi_t^{[n_\psi]}\,]^\top \in \mathbb{R}^{n_\psi}$. The Markov coefficients can be written as

$$h_m \diamond p_t = \mathcal{C}(p_t)\mathcal{A}(p_{t-1})\cdots\mathcal{A}(p_{t-m+1})\mathcal{B}(p_{t-m}) = \sum_{i=0}^{n_\psi}\sum_{j=0}^{n_\psi}\cdots\sum_{k=0}^{n_\psi}\sum_{l=0}^{n_\psi} C_iA_j\cdots A_kB_l\psi^{[i]}\overleftarrow{\psi}_j^{(1)}\cdots\overleftarrow{\psi}_l^{(m)}, \quad (13)$$

where the individual products $C_iA_j\cdots A_kB_l$ are the so-called sub-Markov parameters for $m = 1, 2, \ldots$ and $\overleftarrow{[\cdot]}^{(\tau)}$ denotes the signal shifted τ steps backwards in time, i.e., $\overleftarrow{\psi}_i^{(\tau)} = \psi_{t-\tau}^{[i]}$. The latter notation is used to denote the effect of the time-shift operator in a product form. The Markov coefficients in (12) are independent of the parametrization of the matrix functions and a particular state bases, while the sub-Markov parameters are dependent on the parametrization of the functional dependencies in (3) even if they are independent of the state basis.

3.2 Correlation analysis

As the first step of the proposed identification scheme, i.e., estimating the sub-Markov parameters (13), we will present two methods: 1) *correlation analysis* (CRA) and 2) Bayesian LPV-FIR estimation in Sec. 3.3. CRA, as described in this section, estimates the sub-Markov coefficients individually, opposed to LPV-FIR. CRA results in an estimation procedure which grows linearly in the number of data points and is used to estimate each parameter individually. Hence, the correlation based estimation method has a low computational load. According to the proposed scheme, these sub-Markov parameters will be used to realize an SS form, as given in Sec. 4. The CRA makes use of the stochastic property of u, p, w, v, hence, in this section, u and p are assumed to be sample paths of the stochastic processes $\mathbf{u}$, $\mathbf{p}$, respectively. Note that, in such case, x and y obtained from (1) are sample paths of stochastic processes $\mathbf{x}, \mathbf{y}$ which satisfy $\mathbf{x}_{t+1} = \mathcal{A}(\mathbf{p}_t)\mathbf{x}_t + \mathcal{B}(\mathbf{p}_t)\mathbf{u}_t + \mathcal{G}(\mathbf{p}_t)\mathbf{w}_t$, and $\mathbf{y}_t = \mathcal{C}(\mathbf{p}_t)\mathbf{x}_t + \mathcal{D}(\mathbf{p}_t)\mathbf{u}_t + \mathcal{H}(\mathbf{p}_t)\mathbf{v}_t$. In the section, $\boldsymbol{\psi}_t$ denotes the process $[\,1\ \boldsymbol{\psi}_t^{[i]}\ \cdots\ \boldsymbol{\psi}_t^{[n_\psi]}\,]^\top$ and $\mathbb{I}_s^v$ denotes the set $\{s, s+1, \cdots, v\}$. The first step in the CRA is to define the k-dimensional cross-correlation.

Definition 6 *The k-dimensional cross-correlation function for the stationary signals $(\mathbf{u}, \mathbf{y}, \boldsymbol{\psi})$ is defined as*

$$R_{y\psi_{s_1}\cdots\psi_{s_n}u}(\tau_{s_1}, \ldots, \tau_{s_n}, \tau_u) = \mathbb{E}\left\{\mathbf{y}_t\overleftarrow{\boldsymbol{\psi}}_{s_1}^{(\tau_{s_1})}\cdots\overleftarrow{\boldsymbol{\psi}}_{s_n}^{(\tau_{s_n})}\left(\overleftarrow{\mathbf{u}}^{(\tau_u)}\right)^\top\right\},$$

where s_i is a specific index sequence with $s_1, \ldots, s_n \in \mathbb{I}_0^{n_\psi}$ and $\tau_{s_i} \in \mathbb{Z}_0^+$ is the time shift associated with the specific basis index s_i. □

Theorem 7 *The sub-Markov parameters are given as*

$$C_{s_1}A_{s_2}A_{s_3}\cdots A_{s_{n-1}}B_{s_n} = \frac{R_{y\psi_{s_1}\cdots\psi_{s_n}u}(\tau_{s_1}, \ldots, \tau_{s_n}, \tau_u)}{\sigma^2_{\psi_{s_1}}\ \cdots\ \sigma^2_{\psi_{s_n}}}\Sigma_u^{-2}, \quad (14)$$

and

$$D_{s_1} = \frac{R_{y\psi_{s_1}u}(0,0)}{\sigma^2_{\psi_{s_1}}}\Sigma_u^{-2}, \quad (15)$$

where $\tau_{s_i} = i-1$ and $\tau_u = \tau_{s_n}$ are the time-shifts of the signals $\boldsymbol{\psi}^{[s_1]}, \ldots, \boldsymbol{\psi}^{[s_n]}, \mathbf{u}$ for a specific index sequence $s_1, \ldots, s_n \in \mathbb{I}_0^{n_\psi}$. Eq. (14) *and* (15) *hold in case the following is satisfied:*

- ***C1*** *The output signal is generated by a stable LPV system* (1) *with dependency structure* (3)*.*
- ***C2*** *The noise processes* $\mathbf{w}$, $\mathbf{v}$ *are distributed as in* (2)*.*
- ***C3*** *The input process* $\mathbf{u}$ *is a white noise process with finite variance (*$\mathrm{var}(\mathbf{u}) = \Sigma_u^2$*) and is independent of* $\mathbf{w}$, $\mathbf{v}$*.*
- ***C4*** *Each process $\boldsymbol{\psi}_t^{[i]} \triangleq \psi_t^{[i]}(\mathbf{p}_t)$ is assumed to be a white noise process with finite variance ($\sigma^2_{\psi_0} = 1$, $\mathrm{var}(\boldsymbol{\psi}^{[i]}) = \sigma^2_{\psi_i}$ for $i = \mathbb{I}_1^{n_\psi}$). The processes $\boldsymbol{\psi}^{[i]}$ are mutually independent and $\boldsymbol{\psi}^{[i]}$ is independent of* $\mathbf{u}$*,* $\mathbf{w}$*, and* $\mathbf{v}$*.*

□

PROOF. See Appendix A. ■

Condition C4 is not over restrictive, e.g., if each ψ_i is a function of p_i only, the analytic function ψ_i is odd and bounded with $\psi_i(0) = 0$, and it is driven by a white noise scheduling signal p_i with finite variance, then C4 is satisfied. Note that the sub-Markov parameters in (7) do not depend on the time instant t. An approximation of the cross-correlation and variances in Thm. 7 can be used to estimate the sub-Markov parameters by a finite measured dataset $\mathcal{D}_N$. The variance of the involved signals is estimated by the unbiased sample variance and the k-dimensional cross-correlation is approximated via

$$\hat{\mathrm{R}}_{y\psi_i\cdots\psi_j u}(\tau_i, \ldots, \tau_j, \tau_u) = \frac{1}{N-\tau_u+1}\sum_{t=\tau_u+1}^{N} y\overleftarrow{\psi}_i^{(\tau_i)}\cdots\overleftarrow{\psi}_j^{(\tau_j)}\left(\overleftarrow{u}^{(\tau_u)}\right)^\top. \quad (16)$$

It is assumed that the time series $\mathbf{u}, \boldsymbol{\psi}, \mathbf{x}, \mathbf{y}, \mathbf{w}$ are such that $\lim_{N\to\infty}\hat{\mathrm{R}}_{y\psi_i\cdots\psi_j u}(\cdot) = \mathrm{R}_{y\psi_i\cdots\psi_j u}(\cdot)$. For example, this assumption holds with probability 1 if $\mathbf{u}, \boldsymbol{\psi}, \mathbf{x}, \mathbf{y}$ are jointly ergodic. Joint ergodicity has been proven in case $\boldsymbol{\psi}$ is a random binary noise and $\mathbf{u}$ is white noise [35].

The proposed CRA method may need a large dataset and $N \gg \tau_u$ such that variance of (16) is low enough for an accurate parameter estimate. If the noise process $\boldsymbol{\varpi}$ (11) is a zero mean coloured noise, e.g., under the general noise conditions of (1), the CRA estimation is inefficient [36], i.e., the variance of the estimated parameters does not correspond to the Cramér-Rao bound. Therefore, a larger dataset is required to achieve equivalent parameter estimation variance, comparison to the case that $\boldsymbol{\varpi}$ is a white noise with Gaussian distribution. However, the sub-Markov parameters can be estimated individually and the computational complexity scales with $\mathcal{O}\left(N(2 + n_\mathrm{y}^2 + n_\mathrm{y}n_\mathrm{u}n)\right)$ where n is the amount of specific index sequences $\{s_1, \ldots, s_n\}$ [2]. Hence,

[2] Unbiased sample variance scales with $\mathcal{O}(2N + n_\mathrm{y}^2)$ and (16)

the problem scales linearly in N, n_u, n and quadratic in n_y. For the basis reduced Ho-Kalman method only a subset of the sub-Markov parameters is needed for realization. Hence, the combination of the LPV-SS realization scheme with the CRA significantly reduces the computational demand, as identification of the full impulse response is omitted.

3.3 Bayesian impulse response estimation

As an alternative to CRA, the sub-Markov parameters can be estimated using a Ridge regression based LPV-FIR estimation procedure, where the optimal regularization matrix is determined in a Bayesian way with a Gaussian prior, i.e., *Bayesian LPV-FIR estimation*. How the Bayesian framework and regularized ℓ_2 framework connect, e.g., see [37]. In addition, the Bayesian framework also allows to estimate the functional dependencies $\psi^{[i]}(\cdot)$ in a nonparamtric way [38, 39]. However, for the sake of simplicity, we consider that these functions are known a priori.

3.3.1 The truncated IIR model

In the Bayesian framework, Eq. (11) is approximated by the following finite order truncation:

$$y_t \approx \sum_{i=0}^{n_\mathrm{h}} (h_i \diamond p_t) u_{t-i} + y_t^\mathrm{s}, \tag{17}$$

with $n_\mathrm{h} > 0$. Eq. (17) corresponds to a *finite impulse response* (FIR) model of (11) with order n_h. Due to the convergence of h_i, approximating (11) by (17) is not restrictive. Furthermore, define

$$\mathcal{M}_1 = \begin{bmatrix} B_0 \dots B_{n_\psi} \end{bmatrix}, \;\; \mathcal{M}_j = \begin{bmatrix} A_0 \mathcal{M}_{j-1} \dots A_{n_\psi} \mathcal{M}_{j-1} \end{bmatrix}. \tag{18}$$

Based on (13) and (17), the samples in $\mathcal{D}_N$ satisfy the following relationship:

$$\overline{Y}_N = \overline{\theta}_0 \, \overline{\Phi}_N + \overline{W}_N, \tag{19}$$

with

$$\overline{Y}_N = \begin{bmatrix} y_{n_\mathrm{h}+1} \dots y_N \end{bmatrix}, \qquad \overline{W}_N = \begin{bmatrix} y^\mathrm{s}_{n_\mathrm{h}+1} \dots y^\mathrm{s}_M \end{bmatrix},$$

$$\overline{\theta}_0 = \begin{bmatrix} D_0 \dots D_{n_\psi} \; C_0\mathcal{M}_1 \dots C_{n_\psi}\mathcal{M}_1 \; C_0\mathcal{M}_2 \dots C_{n_\psi}\mathcal{M}_{n_\mathrm{h}} \end{bmatrix},$$

$$\overline{\Phi}_N = \begin{bmatrix} \psi_{n_\mathrm{h}+1} \otimes u_{n_\mathrm{h}+1} & \dots & \psi_N \otimes u_N \\ \psi_{n_\mathrm{h}+1} \otimes \psi_{n_\mathrm{h}} \otimes u_{n_\mathrm{h}} & \dots & \psi_N \otimes \psi_{N-1} \otimes u_{N-1} \\ \vdots & \ddots & \vdots \\ \psi_{n_\mathrm{h}+1} \otimes \dots \otimes \psi_1 \otimes u_1 & \dots & \psi_M \otimes \dots \otimes \psi_{N-n_\mathrm{h}} \otimes u_{N-n_\mathrm{h}} \end{bmatrix},$$

where $M = N - n_\mathrm{h} - 1$, $\overline{Y}_N \in \mathbb{R}^{n_\mathrm{y} \times M}$ are the measured outputs, $\overline{\theta}_0 \in \mathbb{R}^{n_\mathrm{y} \times \sum_{i=1}^{n_\mathrm{h}+1} (1+n_\psi)^i n_\mathrm{u}}$ is the collection of the to-be-estimated sub-Markov parameters, $\overline{\Phi}_N \in \mathbb{R}^{\sum_{i=1}^{n_\mathrm{h}+1} (1+n_\psi)^i n_\mathrm{u} \times M}$ is the regression matrix and $\overline{W}_N \in \mathbb{R}^{n_\mathrm{y} \times M}$ is the cumulative noise process. The resulting output predictor of the MIMO FIR model (19) can be vectorized as

$$\hat{Y}_N = \Phi_N^\top \theta, \tag{20}$$

where $n_\theta = n_\mathrm{y} \sum_{i=1}^{n_\mathrm{h}+1} (1 + n_\psi)^i n_\mathrm{u}$, $\hat{Y}_N \in \mathbb{R}^{n_\mathrm{y} M \times 1}$ is the predicted output, $\Phi_N^\top = \overline{\Phi}_N^\top \otimes \mathrm{I}_{n_\mathrm{y}} \in \mathbb{R}^{n_\mathrm{y} M \times n_\theta}$, and $\theta \in \mathbb{R}^{n_\theta \times 1}$. For notational reasons, also introduce $Y_N = \mathrm{vec}(\overline{Y}_N)$, $\theta_0 = \mathrm{vec}(\overline{\theta}_0)$, and $W_N = \mathrm{vec}(\overline{W}_N)$.

3.3.2 Ridge regression based estimate

Even in the LTI case, a well-know issue in estimation of FIR models via the least-squares approach is the high variance of the estimated parameters, due to the relatively large number of parameters required to adequately represent the process dynamics. ℓ_2 Regularization makes it possible to control the so called *bias-variance trade-off*, i.e., dramatically decrease the variance by introducing a relatively small bias on the estimates [37]. The corresponding *weighted Ridge regression* or Tikhonov regularization problem is given by

$$\min_\theta \; \|\Phi_N^\top \theta - Y_N\|^2_{W_\mathrm{e}} + \|\theta\|^2_{W_\mathrm{r}}, \tag{21}$$

where $\|x\|_W = \sqrt{x^\top W x}$ denotes the *weighted Euclidean norm*, hence, the first term in (21) corresponds to a weighted ℓ_2 norm of the prediction error of (17), while the second term is the weighted ℓ_2 norm of θ. Both $W_\mathrm{e}, W_\mathrm{r} \in \mathbb{R}^{n_\theta \times n_\theta}$ are positive semi-definite (symmetric) regularization matrices and the analytic solution of (21) is

$$\hat{\theta}_\mathrm{RWLS} = \left(\Phi_N W_\mathrm{e} \Phi_N^\top + W_\mathrm{r} \right)^{-1} \Phi_N W_\mathrm{e} Y_N. \tag{22}$$

The regularization matrix W_r is chosen such that $\Phi_N W_\mathrm{e} \Phi_N^\top + W_\mathrm{r}$ is invertible. If $W_\mathrm{r} = 0$, $W_\mathrm{e} = I$, and $\mathbf{y}^\mathrm{s}$ is a white noise process with Gaussian distribution then (22) is the least squares solution, which results in the asymptotically efficient, unbiased, ML estimate.

Analogous to CRA, if $W_\mathrm{r} = 0$, $W_\mathrm{e} = I$, and the additive noise $\mathbf{y}^\mathrm{s}$ is a zero mean coloured noise process, but uncorrelated with the input and scheduling signals, e.g., $\mathbf{w} \neq 0$; then the estimator is still unbiased, although it will result in an inefficient estimator with increased variance. If $\mathbf{y}^\mathrm{s}$ and $\mathbf{u}$ are correlated then an LPV *instrumental variable* (IV) estimator can be used to remove the bias, e.g., see [18].

3.3.3 A Bayesian way of optimizing regularization

In this section, it is assumed that the innovation noise process is zero, i.e., $\mathbf{w} = 0$ in (1). Hence, the output additive noise process $\mathbf{v}$ in (1) is equal to the output additive noise $\mathbf{y}^\mathrm{s}$ in (11) and (17) corresponding to an output error setting. One of the main questions is how to chose the regularization matrix W_r, such that an optimal bias-variance trade-off is found. A recently introduced efficient data-driven approach follows an *empirical Bayes method* [40]. It is assumed that the parameter vector θ_0 is a random variable with *Gaussian distribution*:

$$\boldsymbol{\theta}_0 \sim \mathcal{N}(\theta_\mathrm{a}, P_\alpha), \qquad \theta_\mathrm{a} = 0,$$

where the covariance matrix P_α is a function of some hyper parameters $\alpha \in \mathbb{R}_+^{n_\alpha}$. In the Bayesian setting, under the assumption that u and p are given realizations, Φ_N is deterministic, and, according to (19), the output vector Y_N and the parameters θ_0 are jointly Gaussian variables:

scales with $\mathcal{O}(N n_\mathrm{y} n_\mathrm{u} n)$.

$$\begin{bmatrix}\boldsymbol{\theta}_0\\ \mathbf{Y}_N\end{bmatrix}\sim\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}P_\alpha & P_\alpha\Phi_N\\ \Phi_N^\top P_\alpha & \Phi_N^\top P_\alpha\Phi_N+\mathrm{I}_M\otimes\mathcal{R}\end{bmatrix}\right), \quad (23)$$

with $\mathcal{R}$ as in (2). It can be shown that the maximum *posteriori* estimate $\hat{\theta}_{\mathrm{p}}$ and minimal variance estimate of $\boldsymbol{\theta}_0$ given $\mathbf{Y}_N$ is equivalent to the weighted regularized least squares estimate $\hat{\theta}_{\mathrm{RWLS}}$ (22), e.g., see [37], if the weighting and regularization matrices are chosen as

$$W_{\mathrm{e}} = \mathrm{I}_M\otimes\mathcal{R}^{-1}, \qquad W_{\mathrm{r}} = P_\alpha^{-1}. \quad (24)$$

This connection makes it possible to create an estimate of $\mathcal{R}$ and P_α from data that minimizes the marginal likelihood (23). Notice that covariance matrix P_α, parametrized by α, and the noise covariance matrix $\mathcal{R}$ satisfy

$$\mathbf{Y}_N\sim\mathcal{N}\left(0,\Phi_N^\top P_\alpha\Phi_N+\mathrm{I}_M\otimes\mathcal{R}\right). \quad (25)$$

Hence, the likelihood function of the observation Y_N given α and $\mathcal{R}$ can be used to arrive to their posteriori estimate:

$$\begin{aligned}\hat{\alpha} &= \underset{\alpha}{\operatorname{argmax}}\ f(Y_N|\alpha) = \underset{\alpha}{\operatorname{argmin}}\ -2\log f(Y_N|\alpha)\\ &= \underset{\alpha}{\operatorname{argmin}}\ \log\left(\det\left(\Phi_N^\top P_\alpha\Phi_N+\mathrm{I}_M\otimes\mathcal{R}\right)\right)\\ &\quad + Y_N^\top\left(\Phi_N^\top P_\alpha\Phi_N+\mathrm{I}_M\otimes\mathcal{R}\right)^{-1}Y_N, \quad (26)\end{aligned}$$

where the constant terms are excluded and $f(\cdot)$ is the probability density function of the multivariate normal distribution. For a detailed description of pros and cons of the empirical Bayes method compared to other methods, see [41].

The choice of the parametrization of P_α is of big importance as it governs the "quality" of the estimate. The matrix $P_\alpha=\theta_0^\top\theta_0$ will give the lowest parameter mean-squared-error (MSE) [3] [42]. However, the true system parameters θ_0 are unknown, therefore, P_α should be parametrized to encode system properties such as asymptotic stability. For this purpose, many different kernel functions can be employed, see [43] for a detailed discussion. For the sake of simplicity, in this paper, we aim at Ridge regression, i.e., we will use $P_\alpha=\alpha\mathrm{I}$. Regularized regression, in general, is know to provide estimates with a lower parameter MSE compared to non-regularized methods, like the CRA method. On the other hand, for the regularized regression, the complete model needs to be estimated, from which not all parameters are used in the basis reduced Ho-Kalman realization. Consequently, the combination of regularized regression with LPV-SS realization loses computational efficiency compared to the CRA method with LPV-SS realization, but it is applicable under a much wider set of conditions (e.g., we can relax C2-C4).

4 A basis reduced Ho-Kalman SS realization

The aforementioned IO identification schemes of Sec. 3.2 and 3.3 can consistently estimate the process dynamics of (1) under mild assumptions. However, to achieve our goal; an efficient LPV-SS realization of the estimated IO model is needed. In [21], the well-known Ho-Kalman realization scheme is extended to the LPV case for realizing LPV-SS models with static and affine dependence on the scheduling variable. However, the size of the l-step extended observability and k-step extended reachability matrices grow exponentially in l,k and grow polynomially in the scheduling dimension n_ψ. Recently, a basis reduced Ho-Kalman scheme was proposed [26], where only the non-repetitive parts of the extended Hankel matrix are selected, which drastically decreases the computational load, compared to the full realization scheme of [21, 44]. The proposed scheme does not depend on any approximations, hence, it is an exact, deterministic realization scheme, and will be briefly explained in this section.

We will describe the methodology as the SS realization of a LPV system from a given set of sub-Markov parameters. To indicate which sub-Markov parameters of the involved extended reachability, observability, and Hankel matrices are selected, we need to introduce $[\mathbb{I}_s^v]^n$ as the set of all n-length sequences of the form $(i_1,\ldots,i_n)$ with $i_1,\ldots,i_n\in\mathbb{I}_s^v$. The elements of $\mathbb{I}_s^v$ will be viewed as characters and the finite sequences of elements of $\mathbb{I}_s^v$ will be referred to as strings. Then $[\mathbb{I}_s^v]^n$ is the set of all strings containing exactly n characters. The string $\alpha\in\left[\mathbb{I}_0^{n_\psi}\right]_0^n$ is called a "selection" with $n\geq 0$ where $\left[\mathbb{I}_0^{n_\psi}\right]_0^n=\{\epsilon\}\cup\mathbb{I}_0^{n_\psi}\cup\ldots\cup\left[\mathbb{I}_0^{n_\psi}\right]^n$ and ϵ denotes the empty string. As an example, $\left[\mathbb{I}_0^1\right]_0^2=\{\epsilon,0,1,00,01,10,11\}$. Define by $\#(\alpha)$ the amount of characters of a single string in the set. Applying a sequence α will give the ordering of multiplication of matrices $\{A_i\}_{i=0}^{n_\psi}$, which for $\#(\alpha)\geq 1$, is defined by

$$A_\alpha=\prod_{i=1}^{\#(\alpha)}A_{[\alpha]_i}=A_{[\alpha]_1}A_{[\alpha]_2}\cdots A_{[\alpha]_{\#(\alpha)}}, \quad (27)$$

where $[\alpha]_i$ denotes the i-th character of the string α. Note that $A_\epsilon=I$. To characterize a single sub-Markov parameter, the (i,j)-th element of the matrix $C_\gamma A_\alpha B_\beta\in\mathbb{R}^{n_{\mathrm{y}}\times n_{\mathrm{u}}}$ is denoted by $C_\gamma^{[i]}A_\alpha B_\beta^{[j]}$ for $\alpha\in[\mathbb{I}_0^{n_\psi}]_0^n$, $\beta\in\mathbb{I}_0^{n_\psi}$, $\gamma\in\mathbb{I}_0^{n_\psi}$. Based on the above defined notation, a selection of the extended reachability matrix is denoted by

$$\varsigma=\left\{(\alpha_1^\varsigma,\beta_1,j_1),\ldots,(\alpha_{n_{\mathrm{r}}}^\varsigma,\beta_{n_{\mathrm{r}}},j_{n_{\mathrm{r}}})\right\}, \quad (28)$$

where $\alpha_1^\varsigma,\ldots,\alpha_{n_{\mathrm{r}}}^\varsigma\in\left[\mathbb{I}_0^{n_\psi}\right]_0^n$, $\beta_1,\ldots,\beta_{n_{\mathrm{r}}}\in\mathbb{I}_0^{n_\psi}$, and $j_1,\ldots,j_{n_{\mathrm{r}}}\in\mathbb{I}_0^{n_{\mathrm{u}}}$. The length of the string α_i^ς may vary. Using this basis, a sub-matrix of the extended reachability matrix is selected, defined by

$$\mathcal{R}_\varsigma=\begin{bmatrix}A_{\alpha_1^\varsigma}B_{\beta_1}^{[j_1]} & A_{\alpha_2^\varsigma}B_{\beta_2}^{[j_2]} & \ldots & A_{\alpha_{n_{\mathrm{r}}}^\varsigma}B_{\beta_{n_{\mathrm{r}}}}^{[j_{n_{\mathrm{r}}}]}\end{bmatrix}, \quad (29)$$

where $\mathcal{R}_\varsigma\in\mathbb{R}^{n_{\mathrm{x}}\times n_{\mathrm{r}}}$ and $[j_k]$ denotes the j_k-th column of B_{β_k} for $k=1,\ldots,n_{\mathrm{r}}$. Analogously, a basis of the extended observability matrix is selected by

$$\nu=\left\{(i_1,\gamma_1,\alpha_1^\nu),\ldots,(i_{n_{\mathrm{o}}},\gamma_{n_{\mathrm{o}}},\alpha_{n_{\mathrm{o}}}^\nu)\right\}, \quad (30)$$

where $\alpha_1^\nu,\ldots,\alpha_{n_{\mathrm{o}}}^\nu\in\left[\mathbb{I}_0^{n_\psi}\right]_0^n$, $\gamma_1,\ldots,\gamma_{n_{\mathrm{o}}}\in\mathbb{I}_0^{n_\psi}$, and $i_1,\ldots,i_{n_{\mathrm{o}}}\in\mathbb{I}_0^{n_{\mathrm{y}}}$. This defines the sub-matrix of the extend observability matrix as

$$\mathcal{O}_\nu=\begin{bmatrix}\left(C_{\gamma_1}^{[i_1]}A_{\alpha_1^\nu}\right)^\top & \cdots & \left(C_{\gamma_{n_{\mathrm{o}}}}^{[j_{n_{\mathrm{o}}}]}A_{\alpha_{n_{\mathrm{o}}}^\nu}\right)^\top\end{bmatrix}^\top, \quad (31)$$

[3] The parameter mean-squared-error (MSE) for an estimator is defined as $\mathrm{MSE}(\hat{\theta}_N)=\mathbb{E}\{(\hat{\theta}_N-\theta_0)(\hat{\theta}_N-\theta_0)^\top\}$.

where $\mathcal{O}_\nu \in \mathbb{R}^{n_\mathrm{o} \times n_\mathrm{x}}$ and $[i_k]$ denotes the i_k-th row of C_{γ_k} for $k = 1, \ldots, n_\mathrm{o}$. The sets ς and ν are chosen appropriately, such that $\mathrm{rank}(\mathcal{R}_\varsigma) = n_\mathrm{x}$, $\mathrm{rank}(\mathcal{O}_\nu) = n_\mathrm{x}$, and hence $\mathrm{rank}(\mathcal{O}_\nu \mathcal{R}_\varsigma) = n_\mathrm{x}$. If this condition is satisfied then we call the selection ς and ν a *basis selection*. For such case, define

$$\begin{aligned} \mathcal{H}_{\nu,\varsigma} &= \mathcal{O}_\nu \mathcal{R}_\varsigma, \quad \mathcal{H}_{\nu,\varsigma,k} = \mathcal{O}_\nu A_k \mathcal{R}_\varsigma, \\ \mathcal{H}_{\nu,k} &= \mathcal{O}_\nu B_k, \qquad \mathcal{H}_{k,\varsigma} = C_k \mathcal{R}_\varsigma, \end{aligned} \tag{32}$$

where $\mathcal{H}_{\nu,\varsigma} \in \mathbb{R}^{n_\mathrm{o} \times n_\mathrm{r}}$, $\mathcal{H}_{\nu,\varsigma,k} \in \mathbb{R}^{n_\mathrm{o} \times n_\mathrm{r}}$, $\mathcal{H}_{\nu,k} \in \mathbb{R}^{n_\mathrm{o} \times n_\mathrm{u}}$ and $\mathcal{H}_{k,\varsigma} \in \mathbb{R}^{n_\mathrm{y} \times n_\mathrm{r}}$.

Lemma 8 *Define a column selection ς with $n_\mathrm{r} = n_\mathrm{x}$ and row selection ν with $n_\mathrm{o} \geq n_\mathrm{x}$ such that $\mathrm{rank}(\mathcal{H}_{\nu,\varsigma}) = n_\mathrm{x}$. The set of matrices*

$$\begin{aligned} \hat{A}_k &= \mathcal{H}_{\nu,\varsigma}^\dagger \mathcal{H}_{\nu,\varsigma,k}, \quad \hat{B}_k = \mathcal{H}_{\nu,\varsigma}^\dagger \mathcal{H}_{\nu,k}, \\ \hat{C}_k &= \mathcal{H}_{k,\varsigma}, \end{aligned} \tag{33}$$

for $k \in \mathbb{I}_0^{n_\psi}$ give a joint minimal LPV-SS representation of $\mathcal{S}$ (1) with the dependency structure (3), i.e., $\{\hat{A}_0, \ldots, \hat{C}_{n_\psi}\} \in \mathcal{Q}$. In (33), $\mathcal{H}_{\nu,\varsigma}^\dagger$ denotes the left pseudo inverse of $\mathcal{H}_{\nu,\varsigma}$. □

PROOF. The proof is straightforward by applying the isomorphism $T = \mathcal{R}_\varsigma^{-1}$ and the existence of $\mathcal{H}_{\nu,\varsigma}^\dagger$, as $\mathcal{H}_{\nu,\varsigma}$ has full column rank. ■

From the practical and numerical point of view, a reliable implementation of (33) follows by using *singular value decomposition* (SVD). Define a basis selection $n_\mathrm{r}, n_\mathrm{o} \geq n_\mathrm{x}$ with $\mathrm{rank}(\mathcal{H}_{\nu,\varsigma}) = n_\mathrm{x}$ and compute an economical SVD: $\mathcal{H}_{\nu,\varsigma} = U_{n_\mathrm{x}} \Sigma_{n_\mathrm{x}} V_{n_\mathrm{x}}^\top$. Then a realization of $\mathcal{S}$ is

$$\begin{aligned} \hat{A}_k &= \hat{\mathcal{O}}_\nu^\dagger \mathcal{H}_{\nu,\varsigma,k} \hat{\mathcal{R}}_\varsigma^\dagger, \quad \hat{B}_k = \hat{\mathcal{O}}_\nu^\dagger \mathcal{H}_{\nu,k}, \\ \hat{C}_k &= \mathcal{H}_{k,\varsigma} \hat{\mathcal{R}}_\varsigma^\dagger, \end{aligned} \tag{34}$$

with the pseudo inverses $\hat{\mathcal{R}}_\varsigma^\dagger = V_{n_\mathrm{x}} \Sigma_{n_\mathrm{x}}^{-1/2}$, $\hat{\mathcal{O}}_\nu^\dagger = \Sigma_{n_\mathrm{x}}^{-1/2} U_{n_\mathrm{x}}^\top$ for $k \in \mathbb{I}_0^{n_\psi}$. The realization of (34) gives an LPV-SS representation of $\mathcal{S}$ (1), i.e., $\{\hat{A}_0, \ldots, \hat{C}_{n_\psi}\} \in \mathcal{Q}$. The proof of this methodology can be found in [26].

In case the sub-Hankel matrices (32) are filled with estimated sub-Markov parameters, the state order n_x can be chosen based upon the magnitude of the singular values Σ_{n_x}, i.e., an approximate realization (e.g., see [45]). Note that the realization of (34) does not have any restrictions on the maximum amount of columns chosen $n_\mathrm{r} \geq n_\mathrm{x}$, compared to Lem. 8 where $n_\mathrm{r} = n_\mathrm{x}$. Hence, the rank-revealing property of the SVD of $\mathcal{H}_{\nu,\varsigma}$ allows to find a reliable estimate of n_x.

This bases reduced realization can considerably decrease the size of the Hankel matrix and, therefore, reducing the computational load, compared to realization with the full Hankel matrix [21, Eq. (48)]. In the basis reduced realization, the SVD is only applied on a $n_\mathrm{o} \times n_\mathrm{r}$ matrix instead of a matrix with size $n_\mathrm{y} \sum_{l=1}^{i} (1+n_\psi)^l \times n_\mathrm{u} \sum_{l=1}^{j} (1+n_\psi)^l$ in the full realization case. Note that $n_\mathrm{o}, n_\mathrm{r} = n_\mathrm{x}$ in the ideal case, hence, this is the computational lower bound that can be achieved in the optimal case. The amount of sub-Markov parameters in (32) is $n_\mathrm{o} n_\mathrm{r} + (1+n_\psi)(n_\mathrm{o} n_\mathrm{r} + n_\mathrm{o} n_\mathrm{u} + n_\mathrm{y} n_\mathrm{r})$, which increases linearly in all parameters $n_\psi, n_\mathrm{r}, n_\mathrm{o}, n_\mathrm{u}, n_\mathrm{y}$, compared to $n_\mathrm{y} \sum_{l=1}^{i} (1+n_\psi)^l \cdot n_\mathrm{u} \sum_{l=1}^{j} (1+n_\psi)^l$, which grows exponentially with increasing i and j and polynomially with increasing n_ψ. To illustrate, the realization of a system with input/output dimension $n_\mathrm{y} = n_\mathrm{u} = 2$, state dimension $n_\mathrm{x} = 4$, and scheduling dimension $n_\psi = 5$, the full Hankel matrix $\mathcal{H}_{2,2}$ has 7056 elements, while the sub-Hankel matrices for $n_\mathrm{r} = n_\mathrm{o} = 10$ have only 940 elements.

5 Maximum likelihood refinment

The basis reduced Ho-Kalman realization cannot guarantee that the LPV-SS model realized from the identified sub-Markov parameters is a maximum likelihood estimate, even if the underlying approaches are capable of providing ML estimates. Hence, to reach the maximum likelihood LPV-SS model estimate, two solutions are explored for refinement: 1) the gradient-based (GB) search method, or 2) the expectation maximization (EM) algorithm. Both methods are nonlinear iterative optimization techniques and cannot be used as stand alone methods, as they are prone to local minima. For example, [10, Table III] shows the number of failed model identification iterations for inefficient initial estimates on an LTI-SS identification problem. Hence, Step 1 and Step 2 of our proposed identification scheme, i.e., LPV impulse response estimation with LPV-SS realization, can be seen as a numerical efficient method for initializing GB or EM methods. The efficiency of this combination will be shown in Sec. 6.

5.1 Gradient based PEM

The set of PEM methods aims at minimizing the mean-squared prediction-error criterion w.r.t. the free model parameters (similar to (21) with $W_\mathrm{e} = I$, $W_\mathrm{r} = 0$). The minimazation problem is nonconvex and nonunique for any LPV-SS model based upon $\mathcal{D}_N$. Hence, to find the global optimum computationally efficient is not straight forward. The optimization is usually solved via a gradient-based search strategy such as a Newton or similar type method. In this paper, the enhanced Gauss-Newton based search method of [10] is used. The enhanced Gauss-Newton includes: 1) an automated strategy of regularization and SVD truncation on the Jacobian matrix to obtain a search direction, 2) an Armijo line search backtracking rule, and 3) lowering the dimension of the parameter space by using the *data-driven local coordinate* (DDLC) frame. As isomorphic SS models exist with equivalent IO behaviour, the optimization technique can wander among these sets. The manifold of all SS models with equivalent IO behaviour is known as the *indistinguishable set*. The DDLC frame is the ortho-complement of an affine approximation of the indistinguishable set around the current model parameters. Consequently, the DDLC ensures that the nonlinear optimization does not wander among parameterizations with equivalent IO behaviour and each iteration decreases the prediction error [10]. Additionally, in the LTI case, the DDLC results in a minimal parametrization and, hence, the PEM optimization problem is of minimal dimension. The DDLC has been successfully applied to LTI, bilinear, and LPV systems, e.g., see [9, 46, 47]. The combination of improved gradient-based search strategies

and the DDLC frame increases the computational demand per iteration, however, in general, it increases convergence rate. The original algorithm of [10] provides an identification method for LTI-SS models. The extension to an innovation-form LPV-SS structure with a static, affine Kalman matrix, i.e., $v_x = v_y$ and $\mathcal{G}(\cdot) = I$ (3), does not require additional technical steps and, therefore, is omitted.

5.2 Expectation Maximization

The key element of the EM method is to presume the existence of a complete dataset $Z_N = (Y_N, X_N)$, which contains not only the actual observations Y_N, but also the missing state-sequence X_N. The iterative EM method identifies LPV-SS models by considering the state sequence as the missing data. With this choice, the maximization of the ML is a joint estimation problem and is solved in an alternating manner. The LTI-SS methods are developed in [12,48] and a robust implementation of the EM is provided in [11,49]. For the refinment of the SS model, we will use [11] under the assumption that the data-generating system (1) is with $\mathcal{H}(p) = I$ and $\mathcal{G}(p) = I$ and noise structure (2). Each iteration of the EM consist of two steps: 1) the expectation, and 2) the maximization step. In the expectation step, given the current model estimate, the likelihood of the complete dataset conditional on the data observed is approximated. The likelihood, i.e., obtaining the unknown state trajectory x_t, can be estimated via various approaches, e.g., particle filtering [50], or Kalman filtering [51,52]. In the example section, the Kalman filter approach is used. In the second step, the maximization step, the approximated likelihood is maximized with respect to the model parameters. As the state sequence is known, the estimation problem becomes linear in the parameters with a closed-loop form solution. The EM method is relatively straightforward to implement and the computational load scales linearly with the dataset length. However, the EM algorithm usually converges rapidly in early stages, but its rate of convergence near the maximum is substantially lower than GB maximisation, e.g., see [12,51].

6 Simulation Example

In this section, the performance of the proposed three-step identification procedure is assessed via a Monte-Carlo simulation study using a randomly generated stable LPV-SS model in innovation form with scheduling independent matrix function, i.e., $\mathcal{K}(p_t) = K$. The Monte-Carlo study shows the performance of the methods in the following cases:

(1) Correlation analysis with basis reduced Ho-Kalman LPV-SS realization (without refinement step),
(2) Correlation analysis with basis reduced Ho-Kalman LPV-SS realization and EM or GB refinement step,
(3) Bayesian FIR estimation with basis reduced Ho-Kalman LPV-SS realization (without refinement step),
(4) Bayesian FIR estimation with basis reduced Ho-Kalman LPV-SS realization and EM or GB refinement step.

The proposed procedure is compared to state-of-the-art LPV-SS identification methods, such as the predictor-based subspace identification (PB) [25], successive approximation identification algorithm (SA) [27], and the robust identification/invalidation method (RI) [15]. Furthermore, these approaches' estimated SS model is refined, identical to the CRA and FIR, by using the SS model as initialization for the EM or GB method. Hence, it is assessed which approach can provide better initialization for the ML step and how far the delivered models are from the ML estimate. The case study is performed on a Macbook pro laptop, late 2013 with an 2.6GHz Intel core i5 and Matlab 2014b and, for the existing schemes, the scripts provided by the authors of [15, 25, 27] are used.

6.1 Data-generating system and model structure

The data-generating system is randomly selected in terms of a SS model (8a)-(8b) with input-output dimensions $n_{\mathrm{u}} = n_{\mathrm{y}} = 2$, scheduling dimension $n_\psi = 5$, minimal state dimension $n_{\mathrm{x}} = 4$, and affine dependence, i.e., the known basis functions are $\psi^{[i]} = p^{[i]}$ with $p^{[i]}$ denoting the i^{th} element of p. The SS model represented system has a scheduling independent matrix function, i.e., $\mathcal{K}(p_t) = K$. The innovation form is chosen, because all aforementioned methodologies are able to consistently identify this particular representation. The system was constructed such that (1) and the innovation form based output substituted equation

$$\check{x}_{t+1} = (\mathcal{A}(p_t) - K\mathcal{C}(p_t))\check{x}_t + (\mathcal{B}(p_t) - K\mathcal{D}(p_t))u_t + Ky_t,$$

are asymptotically input to state stable on the domain $p_t \in \mathbb{P} = [-1, 1]^5$, with a quadratic Lyapunov function defined by a constant symmetric matrix [53]. The LPV-SS model is available at www.rolandtoth.us.

6.2 Identification setting

The identification dataset is constructed from white $\mathbf{u}$ with uniform distribution $u_t \sim \mathcal{U}(-1, 1)$, and white $\mathbf{p}$ with random binary distribution on $(-0.9, 0.9)$, each of length $N = 5 \cdot 10^3$. The noise process $\boldsymbol{\xi}$ is taken as a white noise with distribution $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathcal{V})$ where $\mathcal{V}$ is diagonal and it is chosen such that the *signal-to-noise ratio* (SNR)

$$\mathrm{SNR}_y^{[i]} = 10 \log \frac{\sum_{t=1}^{N} (y_t^{[i]})^2}{\sum_{t=1}^{N} (\varpi_t^{[i]})^2},$$

is set for various Monte-Carlo experiments as $\mathrm{SNR}_y^{[i]} = \{40, 25, 10, 0\}$dB for all $i = 1, \ldots, n_{\mathrm{y}}$. The $[i]$ denotes the i-th channel, i.e., element of the vector signal, and $\mathrm{SNR}_y^{[i]}$ is the corresponding SNR on the output $y^{[i]}$. In this setting, the signals are jointly ergodic and the parameters can be consistently identified [35]. The performance of the scheme is tested on a validation dataset $\mathcal{D}_{\mathrm{val}}$ of length $N_{\mathrm{val}} = 200$, with different excitation conditions as the estimation dataset in terms of

$$u_t = \begin{bmatrix} 0.5\cos(0.035t) \\ 0.5\sin(0.035t) \end{bmatrix} + \delta_{t,u}, \tag{35}$$

$$p_t^{[i]} = 0.25 - 0.05i + 0.4\sin\left(0.035t + \frac{2i\pi}{5}\right) + \delta_{t,p_i}, \tag{36}$$

where $\delta_{t,u} \in \mathbb{R}^{n_{\mathrm{u}}}$, $\delta_{t,p_i} \in \mathbb{R}$ are element wise i.i.d. sequences with $\mathcal{U}(-0.15, 0.15)$. To study the statistical properties of the developed identification scheme, a Monte-Carlo

study with $N_{\mathrm{MC}} = 100$ runs is carried out, where in each run a new realization of the input, scheduling, and noise sequences are taken. However, the same $N_{\mathrm{MC}} = 100$ realization are given to all methods. The dataset is available at www.rolandtoth.us. We will asses the performance of the CRA, FIR, and RI model estimates without refinement step by comparing the simulated output $\hat{y}$ of the estimated model to the true output y. In all other cases, the one-step-ahead predicted output $\hat{y}$ of the estimated model is compared to the one-step-ahead predicted output y of the oracle (i.e., the one-step-ahead predicted output y using the original data-generating system). This dichotomy in assessing different signals is caused by the fact that the CRA, FIR, and RI do not identify a noise model, hence, the one-step-ahead predicted output is equal to the simulated output, therefore, comparing it to the measured simulated output is more adequate. On the other hand, the remaining methods include an estimate of a noise model, thus the plant and noise model are assessed by using the one-step-ahead predictor. In this case, it is compared w.r.t. the oracle, as its generated output is the maximum achievable output estimate given the dataset. The performance criterion used is the *best fit rate* (BFR) [4]

$$\mathrm{BFR} = \max\left\{1 - \frac{\frac{1}{N}\sum_{t=1}^{N}\|y_t - \hat{y}_t\|_2}{\frac{1}{N}\sum_{t=1}^{N}\|y_t - \bar{y}\|_2}, 0\right\} \cdot 100\%, \quad (37)$$

using $\mathcal{D}_{\mathrm{val}}$. In (37), $\bar{y}$ defines the mean of the predicted/true output y_t in $\mathcal{D}_{\mathrm{val}}$ and $\hat{y}_t$ is the simulated output w.r.t. (35) and (36) in $\mathcal{D}_{\mathrm{val}}$. Next, we will provide a summary of the used design parameters, and their values are chosen to provided the highest BFR. The FIR model order is chosen $n_{\mathrm{h}} = 2$ and $P_\alpha = \alpha I$. The hyperparameter α is tuned by using the Bayesian MIMO formulation of [54]. In the realization step, the basis reduced Ho-Kalman scheme uses $n_{\mathrm{o}} = n_{\mathrm{r}} = 10$ bases, where the controllability matrix is spanned by $\varsigma = \{(\epsilon, 0, 2), (\epsilon, 1, 2), (\epsilon, 2, 1), (\epsilon, 2, 2), \ldots, (\epsilon, 5, 2)\}$ and the observability is spanned by $\nu = \{(1, 0, \epsilon), \ldots, (2, 1, \epsilon), (2, 2, \epsilon), (1, 3, \epsilon), \ldots, (1, 4, \epsilon), (1, 5, \epsilon), (2, 5, \epsilon)\}$. The basis of the Hankel matrix is selected by using the entries of the full Hankel matrix with the largest absolute value. For the PB method, the future f and past window p are chosen as $f = p = 3$. For the SA method, the number of block rows in the Hankel matrix is chosen to be 4 and the iterative procedure is stopped if the 2-norm of the eigenvalues of the A_0 matrix do not change more then 10^{-6} or if it exceeds 100 iterations. For the RI method only the first 150 data samples are taken into account, as the computational complexity of the problem does not allow to use all data points of $\mathcal{D}_N$. For the EM method, the relative and absolute tolerance on the marginal log likelihood are chosen $2 \cdot 10^{-3}$ and 10^4, respectively, with a maximum of 20 iterations. For the GB method, we use $\beta = 10^{-4}$, $\gamma = 0.75$, $\eta_{\min} = 10^{-5}$, $\alpha_{\min} = 0.001$, $\nu = 0.01$, $\epsilon = 10^{-6}$, and a maximum of 20 iterations.

[4] Usually the BFR is defined per channel. Eq. (37) is the average performance criteria over all channels.

6.3 Analysis of the results

Table 1 shows the mean and the standard deviation of the BFR on $\mathcal{D}_{\mathrm{val}}$ and execution time of the estimation algorithms per Monte Carlo run for different $\mathrm{SNR}_y = \{40, 25, 10, 0\}$dB. Note that the SA method did not often converge for the underlying system of $n_{\mathrm{p}} = 5$ scheduling signals, hence, also a simulation study is done where the system to be identified had $n_{\mathrm{p}} = 2$ scheduling signals. In addition, remark that, the RI method only identifies $\mathcal{C}(\cdot), \mathcal{D}(\cdot)$ and assumes $\mathcal{A}(\cdot), \mathcal{B}(\cdot)$ to be known.

The table shows that the FIR with bases reduced realization outperforms the CRA, PB, SA, and RI methods. The CRA performs worse, because regularized methods, such as FIR, provide estimates with lower parameter MSE compared to the CRA by tuning the bias/variance trade-off. However, the bias/variance trade-off comes with a higher computational cost of around 4 times.

The PB is outperformed by the FIR, as it needs to estimate much more parameters, which is a well known problem [25, Table 1]. The increased amount of parameters to be identified is clearly reflected by the increased computational time. However, PB can identify models where the one-step-ahead predictor dynamics are stable, but not necessary the system dynamics and it can be used in a closed-loop identification setting.

The SA method has, in many cases, problems with convergence. Presumably, this is cased by the LTI subspace method to find the initial guess of the iterative scheme. The subspace method identifies a parameter independent state matrix A. However, the variations of this matrix w.r.t. the scheduling signal p are too severe to be neglected. This conclusion seems to be supported by the substantially higher BFR and the significant more converging trials if the data-generating system has $n_{\mathrm{p}} = 2$ in stead of $n_{\mathrm{p}} = 5$.

It is expected that the RI method outperforms the FIR, as the $\mathcal{A}(\cdot), \mathcal{B}(\cdot)$ matrix functions are known a priori. However, the computational complexity of the RI method does only allow for a small portion of the dataset $\mathcal{D}_N$ to be used for estimation (our case 150 out of 5000), hence, a large decrease in its performance is seen for lower SNRs.

All performance criteria indicate that the additional refinement step, with the EM or GB method, will lead to a better estimate of the model, as expected. Only in case of the $\mathrm{SNR}_y^{[i]} = 0$dB noise scenario, the EM refinement step will not improve the estimate. In this case, the EM method is not able to converge due to the large noise contribution. The GB method outperforms the EM method in all cases. Partially, this might be caused by the additional steps to improve the numerics in the GB method, i.e., the automated strategy of regularization and SVD truncation of the Jacobian matrix and line search backtracking rule, compared to the EM method and it can be caused by the slower rate of convergence of the EM. To improve the EM, similar rules to improve numerics could be implemented. In addition, we would like to highlight that the CRA and FIR are not statistically efficient under these noise scenarios, as they do not identify a noise model. Therefore, the BFR of CRA and FIR without refinement step can be further increased

by also identifying a noise model.

Summarizing, the presented three-step approach results in a maximum-likelihood estimate and accomplish this with a lower computational time and higher performance w.r.t. existing state-of-the-art LPV-SS identification approaches.

7 Conclusion

In this paper, we have presented a computational efficient, modular three-step LPV-SS identification approach, which contains the following three steps: 1) estimation of the Markov coefficient sequence using correlation analysis or a Bayesian FIR estimation, then 2) efficient LPV-SS realization by using a basis reduced Ho-Kalman method, and 3) refinement of the LPV-SS model estimate by a GB or EM optimization methodology. This three-step approach can consistently identify the underlying data-generating system. The effectiveness of the scheme has been demonstrated on a real-world sized MIMO LPV-SS model identification problem under harsh noise conditions and it has been compared to other methods. Any combination of the scheme was able to identify the system within seconds, significantly faster then its competitors with better performance.

Acknowledgements

We would like to thank the authors of [15, 25, 27] for providing their code to make the simulation example possible.

References

[1] M. Groot Wassink, M. Van de Wal, C. Scherer, and O. Bosgra, "LPV control for a wafer stage: beyond the theoretical solution," *Control Engineering Practice*, vol. 13, no. 2, pp. 231–245, 2005.

[2] J. Veenman, C. W. Scherer, and H. Köroğlu, "IQC-Based LPV controller synthesis for the NASA HL20 atmospheric re-entry vehicle," in *Proc. of the AIAA Guidance, Navigation, and Control Conf.*, (Chicago, IL, USA), pp. 1–16, Aug. 2009.

[3] J. W. van Wingerden, I. Houtzager, F. Felici, and M. Verhaegen, "Closed-loop identification of the time-varying dynamics of variable-speed wind turbines," *Int. J. of Robust and Nonlinear Control*, vol. 19, no. 1, pp. 4–21, 2009.

[4] A. A. Bachnas, R. Tóth, A. Mesbah, and J. H. A. Ludlage, "A review on data-driven linear parameter-varying modeling approaches: A high-purity distillation column case study," *J. of Process Control*, vol. 24, no. 4, pp. 272–285, 2014.

[5] J. Mohammadpour and C. Scherer, eds., *Control of linear parameter varying systems with applications*. Springer, 2012.

[6] M. Nemani, R. Ravikanth, and B. A. Bamieh, "Identification of linear parametrically varying systems," in *Proc. of the 34th IEEE Conf. on Decision and Control*, (New Orleans, LA, USA), pp. 2990–2995, 1995.

[7] S. Rizvi, J. Mohammadpour, R. Tóth, and N. Meskin, "An IV-SVM-based approach for identification of state-space LPV models under generic noise conditions," in *Proc. of the 54th IEEE Conf. on Decision and Control*, (Osaka, Japan), pp. 7380–7385, Dec. 2015.

[8] L. H. Lee and K. R. Poolla, "Identifiability issues for parameter-varying and multidimensional linear systems," in *Proc. of the ASME Design, Engineering Technical Conf.*, (Sacramento, California), 1997.

[9] V. Verdult, N. Bergboer, and M. Verhaegen, "Identification of fully parameterized linear and nonlinear state-space systems by projected gradient search," in *Proc. of the 13th IFAC Symposium on System Identification*, (Rotterdam, The Netherlands), pp. 737–742, Aug. 2003.

[10] A. Wills and B. Ninness, "On gradient-based search for multivariable system estimates," *IEEE Trans. on Automatic Control*, vol. 53, pp. 298–306, Feb. 2008.

[11] 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. Lopes dos Santos, T. P. Azevedo Perdicoúlis, C. Novara, J. A. Ramos, and D. E. Rivera, eds.), ch. 11, pp. 295–316, World Scientific, 2011.

[12] M. W. Watson and R. F. Engle, "Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models," *J. of Econometrics*, vol. 23, pp. 385–400, Dec. 1983.

[13] G. Z. Angelis, "System Analysis, Modelling and Control with Polytopic Linear Models," 2001.

[14] P. Gáspár, Z. Szabó, and J. Bokor, "A grey-box identification of an LPV vehicle model for observer-based side-slip angle estimation," in *Proc. of the American Control Conf.*, (New York City, USA), pp. 2961–2965, July 2007.

[15] F. D. Bianchi and R. S. Sánchez-Peña, "Robust identification/invalidation in an LPV framework," *Int. J. of Robust and Nonlinear Control*, vol. 20, no. 3, pp. 301–312, 2009.

[16] C. Novara, "SM Identification of State-Space LPV systems," in *Linear parameter-varying system identification: new developments and trends* (P. Lopes dos Santos, T. P. Azevedo Perdicoúlis, C. Novara, J. A. Ramos, and D. E. Rivera, eds.), ch. 11, pp. 65–93, World Scientific, 2011.

[17] V. Ceronea, D. Piga, and D. Regruto, "A convex relaxation approach to set-membership identification of LPV systems," *Automatica*, vol. 49, p. 28532859, Sep. 2013.

[18] 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.

[19] P. Lopes dos Santos, T.-P. Azevedo-Perdicoúlis, C. Novara, and D. E. R. J. A. Ramos, eds., *Linear parameter-varying system identification: new developments and trends*. World Scientific, 2011.

[20] R. Tóth, *Modeling and identification of linear parameter-varying systems*. Springer, 2010.

[21] R. Tóth, H. S. Abbas, and H. Werner, "On the state-space realization of LPV input-output models: practical approaches," *IEEE Trans. on Control Systems Technology*, vol. 20, pp. 139–153, Jan. 2012.

[22] F. Felici, J. W. van Wingerden, and M. Verhaegen, "Subspace identification of MIMO LPV systems using a periodic scheduling sequence," *Automatica*, vol. 43, no. 10, pp. 1684–1697, 2006.

[23] P. Lopes dos Santos, J. A. Ramos, and J. L. M. de Carvalho, "Identification of linear parameter varying systems using an iterative deterministic-stochastic subspace approach," in *Proc. of the European Control Conf.*, (Kos, Greece), pp. 4867–4873, Jul 2007.

[24] W. E. Larimore, "Identification of nonlinear parameter-varying systems via canonical variate analysis," in *Proc. of the American Control Conf.*, (Washington, DC, USA), pp. 2247–2262, Jun 2013.

[25] 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.

[26] P. B. Cox, R. Tóth, and M. Petreczky, "Estimation of LPV-SS models with static dependency using correlation analysis," in *Proc. of the 1st IFAC Workshop on Linear Parameter Varying Systems*, (Grenoble, France), pp. 91–96, Oct. 2015.

[27] P. Lopes dos Santos, J. A. Ramos, and J. L. Martins de Carvalho, "Subspace identification of linear parameter-varying systems with

Table 1
The mean and the standard deviation (between parentheses) of the BFR and execution time of the estimation algorithms per Monte Carlo run for different $\mathrm{SNR}_y^{[i]} = \{40, 25, 10, 0\}$dB is given. The BFR is based on the one-step-ahead predicted output of the estimated model on the validation dataset, except the methods with an asterisk which are based on the simulated output. The correlation analysis (CRA), finite impulse response (FIR), the predictor-based subspace identification (PB), successive approximation identification (SA), and the robust identification/invalidation (RI) method are used and refined by the expectation-maximization (EM) or gradient based (GB) algorithm. The SA2 indicates the results for the SA method where the system to be identified had $n_{\mathrm{p}} = 2$ scheduling signals. For this table, $N_{\mathrm{MC}} = 100$ Monte Carlo simulations are performed. The number in superscript indicates how many successful trails are taken into account out of the 100 runs.

	BFR [%]				**Time Elapsed** [s]	
	40dB	25dB	10dB	0dB	25dB	0dB
CRA*	81.47 (4.106)	81.44 (4.433)	81.14 (4.504)	75.07 (5.823)	2.269 (0.1481)	2.229 (0.1519)
CRA + EM	99.71 (0.05576)	98.80 (0.1710)	91.31 (0.5744)	74.13 (1.804)	8.297 (0.2722)	8.351 (0.2978)
CRA + GB	99.86 (0.02888)	99.27 (0.1600)	95.81 (0.8313)	87.81 (2.513)	8.626 (0.6105)	12.94 (1.146)
FIR*	99.32 (0.1466)	98.74 (0.2872)	94.19 (1.346)	83.55 (3.868)	10.87 (0.5037)	9.098 (0.4474)
FIR + EM	99.73 (0.05327)	98.80 (0.1679)	91.26 (0.5432)	74.14 (1.788)	16.88 (0.6055)	15.21 (0.3635)
FIR + GB	99.86 (0.02886)	99.27 (0.1600)	95.81 (0.8313)	87.81 (2.513)	17.21 (0.8603)	19.62 (1.236)
PB	95.90 (1.185)	86.25 (2.761)	75.89 (3.051)	61.35 (8.831)	88.47 (0.5274)	88.43 (0.4078)
PB + EM	98.50 (0.2035)	98.02 (0.2849)	92.39 (1.703)	78.80 (4.875)	92.84 (0.5456)	92.79 (0.4136)
PB + GB	99.79 (0.05832)	99.27 (0.1600)	95.81 (0.8312)	87.57 (2.893)	97.62 (1.317)	100.2 (0.4811)
SA2	82.97 (12.90) ^{77}	83.83 (9.771) ^{83}	83.04 (10.70) ^{56}	73.78 (12.04) ^{37}	21.52 (8.165)	26.86 (3.883)
SA2 + GB	97.29 (13.56) ^{77}	99.51 (0.1216) ^{81}	96.68 (2.363) ^{53}	90.49 (2.249) ^{37}	26.94 (8.126)	34.21 (3.687)
SA	65.57 (4.225) ^{18}	65.55 (3.283) ^{13}	66.18 (3.630) ^{24}	53.24 (8.806) ^{29}	108.9 (3.442)	107.0 (0.3076)
SA + GB	97.75 (8.483) ^{16}	99.31 (0.1306) ^{13}	94.51 (6.373) ^{24}	87.51 (2.358) ^{27}	118.0 (3.615)	122.3 (1.793)
RI*	99.25 (0.1876)	95.75 (0.9813)	76.11 (5.991)	26.66 (15.25)	103.4 (5.261)	108.9 (6.406)
RI + EM	99.73 (0.05359)	98.80 (0.1710)	91.29 (0.5441)	74.34 (1.663)	109.3 (5.307)	114.8 (6.438)
RI + GB	99.86 (0.02887)	99.27 (0.1600)	95.81 (0.8313)	87.81 (2.513)	110.3 (5.352)	121.3 (6.465)

innovation-type noise models driven by general inputs and a measurable white noise time-varying parameter vector," *Int. J. of Systems Science*, vol. 39, no. 9, pp. 897–911, 2008.

[28] V. Bogachev, *Measure Theory*. Springer-Verlag Berlin Heidelberg, 2007.

[29] M. Petreczky and G. Mercere, "Affine LPV systems: realization theory, input-output equations and relationship with linear switched systems," in *Proc. of 51th IEEE Conf. on Decision and Control*, (Maui, Hi, USA), pp. 4511–4516, Dec. 2012.

[30] M. Petreczky and R. Vidal, "Realization theory of generalized bilinear systems," *Submitted to IEEE Trans. on Automatic Contron*.

[31] M. Verhaegen and V. Verdult, *Filtering and system identification: a least squares approach*. Cambridge university press, 2007.

[32] P. B. Cox and R. Tóth, "On the connection between different noise structures for LPV-SS models." arXiv: 1610.09173.

[33] V. Verdult and M. Verhaegen, "Subspace identification of multivariable linear parameter-varying systems," *Automatica*, vol. 38, no. 5, pp. 805–814, 2002.

[34] A. G. Dankers, X. B. R. Tóth, P. S. C. Heuberger, and P. M. J. Van den Hof, "Informative data and identifiability in LPV-ARX prediction-Error identification," in *Proc. of the 50th IEEE Conf. on Decision and Control*, (Orlando, FL, USA), pp. 799–804, Dec. 2011.

[35] M. Petreczky and L. Bako, "On the notion of persistence of excitation for linear switched systems," in *Proc. of 50th IEEE Conf. on Decision and Control and European Control Conf.*, (Orlando, FL, USA), pp. 1840–1847, Dec. 2011.

[36] L. Ljung, *System identification: theory for the user*. Springer, 1989.

[37] L. Ljung and T. Chen, "What can regularization offer for estimation of dynamical systems?," in *Proc. of 11th IFAC International Workshop on Adaptation and Learning in Control and Signal Processing*, (Caen, France), pp. 1–8, Jul. 2013.

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

[39] 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), pp. 66–71, Dec. 2015.

[40] B. P. Carlin and T. A. Louis, *Bayes and empirical Bayes methods for data analysis*. London: Chapman & Hall, 2nd ed., 1996.

[41] 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.

[42] Y. C. Eldar, "Uniformly improving the Cramér-Rao bound and maximum-likelihood estimation," *IEEE Trans. on Signal Processing*, vol. 54, no. 8, pp. 2943–2956, 2006.

[43] 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.

[44] B. L. Ho and R. E. Kalman, "Effective construction

of linear state-variable models from input/output functions," *Automatisierungstechnik*, vol. 14, no. 1-12, p. 545548, 1966.

[45] S. Y. Kung, "A new identification and model reduction algorithm via singular value decompositions," in *Proc. 12th Asilomar Conf. on Circuits, Systems and Computers*, (Pacific Grove, CA, USA), pp. 705–714, Nov. 1978.

[46] L. H. Lee and K. Poolla, "Identification of linear parameter-varying systems using nonlinear programming," *J. of Dynamic Systems, Measurement, and Control*, vol. 121, no. 1, pp. 71–78, 1999.

[47] T. McKelvey, A. Helmersson, and T. Ribarits, "Data driven local coordinates for multivariable linear systems and their application to system identification," *Automatica*, vol. 40, no. 9, pp. 1629–1635, 2004.

[48] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *J. Time Series Analysis*, vol. 3, pp. 253–264, Jul. 1982.

[49] S. Gibson and B. Ninness, "Robust maximum-likelihood estimation of multivariable dynamic systems," *Automatica*, vol. 41, pp. 1667–1682, Oct. 2005.

[50] T. B. Schön, A. Wills, and B. Ninness, "System identification of nonlinear state-space models," *Automatica*, vol. 47, no. 1, pp. 39–49, 2011.

[51] S. Gibson, A. Wills, and B. Ninness, "Maximum-likelihood parameter estimation of bilinear systems," *IEEE Trans. on Automatic Control*, vol. 50, no. 10, pp. 1581–1596, 2005.

[52] R. Shumway and D. Stoffer, *Time Series Analysis and its Applications: with R Examples*. Springer Texts in Statistics, 2010.

[53] C. W. Scherer, "Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control for time-varying and linear parametrically-varying systems," *Int. J. of Robust and Nonlinear Control*, vol. 6, pp. 929–952, 1996.

[54] P. B. Cox and R. Tóth, "LPV state-space model identification in the Bayesian setting: A 3-step procedure," in *Proc. of the American Control Conf.*, (Boston, MA, USA), Jul. 2016.

A Proof of Theorem 7

The proof is found by computing the expected value of the cross-correlation between the stationary signals $\mathbf{y}, \boldsymbol{\psi}, \mathbf{u}$, where the signals are assumed to be ergodic. Let us first show the relation of the direct feed through matrices D_{s_1}. Lets substitute the IIR (11) for $\mathbf{y}$ in $\mathrm{R}_{y\psi_{s_1}u}(0,0)$, which gives

$$\begin{aligned}
&\mathrm{R}_{y\psi_{s_1}u}(0,0) = \\
&= \mathbb{E}\left\{(\mathcal{D}(\mathbf{p}_t)\mathbf{u}_t + \mathcal{C}(\mathbf{p}_t)\mathcal{B}(\mathbf{p}_{t-1})\mathbf{u}_{t-1} + \cdots + \mathbf{y}_t^{\mathrm{s}})\boldsymbol{\psi}_t^{[s_1]}\mathbf{u}_t^\top\right\} \\
&= \mathbb{E}\left\{\left(D_0 + \sum_{i=1}^{n_\psi} D_i\boldsymbol{\psi}_t^{[i]}\right)\mathbf{u}_t\boldsymbol{\psi}_t^{[s_1]}\mathbf{u}_t^\top\right\} + \\
&\mathbb{E}\left\{\mathcal{C}(\mathbf{p}_t)\mathcal{B}(\mathbf{p}_{t-1})\left(\mathbf{u}_{t-1}\right)\boldsymbol{\psi}_t^{[s_1]}\mathbf{u}_t^\top\right\} + \cdots + \mathbb{E}\{\mathbf{y}_t^{\mathrm{s}}\boldsymbol{\psi}_t^{[s_1]}\mathbf{u}_t^\top\} \\
&= D_{s_1}\sigma^2_{\psi_{s_1}}\Sigma_u^2. \qquad \text{(A.1)}
\end{aligned}$$

Eq. (A.1) holds due to the whiteness property of the processes $(\mathbf{u}, \boldsymbol{\psi})$ and their independence. Also see that $\mathbb{E}\{\mathbf{y}_t^{\mathrm{s}}\boldsymbol{\psi}_t^{[s_1]}\mathbf{u}_t^\top\} = 0$, as $\mathbf{w}$, $\mathbf{v}$, and $\boldsymbol{\psi}$ are assumed to be independent of $\mathbf{u}$ and $\mathbf{y}^{\mathrm{s}}$ satisfies the relation given in (12), therefore, $\mathbf{y}^{\mathrm{s}}$ is independent from $\mathbf{u}$. Hence, $\mathbb{E}\{\mathbf{y}_t^{\mathrm{s}}\boldsymbol{\psi}_t^{[s_1]}\mathbf{u}_t^\top\} = \mathbb{E}\{\mathbf{y}_t^{\mathrm{s}}\boldsymbol{\psi}_t^{[s_1]}\}\mathbb{E}\{\mathbf{u}_t^\top\} = 0$. For all other sub-Markov parameters, let us consider an ordering of the signal shifts as $\tau_{s_i} = i - 1$, $\tau_u = \tau_{s_n}$, which results in the following formulation

$$\begin{aligned}
&\mathrm{R}_{y\psi_{s_1}\cdots\psi_{s_n}u}(\tau_{s_1}, \ldots, \tau_{s_n}, \tau_u) = \\
&= \mathbb{E}\Big\{ (\mathcal{D}(\mathbf{p}_t)\mathbf{u}_t + \mathcal{C}(\mathbf{p}_t)\mathcal{B}(\mathbf{p}_{t-1})\mathbf{u}_{t-1} + \cdots + \mathbf{y}_t^{\mathrm{s}}) \\
&\qquad \overleftarrow{\boldsymbol{\psi}}_{s_1}^{(\tau_{s_1})} \cdots \overleftarrow{\boldsymbol{\psi}}_{s_n}^{(\tau_{s_n})} \left(\overleftarrow{\mathbf{u}}^{(\tau_u)}\right)^\top \Big\} \\
&= \mathbb{E}\Big\{ C_{s_1}A_{s_2}\cdots A_{s_{n-1}}B_{s_n} \left(\overleftarrow{\boldsymbol{\psi}}_{s_1}^{(\tau_{s_1})}\right)^2 \cdots \\
&\qquad \left(\overleftarrow{\boldsymbol{\psi}}_{s_n}^{(\tau_{s_n})}\right)^2 \overleftarrow{\mathbf{u}}^{(\tau_u)} \left(\overleftarrow{\mathbf{u}}^{(\tau_u)}\right)^\top \Big\} + \\
&\quad \mathbb{E}\Big\{ (\mathcal{D}(\mathbf{p}_t)\mathbf{u}_t + \cdots + \boldsymbol{\varpi}_t) \overleftarrow{\boldsymbol{\psi}}_{s_1}^{(\tau_{s_1})} \cdots \overleftarrow{\boldsymbol{\psi}}_{s_n}^{(\tau_{s_n})} \left(\overleftarrow{\mathbf{u}}^{(\tau_u)}\right)^\top \Big\} \\
&= C_{s_1}A_{s_2}\cdots A_{s_{n-1}}B_{s_n}\sigma^2_{\psi_{s_1}} \cdots \sigma^2_{\psi_{s_n}}\Sigma_u^2. \qquad \text{(A.2)}
\end{aligned}$$

Reordering (A.1) and (A.2) concludes the proof.

Remark 9 *It is possible to get the same sub-Markov parameters with different multiplications of $\boldsymbol{\psi}_{s_i}$ and corresponding shifts, e.g., $\mathrm{R}_{y\psi_{s_1}\psi_{s_2}u}(0,1,1)$ gives the same sub-Markov parameter $C_{s_1}B_{s_2}$ as $\mathrm{R}_{y\psi_{s_1}\ldots\psi_{s_4}u}(0,1,4,4,1)$. In scope of the estimation of these sub-Markov parameters, we impose the above given ordering to keep the multiplications with $\boldsymbol{\psi}^{[s_i]}$ minimal.*