2019 American Control Conference (ACC)
Philadelphia, PA, USA, July 10-12, 2019

Variational Inference for Linear Systems with Latent Parameter Space

Cassiano O. Becker[†] Victor M. Preciado[†]

Abstract— **We present a method to perform identification of systems with external inputs whose parameters are indexed by a lower-dimensional latent space. We apply a variational Bayes inference method to approximate the posterior distribution of the system parameters and latent variables, given input and output measurements. This approach seeks to minimize the Kullback-Leibler divergence between the full (but intractable) posterior distribution of the parameters and an approximating (yet tractable) factorized distribution. The method enables inference for systems whose parameters are subject to latent sources of variation, and therefore constitutes a relevant tool for modeling and control in complex domains, such as biological systems and neuroscience.**

I. Introduction

Estimation of parameters for linear time-invariant (LTI) systems is a problem with well established solutions that comprise, for example, subspace identification [1], Expectation-Maximization algorithms [2], and Bayesian methods [3]. However, in many application settings, system models require the ability to incorporate mechanisms of variation in their parameters [4], thereby demanding the consideration of a larger class of system models and estimation algorithms. In particular, many non-linear systems can be reformulated to belong to specific classes of linear models with *varying* parameters, such as linear time-varying (LTV) systems [5], or linear parameter-varying (LPV) systems [6]. In this respect, while LTV systems benefit from a mature theory for their analysis [7], their identification is statistically challenging due to the large number of parameters to be estimated [8]. To address such difficulty, the problem considered in this paper explores a configuration whereby the variation in parameter values can be described as the *linear span* of an unknown, lower-dimensional *latent space* to be identified. More specifically, this linear span formulation can be conveniently described using the formalism of LPV systems, a variant of which we adopt in this paper. In this case, it is commonly assumed that the parameters scheduling the dynamic variation of the model have values that are known or are fully observed [9], [10]. Conversely, in this paper, this assumption is relaxed to different scenarios where information about those parameters is considered to be noisy or only partially available. Such scenarios can be expected in complex systems occurring, for example, in economics [11] and neuroscience [12], [13]), where effects from large-scale, distributed or latent structure might restrict access to measurements of the scheduling parameters.

This work was supported in part by the US National Science Foundation under the grant CAREER-ECCS-1651433. C.O.B. is supported in part by CAPES, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil.

[†]Department of Electrical and Systems Engineering, School of Engineering and Applied Science, University of Pennsylvania.

In these models, the problem of inference can benefit from a Bayesian approach [14], where the unknown parameters are treated as random variables associated to measurements through conditional probability relationships. The estimation goal is defined as that of performing inference of the posterior distribution over model parameters, conditioned on the available measurements. In particular, the variational Bayesian inference framework adopted in this paper provides a computationally efficient method to approximate posterior distributions [15], [16] in the common case where exact inference in intractable. This approach presents a computational cost that is comparable to (non-Bayesian) Expectation-Maximization methods [17], [18], while yielding more descriptive models (with posterior distribution over parameters) that are also less prone to overfitting [19].

With respect to previous studies, the problem of deterministic system identification for LPV models with known parameters has been addressed, for example, using subspace methods [20] or orthonormal base functions [21], as well as in many recent works [22]–[26]. In the stochastic setting, an Expectation-Maximization method has been proposed in [27]. More recently, Bayesian methods have been applied in [28]–[30]. In [31], the problem of LPV identification with uncertain scheduling variables was addressed, but considering additive error models and applying a Gaussian Processes framework. In [32], LPV identification with noise-corrupted scheduling observations was considered, however the treatment focused on single-input single-output systems. In particular, the approach taken in this paper is inspired by [3] and [33]. The former introduced the application of variational Bayesian inference to linear time-invariant state space systems, while the latter extended the application of variational methods to time-varying systems without considering the effect of external inputs. In this paper, we allow for both time-invariant and parameter-varying input-to-state actuation (a relevant case for control applications) and describe different parameter scheduling scenarios. For clarity of exposition, we simplify the treatment of prior probabilities over hyper-parameters, and provide detailed derivations of the method, hoping to further motivate the use of the variational inference framework in related problems for LTV/LPV system identification.

II. Problem Formulation

We consider a discrete-time linear parameter-varying model evolving according to the set of equations

$$x(k+1) = \mathcal{A}(k)x(k) + \mathcal{B}(k)u(k) + v(k), \tag{1}$$

for $k = 1, 2, \ldots$, where $x_k \equiv x(k) \in \mathbb{R}^n$ is the dynamic state, $u_k \equiv u(k) \in \mathbb{R}^m$ are external inputs, and $v_k \equiv v(k) \in$

$\mathbb{R}^n$ accounts for additive noise. The parameter-varying matrices $\mathcal{A}_k \equiv \mathcal{A}(k) \in \mathbb{R}^{n\times n}$ and $\mathcal{B}_k \equiv \mathcal{B}(k) \in \mathbb{R}^{n\times m}$ present a time-varying behavior induced by the scheduling parameters $\{z_s(k) \in \mathbb{R}\}_{s=0}^{\ell}$ and $\{w_s(k) \in \mathbb{R}\}_{s=0}^{d}$, through the linear parametrization

$$\mathcal{A}_k = \sum_{s=0}^{\ell} z_s(k) A_s, \quad \mathcal{B}_k = \sum_{s=0}^{d} w_s(k) B_s,$$

which is defined in terms of the set of parameters $A = \{A_s\}_{s=0}^{\ell}$ and $B = \{B_s\}_{s=0}^{d}$. In particular, an affine parametrization having $\mathcal{A}_k = A_0 + \sum_{s=1}^{\ell} z_s(k)A_s$ or $\mathcal{B}_k = B_0 + \sum_{s=1}^{d} w_s(k)B_s$ can be recovered by setting $z_0(k) = 1$ or $w_0(k) = 1$ for all k.

To develop our probabilistic treatment in the Bayesian approach, we consider that the parameters in A and B are random variables associated with probability distributions $p(A)$ and $p(B)$. Their distributions are assumed to be component-wise zero-mean normally distributed over each element of A_s and B_s, with variances α^{-1} and β^{-1}, respectively. Therefore,

$$p(A) = \prod_{s=0}^{\ell} p(A_s) \text{ with } p(A_s) = \prod_{r=1}^{n}\prod_{j=1}^{n} \mathcal{N}([A_s]_{r,j}|0, \alpha^{-1}),$$

$$p(B) = \prod_{s=0}^{d} p(B_s) \text{ with } p(B_s) = \prod_{r=1}^{n}\prod_{h=1}^{m} \mathcal{N}([B_s]_{r,h}|0, \beta^{-1}).$$

Further, we note that (1) induces a Markovian-dependence between successive states, which describes the joint probability over the set $X = \{x_k\}_{k=1}^{N+1}$ (conditioned on parameters and latent variables) as

$$p(X|U,Z,W,A,B) = p(x_1)\prod_{k=1}^{N} p(x_{k+1}|x_k, u_k, z_k, w_k, A, B),$$

where we have defined the sets $U = \{u_k\}_{k=1}^{N}$ and $Z = \{z_k\}_{k=1}^{N}$, with $z_k \in \mathbb{R}^{\ell}$ with $z_k = [z_0(k), \dots, z_l(k)]^T$ (and $W = \{w_k\}_{k=1}^{N}$, accordingly).

We now assign a probability distribution to $v(k)$, which we assume to be a zero-mean Gaussian distribution with diagonal covariance, i.e.,

$$v_k \sim \mathcal{N}(v_k|0, \mathrm{diag}(\eta)), \tag{2}$$

where $\eta \in \mathbb{R}_{++}^n$, $\eta = [\eta_1, \dots, \eta_r]^T$. This allows us to write, conditioned on the parameters and latent variables, the state transition probability

$$p(x_{k+1}|x_k, u_k, z_k, w_k, A, B) = \mathcal{N}\Big(x_{k+1}\Big|\sum_{s=0}^{\ell} z_s(k)A_s x_k + \sum_{s=0}^{d} w_s(k) B_s u_k,\ \mathrm{diag}(\eta)\Big).$$

For a more compact treatment, we refer to $a_r^{(s)}$ as the r-th row of A_s and define the matrix parameter $A^{(r)} \in \mathbb{R}^{n\times(\ell+1)}$ where $A^{(r)} = [a_r^{(0)T}|\cdots|a_r^{(\ell)T}]$. Likewise, we define $B^{(r)} \in \mathbb{R}^{m\times(d+1)}$, having $B^{(r)} = [b_r^{(0)T}|\cdots|b_r^{(d)T}]$ with $b_r^{(s)}$ containing the entries of the r-th row of B_s. This allows (1) to be written, for the r-th component of $x(k+1)$, as

$$x_{r,k+1} = (A^{(r)}z_k)^T x_k + (B^{(r)}w_k)^T u_k + v_{r,k}. \tag{3}$$

The corresponding state transition conditional probability, given the parameters and latent variables, can be written as

$$\begin{aligned} p(x_{k+1}&|x_k, u_k, z_k, w_k, A, B) \\ &= \prod_{r=1}^{n} p(x_{r,k+1}|x_k, u_k, z_k, w_k, A, B) \\ &= \prod_{r=1}^{n} \mathcal{N}(x_{r,k+1}|(A^{(r)}z_k)^T x_k + (B^{(r)}w_k)^T u_k, \eta_r), \end{aligned}$$

which, together with $p(x_1)$, assumed to follow $x_1 \sim \mathcal{N}(x_1|\mu_1, P_1)$, fully specifies (2).

Next, we examine the probability of the scheduling parameters Z in three different scenarios:

a) Gaussian i.i.d: In this basic case, we consider the scheduling parameters to be sequentially independent, subject to a parameter controlling their variance. The probability distribution is simply $p(Z) = \prod_{k=1}^{N} p(z_k) = \prod_{k=1}^{N} \mathcal{N}(z_k|0, \tau^{-1}I_{\ell+1})$.

b) Information Profile: We consider that prior information about the scheduling variables is available in the form of a set $\mathcal{Z} = \{\bar{z}_k, \tau_k\}_{k=1}^{N+1}$ representing a per-sample component-wise belief (mean) $\bar{z}_k \in \mathbb{R}^{\ell+1}$ and confidence (precision) $\tau_k \in \mathbb{R}_{++}^{\ell+1}$ over the value of the scheduling parameter. Hence, the distribution of the scheduling parameters is given by $p(Z) = \prod_{k=1}^{N} p(z_k) = \prod_{k=1}^{N} \mathcal{N}(z_k|\bar{z}_k, \mathrm{diag}\,(\tau_k)^{-1})$. This formulation can also be understood as addressing the case with noise or corruption in the measurement of parameters.

c) Random walk: This case introduces a dependence between successive scheduling parameters, whereby the current value of the scheduling parameter conditions the expected value of its subsequent value, i.e., $\mathbb{E}[z_{k+1}] = z_k$, with deviations normally distributed. The joint probability distribution presents a Markovian relationship, and can be written as

$$p(Z) = p(z_1)\prod_{k=1}^{N-1} p(z_{k+1}|z_k) = \prod_{k=1}^{N-1} \mathcal{N}(z_{k+1}|z_k, \mathrm{diag}(\tau)^{-1}).$$

Finally, we note that the scenarios described for the scheduling variables in Z equally apply to the scheduling variables in W, and, therefore the description for the latter is not repeated, for conciseness. Provided with the description of the probability distributions for the LPV model that we address, we can now state our problem.

Problem Statement: Given a set of system measurements $\mathcal{D} = \{X, U\}$, a probabilistic description of the system dynamics $p(X|U,Z,W,A,B)$, and prior distributions for the system and scheduling parameters $p(A)$, $p(B)$ and $p(Z)$, we seek a (possibly approximate) estimate of the posterior distributions for the system and scheduling parameters $\theta = \{A, B, Z, W\}$ when conditioned by $\mathcal{D}$, i.e. $p(\theta|\mathcal{D}) = p(A, B, Z, W|X, U)$.

III. Estimation of Parameters through Bayesian Variational Inference

In this section, we first summarize the general variational inference approach, and then, apply it on the LPV model, arriving at computational expressions for its posterior distribution approximations.

A. Bayesian Variational Inference

Consider a generating model $p(\mathcal{D}, \theta)$ of measurements $\mathcal{D}$, described by a set of parameters (and latent variables) $\theta = \{\theta_j\}_{j=1}^P$. We seek an estimate of the posterior distribution $p(\theta|\mathcal{D})$, obtained by the application of Bayes' rule. Because exact inference of the posterior is intractable (see [3, p.167]), we look for an approximating solution $q^\star(\theta)$ such that

$$q^\star(\theta) = \arg\min_{q(\theta)} \ \mathrm{KL}(q(\theta)||p(\theta|\mathcal{D})), \tag{4}$$

where $\mathrm{KL}(p_1||p_2) := \int p_1(x)\log(p_1(x)/p_2(x))\mathrm{d}x$ denotes the Kullback-Leibler divergence between both densities. If a density $q(\theta) = \prod_{j=1}^P q_j(\theta_j)$ is posited as a factorized approximating distribution, the following known result holds.

Theorem 1 (Mean-field Variational Inference): A solution $q^\star(\theta) = \prod_{j=1}^P q_j^\star(\theta_j)$ to (4) satisfies

$$\log q_j^\star(\theta_j) = \mathbb{E}_{\theta\backslash\theta_j\sim q^\star(\theta\backslash\theta_j)}[\log p(\mathcal{D},\theta)], \tag{5}$$

where $q^\star(\theta\backslash\theta_j) = \prod_{i=1, i\neq j}^P q_i^\star(\theta_i)$.

Proof: Please refer to [19, Section 10.1.1]. ■

The identity in (5) can be further specified by taking in consideration the conditional dependencies between the variables in the model, to yield

$$\log q_j^\star(\theta_j) \propto \ \mathbb{E}_{\theta\backslash\theta_j\sim q(\theta\backslash\theta_j)}\Big[\log p(\theta_j|\mathrm{pa}[\theta_j]) + \sum_{\theta_k\in\mathrm{ch}[\theta_j]} \log p(\theta_k|\mathrm{pa}[\theta_k])\Big], \tag{6}$$

where $\mathrm{pa}[\theta_j]$ denotes the set of 'parents' of θ_j (i.e., the variables that θ_j is conditioned on), and $\mathrm{ch}[\theta_j]$ denotes the set of 'children' of θ_j (i.e., the variables which are conditioned on θ_j). Further, if the probability distributions on the right-hand side of (6) are conditionally conjugate (i.e., have the same functional form), we have that $q^\star(\theta_j)$ will preserve the functional form of the corresponding conditional probabilities appearing in (6). The optimal approximating density $q_j^\star(\theta_j)$ can then be recovered from its log-density in (6) by identifying its moments and reinstating its normalization constant, as we will see in the next section. Finally, we note that the expectations in (6), taken with respect to the approximating distributions $q^\star(\theta\backslash\theta_j)$, depend, in general, on the optimal values of $q_i^\star(\theta_i)$ for $i \neq j$. Therefore, an iterative procedure consisting of sequentially updating $\log q_j(\theta_j), j = 1,\dots,P$, is typically employed, with guaranteed convergence [19].

B. Inference for LPV Parameters and Scheduling variables

We now derive the specific expressions for the approximating distributions for the LPV model considered in this paper, for which $\theta = \{A, B, Z, W\}$ and $\mathcal{D} = \{X, U\}$. Because all prior and conditional distributions in the model are Gaussian-conjugate, the means and covariances of the approximating distributions can be extracted by identifying the quadratic and linear terms in the resulting expressions for their log densities, obtained from (6). For compactness, we denote $\bar{\mathbb{E}}_{\theta_j}[h(\theta)] \equiv \mathbb{E}_{\theta_j\sim q(\theta_j)}[h(\theta)]$ and $\bar{\mathbb{E}}[\theta_j] \equiv \mathbb{E}_{\theta_j\sim q(\theta_j)}[\theta_j]$.

1) Parameter A: We denote by $a_r^{(s)} \in \mathbb{R}^n$ the r-th row of A_s, and let $a^{(r)} = \mathrm{vec}(A^{(r)})$. The joint prior probability of A can then be written as $p(A) = \prod_{s=0}^\ell p(A_s) = \prod_{s=0}^\ell \prod_{r=1}^n \mathcal{N}(a_r^{(s)} \mid \alpha^{-1}I_n)$, having log-density

$$\log p(A) = \frac{n^2(\ell+1)}{2}(\log\alpha - \log(2\pi)) - \frac{\alpha}{2}\sum_{s=0}^{\ell}\sum_{r=1}^{n}(a_r^{(s)})^T a_r^{(s)},$$

such that $\log p(A) \propto -\frac{\alpha}{2}\sum_{r=1}^n (a_r^{(s)})^T a_r^{(s)}$. By applying (6), with $\mathrm{pa}[A] = \emptyset$ and $\mathrm{ch}[A] = \{X\}$, (with X observed, i.e. $\bar{\mathbb{E}}[x_k] = x_k$ and $\bar{\mathbb{E}}[x_k x_k^T] = x_k x_k^T$), we have that

$$\begin{aligned}
&\log q^\star(A) = \log p(A) + \bar{\mathbb{E}}_{Z,W,B}[\log p(X|U,Z,W,A,B)] \\
&\propto -\frac{\alpha}{2}\sum_{s=0}^{\ell}\sum_{r=1}^{n}(a_r^{(s)})^T a_r^{(s)} \\
&-\frac{1}{2}\bar{\mathbb{E}}_{Z,W,B}\left[\sum_{k=1}^{N}\|x_{k+1} - (A^{(r)}z_k)^T x_k - (B^{(r)}w_k)^T u_k\|^2_{\mathrm{diag}(\eta)^{-1}}\right] \\
&\propto -\frac{\alpha}{2}\sum_{s=0}^{\ell}\sum_{r=1}^{n}(a_r^{(s)})^T a_r^{(s)} \\
&-\frac{1}{2}\bar{\mathbb{E}}_{Z,W,B}\Bigg[\sum_{k=1}^{N}\sum_{r=1}^{n}\mathrm{tr}\left(\eta_r(z_k z_k^T)A^{(r)T}(x_k x_k^T)A^{(r)}\right) \\
&\qquad -2\eta_r(x_{r,k+1} - (B^{(r)}w_k)^T u_k)x_k^T A^{(r)} z_k\Bigg] \\
&\propto -\frac{1}{2}\bar{\mathbb{E}}_{Z,W,B}\Bigg[\sum_{r=1}^{n} a^{(r)T}\left(\alpha I_{(\ell+1)n} + \sum_{k=1}^{N}\eta_r x_k x_k^T \otimes z_k z_k^T\right)a^{(r)} \\
&\qquad -2\sum_{r=1}^{n}\eta_r\sum_{k=1}^{N}(x_{r,k+1} - (B^{(r)}w_k)^T u_k)(z_k^T \otimes x_k^T)a^{(r)}\Bigg],
\end{aligned}$$

where we used the cyclic property of the trace, the vectorized variable $a^{(r)} = \mathrm{vec}(A^{(r)})$, and the identity $\mathrm{vec}(LYR) = (R^T \otimes L)\,\mathrm{vec}(Y)$. By identifying quadratic and linear terms, the approximating distribution for A can be expressed as $q^\star(A) = \prod_{r=1}^n \mathcal{N}(a^{(r)}|\mu_a^{(r)}, \Sigma_a^{(r)})$, where $\Sigma_a^{(r)} = \left(\alpha I_{(\ell+1)n} + \sum_{k=1}^N \eta_r x_k x_k^T \otimes \bar{\mathbb{E}}\left[z_k z_k^T\right]\right)^{-1}$ and

$$\begin{aligned}
\mu_a^{(r)} &= \Sigma_a^{(r)}\eta_r\sum_{k=1}^{N}\bar{\mathbb{E}}_{Z,W,B}\left[(x_{r,k+1} - (B^{(r)}w_k)^T u_k)\,\mathrm{vec}(z_k x_k^T)\right] \\
&= \Sigma_a^{(r)}\eta_r\sum_{k=1}^{N}\mathrm{vec}(\bar{\mathbb{E}}[z_k]x_{r,k+1}x_k^T) \\
&\qquad - (\bar{\mathbb{E}}[B^{(r)}]\bar{\mathbb{E}}[w_k])^T u_k\,\mathrm{vec}(\bar{\mathbb{E}}[z_k]x_k^T),
\end{aligned}$$

yielding moments $\bar{\mathbb{E}}[a^{(r)}] = \mu_{a^{(r)}}$ and $\bar{\mathbb{E}}[a^{(r)}a^{(r)^T}] = \mu_{a^{(r)}}\mu_{a^{(r)}}^T + \Sigma_{a^{(r)}}$.

2) Parameter B: Because of their practical relevance, we consider two sub-cases for the set of parameters B: the time-invariant case (B-TI), and the parameter-variant case (B-PV).

a) B-TI: In this case, with some abuse of notation, we redefine B to be a single input-to-state matrix, i.e., $B \in \mathbb{R}^{n\times m}$. We then define $b_r \in \mathbb{R}^m$ to be the r-th row of B, and write the log-density

$$\log p(B) = -\frac{nm}{2}\log(2\pi) + \frac{nm}{2}\log\beta - \frac{\beta}{2}\sum_{r=1}^{n} {b_r}^T b_r$$

By applying (6) with $\text{pa}[B] = \emptyset$ and $\text{ch}[A] = \{X\}$ (with X being observed), we have

$$\begin{aligned}
&\log q^\star(B) = \log p(B) + \bar{\mathbb{E}}_{Z,A}[\log p(X|U,Z,A,B)] \\
&\propto -\frac{\beta}{2}\sum_{r=1}^{n} b_r^T b_r - \frac{1}{2}\bar{\mathbb{E}}_{Z,A}\left[\sum_{k=1}^{N}\|x_{k+1} - \mathcal{A}_k x_k - Bu_k\|^2_{\text{diag}(\eta)^{-1}}\right] \\
&\propto -\frac{1}{2}\sum_{r=1}^{n} {b_r}^T\left(\beta I_m + \eta_r\sum_{k=1}^{N}u_k u_k^T\right)b_r \\
&\quad + \left[\sum_{k=1}^{N}\sum_{r=1}^{n} b_r^T\eta_r u_k\left(x_{r,k+1} - (\bar{\mathbb{E}}[A^{(r)}]\bar{\mathbb{E}}[z_k])^T x_k\right)\right].
\end{aligned}$$

From the quadratic and linear terms above, we have that the approximating distribution $q^\star(B) = \prod_{r=1}^{n}\mathcal{N}(\mu_b^{(r)}|\Sigma_b^{(r)})$, has parameters $\Sigma_{b_r} = \left(\beta I_m + \eta_r\sum_{k=1}^{N}u_k u_k^T\right)^{-1}$ and

$$\mu_{b_r} = \Sigma_{b_r}\sum_{k=1}^{N}\eta_r u_k\left(x_{r,k+1} - (\bar{\mathbb{E}}[A^{(r)}]\bar{\mathbb{E}}[z_k])^T x_k\right), \quad (7)$$

yielding moments $\bar{\mathbb{E}}[b_r] = \mu_{b_r}$ and $\bar{\mathbb{E}}[b_r b_r^T] = \mu_r\mu_r^T + \Sigma_{b_r}$.

b) B-PV: Now, we denote by $b_r^{(s)} \in \mathbb{R}^m$ the r-th row of B_s, and let $b^{(r)} = \text{vec}(B^{(r)})$. The joint prior probability of B can then be written as $p(B) = \prod_{s=0}^{d} p(B_s) = \prod_{s=0}^{d}\prod_{r=1}^{n}\mathcal{N}(b_r^{(s)} \mid \beta^{-1}I_m)$, having log-density

$$\log p(B) = \frac{nm(d+1)}{2}(\log\beta - \log(2\pi)) - \frac{\beta}{2}\sum_{s=0}^{d}\sum_{r=1}^{n}(b_r^{(s)})^T b_r^{(s)},$$

such that $\log p(B) \propto -\frac{\beta}{2}\sum_{s=0}^{d}\sum_{r=1}^{n}(b_r^{(s)})^T b_r^{(s)}$. By applying (6), with $\text{pa}[B] = \emptyset$ and $\text{ch}[B] = \{X\}$, (with X observed), we have that

$$\begin{aligned}
&\log q^\star(B) = \log p(B) + \bar{\mathbb{E}}_{Z,W,A}[\log p(X|U,Z,W,A,B)] \\
&\propto -\frac{\beta}{2}\sum_{s=0}^{d}\sum_{r=1}^{n}(b_r^{(s)})^T b_r^{(s)} \\
&- \frac{1}{2}\bar{\mathbb{E}}_{Z,W,A}\left[\sum_{k=1}^{N}\|x_{k+1} - (A^{(r)}z_k)^T x_k - (B^{(r)}w_k)^T u_k\|^2_{\text{diag}(\eta)^{-1}}\right] \\
&\propto -\frac{1}{2}\bar{\mathbb{E}}_{Z,W,A}\left[\sum_{r=1}^{n} b^{(r)T}\left(\beta I_{(d+1)m} + \sum_{k=1}^{N}\eta_r u_k u_k^T \otimes w_k w_k^T\right)b^{(r)}\right. \\
&\qquad\left. -2\sum_{r=1}^{n}\eta_r\sum_{k=1}^{N}(x_{r,k+1} - (A^{(r)}z_k)^T x_k)(w_k^T \otimes u_k^T)b^{(r)}\right].
\end{aligned}$$

By identifying the quadratic and linear terms, the approximating distribution for B can then be expressed as $q^\star(B) = \prod_{r=1}^{n}\mathcal{N}(b^{(r)}|\mu_b^{(r)}, \Sigma_b^{(r)})$, where

$$\Sigma_b^{(r)} = \left(\beta I_{(d+1)m} + \sum_{k=1}^{N}\eta_r u_k u_k^T \otimes \bar{\mathbb{E}}\left[w_k w_k^T\right]\right)^{-1} \quad \text{and}$$

$$\mu_b^{(r)} = \Sigma_b^{(r)}\eta_r\sum_{k=1}^{N}(x_{r,k+1} - (\bar{\mathbb{E}}[A^{(r)}]\bar{\mathbb{E}}[z_k])^T x_k)\text{vec}(\bar{\mathbb{E}}[w_k]u_k^T),$$

yielding moments $\bar{\mathbb{E}}[b^{(r)}] = \mu_{b^{(r)}}$ and $\bar{\mathbb{E}}[b^{(r)}{b^{(r)}}^T] = \mu_{b^{(r)}}\mu_{b^{(r)}}^T + \Sigma_{b^{(r)}}$.

3) Parameter Z: As discussed, we consider different cases for the latent scheduling variables Z. We begin by noting that the approximating distribution will have the general form

$$\begin{aligned}
&\log q^\star(Z) = \log p(Z) + \bar{\mathbb{E}}_{A,B,W}[\log p(X|U,Z,W,A,B)] \\
&\propto \log p(Z) \\
&- \frac{1}{2}\bar{\mathbb{E}}_{A,B,W}\left[\sum_{k=1}^{N}\sum_{r=1}^{n}\eta_r\left(x_{r,k+1} - (A^{(r)}z_k)^T x_k - (B^{(r)}w_k)^T u_k\right)^2\right],
\end{aligned} \quad (8)$$

differing only in the additive term $\log p(Z)$. We examine each case in detail, next.

a) Gaussian i.i.d: In this case, we have $\log p(Z) \propto -\frac{\tau}{2}\sum_{k=1}^{N} z_k^T z_k$. Applying (6) yields

$$\begin{aligned}
&\log q^\star(Z) = -\frac{\tau}{2}\sum_{k=1}^{N} z_k^T z_k \\
&- \frac{1}{2}\bar{\mathbb{E}}_{A,B,W}\left[\sum_{k=1}^{N}\sum_{r=1}^{n}\eta_r\left(x_{r,k+1} - (A^{(r)}z_k)^T x_k - (B^{(r)}w_k)^T u_k\right)^2\right] \\
&= -\frac{1}{2}\sum_{k=1}^{N} z_k^T\left(\tau I_{\ell+1} + \sum_{r=1}^{n}\eta_r\bar{\mathbb{E}}_A\left[(A^{(r)})^T x_k x_k^T A^{(r)}\right]\right)z_k \\
&\quad + \sum_{k=1}^{N}\sum_{r=1}^{n}\eta_r\left(x_{r,k+1} - (\bar{\mathbb{E}}[B^{(r)}]\bar{\mathbb{E}}[w_z])^T u_k\right)x_k^T\bar{\mathbb{E}}[A^{(r)}]z_k.
\end{aligned}$$

We note that the quadratic term in z_k depends on a transformation of the second moments of $A^{(r)}$, i.e.,

$$\bar{\mathbb{E}}_A\left[(A^{(r)})^T x_k x_k^T A^{(r)}\right] =: \Upsilon_k^{(r)},$$

which can be obtained in terms of its second moments as

$$[\Upsilon_k^{(r)}]_{i,j} = \sum_{s=1}^{n}\sum_{t=1}^{n}[x_k x_k^T]_{t,s}\bar{\mathbb{E}}[a^{(r)}{a^{(r)}}^T]_{\xi(i,t),\xi(j,s)}, \quad (9)$$

with $\xi(i,t) = (i-1)n+t$. Hence, the approximating posterior distribution is given by $q^\star(Z) = \prod_{k=1}^{N}\mathcal{N}(z_k|\mu_z^{(k)}, \Sigma_z^{(k)})$ with

$$\Sigma_z^{(k)} = \left(\tau I_{\ell+1} + \sum_{r=1}^{n}\eta_r\Upsilon_k^{(r)}\right)^{-1},$$

$$\mu_z^{(k)} = \Sigma_z^{(k)}\sum_{r=1}^{n}\eta_r\left(x_{r,k+1} - (\bar{\mathbb{E}}[B^{(r)}]\bar{\mathbb{E}}[w_k])^T u_k\right)x_k^T\bar{\mathbb{E}}[A^{(r)}].$$

b) Information Profile: In this case, we have log-density

$$\log p(Z) \propto -\frac{1}{2}\sum_{k=1}^{N}\|z_k - \bar{z}_k\|^2_{\text{diag}(\tau_k)^{-1}}.$$

Using (9), the approximating posterior distribution is given by $q^\star(Z) = \prod_{k=1}^{N} \mathcal{N}(z_k|\mu_z^{(k)}, \Sigma_z^{(k)})$, with

$$\Sigma_z^{(k)} = \left(\operatorname{diag}(\tau_k) + \sum_{r=1}^{n} \eta_r \Upsilon_k^{(r)}\right)^{-1},$$

$$\mu_z^{(k)} = \Sigma_z^{(k)}\Bigg(\bar{z}_k + \sum_{r=1}^{n} \eta_r \left(x_{r,k+1} - (\bar{\mathbb{E}}[B^{(r)}]\bar{\mathbb{E}}[w_k])^T u_k\right) x_k^T \bar{\mathbb{E}}[A^{(r)}]\Bigg).$$

c) Random walk: In this case, the joint log-density over the variables in Z is given by

$$\log p(Z) \propto -\frac{1}{2} z_1^T \operatorname{diag}(\tau) z_1 - \frac{\tau}{2} \sum_{k=1}^{N-1} \|z_{k+1} - z_k\|_2^2 \quad (10)$$

and presents a dependency between successive variables z_k. By combining the terms depending on z_k in (8) and (10), the joint log-likelihood can be written as an $(\ell+1)N \times (\ell+1)N$ block-quadratic form, yielding a posterior distribution that is also jointly Gaussian. Letting $\Upsilon_k = \sum_{r=1}^{n} \Upsilon_k^{(r)}$, $P = \operatorname{diag}(\tau)$, and $z \in \mathbb{R}^{(\ell+1)N}$ with $z = [z_1^T \cdots z_N^T]^T$, the posterior distribution can be written as

$$p(Z) = \mathcal{N}(z|\mu_z, \Sigma_z) \quad (11)$$

with (inverse) covariance matrix

$$\Sigma_z^{-1} = \begin{bmatrix} Q_1 + \Upsilon_1 + P & -P & \cdots & 0 \\ -P & \Upsilon_2 + 2P & & \vdots \\ \vdots & & \ddots & -P \\ 0 & \cdots & -P & \Upsilon_N + P \end{bmatrix} \quad (12)$$

and mean vector

$$\mu_z = \Sigma_z \lambda_z, \quad (13)$$

where

$$\lambda_z = \begin{bmatrix} \sum_{r=1}^{n} \eta_r (x_{r,2} - (\bar{\mathbb{E}}[B^{(r)}]\bar{\mathbb{E}}[w_k])^T u_1) x_1^T \bar{\mathbb{E}}[A^{(r)}] \\ \sum_{r=1}^{n} \eta_r (x_{r,3} - (\bar{\mathbb{E}}[B^{(r)}]\bar{\mathbb{E}}[w_k])^T u_2) x_2^T \bar{\mathbb{E}}[A^{(r)}] \\ \vdots \\ \sum_{r=1}^{n} \eta_r (x_{r,N+1} - (\bar{\mathbb{E}}[B^{(r)}]\bar{\mathbb{E}}[w_k])^T u_N) x_N^T \bar{\mathbb{E}}[A^{(r)}] \end{bmatrix}.$$

We note that Σ_z is block tri-diagonal, a structure for which there exist efficient inversion algorithms (i.e., scaling linearly with N) [34, Sec. 4.5]. Subsequently, the required moments $\bar{\mathbb{E}}[z_k]$ and $\mathbb{E}_q[z_k z_k^T]$ can be recovered from the corresponding blocks in (13) and (12), respectively.

4) Parameter W: The scheduling parameters W exhibit a structure that mirrors the one presented for the scheduling parameters Z. Because their posterior distributions follow accordingly, the derivations are omitted for the sake of space.

IV. Computational Experiments

To illustrate the Variational Bayesian Inference method for an LPV system, we examine the estimates obtained for an example with scheduling parameters subject to a random walk. The system considered (with $n = 5$ and $m = 2$) is composed of two parameter-varying state transition matrices, A_0 and A_1, and one time-invariant input-to-state matrix B.

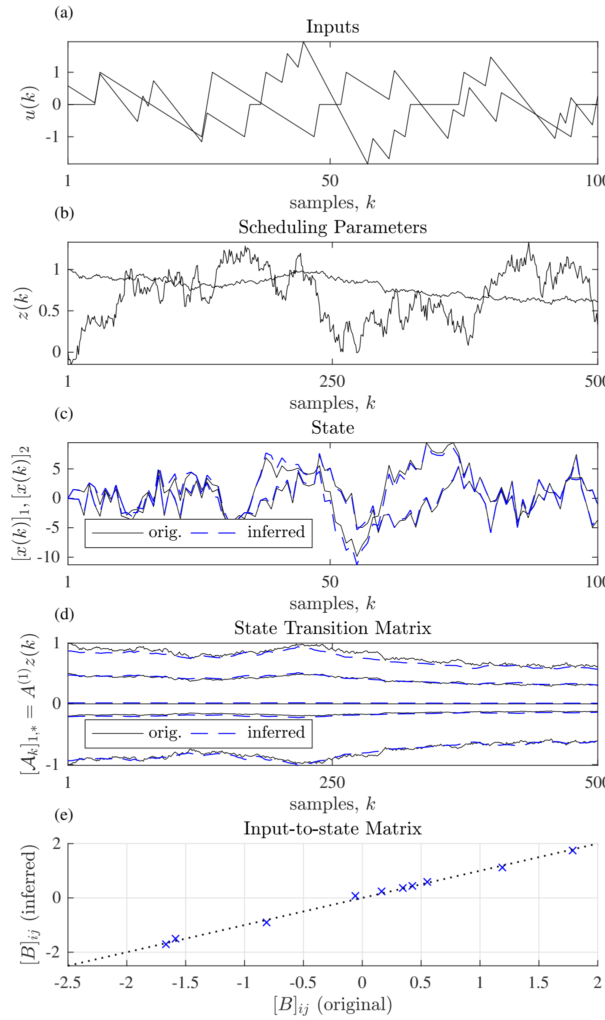

Fig. 1. Variational Bayesian Inference for a simulated LPV system latent parameters $z(k)$ following a random walk. In (a), the inputs $u(k)$ applied on the system are displayed ($m = 2$). In (b), the sample paths of the scheduling parameters are shown, illustrating the effects of low and high variance components. In (c), we show the original and inferred states for the first two components. In (d), we display the entries of the first row ($r = 1, n = 5$) of the the original and inferred parameter-varying state transition matrices, with their behavior induced by the time variation of the scheduling parameters $z(k)$. Finally, in (e), we compare the original and inferred values for the entries of the input-to-state matrix B.

Matrix A_0 is associated with a scheduling parameter $z_0(k)$ presenting a comparatively stable behavior (i.e., low variance ${\tau_0}^{-1} = 1/5000$), while matrix A_1 is associated with a scheduling parameter $z_1(k)$ presenting a higher variance value ($\tau_1^{-1} = 1/150$). The state transition matrices parameters were set as

$$A_0 = \begin{bmatrix} 0.5 & -0.2 & 0 & 1 & -1 \\ -0.2 & 0.3 & 0.7 & 2 & -0.1 \\ 0 & 0.2 & 0.6 & 0 & 0.2 \\ 0.1 & 0 & -0.3 & -0.3 & 0 \\ 0.2 & 0 & 0.4 & -0.33 & -0.33 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & -0.3 & 0 & 0 \\ 0 & 0.4 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0. & 0 \end{bmatrix}.$$

Further, the entries of the input-to-state matrix B were sampled individually from $\mathcal{N}([B]_{ij}|0, \beta^{-1})$, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, with $\beta = 1$. In addition, the state noise precision values were set to $\eta = [5\ \ 10\ \ 2.5\ \ 1\ \ 0.5]^T$. The inputs were defined as a random sequence of piece-wise linear segments, as displayed in Figure 1 (a). The system was simulated to generate a sample path for a random walk of the scheduling parameters, as presented in Figure 1 (b). Together with the inputs and state noise samples, it produced a sequence of states as presented in Figure 1 (c) (only first two components shown, for clarity).

To perform the inference, we defined the initial values of the first moments of the scheduling parameters as $\bar{\mathbb{E}}[z_0(k)]^{(1)} = 1$ and $\bar{\mathbb{E}}[z_1(k)]^{(1)} = 0$, for all k. The initial values for the first moments of the system parameters were defined by letting $A_1^{(1)} = 0_{n\times n}$, and making $\{A_0^{(1)}, B^{(1)}\} = \arg\min_{A_0,B} \sum_{k=1}^{N-1} \|x(k+1) - A_0x(k) - Bu(k)\|_2^2$ (i.e., simple linear regression). The initial values for the second moments of all variables were set as the outer product of their first moments. The results can be seen in Figure 1 (c)-(e). In (c), we present the reconstructed state values using the inferred mean system parameters and the inferred mean scheduling parameters, together with the system inputs (for all k) and the first state measurement ($k = 1$). In (d), we present the estimated first moments of the time varying entries of the first row of the parameter-varying state transition matrix, i.e, $[\mathcal{A}_k]_{1,*} = A^{(1)}z(k)$. Finally, in (e), we compare the original and inferred values for the entries of the input-to-state matrix B. Overall, it can be argued that the inferred mean value of the posterior distribution of the parameters present a good correspondence with their ground truth values, even the presence of variation of the scheduling variables.

V. Conclusion

In this paper, we introduced a Bayesian Variational Inference approach to the estimation of LPV systems with latent scheduling parameters. We derived an algorithmic procedure to generate estimates of the first and second moments of the posterior distribution of system and scheduling parameters, conditioned on measurements of the inputs and state. Further, we specified three different configurations for the behavior of the scheduling parameters, describing the corresponding moment estimation equations.

References

[1] T. Katayama, *Subspace Methods for System Identification*. Springer Science & Business Media, 2006.

[2] S. Gibson and B. Ninness, "Robust Maximum-likelihood Estimation of Multivariable Dynamic Systems," *Automatica*, vol. 41, no. 10, pp. 1667–1682, 2005.

[3] M. J. Beal *et al.*, *Variational Algorithms for Approximate Bayesian Inference*. university of London London, 2003.

[4] M. Niedzwiecki, *Identification of Time-varying Processes*. Wiley New York, 2000.

[5] W. J. Rugh and W. J. Rugh, *Linear System Theory*. Prentice Hall: Upper Saddle River, NJ, 1996, vol. 2.

[6] R. Tóth, *Modeling and Identification of Linear Parameter-varying Systems*. Springer, 2010, vol. 403.

[7] H. Bourlès and B. Marinescu, *Linear Time-Varying Systems: Algebraic-Analytic Approach*. Springer, 2011, vol. 410.

[8] C. O. Becker and V. M. Preciado, "Identification of State-space Linear Time-varying Systems with Sum-of-norms Regularization," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 2833–2838.

[9] A. P. White, G. Zhu, and J. Choi, *Linear Parameter-varying Control for Engineering Applications*. Springer, 2013.

[10] C. Hoffmann and H. Werner, "A Survey of Linear Parameter-varying Control Applications Validated by Experiments or High-fidelity Simulations," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 2, pp. 416–433, 2015.

[11] Y. Zeng and S. Wu, *State-space Models: Applications in Economics and Finance*. Springer, 2013, vol. 1.

[12] L. Paninski, Y. Ahmadian, D. G. Ferreira, S. Koyama, K. R. Rad, M. Vidne, J. Vogelstein, and W. Wu, "A new look at state-space models for neural data," *Journal of computational neuroscience*, vol. 29, no. 1-2, pp. 107–126, 2010.

[13] Z. Chen, *Advanced State Space Methods for Neural and Clinical Data*. Cambridge University Press, 2015.

[14] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. CRC press Boca Raton, FL, 2014, vol. 2.

[15] J. Paisley, D. Blei, and M. Jordan, "Variational bayesian inference with stochastic search," *arXiv preprint arXiv:1206.6430*, 2012.

[16] M. J. Wainwright, M. I. Jordan *et al.*, "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[18] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. John Wiley & Sons, 2007, vol. 382.

[19] C. M. Bishop, *Pattern Recognition and Machine Learning*. springer, 2006.

[20] V. Verdult, "Non Linear System Identification: a State-space Approach," 2002.

[21] R. Tóth, *Modeling and Identification of Linear Parameter-varying Systems, an Orthonormal Basis Function Approach*. Ph. D. dissertation, Delft University of Technology, 2008.

[22] R. P. Aguilera, B. I. Godoy, J. C. Agüero, G. C. Goodwin, and J. I. Yuz, "An EM-based Identification Algorithm for a Class of Hybrid Systems with Application to Power Electronics," *International Journal of Control*, vol. 87, no. 7, pp. 1339–1351, 2014.

[23] W. R. Jacobs, T. Baldacchino, T. J. Dodd, and S. R. Anderson, "Sparse Bayesian Nonlinear System Identification using Variational Inference," *IEEE Transactions on Automatic Control*, 2018.

[24] I. Proimadis, H. Bijl, and J. Van Wingerden, "A Kernel Based Approach for lpv Subspace Identification," *IFAC-PapersOnLine*, vol. 48, no. 26, pp. 97–102, 2015.

[25] C. Hoffmann, A. Isler, and P. Rostalski, "A Factor Graph Approach to Parameter Identification for Affine lpv Systems," in *American Control Conference (ACC), 2017*. IEEE, 2017, pp. 1910–1915.

[26] S. Z. Rizvi, J. Mohammadpour, R. Tóth, and N. Meskin, "A Kernel-based Approach to MIMO LPV State-space Identification and Application to a Nonlinear Process System," *IFAC-PapersOnLine*, vol. 48, no. 26, pp. 85–90, 2015.

[27] A. Wills and B. Ninness, "System Identification of Linear Parameter Varying State-space Models," in *Linear parameter-varying system identification: new developments and trends*. World Scientific, 2012, pp. 295–315.

[28] A. Golabi, N. Meskin, R. Tóth, and J. Mohammadpour, "A Bayesian Approach for Estimation of Linear-regression LPV Models," in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*. IEEE, 2014, pp. 2555–2560.

[29] P. B. Cox and R. Tóth, "LPV State-space Model Identification in the Bayesian Setting: A 3-step procedure," in *American Control Conference (ACC), 2016*. IEEE, 2016, pp. 4604–4610.

[30] A. Golabi, N. Meskin, R. Tóth, and J. Mohammadpour, "A Bayesian approach for LPV model identification and its application to complex processes," *IEEE Transactions on Control Systems Technology*, vol. 25, no. 6, pp. 2160–2167, 2017.

[31] F. Abbasi, J. Mohammadpour, R. Tóth, and N. Meskin, "A Bayesian Approach for Model Identification of LPV Systems with Uncertain Scheduling Variables," in *Decision and Control (CDC), 2015 IEEE 54th Annual Conference on*. IEEE, 2015, pp. 789–794.

[32] D. Piga, P. Cox, R. Toth, and V. Laurain, "LPV System Identification under Noise Corrupted Scheduling and Output Signal Observations," *Automatica*, vol. 53, pp. 329–338, 2015.

[33] J. Luttinen, T. Raiko, and A. Ilin, "Linear State-space Model with Time-varying Dynamics," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 338–353.

[34] G. H. Golub and C. F. Van Loan, *Matrix Computations*. JHU Press, 2012, vol. 3.