Recursive identification of the ARARX model based on the variational Bayes method

DOKOUPIL, J.; VÁCLAVEK, P.

2023 62nd IEEE Conference on Decision and Control (CDC)

eISBN: 979-8-3503-0124-3

DOI: https://doi.org/10.1109/CDC49753.2023.10383518

Accepted manuscript

Recursive Identification of the ARARX Model Based on the Variational Bayes Method

Jakub Dokoupil and Pavel Václavek

Abstract— **Bayesian parameter estimation of autoregressive (AR) with exogenous input (X) systems in the presence of colored model noise is addressed. The stochastic system under consideration is driven by colored noise that arises from passing an initially white noise through an AR filter. Owing to the additional AR filter, the ARARX schema provides more flexibility than the ARX one. The gained flexibility is countered by the fact that the ARARX system is no longer linear-in-parameters unless the white noise components or the AR noise filter are available. This paper analyzes the problem of estimating the unknown coefficients of the ARARX system and the model noise precision under conditions where the AR noise filter is both available and unavailable. While the former condition reduces the estimation problem to standard linear least squares, the latter one gives rise to an analytically intractable estimation problem. The intractability is resolved by the distributional approximation technique based on the variational Bayes (VB) method.**

I. Introduction

Regression-type models are commonly adopted in describing unknown system features based on sequentially observed data. The model parameters rarely have a direct physical interpretation because they usually represent a purposeful approximation to more complex real processes. Consequently, the choice of the model structure is motivated by the intended use of the model rather than strict adherence to the underlying stochastic contexts. However, regression-type models constitute a reasonable compromise between complexity and descriptive capabilities for large collections of real systems [1]. An appealing approach to linear system estimation is to start by estimating a high-order ARX model, whose statistics serve as information-bearing for not only determining low-order ARX models [2] and ARARX models [3] but also approximating ARMAX (with the moving-average (MA) noise filter) structures [4]. The estimation of ARARX systems embodies a difficult problem that is repeatedly addressed in the literature. The difficulty stems from the redundancy elimination of the high-order ARX model to obtain unique ARARX model parameterization, which is the focus of this paper.

In paper [5], the authors constructed a loss function to obtain an asymptotically unbiased estimate of an ARARX model from high-order ARX model statistics, arguing that a reduction is computationally more efficient than the solution given by the direct optimization of the quadratic prediction error (PE) loss function. Such an indirect PE method was further theoretically justified and elaborated into algorithmic details in [6], with the model reduction performed via the Gauss-Newton (GN) method. The paper also explained and advocated the motivation for reducing a high-order ARX model to an ARARX model according to the parsimony principle. Specifically, some accuracy is lost when a complex ARX model is used instead of an ARARX one. A noniterative reduction strategy that results in an inefficient estimator is considered in [3], [7]. The strategy interprets the reduction problem as searching for the greatest common divisor of two discrete polynomials, which can be formulated by means of least squares. Assuming an unknown AR noise filter in the system, an asymptotically unbiased estimate of the ARX part and the corresponding covariance matrix can be obtained with a two-step ARX estimation procedure [8]. The essence of the procedure is to filter and simulate the data by using a high-order noise-free model, which, however, must be built first; this procedure is therefore inconvenient for online learning. To determine the AR noise filter coefficients from the high-order ARX model estimate, the two-step ARX estimation procedure may be further supplemented with a zero-pole cancellation approach [9].

As the MA process is approximable by exploiting an AR filter of a suitably high order, the ARMAX estimation problem can be mapped into the ARX estimation problem. Hence, the reduction concept makes it possible to bypass the stability requirement of the optimal ARMAX predictor, which is a prerequisite for the successful application of the standard PE method. The stability restriction on the MA part is inherently relaxed by the rigorous Bayesian problem formulation. Assuming that the MA part is fixed, the lower-diagonal factorization of the Toeplitz structure correlation matrix acts as a time-varying, prewhitening data filter [10], [11]. Bayesian solutions coherently describe parameter uncertainty, proving vital for both quantifying the risk associated with the data-informed decision-making and probabilistic control design [12], [13]. The Bayesian approach utilizes the probability density function (pdf) to interpret random model parameters, making issues such as the estimator bias, efficiency, and credible interval disappear or irrelevant for any identifiable model. Compared to the PE nonlinear system identification methods, Bayesian learning is generally less prone to overfitting [14]. A certain conceptual advantage of the Bayesian strategies is that they strive for approximations to fit the

This work was supported in part by the Czech Science Foundation under the Project 23-06476S, in part by the infrastructure of RICAIP that has received funding from the European Union's Horizon 2020 research and innovation programme under Grant agreement No, 857306, and in part by Ministry of Education, Youth and Sports under OP RDE Grant agreement No. CZ.02.1.01/0.0/0.0/17 043/0010085.

The authors are with the Faculty of Electrical Engineering and Communication and the Central European Institute of Technology, Brno University of Technology, 612 00 Brno, Czech Republic (e-mail: jakub.dokoupil@ceitec.vutbr.cz; pavel.vaclavek@ceitec.vutbr.cz).

posteriors rather than models to fit the data. The standard PE method relies on the GN search direction, exploiting the first-order Taylor linearization of the system model [15] around the trusted point.

The present paper aims to identify the ARARX model by optimally reducing the high-order ARX system. The reduction is made optimal through approximating the exact posterior pdf by using the product of conditionally independent marginals, in compliance with the VB inference (for a detailed overview, see [16]). We show how the ARARX model is embedded in the ARX structure via adopting the functional form of the dynamic exponential family (DEF) (§6.2.1 in [16]) as a template for the model parameterization, and we also stress that the parametric model is separable in parameters. Consequently, the established model parameterization guarantees a lossless recursive estimation and amenability of the model to the VB method. A similar strategy was recently introduced in [17], [18], where the Hammerstein system is identified by eliminating redundancies in an overparameterized model; this approach too falls within the model reduction issue.

Notation. An $n \times m$ zero matrix is symbolized by $O_{n,m}$; I_n refers to an $n \times n$ identity matrix; ϵ_i^n denotes the ith column of the identity matrix I_n; $\bar{\epsilon}_i^n$ is the ith row of the identity matrix I_n; $\mathbf{1}_n$ is an $n-$dimensional column vector, all of whose components are one; $\otimes$ stands for the Kronecker product; $\circ$ defines the Hadamard product; x' symbolizes the transpose of x; x^* defines the range of x; $\mathring{x}$ refers to the number of members in a countable set x^* or denotes the dimension of a vector x; and $f(x)$ is reserved for the pdf of a random variable x, optionally distinguished by its subscript. Further, $\propto$ means equality up to a normalizing factor; $\mathrm{vec}\,(\cdot)$ represents the vectorization operator; $\equiv$ means equality by definition; the functional derivative of the functional $\mathcal{L}(f(x))$ over $f(x)$ is defined as $\frac{\delta \mathcal{L}(f(x))}{\delta f(x)}$; and the expectation of an arbitrary function $g(x)$ with respect to the pdf $f(x)$ is labeled as $\mathcal{E}_{f(x)}[g(x)] = \int_{x^*} g(x) f(x)\,\mathrm{d}x$.

II. Designing the Algorithms

We assume that the system posits a relationship between a noisy output y_k and the preceding input-output data $\{u_{k-i}, y_{k-i}\}_{i=1}^n$ in the form of the ARARX model (§6.2 in [1]). We then have

$$\begin{cases} y_k = \sum_{i=1}^n b_i u_{k-i} - \sum_{i=1}^n a_i y_{k-i} + v_k, \\ v_k = -\sum_{i=1}^{n_d} d_i v_{k-i} + e_k, \end{cases} \tag{1}$$

where e_k is assumed to be a normally distributed, discrete white noise, $f(e_k|e_{k-1}, \ldots, e_1, d_e) \equiv f(e_k|d_e) \equiv \mathcal{N}(e_k|0, 1/d_e)$, with a zero mean and an unknown precision $d_e \in \mathbb{R}_{>0}$. The output y_k and the noiseless input u_k are both observed on the system at the discrete time instants $k \in k^* \equiv \{k_0, k_0+1, \ldots, \mathring{k}\} \subset \mathbb{Z}$ to form the data record $\mathcal{D}_{1-n-n_d}^k \equiv \{u_i, y_i\}_{i=1-n-n_d}^k$. The lower bound imposed on the data record, $\mathcal{D}_{1-n-n_d}^k$, is chosen to formally secure the indexation of the parametric models in the likelihood function from time $k = 1$, that is, $\prod_{l=1}^k f(y_l|\{a_i, b_i\}_{i=1}^n, \{d_i\}_{i=1}^{n_d}, d_e, \mathcal{D}_{1-n-n_d}^k)$. The ARARX model is parameterized by the set of regression coefficients $\{a_i, b_i\}_{i=1}^n$ and $\{d_i\}_{i=1}^{n_d}$ stacked for clarity into the vectors $\theta \equiv [b_1, \ldots, b_n, a_1, \ldots, a_n]' \in \mathbb{R}^{2n}$ and $\theta_d \equiv [d_1, \ldots, d_{n_d}]' \in \mathbb{R}^{n_d}$ and also by the model noise precision d_e. In the text below, two options with respect to the accessible parameters are examined, and two posteriors which differ in their conditioning information set are constructed. More concretely, we construct $f(\theta, d_e|\theta_d, \mathcal{D}_{1-n-n_d}^k)$ and $f(\theta, \theta_d, d_e|\mathcal{D}_{1-n-n_d}^k)$ to learn the parameters of interest in tandem with the data acquisition.

A. Known coefficients of the AR noise filter

Assume for a moment that an AR process modeling the disturbance is identified by the user. The consistent Bayesian reasoning supported by this knowledge gives rise to an optimal, prewhitening filter which decorrelates the disturbances. The decorrelation then allows for the standard Bayesian estimation of the unknown ARX part.

The ARX part of the model (1) is driven by the colored, normally distributed, discrete noise v_k having a specified mean and a finite correlation span. More explicitly, the disturbance is modeled as an AR process excited by e_k:

$$\begin{aligned} f\big(v_k\big|\{v_{k-i}\}_{i=1}^{n_d}, \theta_d, d_e\big) &= \mathcal{N}\bigg(v_k\bigg| -\sum_{i=1}^{n_d} d_i v_{k-i}, 1/d_e\bigg) \\ &\propto \exp\bigg[-\frac{d_e}{2}\bigg(v_k + \sum_{i=1}^{n_d} d_i v_{k-i}\bigg)^2\bigg]. \end{aligned} \tag{2}$$

To perform the Bayesian recursion, it is necessary to specify the parametric model that incorporates the observed data into the latest posterior. The search for the model requires the disturbance v_k (2) to be transformed to the system output y_k, resulting in

$$\begin{aligned} &f\big(y_k\big|\theta, \theta_d, d_e, \mathcal{D}_{1-n-n_d}^{k-1}\big) \\ &= \mathcal{N}\bigg(y_k\bigg| -\sum_{i=1}^{n_d} d_i y_{k-i} + \underbrace{\bigg[\bar{h}_k' + \sum_{i=1}^{n_d} d_i \bar{h}_{k-i}'\bigg]}_{h_k'}\theta, 1/d_e\bigg) \\ &\propto \exp\big[-\frac{d_e}{2}\big(\tau_k - h_k'\theta\big)^2\big], \end{aligned} \tag{3}$$

where $\tau_k \equiv y_k + \sum_{i=1}^{n_d} d_i y_{k-i}$, and $\bar{h}_k \equiv [u_{k-1}, \ldots, u_{k-n}, -y_{k-1}, \ldots, -y_{k-n}]' \in \mathbb{R}^{2n}$ is an auxiliary regression vector. Hence, the pdf of the normal ARARX model can be viewed as a normal parametric model defined on the filtered data τ_k and h_k. The ARARX introduced to the ARX model transformation in turn opens the door to using the Bayesian estimation effectively. The normality of the parametric model (3) determines the conjugate prior in the form of the normal-Wishart ($\mathcal{NW}$) pdf (§8.1.3 in [13]). This leads to a posterior whose particular factors are defined as

$$\begin{aligned} f\big(\theta\big|\mathcal{S}_k, d_e\big) &= \mathcal{N}\big(\theta\big|\hat{\theta}_k, P_k/d_e\big) \\ &\propto \exp\big[-\big(\theta - \hat{\theta}_k\big)' P_k^{-1}\big(\theta - \hat{\theta}_k\big) d_e/2\big], \end{aligned} \tag{4}$$

$$f\big(d_e\big|\mathcal{S}_k\big) = \mathcal{W}\big(d_e\big|\Sigma_k,\nu_k\big) \propto d_e^{(\nu_k-2)/2}\exp\big[-\Sigma_k d_e/2\big], \tag{5}$$

where $\mathcal{E}_{\mathcal{N}(\theta|\mathcal{S}_k,d_e)}[\theta] = \hat{\theta}_k$ and $\mathcal{E}_{\mathcal{N}(\theta|\mathcal{S}_k,d_e)}[(\theta-\hat{\theta}_k)(\theta-\hat{\theta}_k)'] = P_k/d_e$ represent the particular moments of the multivariate normal distribution (4). The scalars $\Sigma_k > 0$ and $\nu_k > 2$ denote the least squares reminder and the number of degrees of freedom, respectively. It follows from the definition of the Wishart distribution that $\mathcal{E}_{\mathcal{W}(d_e|\Sigma_k,\nu_k)}[d_e] = \nu_k/\Sigma_k$. The set $\mathcal{S}_k \equiv \{s_k,\nu_k\}$ comprises the sufficient statistics for $\{\theta,d_e\}$, with

$$s_k \equiv \operatorname{vec}\left(\begin{bmatrix} P_k^{-1} & -P_k^{-1}\hat{\theta}_k \\ -\hat{\theta}_k' P_k^{-1} & \Sigma_k + \hat{\theta}_k' P_k^{-1}\hat{\theta}_k \end{bmatrix}\right). \tag{6}$$

To initiate the learning procedure, the externally supplied pdf is chosen as

$$f\big(\theta,d_e\big|\hat{\theta}_{k-1},\Xi,\Sigma_0,\nu_0\big) = \mathcal{N}\big(\theta\big|\hat{\theta}_{k-1},\Xi^{-1}/d_e\big) \times \mathcal{W}\big(d_e\big|\Sigma_0,\nu_0\big), \tag{7}$$

where Ξ is a positive definite matrix of an appropriate dimension. Considering (7), the Bayesian update will then smooth the parameter estimate by penalizing the parameter variations from their latest available value rather than from their initial guess [15]. The functional recursion organized with respect to Bayes' rule is as follows:

$$\begin{aligned} f\big(\theta,d_e\big|\mathcal{S}_k\big) \propto \mathcal{N}\left(y_k\middle| -\sum_{i=1}^{n_d} d_i y_{k-i} + h_k'\theta, 1/d_e\right) \\ \times \frac{\mathcal{N}\big(\theta\big|\hat{\theta}_{k-1},\Xi^{-1}/d_e\big)}{\mathcal{N}\big(\theta\big|\hat{\theta}_{k-2},\Xi^{-1}/d_e\big)} \\ \times \mathcal{N}\big(\theta\big|\hat{\theta}_{k-1},P_{k-1}/d_e\big)\mathcal{W}\big(d_e\big|\Sigma_{k-1},\nu_{k-1}\big). \end{aligned} \tag{8}$$

The assignments $\hat{\theta}_0 \equiv \hat{\theta}_{-1}$ and $P_0 \equiv \Xi^{-1}$ made at time $k = 1$ designate pdf (7) to formally initiate the learning procedure. Given the conjugacy of all the pdfs on the right-hand side of (8), the functional recursion is reduced to the least squares-like recursion

$$\varepsilon_{k-1} \equiv \hat{\theta}_{k-1} - \hat{\theta}_{k-2}, \tag{9}$$

$$\hat{\theta}_{c;k-1} \equiv \hat{\theta}_{k-1} + P_{k-1}\Xi\varepsilon_{k-1}, \tag{10}$$

$$\Sigma_{c;k-1} \equiv \Sigma_{k-1} - \varepsilon_{k-1}'\big(I_{\mathring{\theta}} + \Xi P_{k-1}\big)\Xi\varepsilon_{k-1}, \tag{11}$$

$$K_k \equiv P_{k-1}h_k/\big(1 + h_k' P_{k-1}h_k\big), \tag{12}$$

$$\hat{e}_{c;k} \equiv \tau_k - h_k'\hat{\theta}_{c;k-1}, \tag{13}$$

$$\hat{\theta}_k = \hat{\theta}_{c;k-1} + K_k\hat{e}_{c;k}, \tag{14}$$

$$P_k = \big(I_{\mathring{\theta}} - K_k h_k'\big)P_{k-1}\big(I_{\mathring{\theta}} - K_k h_k'\big)' + K_k K_k', \tag{15}$$

$$V_k \equiv P_k^{-1} = V_{k-1} + h_k h_k', \tag{16}$$

$$\Sigma_k = \Sigma_{c;k-1} + \hat{e}_{c;k}^2/\big(1 + h_k' P_{k-1}h_k\big), \tag{17}$$

$$\nu_k = \nu_{k-1} + 1. \tag{18}$$

The recursive solution can be further expanded to operate in a nonstationary environment by means of a data-informed forgetting mechanism, as suggested in [19] or [20], [21]. Having the prespecified precision d_e, one can also consider adopting the adaptive scheme, as designed in [22], [23]. A summary of the estimation procedure for the known noise filter is given by Algorithm 1. Note that the matrix V_k does not have to be propagated to parameterize the posterior (8).

Algorithm 1 The Bayesian parameter estimation procedure for an ARARX model with a known AR noise filter (θ_d).

1: **Initialization phase:**
2: Gather the data set $\mathcal{D}_{1-n-n_d}^1$ to fill the initial filtered regressor vector $h_1 = \bar{h}_1 + \sum_{i=1}^{n_d} d_i\bar{h}_{1-i}$, and $\tau_1 = y_1 + \sum_{i=1}^{n_d} d_i y_{1-i}$, all entering (3).
3: Initialize the statistics $\{\hat{\theta}_0,\Xi,\Sigma_0 > 0,\nu_0 > 2\}$ and execute the assignments $\{\hat{\theta}_{-1} \equiv \hat{\theta}_0, V_0 \equiv P_0^{-1} \equiv \Xi\}$ to obtain, for $k = 1$, the starting point $\{\hat{\theta}_{c;0},\Sigma_{c;0}\}$ (9)–(11) needed to initiate the data update (12)–(18).
4: **Learning phase:**
5: **for** $k \leftarrow 1,\mathring{k}$ **do**
6: **Input:** $\begin{cases} \tau_k, h_k, \Xi, \Sigma_{k-1}, \nu_{k-1}, \\ \hat{\theta}_{k-1}, \hat{\theta}_{k-2}, V_{k-1}, P_{k-1} \end{cases}$
7: Update: $\hat{\theta}_{k-1} \to \hat{\theta}_k$, $P_{k-1} \to P_k$, $V_{k-1} \to V_k$, $\Sigma_{k-1} \to \Sigma_k$, and $\nu_{k-1} \to \nu_k$ ▷ (9)–(18)
8: **Output:** $\hat{\theta}_k$, P_k, V_k, Σ_k, ν_k
9: **end for**

B. Unknown coefficients of the AR *noise filter*

Regrettably, in practice we rarely meet the assumption that the AR process modeling the disturbance is known. Conceptually, Bayesian model comparison can be involved to mitigate the impact of the absence of an explicit noise model [10]. This requires us to evaluate the posterior probabilities on the hypotheses that a specific noise model is the best representative from a finite set of candidates [24]. Although a more refined approach has been designed in this respect [11], it still employs a finite, prespecified mixture of stochastic models with a common ARX part. In the sequel, we expand the range of the ARARX model's practical applications by estimating all its parameters.

The problem is faced by the indirect estimation approach, as no closed-form expression is available to directly propagate the moments of the regression coefficients $\{\theta,\theta_d\}$ coupled within the ARARX model (1). To facilitate the subsequent designing of the inference algorithm, we will show that the parametric model (3) is a member of the dynamic exponential family with separable parameters (DEFS) (§6.3.1 in [16]). To this end, the term in the exponent of (3) is rewritten using the identity $\operatorname{vec}(\mathsf{ACB}) = (\mathsf{B}'\otimes\mathsf{A})\operatorname{vec}(\mathsf{C})$ [25], as indicated below:

$$\begin{aligned} \tau_k - h_k'\theta &= \operatorname{vec}\left(\begin{bmatrix} \bar{h}_k' & -y_k \\ \vdots & \vdots \\ \bar{h}_{k-n_d}' & -y_{k-n_d} \end{bmatrix}\right)' \left(\begin{bmatrix}\theta\\1\end{bmatrix} \otimes \begin{bmatrix}1\\\theta_d\end{bmatrix}\right) \\ &= \begin{bmatrix}\varphi_k\\-y_k\end{bmatrix}' \underbrace{T'\left(\begin{bmatrix}\theta\\1\end{bmatrix} \otimes \left(J\begin{bmatrix}\theta_d\\1\end{bmatrix}\right)\right)}_{\begin{bmatrix}\vartheta' & 1\end{bmatrix}'}, \end{aligned} \tag{19}$$

where $\varphi_k \equiv [u_{k-1}, \ldots, u_{k-n-n_d}, -y_{k-1}, \ldots, -y_{k-n-n_d}]' \in \mathbb{R}^{2(n+n_d)}$ is a high-order ARX model regressor, $J \equiv \begin{bmatrix} O_{1,n_d} & 1 \\ I_{n_d} & O'_{1,n_d} \end{bmatrix}$, and the transformation matrix T possesses the form

$$T \equiv \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & 0 \end{bmatrix}, \tag{20}$$

$$T_{12} = \epsilon^{(n_d+1)(2n+1)-1}_{(n_d+1)2n+1}, \quad T_{21} = \bar{\epsilon}^{2(n+n_d)}_{2n_d+n}, \tag{21}$$

$$T_{11} = \begin{bmatrix} \underbrace{\begin{bmatrix} I_{n_d+1} & O_{n_d+1,2n+n_d-1} \end{bmatrix}}_{E} S^0 \\ \vdots \\ ES^{n-1} \\ ES^{n_d+n} \\ \vdots \\ ES^{n_d+2n-1} \\ O_{1,2(n+n_d)} \\ \begin{bmatrix} I_{n_d-1} & O_{n_d-1,2n+n_d+1} \end{bmatrix} S^{n_d+n} \end{bmatrix}. \tag{22}$$

The matrix $S \equiv \begin{bmatrix} O_{2(n+n_d)-1,1} & I_{2(n+n_d)-1} \\ 0 & O'_{2(n+n_d)-1,1} \end{bmatrix}$ embodies a shifting matrix (§3.7 in [26]), provided we take $S^0 = I_{2(n+n_d)}$ (S^n is the nth power of S). When a given matrix is postmultiplied by S, the columns of that matrix are shifted to the right, invariably by one position, and the first column of the matrix is replaced with a null vector. The form of the corresponding coefficient vector ϑ, which determines the ARARX model dynamics, follows from

$$\vartheta \equiv \begin{bmatrix} T_{11} \\ \bar{\epsilon}^{2(n+n_d)}_{2n_d+n} \end{bmatrix}' \left(\begin{bmatrix} \theta \\ 1 \end{bmatrix} \otimes \left(J \begin{bmatrix} \theta_d \\ 1 \end{bmatrix} \right) \right) \in \mathbb{R}^{2(n+n_d)}. \tag{23}$$

Although the posterior is intractable, the chosen parameterization allows its sufficient statistics to be recursively updated without any information loss. In this scenario, the pdf of the normal ARARX model equals the normal pdf describing the high-order ARX model

$$f\big(y_k\big|\theta, \theta_d, d_e, \mathfrak{D}^{k-1}_{1-n-n_d}\big) = \mathcal{N}\big(y_k\big|\varphi'_k\vartheta, 1/d_e\big). \tag{24}$$

Now, we can proceed to the formal model classification.

Remark 1: The normal ARARX model belongs to the DEFS

$$f\big(y_k\big|\theta, \theta_d, d_e, \mathfrak{D}^{k-1}_{1-n-n_d}\big) \equiv \exp\big[q\big(\vartheta, d_e\big)'\gamma\big(y_k, \varphi_k\big) - \iota_{y_k}\big(\vartheta, d_e\big)\big], \tag{25}$$

under the assignments

$$q\big(\vartheta, d_e\big) = -\frac{d_e}{2}\operatorname{vec}\big([\,\vartheta' \quad 1\,]'[\,\vartheta' \quad 1\,]\big), \tag{26}$$

$$\gamma\big(y_k, \varphi_k\big) = \operatorname{vec}\big([\,\varphi'_k \quad -y_k\,]'[\,\varphi'_k \quad -y_k\,]\big), \tag{27}$$

$$\begin{aligned}\big[\iota_{y_k}\big(\vartheta, d_e\big)\big] &= \ln\left(\int_{-\infty}^{\infty}\exp\big[q\big(\vartheta, d_e\big)'\gamma\big(y_k, \varphi_k\big)\big]\,\mathrm{d}y_k\right) \\ &= \ln\left(\sqrt{2\pi/d_e}\right),\end{aligned} \tag{28}$$

where $\exp[\iota_{y_k}(\vartheta, d_e)] = \sqrt{2\pi/d_e}$ stands for the normalizing factor of the parametric model (25). Making use of the identities $\operatorname{vec}(\mathsf{ACB}) = (\mathsf{B}'\otimes\mathsf{A})\operatorname{vec}(\mathsf{C})$ and $(\mathsf{A}\circ\mathsf{C})\otimes(\mathsf{B}\circ\mathsf{D}) = (\mathsf{A}\otimes\mathsf{B})\circ(\mathsf{C}\otimes\mathsf{D})$ [25] on $\begin{bmatrix}\theta' & 1\end{bmatrix}' \otimes (J\begin{bmatrix}\theta'_d & 1\end{bmatrix}')$ entering (19) yields

$$\begin{aligned}&\begin{bmatrix}\theta' & 1\end{bmatrix}' \otimes \left(J\begin{bmatrix}\theta'_d & 1\end{bmatrix}'\right) \\ &= \underbrace{\left(\big(I_{2n+1}\otimes \mathbf{1}_{n_d+1}\big)\begin{bmatrix}\theta \\ 1\end{bmatrix}\right)}_{\phi} \circ \underbrace{\left(\big(\mathbf{1}_{2n+1}\otimes I_{n_d+1}\big)J\begin{bmatrix}\theta_d \\ 1\end{bmatrix}\right)}_{\phi_d}.\end{aligned} \tag{29}$$

Bearing in mind the previous result (29) and the identity $\operatorname{vec}(xx') = x\otimes x$ [25], we can easily prove from (26) that the model (25) is separable in parameters, satisfying

$$q\big(\vartheta, d_e\big) = \big(T'\otimes T'\big)\big(q_{arx}\big(\theta\big)\circ q_{ar}\big(\theta_d\big)\circ q_e\big(d_e\big)\big), \tag{30}$$

with

$$\begin{cases} q_{arx}\big(\theta\big) = \operatorname{vec}\big(\phi\phi'\big),\ q_{ar}\big(\theta_d\big) = \operatorname{vec}\big(\phi_d\phi'_d\big), \\ q_e\big(d_e\big) = -d_e/2. \end{cases} \tag{31}$$

Thus, we conclude Remark 1 on the ARARX model classification.

Bayesian learning with the high-order ARX model (24) can be approached in the same way as it has been derived for the ARX model defined on the filtered data (3). By introducing the externally supplied pdf

$$\begin{aligned}f\big(\vartheta, d_e\big|\hat{\vartheta}_{k-1}, \bar{\Xi}, \bar{\Sigma}_0, \nu_0\big) = \mathcal{N}\big(\vartheta\big|\hat{\vartheta}_{k-1}, \bar{\Xi}^{-1}/d_e\big) \\ \times \mathcal{W}\big(d_e\big|\bar{\Sigma}_0, \nu_0\big),\end{aligned} \tag{32}$$

which effectively regularizes the data update, the already presented least squares routine (9)-(18) is adopted to perform the algebraic recursion. The only formal difference from the implementation (9)-(18) lies in the use of the substitutions $\{\hat{\vartheta}_k \equiv \hat{\theta}_k, \varphi_k \equiv h_k, y_k \equiv \tau_k, \bar{P}_k \equiv P_k, \bar{V}_k \equiv V_k, \bar{\Sigma}_k \equiv \Sigma_k, \bar{\Xi} \equiv \Xi\}$. An immediate consequence of this treatment is the propagation of the posterior $f(\vartheta, d_e|\bar{\mathcal{S}}_k)$ through its sufficient statistics $\bar{\mathcal{S}}_k \equiv \{\bar{s}_k, \nu_k\}$, where

$$\bar{s}_k \equiv \operatorname{vec}\left(\begin{bmatrix} \bar{V}_k & -\bar{V}_k\hat{\vartheta}_k \\ -\hat{\vartheta}'_k\bar{V}_k & \bar{\Sigma}_k + \hat{\vartheta}'_k\bar{V}_k\hat{\vartheta}_k \end{bmatrix}\right). \tag{33}$$

The posterior $f(\vartheta, d_e|\bar{\mathcal{S}}_k)$ for the nonparsimonious ARARX model representation is constructed at each step to agree with the sequential data retrieval. The second step within the recursive learning is to recover $f(\theta, \theta_d, d_e|\bar{\mathcal{S}}_k)$, namely, the parsimonious ARARX model nested inside the ARX structure. The recovery of the posterior $f(\theta, \theta_d, d_e|\bar{\mathcal{S}}_k)$ is performed using the iterative VB (IVB) algorithm with N consecutive iterations per least squares update. In general, the VB method restores the tractability of the inference problem through approximating the explicit pdf by the product of the marginals, which are forced to be independent of each other. Considering the separation of the natural parameters available for the pdf of the ARARX model (31), the tractable posterior $\breve{f}(\theta, \theta_d, d_e|\bar{\mathcal{S}}_k)$ is restricted to the product

$$\breve{f}\big(\theta, \theta_d, d_e\big|\bar{\mathcal{S}}_k\big) \equiv \breve{f}\big(\theta\big|\bar{\mathcal{S}}_k\big)\breve{f}\big(\theta_d\big|\bar{\mathcal{S}}_k\big)f\big(d_e\big|\bar{\mathcal{S}}_k\big), \tag{34}$$

where the factor $f(d_e|\bar{\mathcal{S}}_k)$ is recognized to be the exact marginal $f(d_e|\bar{\mathcal{S}}_k) = \mathcal{W}(d_e|\bar{\Sigma}_k, \nu_k)$. The proposed factorization (34) stimulated by (31) is in conformity with

the coefficient partitioning, which ensures the conditional conjugacy (§2.1.2 in [27]). This conjugacy stems from the fact that the ARARX model is built by combining the ARX and the AR parts. To optimally infer the two remaining marginals, ($\breve{f}(\theta|\bar{\mathcal{S}}_k)$ and $\breve{f}(\theta_d|\bar{\mathcal{S}}_k)$), a loss functional is constructed to quantify the information loss incurred when the tractable posterior $\breve{f}(\theta, \theta_d, d_e|\bar{\mathcal{S}}_k)$ is used to approximate the explicit posterior $f(\theta, \theta_d, d_e|\bar{\mathcal{S}}_k)$. Let us define the ordered set of parameters $\Theta \equiv \{\theta, \theta_d, d_e\}$ for the sake of brevity and introduce the functional

$$\begin{aligned}\mathcal{L}\big(\breve{f}(\Theta|\bar{\mathcal{S}}_k)\big) &\equiv \mathcal{D}\big(\breve{f}(\Theta|\bar{\mathcal{S}}_k)\big\|f(\Theta|\bar{\mathcal{S}}_k)\big)\\ &+ \eta_\theta\left(\int_{\theta^*}\breve{f}\big(\theta\big|\bar{\mathcal{S}}_k\big)\mathrm{d}\theta - 1\right) + \eta_{\theta_d}\left(\int_{\theta_d^*}\breve{f}\big(\theta_d\big|\bar{\mathcal{S}}_k\big)\mathrm{d}\theta_d - 1\right);\end{aligned} \tag{35}$$

here, $\mathcal{D}(\breve{f}(\Theta|\bar{\mathcal{S}}_k)\|f(\Theta|\bar{\mathcal{S}}_k))$ stands for the Kullback-Leibler divergence (KLD) [28],

$$\mathcal{D}\big(\breve{f}(\Theta|\bar{\mathcal{S}}_k)\big\|f(\Theta|\bar{\mathcal{S}}_k)\big) \equiv \int_{\Theta^*}\breve{f}(\Theta|\bar{\mathcal{S}}_k)\ln\left(\frac{\breve{f}(\Theta|\bar{\mathcal{S}}_k)}{f(\Theta|\bar{\mathcal{S}}_k)}\right)\mathrm{d}\Theta, \tag{36}$$

which is nonnegative, $\mathcal{D}(\breve{f}(\Theta|\bar{\mathcal{S}}_k)\|f(\Theta|\bar{\mathcal{S}}_k)) \geq 0$, with an equality if and only if the two pdfs coincide with each other. The functional (35) is optimized within the calculus of variations approach to yield the two minimizers $\hat{f}(\theta|\bar{\mathcal{S}}_k)$ and $\hat{f}(\theta_d|\bar{\mathcal{S}}_k)$. The terms in (35) scaled by the Lagrange multipliers η_θ and η_{θ_d} rigorously guarantee that each minimizer is a normalized pdf integrating to one. To accomplish the search for the minimizers, the necessary optimality conditions are captured by the lemma below.

Lemma 1: Let $\breve{f}(\theta, \theta_d, d_e|\bar{\mathcal{S}}_k)$ be an approximate posterior restricted to the factorization constraint $\breve{f}(\theta, \theta_d, d_e|\bar{\mathcal{S}}_k) = \breve{f}(\theta|\bar{\mathcal{S}}_k)\breve{f}(\theta_d|\bar{\mathcal{S}}_k)f(d_e|\bar{\mathcal{S}}_k)$, with the marginal $f(d_e|\bar{\mathcal{S}}_k) = \mathcal{W}(d_e|\bar{\Sigma}_k, \nu_k)$ having the fixed functional form. Then, minimizing the functional (35) over the independent marginals $\breve{f}(\theta|\bar{\mathcal{S}}_k)$ and $\breve{f}(\theta_d|\bar{\mathcal{S}}_k)$ yields

$$\hat{f}\big(\theta\big|\bar{\mathcal{S}}_k\big) \propto \exp\Big[\mathcal{E}_{\breve{f}(\theta_d|\bar{\mathcal{S}}_k)\mathcal{W}(d_e|\bar{\Sigma}_k,\nu_k)}\big[\ln\big(f\big(\Theta, \bar{\mathcal{S}}_k\big)\big)\big]\Big], \tag{37}$$

$$\hat{f}\big(\theta_d\big|\bar{\mathcal{S}}_k\big) \propto \exp\Big[\mathcal{E}_{\breve{f}(\theta|\bar{\mathcal{S}}_k)\mathcal{W}(d_e|\bar{\Sigma}_k,\nu_k)}\big[\ln\big(f\big(\Theta, \bar{\mathcal{S}}_k\big)\big)\big]\Big]. \tag{38}$$

Proof: Since the optimization problem can be solved analogously for each factor separately, we will discuss the proof with respect to $\hat{f}(\theta|\bar{\mathcal{S}}_k)$ only. It has proved to be convenient to rewrite the KLD entering (35) as a sum of two parts:

$$\begin{aligned}\mathcal{D}\big(\breve{f}(\Theta|\bar{\mathcal{S}}_k)\big\|f(\Theta|\bar{\mathcal{S}}_k)\big) = &\min_{\breve{f}(\theta|\bar{\mathcal{S}}_k)}\mathcal{L}\big(\breve{f}(\Theta|\bar{\mathcal{S}}_k)\big)\\ &+ \mathcal{D}\big(\breve{f}\big(\theta\big|\bar{\mathcal{S}}_k\big)\big\|\hat{f}\big(\theta\big|\bar{\mathcal{S}}_k\big)\big),\end{aligned} \tag{39}$$

where the part independent of the optimized $\breve{f}(\theta|\bar{\mathcal{S}}_k)$,

$$\ln\left(\int_{\Theta^*}f(\Theta, \bar{\mathcal{S}})\mathrm{d}\Theta\right)$$

$$+ \int_{\theta_d^*}\int_{d_e^*}\breve{f}\big(\theta_d, d_e\big|\bar{\mathcal{S}}\big)\ln\big(\overbrace{\breve{f}\big(\theta_d, d_e\big|\bar{\mathcal{S}}\big)}^{\breve{f}(\theta_d|\bar{\mathcal{S}}_k)\mathcal{W}(d_e|\bar{\Sigma}_k,\nu_k)}\big)\mathrm{d}\theta_d\,\mathrm{d}d_e$$

$$- \ln\left(\int_{\theta_d^*}\int_{d_e^*}\exp\Big[\mathcal{E}_{\breve{f}(\theta_d,d_e|\bar{\mathcal{S}})}\big[\ln\big(f\big(\Theta, \bar{\mathcal{S}}_k\big)\big)\big]\Big]\mathrm{d}\theta_d\,\mathrm{d}d_e\right),$$

absorbs the attained minimum value. Then, direct application of the optimality conditions to (35), the conditions being stipulated by $\frac{\delta\mathcal{L}}{\delta\breve{f}(\theta|\bar{\mathcal{S}}_k)} = \big[\ln(\breve{f}(\theta|\bar{\mathcal{S}}_k)/\hat{f}(\theta|\bar{\mathcal{S}}_k)) + \eta_\theta + 1\big] = 0$ and $\frac{\partial\mathcal{L}}{\partial\eta_\theta} = 0$, identifies the form of the minimizer (37). ■

The conditional conjugacy is the consequence of the normality of each posterior factor $f(\theta|\theta_d, d_e, \mathcal{D}_{1-n-n_d}^k)$ and $f(\theta_d|\theta, d_e, \mathcal{D}_{1-n-n_d}^k)$, and we thus obtain hints to choose the initializer form that starts the iterative optimization at iteration 0 as $\mathcal{N}^{[0]}(\theta_d)$. Consequently, the stationary conditions (37) and (38) constitute tractable VB-learning as the moments of each VB-marginal are available. With the vectors ϕ and ϕ_d introduced in (29), the iterative updating of the VB-marginals, for $i = 1, \ldots, N$, shows as

$$\begin{aligned}\mathcal{N}^{[i]}\big(\theta\big) \propto \exp\Bigg[\mathcal{E}_{\mathcal{N}^{[i-1]}(\theta_d)\mathcal{W}(d_e)}\Bigg[&-\frac{d_e}{2}\operatorname{vec}\big(\phi\phi'\big)'\\ &\times\operatorname{vec}\big(\Phi\circ\big(\phi_d\phi_d'\big)\big)\Bigg]\Bigg],\end{aligned} \tag{40}$$

$$\begin{aligned}\mathcal{N}^{[i]}\big(\theta_d\big) \propto \exp\Bigg[\mathcal{E}_{\mathcal{N}^{[i]}(\theta)\mathcal{W}(d_e)}\Bigg[&-\frac{d_e}{2}\operatorname{vec}\big(\phi_d\phi_d'\big)'\\ &\times\operatorname{vec}\big(\Phi\circ\big(\phi\phi'\big)\big)\Bigg]\Bigg],\end{aligned} \tag{41}$$

where Φ refers to the sufficient statistics (33) of the high-order ARX model via

$$\Phi \equiv T\begin{bmatrix}\bar{V}_k & -\bar{V}_k\hat{\vartheta}_k\\ -\hat{\vartheta}_k'\bar{V}_k & \bar{\Sigma}_k + \hat{\vartheta}_k'\bar{V}_k\hat{\vartheta}_k\end{bmatrix}T'. \tag{42}$$

To acquire the IVB solution, the induced expectations in (40) and (41) are evaluated, and the final forms of the VB-marginals are found upon the completion of squares technique. For these purposes, the matrix Φ is partitioned into blocks

$$\Phi \equiv \begin{bmatrix}\Phi_{11} & \Phi_{21}'\\ \Phi_{21} & \Phi_{22}\end{bmatrix}, \tag{43}$$

where

$$\begin{aligned}\Phi_{11} = &T_{11}\bar{V}_kT_{11}' - T_{12}\hat{\vartheta}_k'\bar{V}_kT_{11}' - T_{11}\bar{V}_k\hat{\vartheta}_kT_{12}'\\ &+ T_{12}\big(\bar{\Sigma}_k + \hat{\vartheta}_k'\bar{V}_k\hat{\vartheta}_k\big)T_{12}',\end{aligned} \tag{44}$$

$$\Phi_{21} = T_{21}\bar{V}_kT_{11}' - T_{21}\bar{V}_k\hat{\vartheta}_kT_{12}', \tag{45}$$

$$\Phi_{22} = T_{21}\bar{V}_kT_{21}'. \tag{46}$$

Further, introduce the substitutions $\Upsilon \equiv I_{2n+1}\otimes\mathbf{1}_{n_d+1}$ and $\Upsilon_d \equiv (\mathbf{1}_{2n+1}\otimes I_{n_d+1})J$. Let the matrix Υ be partitioned according to

$$\Upsilon \equiv \begin{bmatrix}\Upsilon_{11} & \Upsilon_{12}\\ O_{1,2n} & 1\end{bmatrix}, \tag{47}$$

with

$$\Upsilon_{11} = \begin{bmatrix}I_{2n}\otimes\mathbf{1}_{n_d+1}\\ O_{n_d,2n}\end{bmatrix},\ \Upsilon_{12} = \begin{bmatrix}O_{2n(n_d+1),1}\\ \mathbf{1}_{n_d}\end{bmatrix}, \tag{48}$$

and the matrix Υ_d as shown below

$$\Upsilon_d \equiv \begin{bmatrix} \Upsilon_{d11} & \Upsilon_{d12} \\ \bar{\epsilon}_{n_d}^{n_d} & 0 \end{bmatrix}, \tag{49}$$

where

$$\Upsilon_{d11} = \begin{bmatrix} \mathbf{1}_{2n} \otimes \begin{bmatrix} I_{n_d} \\ O_{1,n_d} \end{bmatrix} \\ I_{n_d} \end{bmatrix} \begin{bmatrix} O_{1,n_d} \\ \begin{bmatrix} I_{n_d-1} & O_{n_d-1,1} \end{bmatrix} \end{bmatrix} + \begin{bmatrix} \mathbf{1}_{2n} \otimes \epsilon_{n_d+1}^{n_d+1} \\ O_{n_d,1} \end{bmatrix} \bar{\epsilon}_{n_d}^{n_d}, \tag{50}$$

$$\Upsilon_{d12} = \begin{bmatrix} \mathbf{1}_{2n} \otimes \begin{bmatrix} I_{n_d} \\ O_{1,n_d} \end{bmatrix} \\ I_{n_d} \end{bmatrix} \epsilon_1^{n_d}. \tag{51}$$

In light of the substitutions covered above, (43), (47), and (49), the schema in (40) and (41) is reduced into iterative updating of the statistics of $\mathcal{N}(\theta|\hat{\theta}_k, P_{\theta;k}\bar{\Sigma}_k/\nu_k)$ and $\mathcal{N}(\theta_d|\hat{\theta}_{d;k}, P_{\theta_d;k}\bar{\Sigma}_k/\nu_k)$, for $i = 1, \ldots, N$, in compliance with

$$\mathcal{X}_{12}^i \equiv \Phi_{21}' \circ \big[\Upsilon_{d11}\big(\hat{\theta}_{d;k}^{i-1}\big(\hat{\theta}_{d;k}^{i-1}\big)' + P_{\theta_d;k}^{i-1}\frac{\bar{\Sigma}_k}{\nu_k}\big)\epsilon_{n_d}^{n_d} + \Upsilon_{d12}\big(\hat{\theta}_{d;k}^{i-1}\big)'\epsilon_{n_d}^{n_d}\big], \tag{52}$$

$$\mathcal{X}_{11}^i \equiv \Phi_{11} \circ \big[\Upsilon_{d11}\big(\hat{\theta}_{d;k}^{i-1}\big(\hat{\theta}_{d;k}^{i-1}\big)' + P_{\theta_d;k}^{i-1}\frac{\bar{\Sigma}_k}{\nu_k}\big)\Upsilon_{d11}' + \Upsilon_{d12}\big(\hat{\theta}_{d;k}^{i-1}\big)'\Upsilon_{d11}' + \Upsilon_{d11}\hat{\theta}_{d;k}^{i-1}\Upsilon_{d12}' + \Upsilon_{d12}\Upsilon_{d12}'\big], \tag{53}$$

$$P_{\theta;k}^i = \big(\Upsilon_{11}'\mathcal{X}_{11}^i\Upsilon_{11}\big)^{-1}, \tag{54}$$

$$\hat{\theta}_k^i = P_{\theta;k}^i\big(-\Upsilon_{11}'\mathcal{X}_{11}^i\Upsilon_{12} - \Upsilon_{11}'\mathcal{X}_{12}^i\big), \tag{55}$$

$$\mathcal{Y}_{21}^i \equiv \Phi_{21} \circ \big[\big(\hat{\theta}_k^i\big)'\Upsilon_{11}' + \Upsilon_{12}'\big], \tag{56}$$

$$\mathcal{Y}_{11}^i \equiv \Phi_{11} \circ \big[\Upsilon_{11}\big(\hat{\theta}_k^i\big(\hat{\theta}_k^i\big)' + P_{\theta;k}^i\frac{\bar{\Sigma}_k}{\nu_k}\big)\Upsilon_{11}' + \Upsilon_{12}\big(\hat{\theta}_k^i\big)'\Upsilon_{11}' + \Upsilon_{11}\hat{\theta}_k^i\Upsilon_{12}' + \Upsilon_{12}\Upsilon_{12}'\big], \tag{57}$$

$$P_{\theta_d;k}^i = \big(\Upsilon_{d11}'\mathcal{Y}_{11}^i\Upsilon_{d11} + \epsilon_{n_d}^{n_d}\mathcal{Y}_{21}^i\Upsilon_{d11} + \Upsilon_{d11}'\big(\mathcal{Y}_{21}^i\big)'\bar{\epsilon}_{n_d}^{n_d} + \epsilon_{n_d}^{n_d}\Phi_{22}\bar{\epsilon}_{n_d}^{n_d}\big)^{-1}, \tag{58}$$

$$\hat{\theta}_{d;k}^i = P_{\theta_d;k}^i\big(-\Upsilon_{d11}'\mathcal{Y}_{11}^i\Upsilon_{d12} - \epsilon_{n_d}^{n_d}\mathcal{Y}_{21}^i\Upsilon_{d12}\big), \tag{59}$$

where $\hat{\theta}_{d;k}^0 \equiv \hat{\theta}_{d;k-1}$ and $P_{\theta_d;k}^0 \equiv P_{\theta_d;k-1}$. The recursive cycles of the ordered set of equations above execute the ARX model reduction to yield an optimal approximation of the ARARX model $\{\hat{\theta}_k \equiv \hat{\theta}_k^N, P_{\theta;k} \equiv P_{\theta;k}^N, \hat{\theta}_{d;k} \equiv \hat{\theta}_{d;k}^N, P_{\theta_d;k} \equiv P_{\theta_d;k}^N\}$. The computation procedures to implement the developed method are reported in Algorithm 2.

Algorithm 2 The IVB inference-based parameter estimation procedure for an ARARX model.

Initialization phase:
Gather the data set $\mathcal{D}_{1-n-n_d}^1$ to fill the initial regressor φ_1 entering (24).
Recall that, for the purpose of a high-order ARX model estimation, we implement the least squares method (9)–(18), the sole difference being that we formally relabel the variables. Initialize the statistics $\{\hat{\vartheta}_0, \bar{\Xi}, \bar{\Sigma}_0 > 0, \nu_0 > 2\}$. Use the assignments $\{\hat{\vartheta}_{-1} \equiv \hat{\vartheta}_0, \bar{V}_0 \equiv \bar{P}_0^{-1} \equiv \bar{\Xi}\}$ to obtain, for $k = 1$, the starting point $\{\hat{\vartheta}_{c;0}, \bar{\Sigma}_{c;0}\}$ (9)–(11) needed to initiate the data update (12)–(18).
Initialize the statistics $\{\hat{\theta}_{d;0}, P_{\theta_d;0}\}$ for the VB-marginal modeling the AR noise filter.
Assemble the matrices Υ (47) and Υ_d (49).
Set the desired number of iterations N.
Learning phase:
for $k \leftarrow 1, \mathring{k}$ **do**
  Input: $\begin{cases} y_k, \varphi_k, \bar{\Xi}, \bar{\Sigma}_{k-1}, \nu_{k-1}, \hat{\vartheta}_{k-1}, \hat{\vartheta}_{k-2}, \\ \bar{V}_{k-1}, \bar{P}_{k-1}, \hat{\theta}_{d;k-1}, P_{\theta_d;k-1}, \Upsilon, \Upsilon_d \end{cases}$
  Update: $\hat{\vartheta}_{k-1} \to \hat{\vartheta}_k$, $\bar{P}_{k-1} \to \bar{P}_k$, $\bar{V}_{k-1} \to \bar{V}_k$, $\bar{\Sigma}_{k-1} \to \bar{\Sigma}_k$, and $\nu_{k-1} \to \nu_k$ ▷ (9)–(18)
  Assemble the matrix Φ ▷ (43)
  for $i \leftarrow 1, N$ **do**
    Update: $\{P_{\theta;k}^{i-1}, \hat{\theta}_{d;k}^{i-1}\} \to \{P_{\theta;k}^i, \hat{\theta}_k^i, P_{\theta_d;k}^i, \theta_{d;k}^i\}$ ▷ (52)–(59)
  end for
  Output: $\hat{\vartheta}_k$, $\bar{P}_k$, $\bar{V}_k$, $\bar{\Sigma}_k$, ν_k, $\hat{\theta}_k$, $P_{\theta;k}$, $\hat{\theta}_{d;k}$, $P_{\theta_d;k}$
end for

III. Illustrative Experiments

This section presents a numerical example to provide empirical evidence of the performance of the algorithm. To show its effectiveness, we compare the developed IVB procedure for estimating an ARARX system with the recursive instrumental variable (RIV) method (§9.4 in [1]). The RIV method is in operation with instruments that consist only of delayed inputs. A Bayesian estimation of an ARARX system with a known AR part is included to deliver a reference solution for the ARX part. We simulate the second-order system (1) with a second-order AR noise filter. The coefficients $\{a_i, b_i\}_{i=1}^2$ relate to the discrete transfer function $\mathcal{G}(z) = k_{\mathcal{G}}(z - \exp[T_s])/((z - \exp[T_s p_1])(z - \exp[T_s p_2]))$, and $\{d_i\}_{i=1}^2$ correspond to the discrete polynomial $\mathcal{P}(z) = (z - \exp[-T_s 0.1])(z - \exp[-T_s 0.2])$. The sampling period T_s is chosen as $T_s = 1\,\mathrm{s}$; the poles $p_{1,2} = -0.4 \pm \mathrm{i}0.8$; and the gain $k_{\mathcal{G}} = (1-\exp[T_s p_1])(1-\exp[T_s p_2])/(1-\exp[T_s])$. The input sequence to the system is produced by $u_k = 0.9u_{k-1} + w_k$, using the discrete white noise $w_k \sim \mathcal{N}(0, 1)$. The disturbing white noise sequence $\{e_k\}$ is generated at $d_e = 10$. The simulation is monitored within the time range of $0 - 500\,\mathrm{s}$. All the initial posterior parameter estimates are set to zero vectors. Further, regarding the user-defined input arguments to Algorithms 1 and 2, the learning processes start from $\nu_0 = 10$, $\Xi = I_4$, $\bar{\Sigma}_0 = \Sigma_0 = \frac{1}{10}$, $\bar{\Xi} = I_8$, $P_{\theta_d;0} = 10^6 I_2$, and the number of iterations is $N = 2$.

The result obtained from comparing the method is shown in Fig. 1. As we can observe, the IVB method (Fig. 1(c)) exhibits a comparable estimation quality for the ARX part of the model with the reference analytical Bayesian solution (Fig. 1(b)) and, in addition, the method provides a successful estimate of the AR noise filter (Fig. 1(d)). In this experiment setup with a limited range of observations, the developed IVB method delivers a high disturbance immunity when compared to the RIV method (Fig. 1(a)).

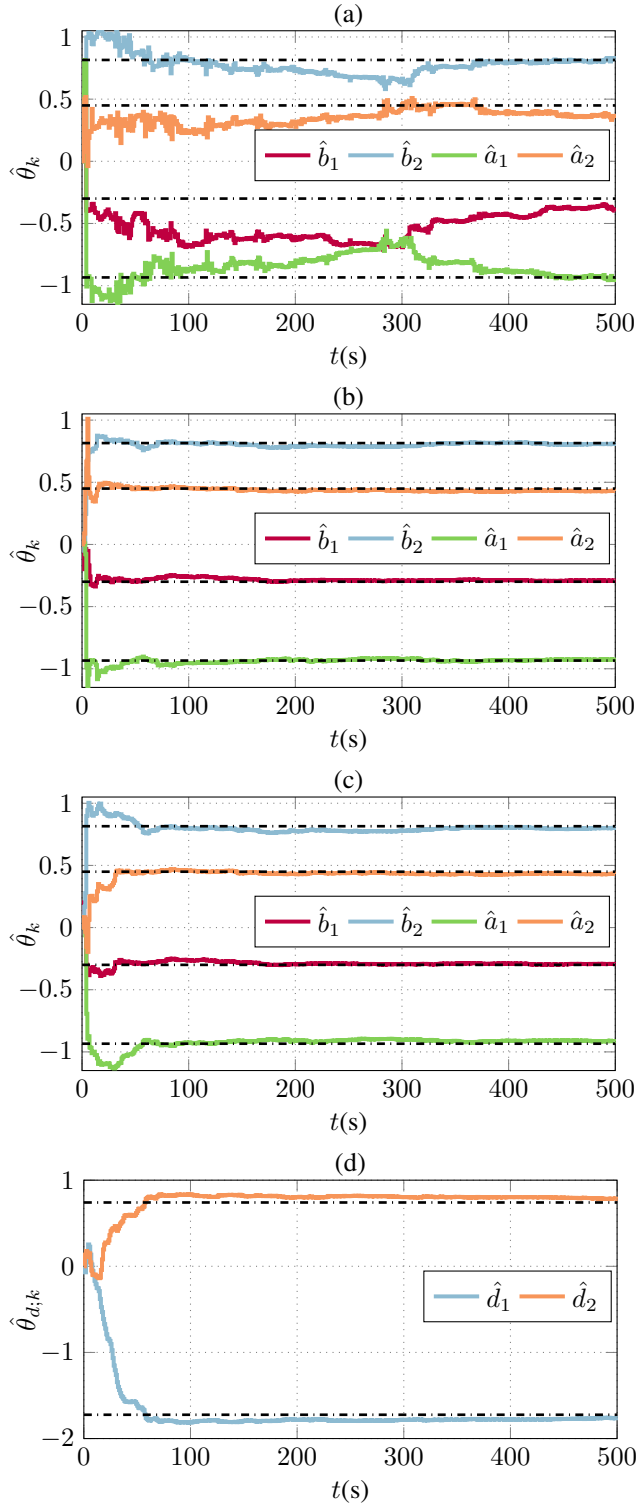

Fig. 1. The trajectories of the ARX model part estimates optimized using (a) the RIV method, (b) the analytical Bayesian solution for the known AR noise filter, and (c) the IVB method. (d) The trajectories of the AR noise filter estimates provided by the IVB method. The true values of the ARARX model coefficients are taken as $\{b_1 \approx -0.3, b_2 \approx 0.82, a_1 \approx -0.93, a_2 \approx 0.45, d_1 \approx -1.72, d_2 \approx 0.74\}$.

IV. Conclusion

The problem of recovering the ARARX model embedded in the high-order ARX model estimate is considered and set into a rigorous probabilistic framework. Respecting the knowledge of the AR noise filter, two algorithm variants are discussed and elaborated on in detail. The ARARX system is classified within the DEFS in Remark 1, offering the construction of an exact posterior and derivation of a systematic procedure for the posterior approximation. The approximation is designed to reduce the model at each step ex-post, after the least squares update of the sufficient statistics has been completed. Lemma 1 converts the reduction problem of the least squares estimate of a high-order ARX model into an optimization problem, tailoring the IVB method to identify the ARARX system. A further theoretical justification concerning the KLD-identifiability [29] of the ARARX model is within the scope of future research.

References

[1] T. Söderström and P. Stoica, *System Identification*. Cambridge, U.K.: Prentice Hall, 1989.

[2] S. S. Niu, L. Ljung, and Å. Björck, "Decomposition methods for solving least-squares parameter estimation," *IEEE Trans. Signal Process.*, vol. 44, no. 11, pp. 2847–2852, Nov. 1996.

[3] R. Diversi, R. Guidorzi, and U. Soverini, "Identification of ARX and ARARX models in the presence of input and output noises," *Eur. J. Control*, vol. 16, no. 3, pp. 242–255, 2010.

[4] R. Diversi, R. Guidorzi, and U. Soverini, "Identification of ARMAX models with additive output noise," in *Proc. 15th IFAC Symp. System Identification*, 2009, pp. 1574–1579.

[5] P. Stoica and T. Söderström, "On reparametrization of loss functions used in estimation and the invariance principle," *Signal Process.*, vol. 17, no. 4, pp. 383–387, Aug. 1989.

[6] P. Stoica, T. Söderström, and B. Friedlander, "An indirect prediction error method for system identification," *Automatica*, vol. 27, no. 1, pp. 183–188, Jan. 1991.

[7] P. Stoica and T. Söderström, "Common factor detection and estimation," *Automatica*, vol. 33, no. 5, pp. 985–989, May 1997.

[8] F. Tjärnström and L. Ljung, "Variance properties of a two-step ARX estimation procedure," *Eur. J. Control*, vol. 9, no. 4, pp. 422–430, 2003.

[9] R. Diversi, "A three-step identification procedure for ARARX models with additive measurement noise," in *Proc. 24th Mediterranean Conf. Control Automation*, 2016, pp. 622–627.

[10] V. Peterka, "Real-time parameter estimation and output prediction for ARMA type system models," *Kybernetika*, vol. 17, no. 6, pp. 526–533, 1981.

[11] L. He and M. Kárný, "Estimation and prediction with ARMMAX model: a mixture of ARMAX models with common ARX part," *Int. J. Adapt. Control Signal Process.*, vol. 17, no. 4, pp. 265–283, May 2003.

[12] M. Kárný, "Towards fully probabilistic control design," *Automatica*, vol. 32, no. 12, pp. 1719–1722, Dec. 1996.

[13] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, and P. Nedoma, *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. London, U.K.: Springer, 2006.

[14] T. P. Minka, "A family of algorithms for approximate Bayesian inference," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2001.

[15] J. Dokoupil, A. Voda, and P. Václavek, "Regularized extended estimation with stabilized exponential forgetting," *IEEE Trans. Autom. Control*, vol. 62, no. 12, pp. 6513–6520, Dec. 2017.

[16] V. Šmídl and A. Quinn, *The Variational Bayes Method in Signal Processing*. Heidelberg, Germany: Springer, 2005.

[17] J. Dokoupil and P. Václavek, "Recursive identification of time-varying Hammerstein systems with matrix forgetting," *IEEE Trans. Autom. Control*, vol. 68, no. 5, pp. 3078–3085, May 2023.

[18] J. Dokoupil and P. Václavek, "Recursive identification of the Hammerstein model based on the variational Bayes method," in *Proc. 60th IEEE Conf. Decision Control*, 2021, pp. 1586–1591.
[19] J. Dokoupil and P. Václavek, "Regularized estimation with variable exponential forgetting," in *Proc. 59th IEEE Conf. Decision Control*, 2020, pp. 312–318.
[20] J. Dokoupil and P. Václavek, "Variable exponential forgetting for estimation of the statistics of the normal-Wishart distribution with a constant precision," in *Proc. 58th IEEE Conf. Decision Control*, 2019, pp. 5094–5100.
[21] J. Dokoupil and P. Václavek, "Design of variable exponential forgetting for estimation of the statistics of the normal-Wishart distribution," in *Proc. Eur. Control Conf.*, 2016, pp. 2565–2570.
[22] J. Dokoupil and P. Václavek, "Data-driven stabilized forgetting design using the geometric mean of normal probability densities," in *Proc. 57th IEEE Conf. Decision Control*, 2018, pp. 1403–1408.
[23] J. Dokoupil and P. Václavek, "Design of variable exponential forgetting for estimation of the statistics of the normal distribution," in *Proc. 55th IEEE Conf. Decision Control*, 2016, pp. 1179–1184.
[24] J. Dokoupil, M. Papež, and P. Václavek, "Comparison of Kalman filters formulated as the statistics of the normal-inverse-Wishart distribution," in *Proc. 54th IEEE Conf. Decision Control*, 2015, pp. 5008–5013.
[25] J. R. Magnus and H. Neudecker, "Symmetry, 0–1 matrices and Jacobians: A review," *Econometr. Theory*, vol. 2, no. 2, pp. 157–190, Aug. 1986.
[26] D. A. Turkington, *Matrix Calculus and Zero-One Matrices*. Cambridge, UK.: Cambridge Univ. Press, 2002.
[27] S. Nakajima, K. Watanabe, and M. Sugiyama, *Variational Bayesian Learning Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2019.
[28] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.
[29] B. Chen, J. Hu, Y. Zhu, and Z. Sun, "Parameter identifiability with Kullback–Leibler information divergence criterion," *Int. J. Adapt. Control Signal Process.*, vol. 23, no. 10, pp. 940–960, Oct. 2009.