Recursive Identification of Time-Varying Hammerstein Systems With Matrix Forgetting

Jakub Dokoupil and Pavel Václavek, *Senior Member, IEEE*

***Abstract*—The real-time estimation of the time-varying Hammerstein system by using a noniterative learning schema is considered and extended to incorporate a matrix forgetting factor. The estimation is cast in a variational-Bayes framework to best emulate the original posterior distribution of the parameters within the set of distributions with feasible moments. The recursive concept we propose approximates the exact posterior comprising undistorted information about the estimated parameters. In many practical settings, the incomplete model of parameter variations is compensated by forgetting of obsolete information. As a rule, the forgetting operation is initiated by the inclusion of an appropriate prediction alternative into the time update. It is shown that the careful formulation of the prediction alternative, which relies on Bayesian conditioning, results in partial forgetting. This article inspects two options with respect to the order of the conditioning in the posterior, which proves vital in the successful localization of the source of inconsistency in the data-generating process. The geometric mean of the discussed alternatives then modifies recursive learning through the matrix forgetting factor. We adopt the decision-making approach to revisit the posterior uncertainty by dynamically allocating the probability to each of the prediction alternatives to be combined.**

***Index Terms*—Hammerstein model, matrix forgetting factor, parameter estimation, variational Bayes.**

I. Introduction

The Hammerstein model consisting of a static nonlinear curve followed by a linear filter provides a capacity to represent a broad class of input nonlinear systems [1], [2]. The list of existing approaches [3] indicates that the Hammerstein model estimation is still dominated by prediction error and maximum likelihood-type methods. The unknown parameters are then obtained by optimizing a certain criterion to best fit the model to the data. This traditional concept is predominantly tied up with point estimation. The available recursive solutions mostly accumulate approximation errors by replacing lossless estimation with one step approximation. This replacement is motivated by updating the latest approximated posterior via a treated parametric model. As a result, approximation errors may accumulate to an extent degrading the estimator performance, making these strategies vulnerable to an inaccurate initial guess.

This article aims to identify the Hammerstein system by approximating the exact posterior probability density function (pdf). The error accumulation is completely avoided by propagating the sufficient statistics of the overparameterized model, which serve as information-bearing for the posterior pdf approximation. The search for the approximate pdf is made optimal by adopting the variational Bayes (VB) method, factorizing the posterior into the product of independent VB-marginals (for a detailed overview, see [4]). The resulting method is designed to account for a hard constraint imposed on the nonlinear curve parameters to uniquely determine the filter gain. The VB method has proven its efficiency when, for instance, tailored to solve the identification for the nonlinear autoregressive with exogenous input (NARX) system [5], to jointly estimate the state and the measurement noise covariance parameters [6], and to iteratively identify the multiple-model based Hammerstein parameter varying systems [7].

In this article, the capability of tracking unmodeled changes in the system dynamics is conceptually achieved via forgetting. At the Bayesian level, a sort of forgetting arises through combining the posterior pdf with its flattened alternative. The combination strategies prominently involve the nonsymmetric Kullback–Leibler divergence (KLD) [8] with different properties depending on the order of the KLD arguments [9]. There is rich literature on the adaptation of a single forgetting factor causing the information about all of the system parameters to be uniformly discounted [10]–[13]. However, the formulation of a matrix forgetting factor capable of providing different forgetting rates for diverse parameter partitions has been neglected. The matrix forgetting algorithms available in [14] and [15] lack any contextualization within the optimization framework, and the solutions thus do not offer an optimal interpretation. Moreover, authors in [14] and [15] numerically search for a symmetric form of the matrix factor that may result in a generally nonsymmetric covariance matrix. In a recent paper [16], a sort of vector forgetting by modifying the least squares criterion is proposed. Importantly, authors in [14]–[16] do not provide any solution on how to apply forgetting differently to various parameters. The Bayesian counterpart to partial forgetting is described in [17], relying on the parallel schema to localize the parameters that are subject to change. On the basis of the work carried out in [11]–[13], [17], we develop a data-informed matrix forgetting factor allowing for tracking a particular parameter subset as well as all the parameters.

Briefly, this article is organized as follows. Section II formally states the estimation problem of the time-varying Hammerstein system from the Bayesian perspective. The relationship between the least squares-like method and the model sufficient statistics is explicated, leaving the question of the choice of the matrix forgetting entities and parameter extraction open. The question related to optimizing the matrix forgetting factor is answered in Section III by adopting the decision-making approach; two variants of the matrix forgetting factor are considered. The optimal extraction of the system parameters from the sufficient statistics in the presence of a hard constraint is discussed

Manuscript received 17 January 2022; revised 2 May 2022; accepted 15 June 2022. Date of publication 5 July 2022; date of current version 26 April 2023. This work was supported in part by Czech Science Foundation under the Project 19-23815S, 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. Recommended by Senior Editor Tetsuya Iwasaki and Guest Editors George J. Pappas, Anuradha M. Annaswamy, Manfred Morari, Claire J. Tomlin, Rene Vidal, and Melanie N. Zeilinger. *(Corresponding author: Jakub Dokoupil.)*

The authors are with the Central European Institute of Technology, Brno University of Technology, 61600 Brno, Czech Republic (e-mail: jakub.dokoupil@ceitec.vutbr.cz; pavel.vaclavek@ceitec.vutbr.cz).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TAC.2022.3188478.

Digital Object Identifier 10.1109/TAC.2022.3188478

in Section IV, relying on the VB method. Section V employs simulated examples to provide empirical evidence of the algorithm performance. Finally, Section VI concludes the article.

Notation: $\mathbf{1}_n$ refers to an n-dimensional vector, all of whose components are one; I_n denotes an $n \times n$ identity matrix; $\mathrm{tr}(\cdot)$ is the matrix trace; $\|\cdot\|_2$ defines the Euclidean vector norm; $|\cdot|$ denotes the determinant; $\otimes$ symbolizes the Kronecker product; $\circ$ denotes the Hadamard product; x^* symbolizes the range of x; $\mathring{x}$ is used to represent the number of members in a countable set x^* or refers to the dimension of a vector x; x' is the transpose of x; and $f(x)$ is reserved for the pdf of a random variable x, optionally distinguished by its subscript. Further, the mathematical expectation of a function $g(x)$ with respect to the pdf $f(x)$ is labeled as $\mathcal{E}_{f(x)}[g(x)] = \int_{x^*} g(x) f(x)\, \mathrm{d}x$; the functional derivative of the functional $\mathcal{L}(f(x))$ over $f(x)$ is defined as $\frac{\delta \mathcal{L}(f(x))}{\delta f(x)}$; $\mathrm{vec}\,(\cdot)$ is the vectorization operator; $\equiv$ stands for equality by definition; and $\propto$ means equality up to a normalizing factor.

II. Problem Statement and Preliminaries

Consider a discrete-time SISO Hammerstein system in which a memoryless nonlinear curve is connected in series with a linear ARX subsystem

$$y_k = \gamma_k' \theta_{a;k} + \theta_{r;k}' G_k \theta_{b;k} + e_k, \tag{1}$$

where the current output y_k depends on the current input u_k and the set of past data through $\gamma_k = [-y_{k-1}, \ldots, -y_{k-n_a}]' \in \mathbb{R}^{n_a}$ and $G_k = \begin{bmatrix} g_1(u_k) & \cdots & g_1(u_{k-n_b}) \\ \vdots & \ddots & \vdots \\ g_{n_r}(u_k) & \cdots & g_{n_r}(u_{k-n_b}) \end{bmatrix} \in \mathbb{R}^{n_r \times (n_b+1)}$. The input u_k and output y_k are both measured 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 $\mathcal{D}_{1-n}^{k} \equiv \{u_i, y_i\}_{i=1-n}^{k}$, with $n \in \mathbb{N}$ referring to the longest time lag appearing in the system. The components of $\theta_{r;k} = [r_{1;k}, \ldots, r_{n_r;k}]'$ combine basis functions $g_i(\cdot)$ to modulate the curve shape, and the components of $\theta_{a;k} = [a_{1;k}, \ldots, a_{n_a;k}]'$ and $\theta_{b;k} = [b_{0;k}, \ldots, b_{n_b;k}]'$ define the dynamics of the ARX subsystem. The unmeasurable model noise e_k is assumed to be white, normally distributed with a zero mean and a nonzero precision d_k, that is, $e_k \sim \mathcal{N}(0, 1/d_k)$. The ordered set $\Theta_k \equiv \{\theta_{a;k}, \theta_{b;k}, \theta_{r;k}, d_k\}$ constitutes the random system parameters to be learned in view of sequential data retrieval. To prevent any information reduction during the data update, the functional form of the dynamic exponential family (DEF) (§6.2.1 in [4]) is adopted as a template for the model parameterization.

Remark 1: The parametric model governed by (1) belongs to the DEF

$$f\left(y_k \middle| \theta_k, d_k, u_k, \mathcal{D}_{1-n}^{k-1}\right) = \exp\left[q\left(\theta_k, d_k\right)' \tau\left(y_k, h_k\right) - \iota_{y_k}\left(\theta_k, d_k\right)\right], \tag{2}$$

under the assignments

$$\theta_k \equiv \begin{bmatrix} \mathbf{1}_{n_a} \\ \left(I_{n_r} \otimes \mathbf{1}_{n_b+1}\right) \theta_{r;k} \end{bmatrix} \circ \begin{bmatrix} \theta_{a;k} \\ \left(\mathbf{1}_{n_r} \otimes I_{n_b+1}\right) \theta_{b;k} \end{bmatrix} \in \mathbb{R}^{\mathring{\theta}}, \quad \mathring{\theta} = n_a + (n_b+1) n_r, \tag{3}$$

$$h_k \equiv \begin{bmatrix} \gamma_k \\ \mathrm{vec}\left(G_k'\right) \end{bmatrix}, \tag{4}$$

$$q\left(\theta_k, d_k\right) = -\frac{d_k}{2} \mathrm{vec}\left([\theta_k' \quad 1]' [\theta_k' \quad 1]\right), \tag{5}$$

$$\tau\left(y_k, h_k\right) = \mathrm{vec}\left([h_k' \quad -y_k]' [h_k' \quad -y_k]\right), \tag{6}$$

with the normalizing factor $\exp[\iota_{y_k}(\theta_k, d_k)] = \int_{-\infty}^{\infty} \exp[q(\theta_k, d_k)' \tau(y_k, h_k)]\, \mathrm{d}y_k = \sqrt{2\pi/d_k}$. Composing the sequence of the pdfs (2) by means of the likelihood function gives rise to the conjugate posterior pdf for the parametric model, having the form

$$f\left(\theta_k, d_k \middle| v_k, \nu_k\right) \propto \exp[q\left(\theta_k, d_k\right)' v_k - (\nu_k + \mathring{\theta} - 2) \iota_{y_k}\left(\theta_k, d_k\right)], \tag{7}$$

where v_k is a vector of a dimension compatible with $q(\theta_k, d_k)$, and the scalar $\nu_k > 2$ is referred to as the number of degrees of freedom. The data compression process is then reduced to the recursive updating of the sufficient statistics $\mathcal{S}_k \equiv \{v_k, \nu_k\}$, which allows for learning about $\{\theta_k, d_k\}$ in tandem with data acquisition.

The correspondence of the parametric model (2) to the DEF determines the conjugate pdf (7) as the normal-Wishart ($\mathcal{NW}$) pdf, with the particular factors defined by

$$f\left(\theta_k \middle| \mathcal{S}_k, d_k\right) = \mathcal{N}\left(\theta_k \middle| \hat{\theta}_k, P_k/d_k\right) \propto \exp\left[-\left(\theta_k - \hat{\theta}_k\right)' P_k^{-1} \left(\theta_k - \hat{\theta}_k\right) d_k/2\right], \tag{8}$$

$$f\left(d_k \middle| \mathcal{S}_k\right) = \mathcal{W}\left(d_k \middle| \Sigma_k, \nu_k\right) \propto d_k^{(\nu_k - 2)/2} \exp\left[-\Sigma_k d_k/2\right], \tag{9}$$

which in turn assigns

$$v_k = \mathrm{vec}\left(\underbrace{\begin{bmatrix} P_k^{-1} & -P_k^{-1}\hat{\theta}_k \\ -\hat{\theta}_k' P_k^{-1} & \Sigma_k + \hat{\theta}_k' P_k^{-1} \hat{\theta}_k \end{bmatrix}}_{\bar{V}_k}\right), \tag{10}$$

where $\Sigma_k > 0$ is referred to as the least squares reminder. As noted earlier, we assume vague knowledge regarding how the parameters actually evolve, and this prevents us from employing the marginalization integral [18]

$$f\left(\theta_k, d_k \middle| \mathcal{S}_{k-1}\right) = \int_{d^*} \int_{\theta^*} f\left(\theta_k, d_k \middle| \theta_{k-1}, d_{k-1}, \mathcal{S}_{k-1}\right) \times f\left(\theta_{k-1}, d_{k-1} \middle| \mathcal{S}_{k-1}\right) \mathrm{d}\theta_{k-1}\, \mathrm{d}d_{k-1}. \tag{11}$$

Owing to such deficiency, we seek a forgetting operation that emerges from rescaling the covariances of $\theta_k|\mathcal{S}_k, d_k$ and $d_k|\mathcal{S}_k$ at each iteration to reinstate the parameter tracking capability. To guarantee a minimal amount of parameter-related information, the additional source to (2) is processed in a way that stabilizes the forgetting and thus compensates for the potential loss of persistency [12]. We have

$$f(\theta_k, d_k | \hat{\theta}_0, \Xi, \Sigma_0, \nu_0) = \mathcal{N}(\theta_k | \hat{\theta}_0, \Xi^{-1}/d_k) \mathcal{W}\left(d_k \middle| \Sigma_0, \nu_0\right), \tag{12}$$

where Ξ is some symmetric positive definite matrix of an appropriate dimension, $\lambda_i \in (0, 1]$ is the forgetting factor, $\Sigma_0 > 0$, and $\nu_0 > 2$. Substituting for the ideal transition operation (11) a forgetting operation, the update of the latest posterior $\mathcal{N}(\theta_{k-1}|\hat{\theta}_{k-1}, P_{k-1}/d_{k-1}) \mathcal{W}(d_{k-1}|\Sigma_{k-1}, \nu_{k-1})$ is organized recursively with respect to Bayes' rule, as follows:

$$f\left(\theta_k, d_k \middle| \mathcal{S}_k\right) \propto \mathcal{N}\left(y_k \middle| h_k' \theta_k, 1/d_k\right) \times \frac{\mathcal{N}\left(\theta_k \middle| \hat{\theta}_0, \Xi^{-1}/d_k\right) \mathcal{W}\left(d_k \middle| \Sigma_0, \nu_0\right)}{\mathcal{N}\left(\theta_k \middle| \hat{\theta}_0, \Omega_{k-1}^{-1} \Xi^{-1}/d_k\right) \mathcal{W}\left(d_k \middle| \lambda_{k-1} \Sigma_0, \lambda_{k-1} \nu_0\right)} \times \mathcal{N}\left(\theta_k \middle| \hat{\theta}_{k-1}, \Lambda_{k-1}^{-1} P_{k-1}/d_k\right) \mathcal{W}\left(d_k \middle| \lambda_{k-1} \Sigma_{k-1}, \lambda_{k-1} \nu_{k-1}\right), \tag{13}$$

where $\Lambda_{k-1} \in \mathbb{R}^{\mathring{\theta} \times \mathring{\theta}}$ is the matrix forgetting factor, and $\Omega_{k-1} \in \mathbb{R}^{\mathring{\theta} \times \mathring{\theta}}$ is constructed so that the pdf $\mathcal{N}(\theta_k|\hat{\theta}_0, \Omega_{k-1}^{-1}\Xi^{-1}/d_k)$ embodies the residual regularization effect of the additional source (12), remaining in the estimator memory after the forgetting has been performed. The application of λ_{k-1} must coincide with a reduction of the degrees of freedom caused by the matrix forgetting to obtain a realistic estimate of d_k, compatible with $\hat{d}_k = \mathcal{E}_{\mathcal{W}(d_k|\Sigma_k, \nu_k)}[d_k] = \nu_k/\Sigma_k$. This coincidence is established in relation to both of the partial forgetting options and is discussed in Remarks 2 and 3. We expand the concept

of a uniform rate of forgetting for all of the parameters by designing the self-tuning factor Λ_k, which is able to localize forgetting on the parameters associated with the system measured input (see Section III).

By setting $\Lambda_0 \equiv \Omega_0 \equiv I_{\mathring{\theta}}$, $\lambda_0 \equiv 1$, and $P_0 \equiv \Xi^{-1}$ in (13) the pdf (12) formally initiates the estimation procedure at the time $k = 1$. The conjugacy inherent between members of the DEF allows us to reduce the functional recursion (13) to the least squares-like algebraic recursion, namely

$$V_{c;k-1} \equiv V_{k-1}\Lambda_{k-1} + \Xi\left(I_{\mathring{\theta}} - \Omega_{k-1}\right), \tag{14}$$

$$P_{c;k-1} \equiv V_{c;k-1}^{-1}, \tag{15}$$

$$\varepsilon_{k-1} \equiv \hat{\theta}_0 - \hat{\theta}_{k-1}, \tag{16}$$

$$\hat{\theta}_{c;k-1} \equiv \hat{\theta}_{k-1} + P_{c;k-1}\Xi\left(I_{\mathring{\theta}} - \Omega_{k-1}\right)\varepsilon_{k-1}, \tag{17}$$

$$\begin{aligned}\Sigma_{c;k-1} \equiv{}& \lambda_{k-1}\Sigma_{k-1} + \varepsilon'_{k-1}\left[I_{\mathring{\theta}} - \Xi\left(I_{\mathring{\theta}} - \Omega_{k-1}\right)P_{c;k-1}\right]\\ &\times \Xi\left(I_{\mathring{\theta}} - \Omega_{k-1}\right)\varepsilon_{k-1} + \left(1-\lambda_{k-1}\right)\Sigma_0,\end{aligned} \tag{18}$$

$$K_k \equiv P_{c;k-1}h_k/\left(1 + h'_k P_{c;k-1}h_k\right), \tag{19}$$

$$\hat{e}_{c;k} \equiv y_k - h'_k\hat{\theta}_{c;k-1}, \tag{20}$$

$$\hat{\theta}_k = \hat{\theta}_{c;k-1} + K_k\hat{e}_{c;k}, \tag{21}$$

$$P_k = \left(I_{\mathring{\theta}} - K_k h'_k\right)P_{c;k-1}\left(I_{\mathring{\theta}} - K_k h'_k\right)' + K_k K'_k, \tag{22}$$

$$V_k = P_k^{-1} = V_{c;k-1} + h_k h'_k, \tag{23}$$

$$\Sigma_k = \Sigma_{c;k-1} + \hat{e}^2_{c;k}/\left(1 + h'_k P_{c;k-1}h_k\right), \tag{24}$$

$$\nu_k = \lambda_{k-1}\nu_{k-1} + \left(1-\lambda_{k-1}\right)\nu_0 + 1. \tag{25}$$

Since θ_k (3) contains product terms coupling the components of $\theta_{b;k}$ and $\theta_{r;k}$, there is no solution to distinguish between $\{\theta_{b;k}, \theta_{r;k}\}$ and $\{\theta_{b;k}/\beta, \beta\theta_{r;k}\}$ scaled with some nonzero and finite constant β. To remove this scaling ambiguity, we fix $r_{1;k} \equiv 1$ by imposing the hard equality constraint on $\theta_{r;k}$ (see Section IV).

III. Data-Informed Matrix Forgetting

Let us now be concerned with the design of the factor Λ_k enabling us to provide different exponential forgetting rates for two partitions of θ_k. In conformity with the announced objective, four distinct alternatives, $\{f_i(\theta_{k+1}, d_{k+1})\}_{i\in\{0,2\}}$ and $\{f_{1,\kappa}(\theta_{k+1}, d_{k+1})\}_{\kappa\in\{\sigma,\rho\}}$, concerning the result of the time update are to be introduced. These alternatives delimit the boundaries on the increase in the parameter uncertainties to allow for a revision of the posterior pdf, reflecting unknown uncertainty about the parameters caused by their variations. The zero alternative $f_0(\theta_{k+1}, d_{k+1})$ corresponds to the latest posterior available, extrapolating all information to describe $\{\theta_{k+1}, d_{k+1}\}$ accumulated so far, that is

$$\begin{aligned}&f_0\left(\theta_{k+1}, d_{k+1}\right)\\ \equiv{}& \int_{d^*}\int_{\theta^*}\delta\left(\theta_{k+1} - \theta_k\right)\delta\left(d_{k+1} - d_k\right)f\left(\theta_k, d_k\middle|\mathcal{S}_k\right)\,\mathrm{d}\theta_k\,\mathrm{d}d_k,\end{aligned} \tag{26}$$

where $\delta(\cdot)$ is the Dirac delta function. The second reference alternative $f_2(\theta_{k+1}, d_{k+1})$ expects that all the parameters $\{\theta_{k+1}, d_{k+1}\}$ have changed within the interval $(k, k+1)$, and increases the uncertainty of the posterior accordingly, through $\alpha \in (0,1)$ to yield

$$\begin{aligned}&f_2\left(\theta_{k+1}, d_{k+1}\right)\\ \equiv{}& \mathcal{N}\left(\theta_{k+1}\middle|\hat{\theta}_k, \alpha^{-1}P_k/d_{k+1}\right)\mathcal{W}\left(d_{k+1}\middle|\alpha\Sigma_k, \alpha\nu_k\right).\end{aligned} \tag{27}$$

The first reference alternative localizes the increase in the uncertainty to a lower dimensional posterior factor. To factorize the normal posterior part into low-dimensional pdfs, let P_k be split in accordance with the splitting of $\theta_k = [\theta'_{a;k}, \theta'_{u;k}]'$, as follows:

$$P_k = \begin{bmatrix} P_{aa;k} & P'_{ua;k}\\ P_{ua;k} & P_{uu;k}\end{bmatrix} \in \mathbb{R}^{\hat{\theta}\times\hat{\theta}} \tag{28}$$

with $P_{aa;k} \in \mathbb{R}^{n_a\times n_a}$. Now, we can proceed to constructing the alternatives $f_{1,\kappa}(\theta_{k+1}, d_{k+1})$ corresponding to the expectation that only the subset of θ_{k+1} ($\theta_{u;k+1}$) associated with the input signal is subject to change. In this respect, two variants of partial forgetting are proposed, and their implications for the least squares routine are discussed. While the first option ($\kappa \equiv \sigma$) relies on the factorization of the normal posterior part according to

$$\begin{aligned}&\mathcal{N}\left(\theta_k\middle|\hat{\theta}_k, P_k/d_k\right)\\ ={}& \mathcal{N}\left(\theta_{a;k}\middle|\hat{\theta}_{a|u;k}, P_{a|u;k}/d_k\right)\mathcal{N}\left(\theta_{u;k}\middle|\hat{\theta}_{u;k}, P_{uu;k}/d_k\right)\end{aligned} \tag{29}$$

with $\hat{\theta}_{a|u;k} = \hat{\theta}_{a;k} - P'_{ua;k}P^{-1}_{uu;k}(\theta_{u;k} - \hat{\theta}_{u;k})$ and $P_{a|u;k} = P_{aa;k} - P'_{ua;k}P^{-1}_{uu;k}P_{ua;k}$, the other ($\kappa \equiv \rho$) builds upon the factorization variant

$$\begin{aligned}&\mathcal{N}\left(\theta_k\middle|\hat{\theta}_k, P_k/d_k\right)\\ ={}& \mathcal{N}\left(\theta_{a;k}\middle|\hat{\theta}_{a;k}, P_{aa;k}/d_k\right)\mathcal{N}\left(\theta_{u;k}\middle|\hat{\theta}_{u|a;k}, P_{u|a;k}/d_k\right),\end{aligned} \tag{30}$$

where $\hat{\theta}_{u|a;k} = \hat{\theta}_{u;k} - P_{ua;k}P^{-1}_{aa;k}(\theta_{a;k} - \hat{\theta}_{a;k})$ and $P_{u|a;k} = P_{uu;k} - P_{ua;k}P^{-1}_{aa;k}P'_{ua;k}$. Considering $f(\theta_a|\theta_u)f(\theta_u) = f(\theta_a)f(\theta_u|\theta_a)$, the results (29) and (30) can be directly obtained by application of Claim 1 from [19] to the normal posterior part.

For the steps that follow, we need to partition V_k and Ξ into blocks, identically to the partitioning performed in (28), namely,

$$V_k = \begin{bmatrix} V_{aa;k} & V'_{ua;k}\\ V_{ua;k} & V_{uu;k}\end{bmatrix} \in \mathbb{R}^{\hat{\theta}\times\hat{\theta}}, \quad \Xi = \begin{bmatrix} \Xi_{aa} & \Xi'_{ua}\\ \Xi_{ua} & \Xi_{uu}\end{bmatrix} \in \mathbb{R}^{\hat{\theta}\times\hat{\theta}}, \tag{31}$$

where $V_{aa;k} \in \mathbb{R}^{n_a\times n_a}$ and $\Xi_{aa} \in \mathbb{R}^{n_a\times n_a}$.

Remark 2: The partial forgetting based on modification of the information submatrix $V_{uu;k}$ [17] employs (29), with the marginal part flattened through α, yielding

$$\begin{aligned}f_{1,\sigma}\left(\theta_{k+1}\middle|d_{k+1}\right) \equiv{}& \mathcal{N}\left(\theta_{a;k+1}\middle|\hat{\theta}_{a|u;k}, P_{a|u;k}/d_{k+1}\right)\\ &\times\mathcal{N}\left(\theta_{u;k+1}\middle|\hat{\theta}_{u;k}, \alpha^{-1}P_{uu;k}/d_{k+1}\right).\end{aligned} \tag{32}$$

The above time update (32) corresponds to the matrix forgetting with $\Lambda_k = \begin{bmatrix} I_{n_a} & (1-\alpha)V^{-1}_{aa;k}V'_{ua;k}\\ 0 & \alpha I_{(n_b+1)n_r}\end{bmatrix}$ and $\Omega_k = \begin{bmatrix} I_{n_a} & (1-\alpha)\Xi^{-1}_{aa}\Xi'_{ua}\\ 0 & \alpha I_{(n_b+1)n_r}\end{bmatrix}$. The Wishart part of the prediction alternative $f_{1,\sigma}(d_{k+1})$ is suggested to reduce the degrees of freedom consistently with (32), which is fulfilled by solving

$$\begin{aligned}f\left(d_{k+1}\middle|\mathcal{D}^k_{1-n}, \mathit{\Pi}, d_{k+1}\right) \propto{}& f\left(y_k\middle|u_k, \mathcal{D}^{k-1}_{1-n}, \mathit{\Pi}, d_{k+1}\right)\\ &\times f\left(d_{k+1}\middle|u_k, \mathcal{D}^{k-1}_{1-n}, \mathit{\Pi}\right),\end{aligned} \tag{33}$$

where $\mathit{\Pi} \equiv \{\hat{\theta}_0, \Xi, \Sigma_0, \nu_0\}$. To solve (33), we consolidate the update (32) within Bayes' rule

$$\begin{aligned}f\left(\theta_{u;k+1}\middle|\mathcal{D}^k_{1-n}, \mathit{\Pi},\ d_{k+1}\right) \propto{}& f\left(\theta_{u;k+1}\middle|\mathcal{D}^{k-1}_{1-n}, \mathit{\Pi}, d_{k+1}\right)\\ &\times f^{\alpha}\left(y_k\middle|u_k, \mathcal{D}^{k-1}_{1-n}, \mathit{\Pi}, \theta_{u;k+1}, d_{k+1}\right).\end{aligned} \tag{34}$$

Since the pdf $f(y_k|u_k, \mathcal{D}^{k-1}_{1-n}, \mathit{\Pi}, d_{k+1})$ from (33) is actually the normalizing factor for (34), it is obtained by integrating out $\theta_{u;k+1}$ from (34), which shows as

$$\begin{aligned}&f\left(y_k\middle|u_k, \mathcal{D}^{k-1}_{1-n}, \mathit{\Pi}, d_{k+1}\right)\\ &\quad= \bar{o}\mathcal{N}^{\alpha}\left(y_k\middle|h'_k\hat{\theta}_{c;k-1}, \left(1 + h'_k P_{c;k-1}h_k\right)/d_{k+1}\right),\end{aligned} \tag{35}$$

where $\bar{o}$ is some positive constant. From the above improper pdf (35), we can directly find the factor $\lambda_k = \alpha$ to write

$$f_{1,\sigma}\left(d_{k+1}\right) = \mathcal{W}\left(d_{k+1}\middle|\alpha\Sigma_k, \alpha\nu_k\right). \tag{36}$$

Remark 3: The partial forgetting based on modification of the covariance submatrix $P_{uu;k}$ follows from (30), with the conditional part flattened through α, resulting in

$$f_{1,\rho}\left(\theta_{k+1}\middle|d_{k+1}\right) \equiv \mathcal{N}\left(\theta_{a;k+1}\middle|\hat{\theta}_{a;k}, P_{aa;k}/d_{k+1}\right) \times \mathcal{N}\left(\theta_{u;k+1}\middle|\hat{\theta}_{u|a;k}, \alpha^{-1}P_{u|a;k}/d_{k+1}\right). \quad (37)$$

The above operation (37) determines the entities of the matrix forgetting as $\Lambda_k = \begin{bmatrix} I_{n_a} & 0 \\ (1-\alpha)P_{ua;k}P_{aa;k}^{-1} & \alpha I_{(n_b+1)n_r} \end{bmatrix}$ and $\Omega_k = \begin{bmatrix} I_{n_a} & 0 \\ (\alpha-1)\Xi_{uu}^{-1}\Xi_{ua} & \alpha I_{(n_b+1)n_r} \end{bmatrix}$. All that remains now is to choose the Wishart part $f_{1,\rho}(d_{k+1})$, which complements the description of the parameters with regard to the impact of Λ_k on λ_k. The inclusion of the forgetting operation (37) into the Bayes update yields

$$f\left(\theta_{a;k+1}\middle|\mathcal{D}_{1-n}^{k}, \Pi, d_{k+1}\right) \propto f\left(\theta_{a;k+1}\middle|\mathcal{D}_{1-n}^{k-1}, \Pi, d_{k+1}\right) \times f\left(y_k\middle|u_k, \mathcal{D}_{1-n}^{k-1}, \Pi, \theta_{a;k+1}, d_{k+1}\right). \quad (38)$$

The normalizing factor for (38) updating the Wishart two steps ahead prediction pdf $f(d_{k+1}|u_k, \mathcal{D}_{1-n}^{k-1}, \Pi)$ is found to be

$$f\left(y_k\middle|u_k, \mathcal{D}_{1-n}^{k-1}, \Pi, d_{k+1}\right) = \mathcal{N}\left(y_k\middle|h_k'\hat{\theta}_{c;k-1}, \left(1 + h_k'P_{c;k-1}h_k\right)/d_{k+1}\right). \quad (39)$$

From (39), it can be deduced that the Wishart part must be treated as

$$f_{1,\rho}\left(d_{k+1}\right) = \mathcal{W}\left(d_{k+1}\middle|\Sigma_k, \nu_k\right), \quad (40)$$

indicating at the same time that no external flattening of the Wishart posterior occurs; therefore, $\lambda_k = 1$.

The optimal design $\hat{f}(\theta_{k+1}, d_{k+1}|\mathcal{S}_k)$ seeks the best approximation of the time-updated posterior $f(\theta_{k+1}, d_{k+1}|\mathcal{S}_k)$ by combining the prediction alternatives $f_{\mathcal{H},\kappa}^{*} \equiv \{f_0, f_{1,\kappa}, f_2\}$ of the arguments $\{\theta_{k+1}, d_{k+1}\}$ into a single pdf. To execute this, a dissimilarity measure between the target pdf $f(\theta_{k+1}, d_{k+1}|\mathcal{S}_k)$ and the particular alternatives is quantified by the KLD. The KLD between any two pdfs f_{T} and f,

$$\mathcal{D}\left(f_{\mathrm{T}}\|f\right) \equiv \int_{\theta^*}\int_{d^*} f_{\mathrm{T}}\left(\theta, d\right) \ln\left(\frac{f_{\mathrm{T}}\left(\theta, d\right)}{f\left(\theta, d\right)}\right) \mathrm{d}\theta\mathrm{d}d, \quad (41)$$

attains its absolute minimum value, which equals zero, at $f_{\mathrm{T}} \equiv f$. To establish a meaningful combination strategy modulating the posterior with regard to the degree of the system nonstationarity, the decision step must respect information about the performances of the prediction alternatives. The required feedback is incorporated into the decision process by considering the nonnegative loss estimates $\varrho_{\mathcal{H},\kappa}^{*} \equiv \{\varrho_0, \varrho_{1,\kappa}, \varrho_2\}$. These are chosen to embody losses incurred at the previous step when selecting each of the prediction alternatives as the best projection of the current posterior

$$\varrho_i \equiv \mathcal{D}\left(\mathcal{N}_\zeta\left(\theta_k\right)\mathcal{W}\left(d_k\middle|\Sigma_k, \nu_k\right)\|f_{i,\zeta}\left(\theta_k\right)f_i\left(d_k\right)\right), \quad (42)$$

$$\mathcal{N}_\zeta\left(\theta_k\right) \equiv \mathcal{N}\left(\theta_k\middle|\hat{\theta}_k, \left(d_k\zeta\right)^{-1}P_k\right), \quad i \in \{0, 2\},$$

$$\varrho_{1,\kappa} \equiv \mathcal{D}\left(\mathcal{N}_\zeta\left(\theta_k\right)\mathcal{W}\left(d_k\middle|\Sigma_k, \nu_k\right)\|f_{1,\kappa,\zeta}\left(\theta_k\right)f_{1,\kappa}\left(d_k\right)\right). \quad (43)$$

Here, the pdfs indexed by ζ refer to the particular normal pdfs $\{f_0, f_{1,\kappa}, f_2\}$ of the argument θ_k, where the precision d_k is additionally multiplied by ζ. The user-defined factor $\zeta \in (0, 1]$ serves to increase the expected level of noise inherent in the normal parts to reduce false detections of parameter changes as a consequence. In particular, the smaller the value of ζ, the more conservative the forgetting.

The randomness about the alternative selection is modeled by assigning a probability φ_i to each of the related pairs from $\{f_{\mathcal{H},\kappa}, \varrho_{\mathcal{H},\kappa}\}$, defining the weights by which the alternatives are combined. The probabilities are arranged into the vector $\varphi \equiv [\varphi_0, \varphi_1, \varphi_2]'$, satisfying $\sum_{i=0}^{2}\varphi_i = 1$. Let us now temporarily omit the time index in $\{\theta_{k+1}, d_{k+1}, \mathcal{S}_k\}$ and the index κ in $\{f_{1,\kappa}, \varrho_{1,\kappa}\}$ for the sake of brevity. The specification of a loss functional coherent with the Bayes principle involves evaluation of the expectation $\mathcal{E}[\mathcal{D}(f(\theta, d)\|f_{\mathcal{H},\kappa}(\theta, d)) - \varrho_{\mathcal{H},\kappa}]$ over $\{f_{\mathcal{H},\kappa}, \varrho_{\mathcal{H},\kappa}\}$. In this connection, we obtain

$$\mathcal{L}_{\mathcal{F}}\left(f\left(\theta, d\middle|\mathcal{S}\right), \varphi\right) = \sum_{i=0}^{2}\varphi_i\left[\mathcal{D}\left(f\left(\theta, d\middle|\mathcal{S}\right)\middle\|f_i\left(\theta, d\right)\right) - \varrho_i\right] + \eta\left(\int_{d^*}\int_{\theta^*} f\left(\theta, d\middle|\mathcal{S}\right)\mathrm{d}\theta\mathrm{d}d - 1\right), \quad (44)$$

where the constraint scaled by the Lagrange multiplier η guarantees that the minimizer $\hat{f}(\theta, d|\mathcal{S})$ integrates to one. The form of the minimizer for an arbitrary φ is inspected by the lemma below:

Lemma 1: The unique constrained minimizer of the functional (44) over $f(\theta, d|\mathcal{S})$ turns out to be given by the geometric mean

$$\hat{f}\left(\theta, d|\mathcal{S}\right) \propto \prod_{i=0}^{2} f_i^{\varphi_i}\left(\theta, d\right). \quad (45)$$

Proof: The summation $\sum_{i=0}^{2}\varphi_i[\mathcal{D}(f\|f_i) - \varrho_i]$ entering (44) can be rearranged into a sum of two parts:

$$\mathcal{D}\left(f\left(\theta, d|\mathcal{S}\right)\middle\|\hat{f}\left(\theta, d|\mathcal{S}\right)\right) + \overbrace{\min_{f(\theta,d|\mathcal{S})}\mathcal{L}_{\mathcal{F}}\left(f\left(\theta, d\middle|\mathcal{S}\right), \varphi\right)}^{\mathcal{K}(\varphi)}, \quad (46)$$

where the part independent of the inspected pdf $f(\theta, d|\mathcal{S})$,

$$\mathcal{K}\left(\varphi\right) = -\sum_{i=0}^{2}\varphi_i\varrho_i - \ln\left(\int_{d^*}\int_{\theta^*}\prod_{i=0}^{2} f_i^{\varphi_i}\left(\theta, d\right)\mathrm{d}\theta\mathrm{d}d\right), \quad (47)$$

absorbs the achieved minimum value. Consequently, the evaluation of the necessary conditions for an extremum $\frac{\delta\mathcal{L}_{\mathcal{F}}}{\delta\hat{f}} = [\ln(f/\hat{f}) + \eta + 1]$ and $\frac{\partial\mathcal{L}_{\mathcal{F}}}{\partial\eta} = 0$ determines the form of the minimizer (45). The uniqueness of the minimizer is given by the strict convexity of the KLD, which is reflected by $\frac{\delta^2\mathcal{L}_{\mathcal{F}}}{\delta\hat{f}^2} = 1/f > 0$. ■

By substituting (46) into (44), one can prove that $\mathcal{K}(\varphi)$ delimits the lower bound on the functional approximation accuracy. To be more explicit,

$$\mathcal{D}\left(f\left(\theta, d|\mathcal{S}\right)\middle\|\hat{f}\left(\theta, d|\mathcal{S}\right)\right) = \mathcal{L}_{\mathcal{F}}\left(f\left(\theta, d\middle|\mathcal{S}\right), \varphi\right) - \mathcal{K}\left(\varphi\right) \geq 0. \quad (48)$$

Hence, taking into account the nonnegativity of the KLD, the best representative of φ is found as the maximizer of $\mathcal{K}(\varphi)$.

Lemma 2: Let the members from $f_{\mathcal{H},\kappa}^{*}$ constitute the set of nonnegative, distinguishable pdfs. Then, the search for the maximizer $\hat{\varphi}$ has a unique solution provided by the necessary and sufficient conditions

$$\begin{cases} \mathcal{D}\left(\hat{f}\left(\theta, d|\mathcal{S}\right)\middle\|f_i\left(\theta, d\right)\right) - \varrho_i = \mu, \text{ all } i \text{ such that } \varphi_i > 0, \\ \mathcal{D}\left(\hat{f}\left(\theta, d|\mathcal{S}\right)\middle\|f_i\left(\theta, d\right)\right) - \varrho_i \leq \mu, \text{ all } i \text{ such that } \varphi_i = 0, \end{cases} \quad (49)$$

where $i \in \{0, 1, 2\}$, and μ is a real-valued scalar.

Proof: Hölder's inequality ([20, §3.1.9]) implies that

$$\int_{x^*}\psi_1^{\vartheta}(x)\psi_2^{(1-\vartheta)}(x)\mathrm{d}x < \left(\int_{x^*}\psi_1(x)\mathrm{d}x\right)^{\vartheta}\left(\int_{x^*}\psi_2(x)\mathrm{d}x\right)^{(1-\vartheta)} \quad (50)$$

holds for any nonnegative, distinguishable functions $\psi_1(x)$, $\psi_2(x)$, and $\vartheta \in (0, 1)$. By invoking (50), the function $-\mathcal{K}(\varphi)$ is recognized as a strictly convex function of φ since it meets Jensen's inequality

$$-\mathcal{K}\left(\omega\vartheta + \varpi\left(1 - \vartheta\right)\right) < \ln\left[\left(\int_{d^*}\int_{\theta^*}\prod_{i=0}^{2} f_i^{\omega_i}\left(\theta, d\right)\mathrm{d}\theta\mathrm{d}d\right)^{\vartheta} \times \left(\int_{d^*}\int_{\theta^*}\prod_{i=0}^{2} f_i^{\varpi_i}\left(\theta, d\right)\mathrm{d}\theta\mathrm{d}d\right)^{(1-\vartheta)}\right] + \sum_{i=0}^{2}\varphi_i\varrho_i, \quad (51)$$

TABLE I
EVALUATION OF THE KKT CONDITIONS FOR $\hat{\varphi}$

Condition	$\hat{\varphi}_0$	$\hat{\varphi}_1$	$\hat{\varphi}_2$
$\epsilon_{01} \leq 1$ and $\epsilon_{02} \leq 1$	1	0	0
$1 \leq \alpha\epsilon_{01}$ and $\epsilon_{12} \leq 1$	0	1	0
$1 \leq \alpha\epsilon_{02}$ and $1 \leq \alpha\epsilon_{12}$	0	0	1
$1 \leq \alpha\epsilon_{01}$	0	$\frac{\alpha\epsilon_{12}-1}{\epsilon_{12}(\alpha-1)}$	$\frac{1-\epsilon_{12}}{\epsilon_{12}(\alpha-1)}$
$\epsilon_{01} \leq \epsilon_{02}$	$\frac{\alpha\epsilon_{02}-1}{\epsilon_{02}(\alpha-1)}$	0	$\frac{1-\epsilon_{02}}{\epsilon_{02}(\alpha-1)}$
$\epsilon_{12} \leq 1$	$\frac{\alpha\epsilon_{01}-1}{\epsilon_{01}(\alpha-1)}$	$\frac{1-\epsilon_{01}}{\epsilon_{01}(\alpha-1)}$	0
Otherwise	$\frac{\alpha\epsilon_{01}-1}{\epsilon_{01}(\alpha-1)}$	$\frac{\epsilon_{12}-\epsilon_{01}}{(\epsilon_{01}\epsilon_{12})(\alpha-1)}$	$\frac{1-\epsilon_{12}}{\epsilon_{12}(\alpha-1)}$

where the components of $\omega \in \varphi^*$ and $\varpi \in \varphi^*$ satisfy $\varphi_i \equiv \omega_i\vartheta + \varpi_i(1-\vartheta)$. Then, the conditions (49) follow from the Karush-Kuhn-Tucker optimality conditions applied to $-\mathcal{K}(\varphi)$ (Theorem 4.4.1 in [21]),

$$\varphi_i \frac{\partial \mathcal{K}(\varphi)}{\partial \varphi_i} = \varphi_i \bar{\mu}, \quad \frac{\partial \mathcal{K}(\varphi)}{\partial \varphi_i} \leq \bar{\mu}, \tag{52}$$

where $i \in \{0,1,2\}$, and $\bar{\mu}$ is a Lagrange multiplier. ■

Recall that the search for the optimal value of φ according to Lemma 2 requires us to express the KLD between two normal-Wishart pdfs. The required analytical form for the KLD is explicated in Theorem 1 in [12]. The formulation of the time update on the basis of the elaborated Bayes principle affects the least squares routine only through the factors $\{\Lambda_k, \Omega_k, \lambda_k\}$ as a consequence of mixing the prediction alternatives. Upon using Lemma 1 and assuming that $\hat{\varphi}$ is known in advance, for $\kappa \equiv \sigma$, the factors $\{\Lambda_k, \Omega_k, \lambda_k\}$ become

$$\begin{cases} \lambda_k = \hat{\varphi}_0(1-\alpha) + \alpha, \\ \Lambda_k = \begin{bmatrix} (\lambda_k + \hat{\varphi}_1(1-\alpha))I_{\mathring{\theta}_a} & \hat{\varphi}_1(1-\alpha)V_{aa;k}^{-1}V_{ua;k}' \\ 0 & \lambda_k I_{\mathring{\theta}_u} \end{bmatrix}, \\ \Omega_k = \begin{bmatrix} (\lambda_k + \hat{\varphi}_1(1-\alpha))I_{\mathring{\theta}_a} & \hat{\varphi}_1(1-\alpha)\Xi_{aa}^{-1}\Xi_{ua}' \\ 0 & \lambda_k I_{\mathring{\theta}_u} \end{bmatrix}, \end{cases} \tag{53}$$

and, for $\kappa \equiv \rho$, the factors are formulated as

$$\begin{cases} \lambda_k = (\hat{\varphi}_0 + \hat{\varphi}_1)(1-\alpha) + \alpha, \\ \Lambda_k = \begin{bmatrix} \lambda_k I_{\mathring{\theta}_a} & 0 \\ \hat{\varphi}_1(1-\alpha)P_{ua;k}P_{aa;k}^{-1} & (\hat{\varphi}_0(1-\alpha)+\alpha)I_{\mathring{\theta}_u} \end{bmatrix}, \\ \Omega_k = \begin{bmatrix} \lambda_k I_{\mathring{\theta}_a} & 0 \\ \hat{\varphi}_1(\alpha-1)\Xi_{uu}^{-1}\Xi_{ua} & (\hat{\varphi}_0(1-\alpha)+\alpha)I_{\mathring{\theta}_u} \end{bmatrix}. \end{cases} \tag{54}$$

The next step is to find the optimal value of φ provided by the KKT conditions (49). This task requires us to examine whether the solution lies inside the feasible region $S = \{\varphi \in \varphi^* \subseteq \mathbb{R}^3_{\geq 0} : \varphi_i \geq 0, i = 0,1,2, \sum_{i=0}^2 \varphi_i = 1\}$ or on the constraint boundaries.

In order to perform an exhaustive description of $\hat{\varphi}$, captured in Table I, we introduce the auxiliary variables $\epsilon_{02}, \epsilon_{12}$, and ϵ_{01}. By denoting $\bar{r}_k \equiv \hat{\theta}_k - \hat{\theta}_{k-1}$ and letting $z_k \equiv \nu_{k-1}\ln(\hat{d}_{k-1}/\hat{d}_k) + \hat{d}_k\Sigma_{k-1} + \nu_{k-1}/\nu_k - \nu_{k-1}$ with $\hat{d}_k = \nu_k/\Sigma_k$, we are ready to specify

$$\epsilon_{02} \equiv \left[tr\left(V_{k-1}P_k\right) + \hat{d}_k\zeta\bar{r}_k'V_{k-1}\bar{r}_k + z_k\right]/\left(1+\mathring{\theta}\right). \tag{55}$$

If $f_{1,\sigma}(\theta_{k+1}, d_{k+1})$ (the case $\kappa \equiv \sigma$) is instated into $\mathcal{K}(\varphi)$ (47), the other variables are represented in Table I by

$$\begin{cases} \epsilon_{12} \equiv \left[tr\left(\underbrace{(V_{k-1}-U_{k-1})}_{\mathcal{X}_{k-1}}P_k\right) + \hat{d}_k\zeta\bar{r}_k'\mathcal{X}_{k-1}\bar{r}_k\right]/\mathring{\theta}_a, \\ \epsilon_{01} \equiv \left[tr\left(U_{k-1}P_k\right) + \hat{d}_k\zeta\bar{r}_k'U_{k-1}\bar{r}_k + z_k\right]/(1+\mathring{\theta}_u), \end{cases} \tag{56}$$

where $U_{k-1} \equiv \begin{bmatrix} 0 & 0 \\ 0 & P_{uu;k-1}^{-1} \end{bmatrix} \in \mathbb{R}^{\mathring{\theta}\times\mathring{\theta}}$. Otherwise, if $\kappa \equiv \rho$ is the selected option, we operate with the pair

$$\begin{cases} \epsilon_{12} \equiv \left[tr\left(A_{k-1}P_k\right) + \hat{d}_k\zeta\bar{r}_k'A_{k-1}\bar{r}_k + z_k\right]/(1+\mathring{\theta}_a), \\ \epsilon_{01} \equiv \left[tr\left(\underbrace{\mathcal{Y}_{k-1}}_{(V_{k-1}-A_{k-1})}P_k\right) + \hat{d}_k\zeta\bar{r}_k'\mathcal{Y}_{k-1}\bar{r}_k\right]/\mathring{\theta}_u, \end{cases} \tag{57}$$

where $A_{k-1} \equiv \begin{bmatrix} P_{aa;k-1}^{-1} & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{\mathring{\theta}\times\mathring{\theta}}$. This completes the derivation for the data-informed tuning of the matrix factors of the claimed forms.

IV. ESTIMATING THE HAMMERSTEIN MODEL BASED ON THE VB METHOD

The VB method is employed to restore the tractability of the inference problem via approximating the posterior pdf $f(\Theta_k|\mathcal{S}_k)$ by the product of conditionally independent posteriors. In this article, the target pdf $\breve{f}(\Theta_k|\mathcal{S}_k)$ is restricted to the product of the marginal pdfs for $\theta_{L;k} \equiv [\theta_{a;k}', \theta_{b;k}']' \in \mathbb{R}^{n_a+n_b+1}$ and $\{\theta_{r;k}, d_k\}$:

$$\breve{f}\left(\Theta_k|\mathcal{S}_k\right) \equiv f\left(\theta_{L;k}|\mathcal{S}_k\right)\breve{f}\left(\theta_{r;k}, d_k|\mathcal{S}_k\right). \tag{58}$$

Under the constraint assumption, $r_{1;k} \equiv 1$, the first factor on the right-hand side of (58) is recognized to be the exact marginal

$$f\left(\theta_{L;k}|\mathcal{S}_k\right) = \mathcal{T}\left(\theta_{L;k}|\hat{\theta}_{L;k}, \chi_{\theta_L}, \nu_k\right) \tag{59}$$

$$\propto \left[1 + \left(\theta_{L;k} - \hat{\theta}_{L;k}\right)'\chi_{\theta_L}^{-1}\left(\theta_{L;k} - \hat{\theta}_{L;k}\right)\right]^{-(\nu_k+\mathring{\theta}_L)/2}$$

with $\chi_{\theta_L} \equiv \Sigma_k P_{L;k}$, designated by the Student's t ($\mathcal{T}$) pdf. Let us note that the estimate $\hat{\theta}_{L;k}$ represents the first $\mathring{\theta}_L$ rows of $\hat{\theta}_k$ (21), and $P_{L;k}$ is nested in the first $\mathring{\theta}_L$ rows and columns of P_k (22). Since the exact marginals describing $\theta_{L;k}|\mathcal{S}_k$ and $d_k|\mathcal{S}_k$ are accessible, it only remains to infer $\theta_{r;k}$ given $\{\mathcal{S}_k, d_k\}$ by eliminating the redundancies in $\hat{\theta}_{u;k}$. To formalize the concept covered above, a loss functional quantifying the information loss incurred when moving from $f(\Theta_k|\mathcal{S}_k)$ to $\breve{f}(\Theta_k|\mathcal{S}_k)$ is constructed and optimized within the calculus of variations approach to yield $\hat{f}(\theta_{r;k}|\mathcal{S}_k, d_k)$. The loss functional takes the form

$$\begin{aligned} \mathcal{L}_{\mathcal{V}}\left(\breve{f}\left(\Theta_k|\mathcal{S}_k\right)\right) \equiv{}& \mathcal{D}\left(\breve{f}\left(\Theta_k|\mathcal{S}_k\right)\|f\left(\Theta_k|\mathcal{S}_k\right)\right) \\ &+ \eta_\iota\left(\int_{d^*}\int_{\theta_r^*}\breve{f}\left(\theta_{r;k}, d_k|\mathcal{S}_k\right)\mathrm{d}\theta_{r;k}\,\mathrm{d}d_k - 1\right) \\ &+ \eta_l\left(\int_{d^*}\int_{\theta_r^*}(l'\theta_{r;k} - 1)\breve{f}\left(\theta_{r;k}, d_k|\mathcal{S}_k\right)\mathrm{d}\theta_{r;k}\,\mathrm{d}d_k\right). \end{aligned} \tag{60}$$

The expressions in (60) scaled by the Lagrange multipliers η_ι and η_l activate the normalization and mean value hard equality constraints, respectively. The normalization constraint forces the VB-marginal $\hat{f}(\theta_{r;k}, d_k|\mathcal{S}_k)$ to be a proper pdf. By introducing an $\mathring{\theta}_r-$dimensional vector of the form $l \equiv [1, 0, \ldots, 0]'$, the mean value constraint rigorously sets the estimate of $r_{1;k}$ to equal one. The conditions for the global optimality of $\hat{f}(\theta_{r;k}, d_k|\mathcal{S}_k)$ are the statements reported by the lemma below:

Lemma 3: Let $\breve{f}(\Theta_k|\mathcal{S}_k)$ be established as an approximation of $f(\Theta_k|\mathcal{S}_k)$, with the restriction that the factorization $\breve{f}(\Theta_k|\mathcal{S}_k) = f(\theta_{L;k}|\mathcal{S}_k)\breve{f}(\theta_{r;k}, d_k|\mathcal{S}_k)$ is the product of the independent marginals. Then, having a fixed functional form for the factor $f(\theta_{L;k}|\mathcal{S}_k) = \mathcal{T}(\theta_{L;k}|\hat{\theta}_{L;k}, \chi_{\theta_L}, \nu_k)$, the unique minimum of (60) is reached for

$$\begin{aligned} \hat{f}\left(\theta_{r;k}, d_k|\mathcal{S}_k\right) = {}& \hat{f}_\xi\left(\theta_{r;k}, d_k|\mathcal{S}_k\right) \\ &\times \exp\left[-\left(1 + \eta_\iota + \eta_l\left(l'\theta_{r;k} - 1\right)\right)\right], \end{aligned} \tag{61}$$

where $\hat{f}_\xi(\theta_{r;k}, d_k|\mathcal{S}_k)$ denotes the VB-marginal that is unconstrained in the mean value of the estimand $\theta_{r;k}$

$$\hat{f}_\xi\left(\theta_{r;k}, d_k\middle|\mathcal{S}_k\right) \propto \exp\left[\mathcal{E}_{\mathcal{T}(\theta_{L;k}|\hat{\theta}_{L;k},\chi_{\theta_L},\nu_k)}\left[\ln\left(f\left(\Theta_k, \mathcal{S}_k\right)\right)\right]\right]. \tag{62}$$

The multiplier η_l is obtained by solving the integral equation

$$\int_{d^*}\int_{\theta_r^*} l'\theta_{r;k}\hat{f}_\xi\left(\theta_{r;k}, d_k\middle|\mathcal{S}_k\right)\exp\left[-\eta_l l'\theta_{r;k}\right]\mathrm{d}\theta_{r;k}\,\mathrm{d}d_k$$

$$= \int_{d^*}\int_{\theta_r^*} \hat{f}_\xi\left(\theta_{r;k}, d_k\middle|\mathcal{S}_k\right)\exp\left[-\eta_l l'\theta_{r;k}\right]\mathrm{d}\theta_{r;k}\,\mathrm{d}d_k \tag{63}$$

and the constant $\exp[1+\eta_\iota]$ independent of Θ_k substitutes

$$\exp\left[1+\eta_\iota\right]$$

$$= \int_{d^*}\int_{\theta_r^*} \hat{f}_\xi\left(\theta_{r;k}, d_k\middle|\mathcal{S}_k\right)\exp\left[-\eta_l\left(l'\theta_{r;k}-1\right)\right]\mathrm{d}\theta_{r;k}\,\mathrm{d}d_k. \tag{64}$$

Proof: It proves convenient to rewrite the KLD entering (60) into a sum of two parts

$$\mathcal{D}\left(\breve{f}\left(\Theta_k\middle|\mathcal{S}_k\right)\middle\|f\left(\Theta_k\middle|\mathcal{S}_k\right)\right) = \min_{\breve{f}(\theta_{r;k},d_k|\mathcal{S}_k)} \mathcal{L}_{\mathcal{V}}\left(\breve{f}\left(\Theta_k\middle|\mathcal{S}_k\right)\right)$$

$$+\mathcal{D}\left(\breve{f}\left(\theta_{r;k}, d_k\middle|\mathcal{S}_k\right)\middle\|\hat{f}_\xi\left(\theta_{r;k}, d_k\middle|\mathcal{S}_k\right)\right), \tag{65}$$

where the part independent of the optimized $\breve{f}(\theta_{r;k}, d_k|\mathcal{S}_k)$,

$$\int_{\theta_L^*} f\left(\theta_{L;k}\middle|\mathcal{S}_k\right)\ln\left(f\left(\theta_{L;k}\middle|\mathcal{S}_k\right)\right)\mathrm{d}\theta_{L;k} + \ln\left(\int_{\Theta^*} f_{\Theta\mathcal{S}}\mathrm{d}\Theta_k\right)$$

$$-\ln\left(\int_{d^*}\int_{\theta_r^*}\exp\left[\mathcal{E}_{\mathcal{T}(\theta_{L;k}|\hat{\theta}_{L;k},\chi_{\theta_L},\nu_k)}\left[\ln\left(f_{\Theta\mathcal{S}}\right)\right]\right]\mathrm{d}\theta_{r;k}\mathrm{d}d_k\right)$$

with $f_{\Theta\mathcal{S}} \equiv f(\Theta_k, \mathcal{S}_k)$, is the minimum attained by the functional optimization. Bearing in mind the arrangement (65), the results (61), (63), and (64) are directly obtained by applying the optimality conditions designated by $\frac{\delta\mathcal{L}_{\mathcal{V}}}{\delta\breve{f}(\theta_{r;k},d_k|\mathcal{S}_k)} = \ln(\breve{f}(\theta_{r;k}, d_k|\mathcal{S}_k)/(\mathcal{C}\hat{f}_\xi))$, $\mathcal{C} \equiv \exp[-(1+\eta_\iota+\eta_l(l'\theta_{r;k}-1))]$, $\frac{\partial\mathcal{L}_{\mathcal{V}}}{\partial\eta_\iota} = 0$, and $\frac{\partial\mathcal{L}_{\mathcal{V}}}{\partial\eta_l} = 0$. The uniqueness of the solution (61) confirms $\frac{\delta^2\mathcal{L}_{\mathcal{V}}}{\delta\breve{f}^2(\theta_{r;k},d_k|\mathcal{S}_k)} > 0$. ■

Now, with Lemma 3, we are in the position to specify the pdf $\hat{f}(\theta_{r;k}, d_k|\mathcal{S}_k)$. To this end, let us introduce the auxiliary variables

$$\mathcal{Z} \equiv I_{\mathring{\theta}_r} \otimes \mathbf{1}_{\mathring{\theta}_b}, \quad \mathcal{U} \equiv \mathbf{1}_{\mathring{\theta}_r} \otimes I_{\mathring{\theta}_b} \tag{66}$$

and assume that $P_{L;k}$ is partitioned in compliance with the partitioning of $\theta_{L;k} \equiv [\theta'_{a;k}, \theta'_{b;k}]' \in \mathbb{R}^{\mathring{\theta}_L}$, as follows:

$$P_{L;k} \equiv \begin{bmatrix} P_{aa;k} & P'_{ba;k} \\ P_{ba;k} & P_{bb;k} \end{bmatrix} \in \mathbb{R}^{\mathring{\theta}_L\times\mathring{\theta}_L}, \tag{67}$$

where $P_{aa;k} \in \mathbb{R}^{\mathring{\theta}_a\times\mathring{\theta}_a}$. After some algebra (see Appendix A), the optimal approximation is found to be

$$\hat{f}\left(\theta_{r;k}\middle|\mathcal{S}_k, d_k\right) = \mathcal{N}\left(\theta_{r;k}\middle|\hat{\theta}_{r;k}, P_{r;k}/d_k\right), \tag{68}$$

$$\hat{\theta}_{r;k} = \hat{\xi}_k - P_{r;k}l\left(l'\hat{\xi}_k - 1\right)/(l'P_{r;k}l), \tag{69}$$

where $\hat{\xi}_k$ is the mean value of $\theta_{r;k}$, on which the equality constraint is not imposed. The normalized (by d_k) covariance matrix $P_{r;k}$ and the mean $\hat{\xi}_k$ are calculated as

$$P_{r;k} = \left(\mathcal{Z}'\left(V_{uu;k} \circ \left[\mathcal{U}\left(\hat{\theta}_{b;k}\hat{\theta}'_{b;k} + P_{bb;k}\frac{\Sigma_k}{\nu_k-2}\right)\mathcal{U}'\right]\right)\mathcal{Z}\right)^{-1}, \tag{70}$$

$$\hat{\xi}_k = P_{r;k}\mathcal{Z}'\Bigg[\left([V_{ua;k}, \quad V_{uu;k}]\hat{\theta}_k\right) \circ \left(\mathcal{U}\hat{\theta}_{b;k}\right)$$

$$-\left(V_{ua;k} \circ \left(\mathcal{U}\left(\hat{\theta}_{b;k}\hat{\theta}'_{a;k} + P_{ba;k}\frac{\Sigma_k}{\nu_k-2}\right)\right)\right)\mathbf{1}_{\mathring{\theta}_a}\Bigg]. \tag{71}$$

Algorithm 1: The VB Inference-based Estimation Procedure for the Time-Varying Hammerstein System.

Initialization phase:

1. Make the assignment $\hat{\varphi}_0 \equiv 1$, $\hat{\varphi}_1 \equiv 0$, and $\hat{\varphi}_2 \equiv 0$ to obtain $\Lambda_0 = \Omega_0 = I_{\mathring{\theta}}$ and $\lambda_0 = 1$.
2. Set the statistics $\{\hat{\theta}_0, \Xi, \Sigma_0 > 0, \nu_0 > 2\}$ and make the assignment $\{\hat{\theta}_{-1} \equiv \hat{\theta}_0, V_0 \equiv P_0^{-1} \equiv \Xi\}$ to ultimately obtain, for $k = 1$, the starting point $\{V_{c;0}, P_{c;0}, \hat{\theta}_{c;0}, \Sigma_{c;0}, \nu_0\}$ needed to initiate the data update (19)–(25).
3. Set the upper bound on the parameter uncertainty increase $\alpha \in (0,1)$, choose the heuristic factor $\zeta \in (0,1]$, and select the forgetting strategy to be applied $\kappa \in \{\sigma, \rho\}$.
4. Collect consecutive data to form the regressor h_1 (4).

Learning phase:

5. **while** $k \leftarrow 1, \mathring{k}$ **do**
   - **input :** $\begin{cases} y_k, h_k, \Xi, \Sigma_{k-1}, \nu_{k-1}, \sigma \text{ or } \rho, \\ \hat{\theta}_{k-1}, \hat{\theta}_0, V_{k-1}, P_{k-1}, \Lambda_{k-1}, \Omega_{k-1}, \lambda_{k-1} \end{cases}$
6. Update ▷ (14)–(25) $\{\hat{\theta}_{k-1}, V_{k-1}, \Sigma_{k-1}, \nu_{k-1}\} \rightarrow \{\hat{\theta}_k, V_k, \Sigma_k, \nu_k\}$
7. Calculate ϵ_{02} ▷ (55)
8. Calculate $\{\epsilon_{12}, \epsilon_{01}\}$ ▷ (56) **or** (57)
9. Evaluate $\hat{\varphi}$ ▷ Table I
10. Calculate $\{\lambda_k, \Lambda_k, \Omega_k\}$ ▷ (53) **or** (54)
11. Select $\hat{\theta}_{L;k}$ and $P_{L;k}$ nested in $\hat{\theta}_k$ and P_k, respectively, to establish $\mathcal{N}(\theta_{L;k}|\hat{\theta}_{L;k}, P_{L;k}/d_k)$.
12. Calculate $P_{r;k}$ (70), $\hat{\xi}_k$ (71), and subsequently $\hat{\theta}_{r;k}$ (69) to parameterize $\mathcal{N}(\theta_{r;k}|\hat{\theta}_{r;k}, P_{r;k}/d_k)$.
13. Use $\mathcal{W}(d_k|\Sigma_k, \nu_k)$ to describe d_k. ▷ (9)
14. **end**

To round out the implementation issues, the computation procedures for the proposed estimator are summarized by Algorithm 1.

V. Simulation Studies

This section presents several numerical examples to illustrate the behavior of the developed methods. To show its advantages, the suggested VB strategy to identify the Hammerstein system is compared with the simple averaging (AV) approach [7]. To demonstrate the main feature of the matrix forgetting, the change localization capabilities are tested.

The Hammerstein system with a polynomial input nonlinearity is estimated

$$y_k + \sum_{i=1}^{2} a_i y_{k-i} = \sum_{i=0}^{2}\sum_{j=1}^{3} b_i r_j u_{k-i}^j + e_k,\ e_k \sim \mathcal{N}(0, 1/d).$$

The parameters of the linear filter $\{a_i, b_i\}$ are chosen to correspond to the discretized transfer function $\mathcal{G}(s) = 1.5(-0.5s+1)(s+1)/(T^2s^2 + 2\xi_g Ts + 1)$ sampled with the period of 1 s, where $\xi_g = 0.6$, and T is set as $T = 5$ s. The polynomial coefficients, if not stated otherwise, are specified as $r_1 = 1$, $r_2 = -0.4$, and $r_3 = \frac{1}{20}$. The input sequence $\{u_k\}$ is produced by the autoregressive model $u_k = 0.9u_{k-1} + w_k$ driven by a discrete white noise, $w_k \sim \mathcal{N}(0,1)$. All the experiments are monitored within the time span of $0 - 500$ s. The fit between the model and its estimate is evaluated through the measure

$$\Delta_k \equiv \sqrt{\bar{\Delta}_k/\left(\|\theta_{a;k}\|_2^2 + \|\theta_{b;k}/\beta\|_2^2 + \|\beta\theta_{r;k}\|_2^2\right)}, \tag{72}$$

where $\bar{\Delta}_k \equiv \left\|\theta_{a;k} - \hat{\theta}_{a;k}\right\|_2^2 + \left\|\frac{1}{\beta}\theta_{b;k} - \hat{\theta}_{b;k}\right\|_2^2 + \left\|\beta\theta_{r;k} - \hat{\theta}_{r;k}\right\|_2^2$ and $\beta \equiv \frac{1}{r_{1;k}}$. In view of the user-defined input arguments to Algorithm 1, the estimation starts from $\Xi = 10^{-1}I_{11}$, $\Sigma_0 = 1$, $\nu_0 = 10$, and $\hat{\theta}_0$ is chosen to be an $11-$D vector, all of whose components are zero.

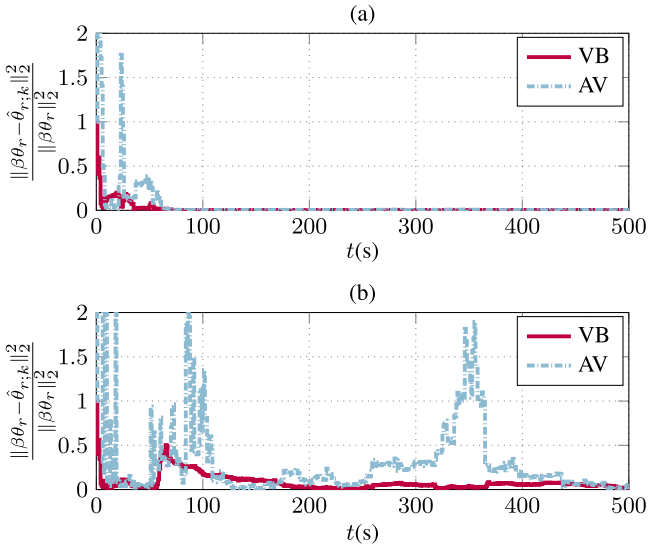

Fig. 1. Comparison of the coefficient product decoupling strategies.

A. Estimation Quality Achieved With the Recursive Strategies

The following example compares the proposed solution (VB) with the averaging strategy (AV) to decouple the coefficient products between the linear and the nonlinear subsystems paired via the least squares estimate $\hat{\theta}_k$. Similarly to our solution, the AV strategy builds on the assumption that the first polynomial coefficient is set as one ($r_1 = 1$) but it eliminates the redundancies by computing their average value

$$\hat{\theta}_{r;k}[j] = \frac{1}{\mathring{\theta}_b}\sum_{i=1}^{\mathring{\theta}_b}\hat{\theta}_k\left[\mathring{\theta}_a + (j-1)\mathring{\theta}_b + i\right]/\hat{\theta}_{b;k}[i], \tag{73}$$

where $j = 2, \ldots, \mathring{\theta}_r$ and $\theta[i]$ denotes the ith entry of the vector. Note that the forgetting operation is not contemplated in this example ($\Lambda_k \equiv \Omega_k \equiv I_{11}$, $\lambda_k \equiv 1$). The results attained for the decoupling strategies are shown in Fig. 1. The impact of the model noise precision $d \in \{10^4, 10\}$ on the identification quality, judged by $\frac{\|\beta\theta_r-\hat{\theta}_{r;k}\|_2^2}{\|\beta\theta_r\|_2^2}$, is tested. From Fig. 1(a), we can derive that both of the strategies exhibit similar steady-state responses at the low noise level ($d = 10^4$). If we increase the noise intensity to $d = 10$ [see Fig. 1(b)], the proposed VB strategy provides more robust, uncertainty-aware averaging than that given in (73).

B. Tracking Quality Achieved With the Forgetting Strategies

The simulation runs are performed with either the least squares with stabilized matrix forgetting proposed herein or the standard exponential least squares endowed with the adaptation rule by Ydstie and Sargent [10],

$$\lambda_{k-1} = \left(\bar{\eta}_k + \sqrt{\bar{\eta}_k^2 + 4h_k' P_{k-1} h_k}\right)/2, \tag{74}$$

where $\bar{\eta}_k \equiv 1 - h_k' P_{k-1} h_k - \frac{\nu_{k-1}(y_k - h_k'\hat{\theta}_{k-1})^2}{\Sigma_{k-1} Q}$ with the tuning knob Q representing the effective number of degrees of freedom.

The experiment that follows is undertaken to show the advantages of matrix forgetting when only the polynomial curve coefficients are subject to change. The initial coefficient values satisfy $\theta_{r;0} = [1, -0.2, 0.1]$ and there is a sudden change at the time $t = 250\,\mathrm{s}$, which results in $\theta_{r;250} = [0.4, -0.4, \frac{1}{8}]$. The matrix forgetting strategies employing $f_{1,\sigma}$ and $f_{1,\rho}$ will be labeled as MF_σ and MF_ρ, respectively. The scalar forgetting driven by the rule (74) is referred to as SF. For this scenario, the settings $\{\alpha = 0.5, \zeta = 0.5\}$ are assigned to both the matrix forgetting strategies, and SF is in operation with $Q = 20$. The information matrix is initiated from $\Xi = 10^{-2} I_{11}$, and the noise precision d is assigned as 10^3.

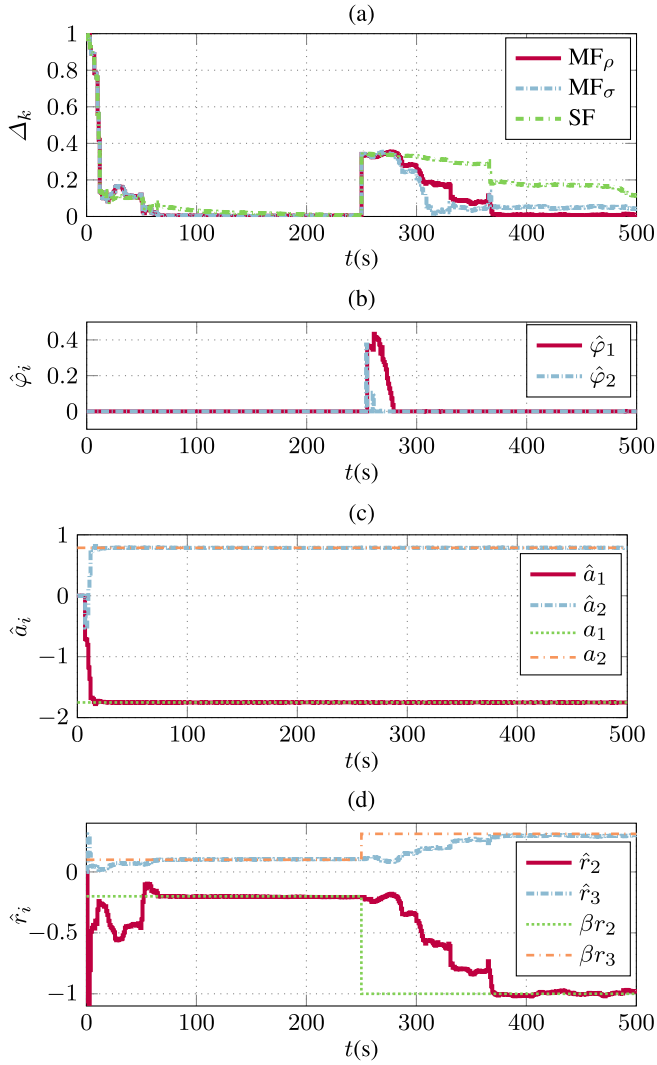

Fig. 2. (a) Estimation quality obtained with the three algorithms; (b) the time courses of $\{\hat{\varphi}_1, \hat{\varphi}_2\}$ optimized for MF_ρ; and the corresponding trajectories of the estimates of $\theta_{a;k}$ (c) and $\theta_{r;k}$ (d).

The result of the experiment is captured in Fig. 2. The MF_ρ technique attains the smaller Δ_k [see Fig. 2(a)] than the other forgetting strategies, as it effectively exploits the probability $\hat{\varphi}_1$ [see Fig. 2(b)] indicating the source of inconsistency. Note that $\hat{\varphi}_2$ also receives some support [see Fig. 2(b)] since the changing parameters are a subset of all the parameters. Importantly, the MF_ρ algorithm has exhibited its ability to selectively track the changing parameters [see Fig. 2(d)], leaving the estimate of the filter parameters $\theta_{a;k}$ unaffected [see Fig. 2(c)].

VI. Conclusion

The main aim of this article have been to resolve the problems of selective forgetting along with the recovering of the Hammerstein model parameters (lost in the overparameterization step) within a rigorous probabilistic framework.

The novelty of the research rests in theoretical results leading to modifications of the recursive least squares method. Remarks 2 and 3 present two partial forgetting strategies: While the former Remark introduces the matrix factor that discounts the information submatrix [17], the latter one proposes a concept modifying the covariance submatrix. This modification is essentially justified by the Kalman filtering-based estimation view, where only a subset of the parameters is time-varying, driven by a random walk with a known covariance matrix of parameter increments. Lemmas 1 and 2 (elaborate on the results stated in [12]) allow us to derive a novel and formal approach for dealing with automated selective forgetting based on the geometric mean of the pdfs. Remark 1 classifies the Hammerstein model as a member of the DEF, enabling a closed-form expression for the propagation of the sufficient statistics of the overparameterized model. Lemma 3 converts the problem of eliminating the redundancies in the least squares estimate of the overparameterized model into an optimization problem, tailoring the VB method to identify the Hammerstein systems. Consequently, the exact posterior is approximated at each step ex post, after the data update has been completed, avoiding any transmission of the VB-moments through iterative cycles.

Appendix A

To present the systematic procedure for deriving the final form of the part of the VB-marginal (68), let us express the natural parameters of the DEF (5) as

$$q\left(\theta_k, d_k\right) = q_L\left(\theta_{L;k}\right) \circ q_r\left(\theta_{r;k}, d_k\right), \tag{A.1}$$

where the particular vector functions possess the form

$$q_L\left(\theta_{L;k}\right) \equiv \operatorname{vec}\left(\phi_L \phi_L'\right), \tag{A.2}$$

$$\phi_L \equiv [\theta_{a;k}' \quad \theta_{b;k}' \mathcal{U}' \quad 1]', \tag{A.3}$$

$$q_r\left(\theta_{r;k}, d_k\right) \equiv -\frac{d_k}{2} \operatorname{vec}\left(\phi_r \phi_r'\right), \tag{A.4}$$

$$\phi_r \equiv [\mathbf{1}_{\hat{\theta}_a}' \quad \theta_{r;k}' \mathcal{Z}' \quad 1]'. \tag{A.5}$$

The result (A.1) can be verified from (5) by virtue of the identities $\operatorname{vec}(xx') = x \otimes x$ and $(E \circ C) \otimes (B \circ D) = (E \otimes B) \circ (C \otimes D)$, which are proven in [22]. With the augmented information matrix $\bar{V}_k$ given in (10) and the relations

$$\phi_L = \underbrace{\begin{bmatrix} I_{\hat{\theta}_a} & 0 & 0 \\ 0 & \mathcal{U} & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\Upsilon_L} \begin{bmatrix} \theta_{a;k} \\ \theta_{b;k} \\ 1 \end{bmatrix}, \quad \phi_r = \underbrace{\begin{bmatrix} 0 & \mathbf{1}_{\hat{\theta}_a} \\ \mathcal{Z} & 0 \\ 0 & 1 \end{bmatrix}}_{\Upsilon_r} \begin{bmatrix} \theta_{r;k} \\ 1 \end{bmatrix},$$

the term in the exponent of the joint pdf $f(\Theta_k, \mathcal{S}_k)$ in (62) becomes

$$\begin{aligned} &[\theta_k' \quad 1] \bar{V}_k [\theta_k' \quad 1]' = \operatorname{vec}\left(\phi_r \phi_r'\right)' \operatorname{vec}\left(\bar{V}_k \circ \left(\phi_L \phi_L'\right)\right) \\ &= \begin{bmatrix} \theta_{r;k} \\ 1 \end{bmatrix}' \Upsilon_r' \left(\bar{V}_k \circ \left(\Upsilon_L \begin{bmatrix} \theta_{L;k} \theta_{L;k}' & \theta_{L;k} \\ \theta_{L;k}' & 1 \end{bmatrix} \Upsilon_L' \right) \right) \Upsilon_r \begin{bmatrix} \theta_{r;k} \\ 1 \end{bmatrix}. \end{aligned} \tag{A.6}$$

By employing (A.6), evaluating the expectation, and subsequently completing the square in (62), the VB-marginal unconstrained in the mean value shows as

$$\begin{aligned} \hat{f}_\xi\left(\theta_{r;k}, d_k \middle| \mathcal{S}_k\right) &\propto \exp\left[-\Sigma_{\xi;k} d_k/2\right] d_k^{\left(\nu_k + \mathring{\theta}_r - 2\right)/2} \\ &\times \exp\left[-\left(\theta_{r;k} - \hat{\xi}_k\right)' P_{r;k}^{-1} \left(\theta_{r;k} - \hat{\xi}_k\right) d_k/2\right], \end{aligned} \tag{A.7}$$

where $\Sigma_{\xi;k}$ is the least squares reminder associated with the quadratic form (A.6). For the solution to respect the equality constraint imposed on $\theta_{r;k}$, it is necessary to acquire the analytical definitions for η_l and $\exp[1 + \eta_\iota]$. We found η_l by solving (63), as follows:

$$\eta_l = d_k \left(l' \hat{\xi}_k - 1\right) / \left(l' P_{r;k} l\right). \tag{A.8}$$

Calculating the integral (64) with $\hat{f}_\xi(\theta_{r;k}, d_k | \mathcal{S}_k) = \mathcal{N}(\theta_{r;k} | \hat{\xi}_k, P_{r;k}/d_k) \mathcal{W}(d_k | \Sigma_{\xi;k}, \nu_k)$ (A.7) identifies $\exp[1 + \eta_\iota]$, as

$$\exp\left[1 + \eta_\iota\right] = \left(\frac{\Sigma_{\xi;k}}{\Sigma_{\xi;k} + \left(l' \hat{\xi}_k - 1\right)^2 / \left(l' P_{r;k} l\right)} \right)^{\nu_k/2}. \tag{A.9}$$

Combining (A.7), (A.8), and (A.9) into (61) generates the form of the constrained VB-marginal in congruence with (68), which concludes our derivation.

References

[1] F. R. Ávila, H. T. Carvalho, and L. W. P. Biscainho, "Bayesian blind identification of nonlinear distortion with memory for audio applications," *IEEE Signal Process. Lett.*, vol. 23, no. 4, pp. 414–418, Apr. 2016.

[2] S. Chandra, M. Hayashibe, and A. Thondiyath, "Muscle fatigue induced hand tremor clustering in dynamic laparoscopic manipulation," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 50, no. 12, pp. 5420–5431, Dec. 2020.

[3] M. Schoukens and K. Tiels, "Identification of block-oriented nonlinear systems starting from linear approximations: A survey," *Automatica*, vol. 85, pp. 272–292, Nov. 2017.

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

[5] W. R. Jacobs, T. Baldacchino, T. Dodd, and S. R. Anderson, "Sparse Bayesian nonlinear system identification using variational inference," *IEEE Trans. Autom. Control*, vol. 63, no. 12, pp. 4172–4187, Dec. 2018.

[6] S. Särkkä and A. Nummenmaa, "Recursive noise adaptive Kalman filtering by variational Bayesian approximations," *IEEE Trans. Autom. Control*, vol. 54, no. 3, pp. 596–600, Mar. 2009.

[7] J. Ma, B. Huang, and F. Ding, "Iterative identification of Hammerstein parameter varying systems with parameter uncertainties based on the variational Bayesian approach," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 3, pp. 1035–1045, Mar. 2020.

[8] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[9] M. Kárný and J. Andrýsek, "Use of Kullback–Leibler divergence for forgetting," *Int. J. Adapt. Control Signal Process.*, vol. 23, no. 10, pp. 961–975, Oct. 2009.

[10] B. E. Ydstie and R. W. H. Sargent, "Convergence and stability properties of an adaptive regulator with variable forgetting factor," *Automatica*, vol. 22, no. 6, pp. 749–751, Nov. 1986.

[11] R. Kulhavý, "Restricted exponential forgetting in real-time identification," *Automatica*, vol. 23, no. 5, pp. 589–600, Sep. 1987.

[12] 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.

[13] J. Dokoupil and P. Václavek, "Regularized estimation with variable exponential forgetting," in *Proc. 59th IEEE Conf. Decis. Control*, 2020, pp. 312–318.

[14] A. S. Poznyak and J. J. Medel Juarez, "Matrix forgetting factor with adaptation," *Int. J. Syst. Sci.*, vol. 30, no. 8, pp. 865–878, 1999.

[15] J. Li, Y. Zheng, and Z. Lin, "Recursive identification of time-varying systems: Self-tuning and matrix RLS algorithms," *Syst. Control Lett.*, vol. 66, no. 1, pp. 104–110, Apr. 2014.

[16] A. L. Bruce, A. Goel, and D. S. Bernstein, "Recursive least squares with matrix forgetting," in *Proc. Amer. Control Conf.*, 2020, pp. 1406–1410.

[17] K. Dedecius, I. Nagy, and M. Kárný, "Parameter tracking with partial forgetting method," *Int. J. Adapt. Control Signal Process.*, vol. 26, no. 1, pp. 1–12, Jan. 2012.

[18] V. Peterka, "Bayesian approach to system identification," in *Trends and Progress in System Identification*. P. Eykhoff, Ed. Oxford, U.K.: Pergamon, 1981, pp. 239–304.

[19] J. Dokoupil and P. Václavek, "Forgetting factor Kalman filter with dependent noise processes," in *Proc. 58th IEEE Conf. Decis. Control*, 2019, pp. 1809–1815.

[20] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[21] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.

[22] 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.