Total Least Squares from a Bayesian Perspective: Incorporating Data-Informed Forgetting

DOKOUPIL, J.; VÁCLAVEK, P.

2024 IEEE 63rd Conference on Decision and Control (CDC)

pages: 5737-5744

eISBN: 979-8-3503-1633-9

DOI: https://doi.org/10.1109/CDC56724.2024.10885920

Accepted manuscript

Total Least Squares from a Bayesian Perspective: Incorporating Data-Informed Forgetting

Jakub Dokoupil and Pavel Václavek

Abstract— **The real-time estimation of error-in-variables (EIV) models with unknown time-varying parameters is considered and resolved using a Bayesian framework. The stochastic model under consideration is a regression-type model that accounts for inherently inaccurate measurements, which are corrupted by the normal noise. The EIV model identification is traditionally performed via total least squares (TLS), relying on computationally intensive methods to numerically obtain a point estimate. Such a concept, despite its theoretical appeal, nevertheless lacks the ability to quantify the uncertainty associated with the parameter estimates. Thus, this limitation hinders the concept from being combined with the statistical decision-making strategies. The paper opens the way towards enriching the standard TLS in this respect. The enrichment is achieved by projecting the unnormalized posterior generated by the EIV parametric models onto the normal-Wishart distribution. This projection is made optimal by minimizing the Kullback-Leibler distance between the unnormalized and the normal-Wishart posteriors while imposing a hard equality constraint on the mean parameter scalar product. By establishing credible intervals for both the regression parameters and the noise precision, the resultant procedure is additionally endowed with Bayesian data-informed forgetting, which allows for effective operation in nonstationary environments.**

I. Introduction

Real-world data are widely acknowledged to be contaminated by errors, and the presence of measurement errors can substantially affect the outcomes of analyses. The parametric identification of equation-error models is well documented in the literature, with numerous solutions proposed from both the deterministic [1] and the Bayesian [2] perspectives. Identifying system parameters from noise-corrupted measurements at both the output and the input is regarded as an equally relevant but more intricate problem. The challenge arises from the uncertainty inherent in the accessible data, leading to the treatment of noise-free data as hidden variables. Representations where the inputs and the outputs are observed on a system with measurement errors are usually referred to as error-in-variables (EIV) models. Regrettably, the literature available on relevant Bayesian solutions is rather scarce. Due to the absence of a Bayesian problem formulation, the integration of EIV models with the existing strategies for the Bayesian nonlinear identification [3], [4], inclusion of prior information, probabilistic control design [5], transferring knowledge between model hypotheses [6], and risk assessment for data-informed decision-making [2] becomes unattainable. The paper aims to fill this gap, providing a constructive solution to address this important issue.

The application of EIV models can be motivated by the effort to elucidate the underlying system dynamics, namely, the relationship between the noise-free input and noise-free output. The standard deterministic approaches to identify the EIV models adjust the least squares (LS) solutions to avoid biased estimates. These procedures necessitate some form of preprocessing to compress the data into sufficient statistics, which subsequently facilitate the final computation of the parameter estimates. The particular methods then differ in the way the data compression is executed. Computationally relatively simple methods exploit instrumental variables that are uncorrelated with disturbances but highly correlated with data regressors. In [7], the concept was further enhanced by being combined with a weighted subspace fitting approach. An alternative strategy involves determining the relationship between the biased LS estimate and its unbiased counterpart and retrieving the desired estimate by solving a set of bilinear equations [8]. An attempt to face the combined equation-error and EIV model using the bias-compensated LS strategy is presented in [9]. The solution to an overdetermined set of linear equations where both the right-hand side and the coefficient matrix are subject to errors is fundamentally provided by total least squares (TLS). The effects of applying TLS as opposed to LS, their algebraic connections, and the computational sensitivity aspects are discussed in [10]. Utilizing TLS also enables us to solve the problem of jointly identifying the hidden variables and the parameter vector. In the context of the regression-type models, it becomes imperative to consider the couplings between the various elements in the coefficient matrix and the right-hand side outputs. In these cases, where the coefficient matrix may exhibit either Toeplitz or Hankel characteristics, one can employ a structured TLS that takes the couplings into account (see workshop proceedings [11]). The survey paper by Söderström [12] offers a comprehensive overview of the diverse approaches to EIV model identification, incorporating an analysis of identifiability based on the spectral factorization of the input-output data.

Considerable effort has been dedicated to advancing the maximum likelihood (ML) methods in the domain of EIV models, with the findings pointing to the provision of unbi-

This work was supported in part by the Czech Science Foundation under Grant 23-06476S, in part by the European Union through the project Robotics and Advanced Industrial Production under Grant No. CZ.02.01.01/00/22_008/0004590, in part by the infrastructure of RICAIP that has received funding from the European Union's Horizon 2020 research and innovation program 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).

ased parameter estimates via an appropriately deregularized LS solution. These findings confirm a coincidence between the minimization of the TLS criterion and the maximization of the likelihood function formulated for the EIV model [13], [14]. An alternative to relying on the deregularized LS is to resort to the expectation-maximization [13] or quasi linear LS [15] algorithms, which iteratively converge to a stationary point of the total ML objective. The origin of the TLS estimator, rooted in the methodological reasoning of total ML, allows us in turn to obtain the confidence bounds on the TLS estimates. Conceptually, the given interval is determined by the inverse of the Fisher information matrix, whose exact mathematical construction demands knowledge of the hidden variables and the precision of the noise terms. In practical scenarios, the necessary knowledge is typically not at our disposal. The issue is reflected on in articles [13], [14], [16], [17], the authors delving into the performance degradation inflicted by a lack of information regarding the unknown nuisance parameters. In contrast to the frequentist approach, the Bayesian methodology treats the unknown parameters as random variables and constructs the posterior probability density function (pdf) to describe these parameters. One of the earliest studies addressing the EIV model from the Bayesian perspective is outlined by Zellner in [18]. Zellner's analysis demonstrates that a simple EIV model can generate the posterior in the form of the bivariate Student's t pdf. Regrettably, this conclusion cannot be further extended to a complex, multivariable EIV model, as the normalizing integral becomes analytically intractable. Zellner also identified some basic difficulties with the EIV model analysis when all the parameters are treated as unknown. These difficulties arise particularly when the model includes parameters such as the ratio between the precisions of the noises. This finding was later corroborated by Söderström in his examination of the EIV model identifiability [12]. However, for identification purposes, the aforementioned frequentist and Bayesian approaches are not of much help, as they do not consider the couplings within the overdetermined system of equations. Moreover, these approaches are primarily tailored for batch data processing. Our aim, generally, is to expand on the attained results and to demonstrate how solving the constrained optimization problem of the fixed-form variational Bayes (FFVB) explains the TLS method as a recursive propagation of the sufficient statistics of the normal-Wishart ($\mathcal{NW}$) distribution (§8.1.3 in [2]).

Given the uncertainty associated with the EIV model parameters, the solution can be seamlessly integrated with the Bayesian tracking techniques [19]. The Bayesian compensation for the unspecified model of the parameters' variations is conceptually achieved via forgetting obsolete information. As a rule, the forgetting operation is initiated by combining the posterior with its flattened alternative(s): The more flattened the pdf, the greater the uncertainty about the parameters, and a higher variability of the parameters is expected. To optimally reassess the posterior to effectively control the relative significance of new data as compared to those already obtained, we employ the statistical decision-making approach [20]–[22].

Notation. An $n \times m$ zero matrix is defined by $O_{n,m}$; I_n stands for an $n \times n$ identity matrix; $\otimes$ refers to the Kronecker product; x' means the transpose of x; x^* symbolizes 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; $\mathrm{tr}(\cdot)$ is the trace operator; and $|\cdot|$ stands for the absolute value of the determinant. Let us define a symmetric $n \times n$ matrix X. Then, $\mathrm{vec}\,(X)$ (§1.4.1 in [23]) is an $n^2 \times 1$ vector stacking columns of X one beneath the other, while $\mathrm{vech}\,(X)$ (§1.4.3 in [23]) denotes an $\frac{1}{2}n(n+1) \times 1$ vector stacking the elements of X both on and below the main diagonal, one beneath the other. In this way, $\mathrm{vech}\,(X)$ is obtained from $\mathrm{vec}\,(X)$ by eliminating all elements above the diagonal of X to contain only the generically distinct elements of X. The transformation of $\mathrm{vech}\,(X)$ into $\mathrm{vec}\,(X)$ is achieved by using an $n^2 \times \frac{1}{2}n(n+1)$ zero-one duplication matrix D_n (§3.3.1 in [23]), such that $D_n \,\mathrm{vech}\,(X) = \mathrm{vec}\,(X)$. Further, $f(x)$ is reserved for the pdf of a random variable x, optionally distinguished by its subscript; $\propto$ means equality up to a normalizing factor; $\equiv$ signifies equality by definition; 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 Algorithm

As noted earlier, the algebraic connections between TLS and LS suggest that TLS can be approached as deregularized LS (§6 in [10]). In our derivation, the LS statistics serve to condense the observed data into finite-dimensional entities, subsequently utilized for the final computation. Before we turn to explicating the Bayesian interpretation of TLS, it is instructive to show how the observations drawn from the normal regression-type model are incorporated into the inference mechanism. Initially, we assume that no perturbations affect the data in the regressor, reducing the Bayes update to the LS update of the $\mathcal{NW}$ pdf statistics. Later, the scenario is revisited to explore the effect of the presence of disturbances in the regressor, forcing us to search for the optimal $\mathcal{NW}$ pdf within the approximate inference methods. More concretely, the search is converted into a constrained optimization problem of the FFVB. Having the best representative of the posterior in this context, the design of the data-informed forgetting is cast in the decision making framework.

A. Recursive LS from the Bayesian viewpoint

To condense the data measurable on the system, the normal regression-type model with a deterministic regressor vector is considered. The system is modeled as

$$y_k = h_k' \theta_k + e_k, \qquad e_k \sim \mathcal{N}\left(0, 1/d_k\right), \tag{1}$$

where $y_k \in \mathbb{R}$ and $u_k \in \mathbb{R}$ are an indirectly affected output and a directly manipulated input, respectively. The regressor $h_k \equiv [u_{k-1}, \ldots, u_{k-n}, -y_{k-1}, \ldots, -y_{k-n}]' \in \mathbb{R}^{2n}$ consists of lagged inputs and outputs. The noise term e_k is assumed to be normally distributed discrete white noise

with a zero mean and an unknown precision $d_k \in \mathbb{R}_{>0}$. The model is parameterized by the set of unknown regression coefficients stacked into the vector $\theta_k \in \mathbb{R}^{2n}$ and by the model noise precision d_k. The input and the output are both collected on the system at the discrete time instants $k \in k^* \equiv \{k_0, k_0+1, \dots, \mathring{k}\} \subset \mathbb{Z}$ to form the data record $\mathcal{D}_{1-n}^{k} \equiv \{u_i, y_i\}_{i=1-n}^{k}$. The lower bound in $\mathcal{D}_{1-n}^{k}$ is chosen to formally index the parametric models within the likelihood function starting from time $k=1$, that is, $\prod_{l=1}^{k} f(y_l|\theta_k, d_k, \mathcal{D}_{1-n}^{l-1})$.

To perform conditional joint inference on the parameters of interest, θ_k and d_k, the parametric model is constructed. The construction entails transforming e_k into the system output y_k, thereby resulting in

$$f\big(y_k\big|h_k, \theta_k, d_k\big) = \mathcal{N}\left(h_k'\theta_k, 1/d_k\right). \tag{2}$$

The normal pdf (2) is proportional to the exponential of a quadratic form, specified by

$$\mathcal{N}\left(h_k'\theta_k, 1/d_k\right) \propto \exp\big[-\begin{bmatrix}\theta_k' & 1\end{bmatrix}\psi_k\psi_k'\begin{bmatrix}\theta_k' & 1\end{bmatrix}' d_k/2\big],$$

with the augmented regressor vector $\psi_k' \equiv [h_k', -y_k]$. The normality assumption of the parametric model (2) establishes a conjugate prior in the form of the $\mathcal{NW}$ pdf. In light of the conjugacy principle, updating the prior with the parametric model yields a posterior whose particular factors are described as

$$\begin{aligned} f\big(\theta_k\big|\mathcal{S}_k, d_k\big) &= \mathcal{N}\big(\theta_k\big|\hat{\theta}_{ls;k}, P_{ls;k}/d_k\big) \qquad (3)\\ &\propto \exp\big[-\big(\theta_k-\hat{\theta}_{ls;k}\big)'P_{ls;k}^{-1}\big(\theta_k-\hat{\theta}_{ls;k}\big)d_k/2\big],\\ f\big(d_k\big|\mathcal{S}_k\big) &= \mathcal{W}\big(d_k\big|\Sigma_{ls;k}, \nu_{ls;k}\big) \qquad (4)\\ &\propto d_k^{(\nu_{ls;k}-2)/2}\exp\big[-\Sigma_{ls;k}d_k/2\big], \end{aligned}$$

where $\mathcal{S}_k$ denotes the sufficient statistics for $\{\theta_k, d_k\}$, which will be specified later in this paragraph. The expectations $\mathcal{E}_{\mathcal{N}(\theta_k|\mathcal{S}_k,d_k)}[\theta_k] = \hat{\theta}_{ls;k}$ and $\mathcal{E}_{\mathcal{N}(\theta_k|\mathcal{S}_k,d_k)}[(\theta_k-\hat{\theta}_{ls;k})(\theta_k-\hat{\theta}_{ls;k})'] = P_{ls;k}/d_k$ correspond with the particular moments of the multivariate normal distribution (3). The scalars $\Sigma_{ls;k} > 0$ and $\nu_{ls;k} > 2$ represent the least squares reminder and the number of degrees of freedom, respectively. As for the mean value of d_k, the Wishart distribution's definition implies that $\mathcal{E}_{\mathcal{W}(d_k|\Sigma_{ls;k},\nu_{ls;k})}[d_k] = \nu_{ls;k}/\Sigma_{ls;k}$. The sufficient statistics $\mathcal{S}_k$ are denoted by the subsets $\mathcal{S}_k \equiv \{s_k, \nu_{ls;k}\}$, where s_k conforms to the previously defined posteriors (3), (4) when assembled according to [22]

$$s_k \equiv \operatorname{vec}\Bigg(\underbrace{\begin{bmatrix} P_{ls;k}^{-1} & -P_{ls;k}^{-1}\hat{\theta}_{ls;k} \\ -\hat{\theta}_{ls;k}'P_{ls;k}^{-1} & \Sigma_{ls;k}+\hat{\theta}_{ls;k}'P_{ls;k}^{-1}\hat{\theta}_{ls;k}\end{bmatrix}}_{\bar{V}_k}\Bigg). \tag{5}$$

Consistent with the prior discussions, our analysis operates under the assumption of vague knowledge concerning the time evolution model $f(\theta_k, d_k|\theta_{k-1}, d_{k-1}, \mathcal{S}_{k-1})$, which impedes the application of the marginalization integral

$$\begin{aligned} f\big(\theta_k, d_k\big|\mathcal{S}_{k-1}\big) = \int_{d^*}\int_{\theta^*} & f\big(\theta_k, d_k\big|\theta_{k-1}, d_{k-1}, \mathcal{S}_{k-1}\big) \qquad (6)\\ & \times f\big(\theta_{k-1}, d_{k-1}\big|\mathcal{S}_{k-1}\big)\,\mathrm{d}\theta_{k-1}\,\mathrm{d}d_{k-1}. \end{aligned}$$

To address such deficiency, the covariances of $\theta_k|\mathcal{S}_k, d_k$ and $d_k|\mathcal{S}_k$ are rescaled at each iteration, effectively converting Bayesian filtering into tracking. To prevent the forgetting blow-up effect and to compensate for the potential loss of persistency during poor excitation, a minimal amount of parameter-related information is guaranteed by the externally supplied (ES) pdf

$$\begin{aligned} f\big(\theta_k, d_k\big|\hat{\theta}_{ls;k-1}, \Xi, \Sigma_{ls;0}, \nu_{ls;0}\big) = {} & \mathcal{N}\big(\theta_k\big|\hat{\theta}_{ls;k-1}, \Xi^{-1}/d_k\big)\\ & \times \mathcal{W}\big(d_k\big|\Sigma_{ls;0}, \nu_{ls;0}\big), \qquad (7) \end{aligned}$$

where Ξ is a symmetric positive definite matrix of an appropriate dimension, $\Sigma_{ls;0} > 0$, and $\nu_{ls;0} > 2$. Considering (7), the Bayes update will impose the soft constraint on the parameters to smoothen the parameter estimates, thereby preventing them from changing too rapidly [21]. By substituting the conceptually correct transition operation (6) with a forgetting one, the latest available posterior is flattened through the forgetting factor $\lambda_{k-1} \in (0, 1]$ to yield $\mathcal{N}(\theta_k|\hat{\theta}_{ls;k-1}, \frac{P_{ls;k-1}}{d_k\lambda_{k-1}})\mathcal{W}(d_k|\lambda_{k-1}\Sigma_{ls;k-1}, \lambda_{k-1}\nu_{ls;k-1})$. The data correction and regularization of the flattened posterior is organized recursively with respect to Bayes' rule, as follows:

$$\begin{aligned} f\big(\theta_k, d_k\big|\mathcal{S}_k\big) &\propto \mathcal{N}\big(y_k\big|h_k'\theta_k, 1/d_k\big) \qquad (8)\\ &\times \frac{\mathcal{N}\big(\theta_k\big|\hat{\theta}_{ls;k-1}, \Xi^{-1}/d_k\big)\mathcal{W}\big(d_k\big|\Sigma_{ls;0}, \nu_{ls;0}\big)}{\mathcal{N}\big(\theta_k\big|\hat{\theta}_{ls;k-2}, \frac{1}{d_k\lambda_{k-1}}\Xi^{-1}\big)\mathcal{W}\big(d_k\big|\lambda_{k-1}\Sigma_{ls;0}, \lambda_{k-1}\nu_{ls;0}\big)}\\ &\times \mathcal{N}\Big(\theta_k\big|\hat{\theta}_{ls;k-1}, P_{ls;k-1}/(d_k\lambda_{k-1})\Big)\\ &\times \mathcal{W}\big(d_k\big|\lambda_{k-1}\Sigma_{ls;k-1}, \lambda_{k-1}\nu_{ls;k-1}\big). \end{aligned}$$

By setting $\lambda_0 \equiv 1$, $\hat{\theta}_{ls;0} \equiv \hat{\theta}_{ls;-1}$, and $P_{ls;0} \equiv \Xi^{-1}$ at time $k=1$, the pdf (7) is appointed to formally initiate the estimation routine. By leveraging the conjugacy of all the pdfs on the right-hand side of (8), the functional recursion is reduced to the LS recursion

$$\begin{aligned} V_{c;k-1} &\equiv \lambda_{k-1}V_{ls;k-1} + \big(1-\lambda_{k-1}\big)\Xi, \qquad (9)\\ P_{c;k-1} &\equiv V_{c;k-1}^{-1}, \qquad (10)\\ \varepsilon_{k-1} &\equiv \hat{\theta}_{ls;k-1} - \hat{\theta}_{ls;k-2}, \qquad (11)\\ \hat{\theta}_{c;k-1} &\equiv \hat{\theta}_{ls;k-1} + \lambda_{k-1}P_{c;k-1}\Xi\varepsilon_{k-1}, \qquad (12)\\ \Sigma_{c;k-1} &\equiv \lambda_{k-1}\big[\Sigma_{ls;k-1} - \varepsilon_{k-1}'\big(I_{\mathring{\theta}} + \lambda_{k-1}\Xi P_{c;k-1}\big) \qquad (13)\\ &\quad \times \Xi\varepsilon_{k-1}\big] + \big(1-\lambda_{k-1}\big)\Sigma_{ls;0},\\ K_k &\equiv P_{c;k-1}h_k/\big(1+h_k'P_{c;k-1}h_k\big), \qquad (14)\\ \hat{e}_{c;k} &\equiv y_k - h_k'\hat{\theta}_{c;k-1}, \qquad (15)\\ \hat{\theta}_{ls;k} &= \hat{\theta}_{c;k-1} + K_k\hat{e}_{c;k}, \qquad (16)\\ P_{ls;k} &= \big(I_{\mathring{\theta}} - K_kh_k'\big)P_{c;k-1}\big(I_{\mathring{\theta}} - K_kh_k'\big)' + K_kK_k', \qquad (17)\\ V_{ls;k} &= P_{ls;k}^{-1} = V_{c;k-1} + h_kh_k', \qquad (18)\\ \Sigma_{ls;k} &= \Sigma_{c;k-1} + \hat{e}_{c;k}^2/\big(1+h_k'P_{c;k-1}h_k\big), \qquad (19)\\ \nu_{ls;k} &= \underbrace{\lambda_{k-1}\nu_{ls;k-1} + \big(1-\lambda_{k-1}\big)\nu_{ls;0}}_{\nu_{c;k-1}} + 1. \qquad (20) \end{aligned}$$

In the following, the modification of the LS to account for uncertainty in the regressor vector will be discussed, and

a refinement of the forgetting mechanism to automatically comply with the degree of the process nonstationarity will be proposed. Further insights into the justification for employing the ordinary LS method, emphasizing its Bayesian interpretation, along with that of the Kalman filter, can be found in Peterka [24].

B. Recursive TLS *from the Bayesian viewpoint*

According to the EIV approach, the system posits a relationship between the noisy outputs and inputs in the form

$$\begin{cases} h_k = h_{0;k} + \xi_k, \quad \xi_k \sim \mathcal{N}\left(O_{\mathring{\theta},1}, I_{\mathring{\theta}}/d_k\right), \\ y_k = \overbrace{h'_{0;k}\theta_k}^{y_{0;k}} + e_{y;k}, \quad e_{y;k} \sim \mathcal{N}\left(0, 1/d_k\right). \end{cases} \tag{21}$$

Now, in the context of the EIV analysis, the regressor h_k is conceptualized as a random vector contaminated with additive measurement noises. These noises are arranged into the vector $\xi_k \equiv [e_{u;k-1}, \ldots, e_{u;k-n}, -e_{y;k-1}, \ldots, -e_{y;k-n}]'$. Conveniently, the regressor maintains the same structure as that in model (1) and likewise contains directly observed data, $\{u_i, y_i\}_{i=k-n}^{k-1}$. The noise-free input $u_{0;k}$ and the undisturbed output $y_{0;k}$ represent hidden variables stacked into the vector $h_{0;k} \equiv [u_{0;k-1}, \ldots, u_{0;k-n}, -y_{0;k-1}, \ldots, -y_{0;k-n}]'$. The noise components $e_{u;k}$ and $e_{y;k}$ are assumed to be mutually independent, zero-mean, normally distributed white noise sequences, both sharing the same precision d_k.

We intend again to infer the parameters θ_k and d_k by constructing the conditional posterior, complicated in this scenario due to the uncertainty inherent in the accessible data. To formulate the parametric model in general, we consider the joint pdf of y_k and $h_{0;k}$, conditioned on θ_k, d_k, and $\mathcal{D}_{1-n}^{k-1}$. The joint pdf is obtained via an affine transformation of the normal vector $[e_{y;k},\, \xi_k']'$ into $[y_k,\, h_{0;k}']'$, driven by

$$\left| \frac{\partial}{\partial \begin{bmatrix} y_k \\ h_{0;k} \end{bmatrix}'} \left(\underbrace{\begin{bmatrix} 1 & -\theta_k' \\ O_{\mathring{\theta},1} & -I_{\mathring{\theta}} \end{bmatrix} \begin{bmatrix} y_k \\ h_{0;k} \end{bmatrix} + \begin{bmatrix} 0 \\ h_k \end{bmatrix}}_{\mathcal{X}} \right) \right| \times \mathcal{N}\left(\mathcal{X} \middle| O_{\mathring{\theta}+1,1}, I_{\mathring{\theta}+1}/d_k\right).$$

The resulting joint pdf shows as

$$f\left(\begin{bmatrix} y_k \\ h_{0;k} \end{bmatrix} \middle| \theta_k, d_k, \mathcal{D}_{1-n}^{k-1} \right) = \mathcal{N}\left(\begin{bmatrix} y_k \\ h_{0;k} \end{bmatrix} \middle| \begin{bmatrix} h_k'\theta_k \\ h_k \end{bmatrix}, \frac{1}{d_k} \begin{bmatrix} 1+\theta_k'\theta_k & \theta_k' \\ \theta_k & I_{\mathring{\theta}} \end{bmatrix} \right). \tag{22}$$

In line with the Bayesian approach, however, only the pdf for $y_k|\theta_k, d_k, \mathcal{D}_{1-n}^{k-1}$ will exert an influence on the posterior inferences. The sought parametric model is derivable from (22) by integrating with respect to the hidden variables $h_{0;k}$. Analogously, the joint pdf (22) can be factorized into the product of a marginal and a conditional pdf by employing Claim 1 from [25]. The marginal pdf is given by

$$f\left(y_k \middle| \theta_k, d_k, \mathcal{D}_{1-n}^{k-1}\right) = \mathcal{N}\left(y_k \middle| h_k'\theta_k, \left(1+\theta_k'\theta_k\right)/d_k\right), \tag{23}$$

and, for completeness, the corresponding conditional pdf is represented by

$$f\left(h_{0;k} \middle| \theta_k, d_k, y_k, \mathcal{D}_{1-n}^{k-1}\right) = \mathcal{N}\left(h_{0;k} \middle| h_k + \frac{\theta_k\left(y_k - h_k'\theta_k\right)}{1+\theta_k'\theta_k}, \frac{1}{d_k}\left(I_{\mathring{\theta}} - \frac{\theta_k\theta_k'}{1+\theta_k'\theta_k}\right) \right).$$

To allow the recursive evaluation of the posterior, the parametric model is used as a template for the ES pdf, ensuring uniform interpretation of the information. This self-reproducibility requirement necessitates conditioning the supplied $\mathcal{NW}$ pdf parameters on $\omega_k^2 \equiv 1/(1+\theta_k'\theta_k)$. Let us employ the following ES pdf:

$$\begin{aligned} &f\left(\theta_k, d_k \middle| \hat{\theta}_{ls;k-1}, \Xi, \Sigma_{ls;0}, \nu_{ls;0}, \omega_k\right) \\ &= \mathcal{N}\left(\theta_k \middle| \hat{\theta}_{ls;k-1}, \Xi^{-1}/(d_k\omega_k^2)\right) \mathcal{W}\left(d_k \middle| \Sigma_{ls;0}\omega_k^2, \nu_{ls;0}\right). \end{aligned} \tag{24}$$

The form of the pdf (24) guarantees the variable regularization-based stabilization in the forgetting and serves as the starting point for the estimation routine. On overlapping the ES pdf by the sequentially retrieved data, the posterior becomes

$$\begin{aligned} &f\left(\theta_k, d_k \middle| \mathcal{S}_k, \omega_k\right) \propto \mathcal{N}\left(y_k \middle| h_k'\theta_k, 1/(d_k\omega_k^2)\right) \\ &\times\, d_k^{(\nu_{c;k-1}+\mathring{\theta}-2)/2} \omega_k^{(\nu_{c;k-1}+\mathring{\theta})} \\ &\times\, \exp\left[-\frac{d_k\omega_k^2}{2} \begin{bmatrix} \theta_k \\ 1 \end{bmatrix}' \begin{bmatrix} P_{c;k-1}^{-1} & \mathcal{M}_{21}' \\ \mathcal{M}_{21} & \mathcal{M}_{22} \end{bmatrix} \begin{bmatrix} \theta_k \\ 1 \end{bmatrix} \right], \end{aligned} \tag{25}$$

where $\mathcal{M}_{21} = -\hat{\theta}_{c;k-1}' P_{c;k-1}^{-1}$ and $\mathcal{M}_{22} = \Sigma_{c;k-1} + \hat{\theta}_{c;k-1}' P_{c;k-1}^{-1} \hat{\theta}_{c;k-1}$. Although the assumptions embodied in (25) allow for the posterior inferences on the parameters, this is countered by the necessity of knowing ω_k, which is unavailable in practical scenarios. It is worth noting that with a noninformative prior Bayes' rule leads to a functional form of the posterior in accordance with (25); in such a case, similarly, analytical normalization is intractable without the knowledge of ω_k.

The objective thus is to construct a posterior that no longer relies on the knowledge of the function over the investigated parameters. The EIV identification can also be tackled through the free-form variational Bayes (VB) method, which has proved its efficiency when, for instance, tailored to solve the indirect identification problems [26], [27]. The free-form VB strategy entails introducing hidden variables into the parametric model, followed by their VB marginalization to ensure their absence in the posterior (see §6.3.3 in [28] for guidance on handling hidden variables). The adopted optimal design seeks the best representation for the pdf (25) under the $\mathcal{NW}$ distribution, subject to the constraint assumption $\omega_k^2 = 1/(1+\theta_k'\theta_k)$. To execute this, a dissimilarity measure between the trial pdf, $f_{\mathrm{T}}(\theta, d)$, and the explicit pdf, $f_{\mathrm{E}}(\theta, d)$, is quantified by the Kullback-Leibler divergence (KLD) [29]

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

The KLD reaches its absolute minimum value, which is zero, at $f_{\mathrm{T}} \equiv f_{\mathrm{E}}$. As previously considered, the loss function

is established to quantify the loss incurred when the $\mathcal{NW}$ posterior, unconditioned on ω_k, is used to fit the data update process (25). To work with f_{T}, we employ

$$\mathcal{L}_\eta\big(\hat{\theta}_k, P_k, \Sigma_k, \nu_k, \omega_k\big) \equiv \mathcal{D}\big(f_{\mathrm{T}}\big\|\omega_k^{-\nu_{ls;k}} f_{\mathrm{E}}\big) \tag{26}$$
$$- \frac{\eta}{2}\int_{\theta^*}\int_{d^*} d_k\left(\omega_k^2\big(\theta_k'\theta_k + 1 - \mathrm{tr}\left(P_k/d_k\right)\big) - 1\right) f_{\mathrm{T}}\mathrm{d}d_k\mathrm{d}\theta_k,$$

where $f_{\mathrm{T}} \equiv \mathcal{N}(\theta_k|\hat{\theta}_k, \frac{1}{d_k}P_k)\mathcal{W}(d_k|\Sigma_k, \nu_k)$ and the explicit pdf corresponds to the right-hand side of (25)

$$f_{\mathrm{E}} \equiv \mathcal{N}\big(y_k\big|h_k'\theta_k, 1/(d_k\omega_k^2)\big)\mathcal{N}\big(\theta_k\big|\hat{\theta}_{c;k-1}, P_{c;k-1}/(d_k\omega_k^2)\big)$$
$$\times\, \mathcal{W}\big(d_k\big|\Sigma_{c;k-1}\omega_k^2, \nu_{c;k-1}\big)/\iota,$$

with the normalizing factor ι. The factor ι is independent of the optimized parameters and follows the Student's t distribution; its evaluation for this type of problem can be found in [30]. The expression in (26) scaled by the Lagrange multiplier η activates the mean value hard equality constraint. The constraint function is normalized by d_k to be consistent with the quadratic forms in the exponential functions entering the pdfs. The requirement for approximating the Wishart part was relaxed by using its nonnormalized variant, $\omega_k^{-\nu_{ls;k}} f_{\mathrm{E}}$; otherwise, the method of Lagrange multipliers fails to determine the extrema. The evaluation of the loss function (26) can be decomposed into a sum of specific integrals involving the gamma $\Gamma(\cdot)$ and digamma $\Psi(\cdot)$ functions. By considering the integral below [21],

$$\mathcal{E}_{\mathcal{W}(d_k|\Sigma_k,\nu_k)}\big[\ln\left(d_k\right)\big] = \Psi\big(\nu_k/2\big) + \ln\left(2\right) - \ln\left(\Sigma_k\right),$$

which facilitates evaluating the expectation

$$\mathcal{E}_{\mathcal{N}(\theta_k|\hat{\theta}_k,\frac{1}{d_k}P_k)\mathcal{W}(d_k|\Sigma_k,\nu_k)}\big[\ln\big(\mathcal{W}\big(d_k\big|\Sigma_{c;k-1}\omega_k^2, \nu_{c;k-1}\big)$$
$$\times\,\mathcal{N}\big(\theta_k\big|\hat{\theta}_{c;k-1}, P_{c;k-1}/(d_k\omega_k^2)\big)\big)\big] = \big(\mathring{\theta} + \nu_{c;k-1}\big)\ln\left(\omega_k\right)$$
$$+ \ln\left((2\pi)^{-\frac{\mathring{\theta}}{2}}\left|V_{c;k-1}\right|^{\frac{1}{2}}\left(\frac{\Sigma_{c;k-1}}{2}\right)^{\frac{\nu_{c;k-1}}{2}}\frac{1}{\Gamma\left(\frac{\nu_{c;k-1}}{2}\right)}\right)$$
$$+ \frac{\nu_{c;k-1} + \mathring{\theta} - 2}{2}\left[\Psi\left(\frac{\nu_k}{2}\right) + \ln\left(\frac{2}{\Sigma_k}\right)\right]$$
$$- \frac{\omega_k^2}{2}\big(\hat{\theta}_k - \hat{\theta}_{c;k-1}\big)'V_{c;k-1}\big(\hat{\theta}_k - \hat{\theta}_{c;k-1}\big)\frac{\nu_k}{\Sigma_k}$$
$$- \frac{\omega_k^2}{2}\left(\frac{\Sigma_{c;k-1}\nu_k}{\Sigma_k} + \mathrm{tr}\left(P_kV_{c;k-1}\right)\right),$$

all the necessary templates are readily available to construct the final form of the loss function. The resulting function is identified as

$$\mathcal{L}_\eta\big(\hat{\theta}_k, P_k, \Sigma_k, \nu_k, \omega_k\big)$$
$$= \frac{1}{2}\ln\left(\frac{\left|P_k\right|^{-1}}{\left|V_{ls;k}\right|}\right) + \frac{\omega_k^2}{2}\,\mathrm{tr}\left(V_{ls;k}P_k\right) + \ln\left(\frac{\Gamma\big(\nu_{ls;k}/2\big)}{\Gamma\big(\nu_k/2\big)}\right)$$
$$- \frac{\mathring{\theta}}{2} + \frac{\nu_{ls;k}}{2}\ln\left(\frac{\Sigma_k}{\Sigma_{ls;k}}\right) - \frac{\nu_k}{2}\left(1 - \frac{\omega_k^2\Sigma_{ls;k}}{\Sigma_k}\right)$$
$$+ \frac{\nu_k\omega_k^2}{2\Sigma_k}\big(\hat{\theta}_k - \hat{\theta}_{ls;k}\big)'V_{ls;k}\big(\hat{\theta}_k - \hat{\theta}_{ls;k}\big) - \mathring{\theta}\ln\left(\omega_k\right)$$
$$+ \frac{\nu_k - \nu_{ls;k}}{2}\Psi\left(\frac{\nu_k}{2}\right) - \frac{\eta\nu_k}{2\Sigma_k}\left(\omega_k^2\hat{\theta}_k'\hat{\theta}_k + \omega_k^2 - 1\right).$$

The equivalence of the TLS solution with f_{T}'statistics minimizing the loss function $\mathcal{L}_\eta$ is reported by the lemma below.

Lemma 1: Let f_{T} be established as an approximation of f_{E}, with the restriction that the functional form of f_{T} is given by the $\mathcal{NW}$ pdf. Then, searching for the best representative of f_{T} by minimizing the loss function (26) yields

$$\hat{\theta}_{tls;k} = \big(V_{ls;k} - \hat{\eta}I_{\mathring{\theta}}\big)^{-1}V_{ls;k}\hat{\theta}_{ls;k}, \tag{27}$$
$$\hat{\omega}_k^2 = 1/(\hat{\theta}_{tls;k}'\hat{\theta}_{tls;k} + 1), \tag{28}$$
$$P_{tls;k}^{-1} \equiv V_{tls;k} = \hat{\omega}_k^2V_{ls;k}, \tag{29}$$
$$\Sigma_{tls;k} = \hat{\eta} \tag{30}$$
$$= \hat{\omega}_k^2\Sigma_{ls;k} + \hat{\omega}_k^2\big(\hat{\theta}_{tls;k} - \hat{\theta}_{ls;k}\big)'V_{ls;k}\big(\hat{\theta}_{tls;k} - \hat{\theta}_{ls;k}\big),$$
$$\nu_{tls;k} = \nu_{ls;k}, \tag{31}$$

where $\{\hat{\theta}_{tls;k}, P_{tls;k}, \Sigma_{tls;k}, \nu_{tls;k}, \hat{\omega}_k^2\}$ embody the unique, globally optimal values of the arguments of the loss function $\mathcal{L}_\eta$. The optimized Lagrange multiplier $\hat{\eta}$ represents the smallest eigenvalue of the augmented information matrix $\bar{V}_k$ introduced in (5).

Proof: To confirm the unique global optimality of the solution, the first-order necessary and second-order sufficient conditions are elaborated and discussed in the Appendix. ■

The optimal representation of the posterior $\hat{f}_{\mathrm{T}}(\theta_k, d_k) = \mathcal{N}(\theta_k|\hat{\theta}_{tls;k}, \frac{1}{d_k}P_{tls;k})\mathcal{W}(d_k|\Sigma_{tls;k}, \nu_{tls;k})$ for the EIV model is constructed at each step to agree with the sequential data retrieval. The expansion of the model to incorporate uncertainties in the regressor employs data compression via the LS into $\bar{V}_k$, followed by the ex post deregularization of the LS solution to conduct the EIV model estimation. We suggest that the estimation routine be initialized by setting Ξ, $\Sigma_{ls;0}$, and $\nu_{ls;0}$, as they have a clear interpretation. Where we intend to initiate the learning from the nonzero vector $\hat{\theta}_{tls;0}$, it is necessary to determine $\Sigma_{tls;0}$, if possible, according to $\Sigma_{tls;0} = \frac{1-\sqrt{1-4\hat{\omega}_0^4\Sigma_{ls;0}\hat{\theta}_{tls;0}'\Xi^{-1}\hat{\theta}_{tls;0}}}{2\hat{\omega}_0^2\hat{\theta}_{tls;0}'\Xi^{-1}\hat{\theta}_{tls;0}}$, and subsequently to recalculate $\hat{\theta}_{ls;0} = \Xi^{-1}(\Xi - \Sigma_{tls;0}I_{\mathring{\theta}})\hat{\theta}_{tls;0}$.

C. Automatic forgetting factor selection

The aim of this section is to briefly justify the emergence of the forgetting factor, $\lambda_k \in (0, 1]$, in the LS method and to establish its automatic adjustment from the perspective of optimization theory. Temporarily, the time index for variables θ_{k+1} and d_{k+1} will be omitted for the sake of brevity. Our starting point now is the loss functional [22], [30]

$$\mathcal{L}_{\mathcal{F}}\big(f\big(\theta, d\big), \varphi\big) = \sum_{i=1}^{2}\varphi_i\big[\mathcal{D}\big(f\big(\theta, d\big)\big\|f_i\big(\theta, d\big)\big) - \varrho_i\big]$$
$$+ \eta_\iota\left(\int_{d^*}\int_{\theta^*} f\big(\theta, d\big)\mathrm{d}\theta\mathrm{d}d - 1\right), \tag{32}$$

where $f(\theta, d)$ denotes the time updated posterior subject to correction and $f_i(\theta, d)$ refers to its ith prediction alternative. The probability vector $\varphi \equiv [\varphi_1, \varphi_2]'$, satisfying $\sum_{i=1}^{2}\varphi_i = 1$, defines the weights by which each alternative is combined into a single pdf. The nonnegative loss estimates $\varrho \equiv \{\varrho_1, \varrho_2\}$ provide feedback for the decision-making,

ensuring that the resulting correction is substantiated in response to the empirically confirmed performance. The term scaled by the Lagrange multiplier η_ι ensures that the minimizer $\hat{f}(\theta, d)$ integrates to one. In our setup, the optional components in (32) are selected as $f_1(\theta, d) \equiv \mathcal{N}(\theta|\hat{\theta}_{tls;k}, \frac{1}{d}P_{tls;k})\mathcal{W}(d|\Sigma_{tls;k}, \nu_{tls;k})$, and $f_2(\theta, d) \equiv \mathcal{N}(\theta|\hat{\theta}_{tls;k}, \frac{1}{\alpha d}P_{tls;k})\mathcal{W}(d|\alpha\Sigma_{tls;k}, \alpha\nu_{tls;k})$, with $\alpha \in (0, 1)$ setting an upper bound on the parameter uncertainty increase. Further, to reduce false detections in the parameter variations, the expected disturbing noise level is artificially increased via the factor $\zeta \in (0, 1]$. This is reflected by the loss estimates, which penalize the distance of each prediction alternative from the scaled posterior $f_\zeta(\theta_k, d_k) \equiv \mathcal{N}(\theta_k|\hat{\theta}_{tls;k}, \frac{1}{d_k\zeta}P_{tls;k})\mathcal{W}(d_k|\Sigma_{tls;k}, \nu_{tls;k})$. Specifically, we have $\varrho_1 \equiv \mathcal{D}(f_\zeta(\theta_k, d_k)\|f_{1,\zeta}(\theta_k|d_k)f_1(d_k))$, with $f_{1,\zeta}(\theta_k|d_k) \equiv \mathcal{N}(\theta_k|\hat{\theta}_{tls;k-1}, \frac{1}{d_k\zeta}P_{tls;k-1})$ and $\varrho_2 \equiv \mathcal{D}(f_\zeta(\theta_k, d_k)\|f_{2,\zeta}(\theta_k|d_k)f_2(d_k))$, where $f_{2,\zeta}(\theta_k|d_k) \equiv \mathcal{N}(\theta_k|\hat{\theta}_{tls;k-1}, \frac{1}{d_k\zeta}\alpha^{-1}P_{tls;k-1})$. Let us add that the smaller the heuristic factor ζ, the less sensitive the algorithm to inconsistencies in the data-generating process. Readers interested in a more detailed explanation of the construction of the loss functional $\mathcal{L}_\mathcal{F}$ and the selection of its optional components are directed to [22].

We again omit the time index in the variables θ_{k+1} and d_{k+1} for the sake of clarity. The functional (32) yields a unique minimizer for $f(\theta, d)$, which is identified as the geometric mean (Lemma 1 in [22]):

$$\hat{f}\big(\theta, d\big) \propto \prod_{i=1}^{2} f_i^{\varphi_i}\big(\theta, d\big). \tag{33}$$

The best representative of φ is found by solving the necessary and sufficient optimality conditions (Lemma 2 in [22])

$$\begin{cases} \mathcal{D}\big(\hat{f}\big(\theta, d|\big)\big\|f_i\big(\theta, d\big)\big) - \varrho_i = \mu, & \text{all } i \text{ such that } \varphi_i > 0, \\ \mathcal{D}\big(\hat{f}\big(\theta, d|\big)\big\|f_i\big(\theta, d\big)\big) - \varrho_i \leq \mu, & \text{all } i \text{ such that } \varphi_i = 0, \end{cases} \tag{34}$$

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

The formulation of the time updated posterior on the basis of the geometric mean justifies the presence of the forgetting factor $\lambda_k = \hat{\varphi}_1(1-\alpha)+\alpha$ in the LS routine, as it arises from merging the prediction alternatives. To implement data-informed forgetting, the optimality conditions (34) are evaluated, leading to

$$\lambda_k = 1/\Omega_\zeta, \tag{35}$$

where

$$\begin{aligned} \Omega_\zeta = \Bigg[& \operatorname{tr}\left(V_{tls;k-1}P_{tls;k}\right) + \nu_{tls;k-1}\ln\left(\frac{\Sigma_{tls;k}\nu_{tls;k-1}}{\Sigma_{tls;k-1}\nu_{tls;k}}\right) \\ & + \frac{\nu_{tls;k}}{\Sigma_{tls;k}}\left(\Sigma_{tls;k-1} + (\underbrace{\hat{\theta}_{tls;k} - \hat{\theta}_{tls;k-1}}_{\gamma_k})'\zeta V_{tls;k-1}\gamma_k\right) \\ & + \frac{\nu_{tls;k-1}}{\nu_{tls;k}} - \nu_{tls;k-1}\Bigg]\frac{1}{\mathring{\theta}+1}. \end{aligned} \tag{36}$$

To prevent violating the constraint boundaries, set $\lambda_k = \alpha$ when $1 \leq \alpha\Omega_\zeta$, and $\lambda_k = 1$ whenever $1 \geq \Omega_\zeta$. The results for λ_k rely on the asymptotic approximations, $\Gamma(\nu) \simeq \exp[-\nu]\nu^{(\nu-1/2)}(2\pi)^{1/2}$ and $\Psi(\nu) \equiv \partial[\ln(\Gamma(\nu))]/\partial\nu \simeq \ln(\nu) - 1/(2\nu)$, used in the assessed KLDs.

Algorithm 1 The Bayesian estimation procedure for a time-varying EIV model.

Initialization phase:
Gather the starting data set $\mathcal{D}_{1-n}^1$ to fill h_1.
Set the lower bound on the forgetting factor $\alpha \in \big(0, 1\big)$ and set the heuristic factor $\zeta \in (0, 1]$.
Set the initial value of the forgetting factor to $\lambda_0 \equiv 1$.
Initialize the statistics $\{\hat{\theta}_{ls;0}, \Xi, \Sigma_{ls;0} > 0, \nu_{ls;0} > 2\}$ and execute the assignments $\{\hat{\theta}_{ls;-1} \equiv \hat{\theta}_{ls;0}, V_{ls;0} \equiv P_{ls;0}^{-1} \equiv \Xi\}$ to obtain, for $k = 1$, the starting point $\{V_{c;0}, P_{c;0}, \hat{\theta}_{c;0}, \Sigma_{c;0}, \nu_{c;0}\}$ needed to initiate the data LS correction (14)–(20).
Learning phase:
for $k \leftarrow 1, \mathring{k}$ **do**
  Input: $\begin{cases} y_k, h_k, \Xi, \Sigma_{ls;k-1}, \nu_{ls;k-1}, \\ \hat{\theta}_{ls;k-1}, \hat{\theta}_{ls;k-2}, V_{ls;k-1}, \lambda_{k-1}, \\ \hat{\theta}_{tls;k-1}, V_{tls;k-1}, \Sigma_{tls;k-1} \end{cases}$
  LS update: $\{V_{ls;k-1}, \hat{\theta}_{ls;k-1}, \Sigma_{ls;k-1}, \nu_{ls;k-1}\} \rightarrow \{P_{ls;k}, V_{ls;k}, \hat{\theta}_{ls;k}, \Sigma_{ls;k}, \nu_{ls;k}\}$ ▷ (9)–(20)
  Assemble the matrix $\bar{V}_k$ ▷ (5)
  Determine $\hat{\eta}$ as the smallest eigenvalue of $\bar{V}_k$
  TLS update: $\{P_{ls;k}, V_{ls;k}, \hat{\theta}_{ls;k}, \nu_{ls;k}, \hat{\eta}\} \rightarrow$
  $\{P_{tls;k}, V_{tls;k}, \hat{\theta}_{tls;k}, \Sigma_{tls;k}, \nu_{tls;k}\}$ ▷ (27)–(31)
  Calculate Ω_ζ ▷ (36)
  Inspect whether the factor λ_k lies inside the feasible region and evaluate it ▷ (35)
  Output: $\hat{\theta}_{ls;k}$, $P_{ls;k}$, $V_{ls;k}$, $\Sigma_{ls;k}$, $\nu_{ls;k}$, $\hat{\theta}_{tls;k}$, $P_{tls;k}$, $V_{tls;k}$, $\Sigma_{tls;k}$, $\nu_{tls;k}$, λ_k
end for

To elucidate the implementation details, the proposed procedure is summarized in Algorithm 1.

III. Illustrative Experiments

The numerical illustration of the developments introduced in this paper is provided through simulating a second-order EIV model (21), $y_{0;k} + \sum_{i=1}^{2} a_i y_{0;k-i} = \sum_{i=1}^{2} b_i u_{0;k-i}$, where the available signals $y_k = y_{0;k} + e_{y;k}$ and $u_k = u_{0;k} + e_{u;k}$ are measured with normally distributed, mutually uncorrelated errors. The coefficients $\{a_i, b_i\}_{i=1}^2$ correspond to the discrete transfer function $\mathcal{G}(z) = k_\mathcal{G}(z - \exp[T_s n_0])/((z - \exp[T_s p_1])(z - \exp[T_s p_2]))$. The sampling period T_s is set to $T_s = 1\,\text{s}$; the gain equals $k_\mathcal{G} = (1 - \exp[T_s p_1])(1 - \exp[T_s p_2])/(1 - \exp[T_s n_0])$; and the white measurement error sequences $\{e_{u;k}, e_{y;k}\}$ are generated at $d = 10^3$ or $d = 10^2$. The initial zero and poles' values satisfy $n_0 = 1$ and $p_{1,2} = -0.4 \pm \text{i}0.8$, and the system experiences a sudden change at the time $t = 250\,\text{s}$, which causes the zero and the poles to switch to $n_0 = -1$ and $p_{1,2} = -0.2 \pm \text{i}0.4$. The input fed into the system is driven by $u_{0;k} = 0.9u_{0;k-1} + w_k$, where w_k is the discrete white noise $w_k \sim \mathcal{N}(0, 1)$. Further, regarding the user-defined input arguments to Algorithm 1,

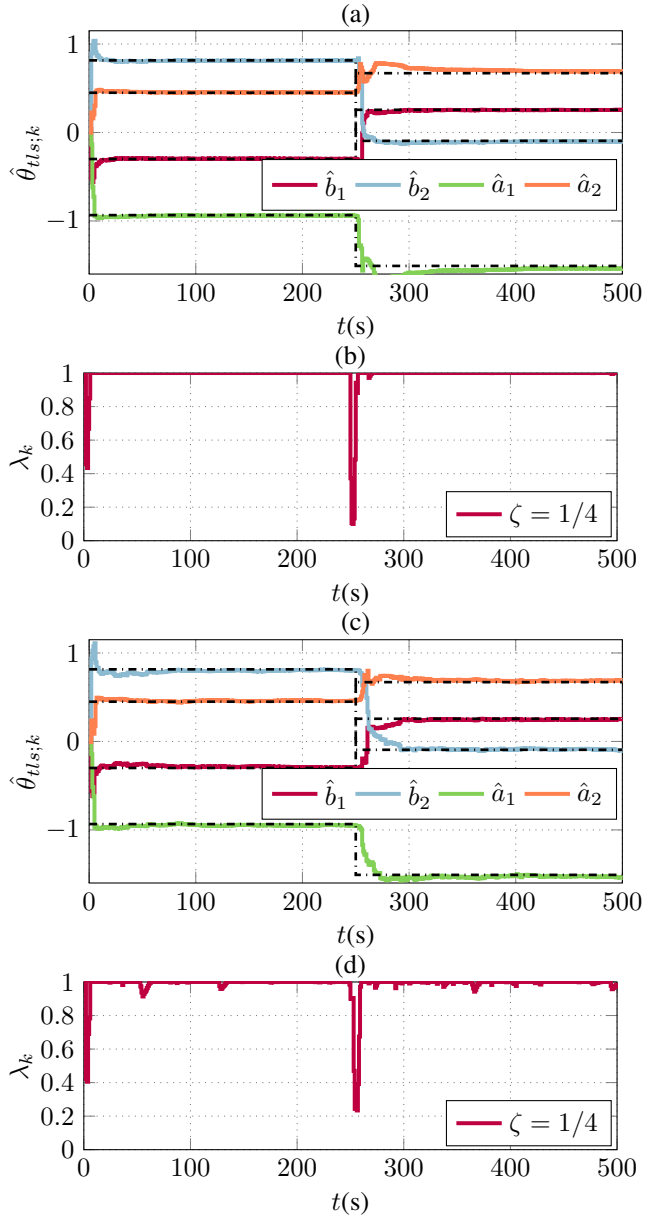

Fig. 1. The trajectories of the EIV model estimates, optimized using the adaptive variant of the TLS, at $d = 10^3$ (a) and at $d = 10^2$ (c). Plot (b) shows the corresponding time course of the forgetting factor λ_k at $d = 10^3$, while plot (d) represents the same at $d = 10^2$. The true values of the EIV model coefficients are initially represented by $\{b_1 \approx -0.3, b_2 \approx 0.82, a_1 \approx -0.93, a_2 \approx 0.45\}$, and after 250 s, they are switched to $\{b_1 \approx 0.26, b_2 \approx -0.09, a_1 \approx -1.51, a_2 \approx 0.67\}$.

the learning initiates with $\hat{\theta}_{ls;0} = O_{4,1}$, $\Xi = I_4$, $\Sigma_{ls;0} = 1$, $\nu_{ls;0} = 10$, $\alpha = 0.1$ and $\zeta = 1/4$. The results of the experiments are plotted in Fig. 1. In this experiment setup, the results indicate the algorithm's ability to combat noise while effectively tracking parameter changes.

IV. Conclusion

The problem of estimating a regression-type EIV model with time-varying parameters has been faced within a rigorous probabilistic framework. The estimator concept formulation presented herein takes into account the uncertainty inherent in the accessible data and expands upon the solution relying on a noise-free, deterministic regressor. The approach provides a theoretical foothold for a broader objective in probabilistic designs, aimed at systematically learning generally nonlinear EIV systems from data streams and data-informed risk assessment. Although the solution to a linear-in-parameters problem is algebraically equivalent to the TLS method, it provides a useful metric for quantifying the uncertainty of the parameter estimates in the presence of an unknown precision of the measurement noises. The Bayesian interpretation of the TLS solution was justified by solving the equality-constrained fixed-form variational Bayes problem, supported by an application of optimization theory relying on the method of Lagrange multipliers. The benefit of the original problem formulation was demonstrated through its integration into the statistical decision-making, resulting in the development of an original adaptive estimator for an EIV model.

Appendix

By computing the gradient of the loss function $\mathcal{L}_\eta$ with respect the design variables $\Theta = [\mathrm{vech}(P_k)', \hat{\theta}_k', \omega_k, \Sigma_k, \nu_k]'$ and the multiplier η at $\hat{\Theta} = [\mathrm{vech}(P_{tls;k})', \hat{\theta}_{tls;k}', \hat{\omega}_k, \Sigma_{tls;k}, \nu_{tls;k}]'$ and $\hat{\eta}$, and subsequently equating the result to zero, we arrive at the first-order necessary conditions

$$\left.\frac{\partial \mathcal{L}_\eta}{\partial\, \mathrm{vech}\,(P_k)}\right|_{\substack{\hat{\Theta}\\ \hat{\eta}}} = -\frac{1}{2} D_{\hat{\theta}}' \,\mathrm{vec}\,(V_{tls;k}) + \frac{\hat{\omega}_k^2}{2} D_{\hat{\theta}}' \,\mathrm{vec}\,(V_{ls;k}) = 0, \tag{A.1}$$

$$\left.\frac{\partial \mathcal{L}_\eta}{\partial \hat{\theta}_k}\right|_{\substack{\hat{\Theta}\\ \hat{\eta}}} = \hat{\omega}_k^2 \frac{\nu_{tls;k}}{\Sigma_{tls;k}} V_{ls;k}\big(\hat{\theta}_{tls;k} - \hat{\theta}_{ls;k}\big) - \hat{\eta}\hat{\omega}_k^2 \frac{\nu_{tls;k}}{\Sigma_{tls;k}} \hat{\theta}_{tls;k} = 0, \tag{A.2}$$

$$\left.\frac{\partial \mathcal{L}_\eta}{\partial \omega_k}\right|_{\substack{\hat{\Theta}\\ \hat{\eta}}} = 0 \implies -\hat{\eta}\big(\hat{\omega}_k^2 \hat{\theta}_{tls;k}' \hat{\theta}_{tls;k} + \hat{\omega}_k^2\big) + \Sigma_{tls;k} = 0, \tag{A.3}$$

$$\left.\frac{\partial \mathcal{L}_\eta}{\partial \Sigma_k}\right|_{\substack{\hat{\Theta}\\ \hat{\eta}}} = \frac{\nu_{ls;k}}{2\Sigma_{tls;k}} - \hat{\omega}_k^2 \frac{\nu_{tls;k}}{2\Sigma_{tls;k}^2} \begin{bmatrix} \hat{\theta}_{tls;k} \\ 1 \end{bmatrix}' \bar{V}_k \begin{bmatrix} \hat{\theta}_{tls;k} \\ 1 \end{bmatrix} = 0, \tag{A.4}$$

$$\left.\frac{\partial \mathcal{L}_\eta}{\partial \nu_k}\right|_{\substack{\hat{\Theta}\\ \hat{\eta}}} = \frac{\nu_{tls;k} - \nu_{ls;k}}{2} \left.\frac{\partial}{\partial \nu_k} \Psi\left(\frac{\nu_k}{2}\right)\right|_{\nu_{tls;k}} = 0, \tag{A.5}$$

$$\left.\frac{\partial \mathcal{L}_\eta}{\partial \eta}\right|_{\substack{\hat{\Theta}\\ \hat{\eta}}} = -\frac{\nu_{tls;k}}{2\Sigma_{tls;k}} \Big(\hat{\omega}_k^2 \hat{\theta}_{tls;k}' \hat{\theta}_{tls;k} + \hat{\omega}_k^2 - 1\Big) = 0. \tag{A.6}$$

From the condition (A.6), it follows for the scalar product that $[\hat{\omega}_k \hat{\theta}_{tls;k}', \hat{\omega}_k][\hat{\omega}_k \hat{\theta}_{tls;k}', \hat{\omega}_k]' = 1$. Thus, directly from (A.3), we obtain the identity $\Sigma_{tls;k} = \hat{\eta}$. Given the objective of minimizing $\mathcal{L}_\eta$ as much as possible and the fact

that $\mathcal{L}_\eta$ decreases with decreasing $\Sigma_{tls;k}$, we capitalize on the inherent properties of the Rayleigh quotient (§11.5 in [31]). Specifically, we use the lower bound of the Rayleigh quotient, $\hat{\eta} \leq \frac{[\hat{\omega}_k \hat{\theta}'_{tls;k}, \hat{\omega}_k] \bar{V}_k [\hat{\omega}_k \hat{\theta}'_{tls;k}, \hat{\omega}_k]'}{[\hat{\omega}_k \hat{\theta}'_{tls;k}, \hat{\omega}_k][\hat{\omega}_k \hat{\theta}'_{tls;k}, \hat{\omega}_k]'}$, represented by the smallest eigenvalue of $\bar{V}_k$ when $[\hat{\omega}_k \hat{\theta}'_{tls;k}, \hat{\omega}_k]'$ is the corresponding eigenvector. Choosing $\hat{\eta}$ and $\hat{\theta}_{tls;k}$ according to the aforementioned lower bound automatically satisfies (A.2), (A.4), and the objective of globally minimizing $\mathcal{L}_\eta$. Considering the positivity of the polygamma function $\frac{\partial}{\partial \nu_k} \Psi(\frac{\nu_k}{2})$ in (A.5), it is evident that $\nu_{tls;k} = \nu_{ls;k}$.

To prove that the global minimum $\hat{\Theta}$ is unique, the second-order sufficient conditions are inspected. Recall that the unique minimum is achieved when both $\frac{\partial (\nu_k(\omega_k^2 \hat{\theta}'_k \hat{\theta}_k + \omega_k^2 - 1)/(2\Sigma_k))}{\partial \Theta'}\big|_{\hat{\Theta}} \beta = 0$ and $\beta' \frac{\partial^2 \mathcal{L}_\eta}{\partial \Theta \partial \Theta'}\big|_{\substack{\hat{\Theta} \\ \hat{\eta}}} \beta > 0$, where β is a nonzero vector. The Hessian of the loss function $\mathcal{L}_\eta$ at $\hat{\Theta}$ and $\hat{\eta}$ is expressed as

$$\frac{\partial^2 \mathcal{L}_\eta}{\partial \Theta \partial \Theta'}\bigg|_{\substack{\hat{\Theta} \\ \hat{\eta}}} = \begin{bmatrix} \mathcal{H}_{11} & O'_{2,\frac{1}{2}\mathring{\theta}(\mathring{\theta}+1)+\mathring{\theta}+1} \\ O & \begin{bmatrix} \frac{1}{2}\frac{\nu_{tls;k}}{\Sigma^2_{tls;k}} & -\frac{1}{2}\frac{1}{\Sigma_{tls;k}} \\ -\frac{1}{2}\frac{1}{\Sigma_{tls;k}} & \frac{1}{2}\frac{\partial}{\partial \nu_k}\Psi\left(\frac{\nu_k}{2}\right)\Big|_{\nu_{tls;k}} \end{bmatrix} \end{bmatrix},$$

where the nested submatrix $\mathcal{H}_{11}$ is defined as

$$\mathcal{H}_{11} = \begin{bmatrix} \frac{1}{2} D'_{\mathring{\theta}} \left(V_{tls;k} \otimes V_{tls;k}\right) D_{\mathring{\theta}} & O & \hat{\omega}_k D'_{\mathring{\theta}} \operatorname{vec}\left(V_{ls;k}\right) \\ O & \mathcal{H}_s & O \\ \hat{\omega}_k \operatorname{vec}\left(V_{ls;k}\right)' D_{\mathring{\theta}} & O & \mathring{\theta}\frac{2}{\hat{\omega}_k^2} \end{bmatrix},$$

with $\mathcal{H}_s = \hat{\omega}_k^2 \frac{\nu_{tls;k}}{\Sigma_{tls;k}} \left(V_{ls;k} - \hat{\eta} I_{\hat{\theta}}\right)$. By evaluating the determinant of the Hessian matrix, we can demonstrate that it is positive-semidefinite, as all its principal minors are nonnegative. We found the nonzero vector β that satisfies $\frac{\partial (\nu_k(\omega_k^2 \hat{\theta}'_k \hat{\theta}_k + \omega_k^2 - 1)/(2\Sigma_k))}{\partial \Theta'}\big|_{\hat{\Theta}} \beta = 0$, giving rise to the identity $\hat{\omega}_k^2 \hat{\theta}'_{tls;k} \beta_\theta = \frac{\beta_\omega}{\omega}$, where β_θ represents an arbitrary nonzero vector and β_ω denotes a nonzero scalar, by definition. Then, the sufficiency condition results in

$$\beta' \left.\frac{\partial^2 \mathcal{L}_\eta}{\partial \Theta \partial \Theta'}\right|_{\substack{\hat{\Theta} \\ \hat{\eta}}} \beta \geq \beta'_\theta \mathcal{H}_s \beta_\theta + \mathring{\theta} \hat{\omega}_k^4 \left(\beta'_\theta \hat{\theta}_{tls;k}\right)^2 > 0. \quad \text{(A.7)}$$

Given that $V_{ls;k}$ represents the $\mathring{\theta} \times \mathring{\theta}$ principal submatrix of $\bar{V}_k$, we can conclude, according to the Poincaré separation theorem (§11.10 in [31]), that $\mathcal{H}_s$ is positive-semidefinite as $\hat{\eta}$ is less or equal to all the eigenvalues of $V_{ls;k}$. Therefore, for nonzero $\hat{\theta}_{tls;k}$, the inequality (A.7) holds true, and the solution $\hat{\Theta}$ is a unique minimum point.

REFERENCES

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

[2] 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.

[3] T. Salimans and D. A. Knowles, "Fixed-form variational posterior approximation through stochastic linear regression," *Bayesian Anal.*, vol. 8, no. 4, pp. 837–882, Dec. 2013.

[4] 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.

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

[6] M. Kárný, "On assigning probabilities to new hypotheses," *Pattern Recognit. Lett.*, vol. 150, pp. 170–175, 2021.

[7] P. Stoica, M. Cedervall, and A. Eriksson, "Combined instrumental variable and subspace fitting approach to parameter estimation of noisy input-output systems," *IEEE Trans. Signal Process.*, vol. 43, no. 10, pp. 2386–2397, Oct. 1995.

[8] W. X. Zheng, "Transfer function estimation from noisy input and output data," *Int. J. Adapt. Control Signal Process.*, vol. 12, no. 4, pp. 365–380, 1998.

[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] S. V. Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*. Philadelphia, USA: SIAM, 1991.

[11] S. V. Huffel and P. Lemmerling, Eds., *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications*. Dordrecht, The Netherlands: Kluwer, 2002.

[12] T. Söderström, "Errors-in-variables methods in system identification," *Automatica*, vol. 43, no. 6, pp. 939–958, 2007.

[13] A. Wiesel, Y. C. Eldar, and A. Yeredor, "Linear regression with Gaussian model uncertainty: Algorithms and bounds," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2194–2205, June 2008.

[14] O. Nestares, D. J. Fleet, and D. J. Heeger, "Likelihood functions and confidence bounds for total-least-squares problems," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2000, pp. 523–530.

[15] X. Fang, B. Li, H. Alkhatib, W. Zeng, and Y. Yao, "Bayesian inference for the errors-in-variables model," *Stud. Geophys. Geod.*, vol. 61, pp. 35–52, Jan. 2017.

[16] J. L. Crassidis and Y. Cheng, "Maximum likelihood analysis of the total least squares problem with correlated errors," *J. Guid. Control Dyn.*, vol. 42, no. 6, pp. 1204–1217, Jun. 2019.

[17] J. L. Crassidis and Y. Cheng, "Error-covariance analysis of the total least-squares problem," *J. Guid. Control Dyn.*, vol. 37, no. 4, pp. 1053–1063, July 2014.

[18] A. Zellner, *An Introduction to Bayesian Inference in Econometrics*. New York, USA: John Wiley & Sons, 1971.

[19] R. Kulhavý and M. B. Zarrop, "On a general concept of forgetting," *Int. J. Control*, vol. 58, no. 4, pp. 905–924, 1993.

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

[21] 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.

[22] 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.

[23] D. A. Turkington, *Matrix Calculus and Zero-One Matrices*. Cambridge, UK.: Cambridge Univ. Press, 2002.

[24] 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.

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

[26] 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.

[27] J. Dokoupil and P. Václavek, "Recursive identification of the ARARX model based on the variational Bayes method," in *Proc. 62th IEEE Conf. Decision Control*, 2023, pp. 4215–4222.

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

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

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

[31] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus With Applications in Statistics and Econometrics*. Hoboken, NJ, USA: John Wiley & Sons, 2019.