Explicit construction of the minimum error variance estimator for stochastic LTI-ss systems

Eringis, Deividas; Leth, John; Tan, Zheng Hua; Wisniewski, Rafal; Petreczky, Mihaly

Published in:
Automatica

DOI (link to publication from Publisher):
10.1016/j.automatica.2023.111018

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Automatica 153 (2023) 111018

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Technical communique

Explicit construction of the minimum error variance estimator for stochastic LTI-ss systems☆

Deividas Eringis [a,*], John Leth [a], Zheng-Hua Tan [a], Rafal Wisniewski [a], Mihaly Petreczky [b]

[a] Dept. of Electronic Systems, Aalborg University, Aalborg, Denmark
[b] Laboratoire Signal et Automatique de Lille (CRIStAL), Lille, France

ARTICLE INFO

Article history:
Received 9 February 2022
Received in revised form 3 January 2023
Accepted 27 February 2023
Available online 21 April 2023

Keywords:
Realisation theory
Estimation theory
Synthesis of stochastic systems

ABSTRACT

We showcase the derivation of the optimal (minimum error variance) estimator, when one part of the stochastic LTI system outputs is not measured but is able to be predicted from the measured system outputs.

1. Introduction

In system identification it is often assumed that the joint process $(\mathbf{y}, \mathbf{w})$ has a realisation by an autonomous stochastic linear time invariant state space system (LTI-ss) driven by white noise. It is then natural to ask the question how to construct a minimal stochastic LTI-ss realisation of $\mathbf{y}$ with input $\mathbf{w}$, from an LTI-ss realisation of the joint process $(\mathbf{y}, \mathbf{w})$, instead of computing a realisation of $\mathbf{y}$ using oblique projections as described in Katayama (2005) and Lindquist and Picci (2015).

This paper presents an explicit construction of a minimal stochastic LTI-ss representation of $\mathbf{y}$ in innovation form with an exogenous input $\mathbf{w}$ from an autonomous stochastic LTI-ss representation of the joint process $(\mathbf{y}, \mathbf{w})$. The construction assumes that $(\mathbf{y}, \mathbf{w})$ is a stationary, zero-mean, jointly Gaussian stochastic process, and there is no feedback from $\mathbf{y}$ to $\mathbf{w}$. It uses the result of Jozsa, Petreczky, and Camlibel (2018) stating that there exists a minimal LTI-ss realisation of $(\mathbf{y}, \mathbf{w})$ with matrices which admit an upper-triangular form. This allows us to separate out part of the innovation noise of $(\mathbf{y}, \mathbf{w})$, which purely drives $\mathbf{w}$, thus allowing us to formulate this construction. Note that LTI-ss representation of $\mathbf{y}$ in innovation form with input $\mathbf{w}$ can naturally be identified with the best (smallest variance) linear estimator of $\mathbf{y}(t)$ given past $(s < t)$ and present $(s = t)$ measurements of $\mathbf{w}(s)$ (Katayama, 2005; Lindquist & Picci, 2015). In turn, the problem of finding such a predictor can also be thought of as trying to estimate non-measurable quantities of a system from measurable quantities.

Our motivation for developing an explicit construction of an LTI-ss realisation of $\mathbf{y}$ with input $\mathbf{w}$ from a LTI-ss realisation of $(\mathbf{y}, \mathbf{w})$ was that this construction is useful for constructing parametrisations of LTI-ss predictors of $\mathbf{y}$ which are driven by $\mathbf{w}$. Such parametrisations are then useful for formulating system identification algorithms and for developing PAC-Bayesian type error bounds (Alquier, Ridgway, & Chopin, 2016) for LTI-ss systems (Eringis, Leth, Tan, Wisniewski, Esfahan et al., 2021; Eringis, Leth, Tan, Wisniewski, & Petreczky, 2022). In particular, the construction of the paper leads to alternative system identification algorithms.

As it was pointed out above, stochastic realisation theory with inputs is a mature topic with several publications, see the monographs (Caines, 1988; Katayama, 2005; Lindquist & Picci, 2015) and the references therein. However, we have not found in the literature an explicit procedure for constructing a stochastic LTI-ss realisation in innovation form of $\mathbf{y}$ with input $\mathbf{w}$ from the joint stochastic LTI-ss realisation of $(\mathbf{y}, \mathbf{w})$. The current note is intended to fill this gap. In frequency domain one can use causal real rational transfer function matrices to describe processes $\mathbf{y}$ and $\mathbf{w}$, and analysing these processes with feedback free assumption, yields a straightforward construction of estimator of $\mathbf{y}$ given $\mathbf{w}$, see Caines and Chan (1975) and Gevers and Anderson (1982). In this paper we study this problem in time domain, using LTI-ss representations. Note that the problem considered in this paper is superficially similar to that of Wiener-filtering (van Schuppen, 2021). However, the details of the Wiener-filtering problem are

☆ The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Sandip Roy under the direction of Editor André L. Tits.

* Corresponding author.
E-mail addresses: der@es.aau.dk (D. Eringis), jjl@es.aau.dk (J. Leth), zt@es.aau.dk (Z.-H. Tan), raf@es.aau.dk (R. Wisniewski), mihaly.petreczky@centralelille.fr (M. Petreczky).

https://doi.org/10.1016/j.automatica.2023.111018

different, and the existing results do not seem to be directly applicable to problem studied in the paper. An extended version of this paper is available in the report (Eringis, Leth, Tan, Wisniewski and Petreczky, 2021).

This paper is organised as follows. We start by defining the notation and terminology used in this paper. In Section 2 we state the main assumptions of the paper. In Section 3 we present the main result of the paper, in Section 4 we present its proof. Finally, in Section 5 we discuss potential applications and present a number of examples.

Notation and terminology Let $\mathbf{F}$ denote a σ-algebra on the set Ω and $\mathbf{P}$ be a probability measure on $\mathbf{F}$. Unless otherwise stated all probabilistic considerations will be with respect to the probability space $(\Omega, \mathbf{F}, \mathbf{P})$. In this paragraph let $\mathbb{E}$ denote some euclidean space. We associate with $\mathbb{E}$ the topology generated by the 2-norm $\|\cdot\|_2$, and the Borel σ-algebra generated by the open sets of $\mathbb{E}$. The closure of a set M is denoted clM. For $S \subseteq \mathbb{N}$ and stochastic variables $\mathbf{y}, \mathbf{z}_1, \mathbf{z}_2, \ldots$ with values in $\mathbb{R}$ we denote by $\mathbf{E}(\mathbf{y} \mid \{\mathbf{z}_i\}_{i\in S})$ the conditional expectation of $\mathbf{y}$ with respect to the σ-algebra $\sigma(\{\mathbf{z}_i\})$ generated by the family $\{\mathbf{z}_i\}_{i\in S}$. Recall that $\mathbf{E}(\mathbf{zx})$ define an inner product in $L^2(\Omega, \mathbf{F}, \mathbf{P})$ and that $\mathbf{E}(\mathbf{y} \mid \{\mathbf{z}_i\}_{i\in S})$ can be interpreted as the orthogonal projection onto the closed subspace $L^2(\Omega, \sigma(\{\mathbf{z}_i\}_{i\in S}), \mathbf{P})$ which also can be identified with the closure of the subspace generated by $\{\mathbf{z}_i\}_{i\in S}$. Moreover, for a closed subspace H of $L^2(\Omega, \mathbf{F}, \mathbf{P})$ and a stochastic variable $\mathbf{y}$ with values in $\mathbb{E}$ and $\mathbf{E}(\|\mathbf{y}\|_2^2) < \infty$, we let $\mathbf{E}(\mathbf{y} \mid H)$ denote the $\dim(\mathbb{E})$-dimensional vector with ith coordinate equal to $\mathbf{E}(\mathbf{y}_i \mid H)$ with $\mathbf{y}_i$ denoting the ith coordinate of $\mathbf{y}$.

There are two closed subspaces of particular importance. Following Lindquist and Picci (2015), for a discrete time stochastic process $\mathbf{z}(t)$ with values in $\mathbb{E}$ and $\mathbf{E}(\|z(t)\|_2^2) < \infty$, we write $H_t^-(\mathbf{z})$ for the closure of the linear subspace in $L^2(\Omega, \mathbf{F}, \mathbf{P})$ generated by the coordinate functions $\mathbf{z}_i(s)$ of $\mathbf{z}(s)$ for all $s < t$. We let $H_t^+(\mathbf{z})$ denote the closure of the linear subspace in $L^2(\Omega, \mathbf{F}, \mathbf{P})$ generated by coordinate functions $\mathbf{z}_i(s)$ of $\mathbf{z}(s)$ for all $s \geq t$, and $H(\mathbf{z})$ denote the closure of the linear subspace in $L^2(\Omega, \mathbf{F}, \mathbf{P})$ generated by coordinate functions $\mathbf{z}_i(s)$, $s \in \mathbb{Z}$.

Let A, B and C be closed subspaces of $L^2(\Omega, \mathbf{F}, \mathbf{P})$. We then define $A \vee B = cl\{a + b \mid a \in A,\ b \in B\}$, and say that A and B are orthogonal given C, denoted $A \perp B \mid C$, if $\mathbf{E}\Big((a - \mathbf{E}(a \mid C))\big(b - \mathbf{E}(b \mid C)\big)\Big) = 0$, for all $a \in A$ and $b \in B$. We use the following notation, $\mathcal{Y} = \mathbb{R}^p$, $\mathcal{W} = \mathbb{R}^q$.

2. Assumptions

Suppose we want to construct an estimator of the stochastic process $\mathbf{y}(t) : \Omega \to \mathcal{Y}$ given a sequence of measurements as inputs obtained from the stochastic process $\mathbf{w}(t) : \Omega \to \mathcal{W}$. In order to narrow down and formally describe the estimation problem, we assume that the processes $\mathbf{y}(t)$ and $\mathbf{w}(t)$ can be represented as outputs of an LTI system in forward innovation form:

Assumption 1. The processes $\mathbf{y}(t)$ and $\mathbf{w}(t)$ can be generated by a stochastic discrete-time minimal LTI system of the form

$$\mathbf{x}(t+1) = A_g\mathbf{x}(t) + K_g\mathbf{e}_g(t) \tag{1a}$$

$$\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{w}(t)\end{bmatrix} = C_g\mathbf{x}(t) + \mathbf{e}_g(t), \quad Q = \mathbf{E}[\mathbf{e}_g^T(t)\mathbf{e}_g(t)] \tag{1b}$$

where $A_g \in \mathbb{R}^{n\times n}, K_g \in \mathbb{R}^{n\times m}, C_g = [C_{\mathbf{y}}^T, C_{\mathbf{w}}^T]^T \in \mathbb{R}^{(p+q)\times n}$ for $n \geq 0$, $m, p > 0$ and $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^p$, $\mathbf{w} \in \mathbb{R}^q$ and $\mathbf{e}_g$ are stationary, square-integrable, zero-mean, and jointly Gaussian stochastic processes. The processes $\mathbf{x}$ and $\mathbf{e}_g$ are called state and noise process, respectively. Furthermore, we require that A_g is stable (all its eigenvalues are inside the open unit circle) the stationary Gaussian process $\mathbf{e}_g(t)$ is white noise and uncorrelated with $\mathbf{x}(t-k)$. We identify the system (1) with the tuple $(A_g, K_g, C_g, I, \mathbf{e}_g)$; note that the state process $\mathbf{x}$ is uniquely defined by the infinite sum $\mathbf{x}(t) = \sum_{k=1}^{\infty} A_g^{k-1}K_g\mathbf{e}_g(t-k)$.

Before we can continue we have to consider the relationship between $\mathbf{y}$ and $\mathbf{w}$. For technical reasons we cannot have feedback from $\mathbf{y}$ to $\mathbf{w}$, as $\mathbf{w}$ would then be determined by a dynamical relation involving the past of the process $\mathbf{y}$. As such we have Assumption 2

Assumption 2. There is no feedback from $\mathbf{y}$ to $\mathbf{w}$, following definition 17.1.1. from Lindquist and Picci (2015), i.e., $H_t^-(\mathbf{y}) \perp H_t^+(\mathbf{w}) \mid H_t^-(\mathbf{w})$ holds. Furthermore, we assume that the spectral density of $\mathbf{w}$ is a coercive (Lindquist & Picci, 2015, Definition 9.4.1).

The no feedback assumption is equivalent to weak feedback free assumption (Caines, 1976) or Granger non-causality (Granger, 1963).

3. Result

Under Assumption 2, from Jozsa et al. (2018) it follows that there exists a similarity transformation T of (1) such that $\bar{A}_g = TA_gT^{-1}$, $\bar{K}_g = TK_g$ and $\bar{C}_g = C_gT^{-1}$ are upper block triangular, specifically (1) can be represented as

$$\bar{\mathbf{x}}(t) = \begin{bmatrix}A_{1,1} & A_{1,2}\\ 0 & A_{2,2}\end{bmatrix}\bar{\mathbf{x}}(t) + \begin{bmatrix}K_{1,1} & K_{1,2}\\ 0 & K_{2,2}\end{bmatrix}\begin{bmatrix}\mathbf{e}_1(t)\\ \mathbf{e}_2(t)\end{bmatrix} \tag{2a}$$

$$\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{w}(t)\end{bmatrix} = \begin{bmatrix}C_{1,1} & C_{1,2}\\ 0 & C_{2,2}\end{bmatrix}\bar{\mathbf{x}}(t) + \begin{bmatrix}\mathbf{e}_1(t)\\ \mathbf{e}_2(t)\end{bmatrix} \tag{2b}$$

where $[\mathbf{e}_1^T(t)\ \ \mathbf{e}_2^T(t)]^T = \mathbf{e}_g(t)$, and such that $(A_{2,2}, C_{2,2})$ is observable, and $A_{22} - K_{22}C_{22}$ is a stable (Schur-) matrix. Moreover, $A_{i,j} \in \mathbb{R}^{p_i\times p_j}$, $K_{i,j} \in \mathbb{R}^{p_i\times r_j}$, $C_{i,j} \in \mathbb{R}^{r_i\times p_j}$, with $r_1 = p$ and $r_2 = q$.

Then $\mathbf{y}$ can be represented as the output of the following stochastic LTI-ss driven by input $\mathbf{w}$:

$$\begin{aligned}\bar{\mathbf{x}}(t+1) &= \tilde{A}\bar{\mathbf{x}}(t) + \tilde{K}\mathbf{w}(t) + \tilde{K}_e\mathbf{e}_s(t)\\ \mathbf{y}(t) &= \tilde{C}\bar{\mathbf{x}}(t) - D_0\mathbf{w}(t) + \mathbf{e}_s(t)\end{aligned} \tag{3}$$

where

$$\begin{aligned}\tilde{A} &= \begin{bmatrix}A_{1,1} & A_{1,2} - (K_{1,2} + K_{1,1}D_0)C_{2,2}\\ 0 & A_{2,2} - K_{2,2}C_{2,2}\end{bmatrix},\\ \tilde{K} &= \begin{bmatrix}K_{1,2} + K_{1,1}D_0\\ K_{2,2}\end{bmatrix}, \quad \tilde{K}_e = \begin{bmatrix}K_{1,1}\\ 0\end{bmatrix}\\ \tilde{C} &= \begin{bmatrix}C_{1,1} & C_{1,2} - D_0C_{2,2}\end{bmatrix}, \quad D_0 = Q_{1,2}Q_{2,2}^{-1}.\end{aligned} \tag{4}$$

$$\mathbf{e}_s(t) = \mathbf{y}(t) - \mathbf{E}[\mathbf{y}(t)|H_t^-(\mathbf{y}) \vee H_{t+1}^-(\mathbf{w})]. \tag{5}$$

with the covariance $Q = \mathbf{E}[\mathbf{e}_g^T(t)\mathbf{e}_g(t)]$ partitioned according to (2). Here $\mathbf{e}_s(t)$ is the innovation process of $\mathbf{y}$ with respect to $\mathbf{w}$, i.e., it is the difference between $\mathbf{y}$ and the best possible prediction of $\mathbf{y}(t)$ based on past values of $\mathbf{y}$ and past and present values of $\mathbf{w}$. Under Assumption 2 from Lindquist and Picci (2015, Ch. 17, Proposition 17.1.3) it follows that

$$\mathbf{e}_s(t) = \mathbf{y}_s(t) - \mathbf{E}[\mathbf{y}_s(t)|H_t^-(\mathbf{y}_s) \vee H(\mathbf{w})]. \tag{6a}$$

In particular, by using Lindquist and Picci (2015, Chapter 17, (17.31)), it follows that the LTI-ss (4) in innovation form with input $\mathbf{w}$ gives rise to the following optimal, **in the least squared sense**, predictor of $\mathbf{y}$ based on $\{\mathbf{w}(s)\}_{s\leq t}$:

$$\hat{\mathbf{x}}(t+1) = \tilde{A}\hat{\mathbf{x}}(t) + \tilde{K}\mathbf{w}(t) \tag{7a}$$

$$\hat{\mathbf{y}}(t) = \tilde{C}\hat{\mathbf{x}}(t) - D_0\mathbf{w}(t) \tag{7b}$$

where, $\hat{\mathbf{y}}(t) = \mathbf{E}[\mathbf{y}(t) \mid H_{t+1}^-(\mathbf{w})]$, and $\hat{\mathbf{x}}(t) = \mathbf{E}[\bar{\mathbf{x}}(t) \mid H_{t+1}^-(\mathbf{w})]$.

4. Proof

Now consider a similarity transformation T of (1) such that $\bar{A}_g = TA_gT^{-1}$, $\bar{K}_g = TK_g$ and $\bar{C}_g = C_gT^{-1}$ are upper block triangular, see (2). From Jozsa et al. (2018) it then follows that $(A_{2,2}, K_{2,2}, C_{2,2}, \mathbf{e}_2)$ is a minimal LTI-ss representation of $\mathbf{w}$ in innovation form. Hence $\mathbf{e}_2(t)$ is the innovation process of $\mathbf{w}$ i.e., $\mathbf{e}_2(t) = \mathbf{w}(t) - \mathbf{E}[\mathbf{w}(t) \mid H_t^-(\mathbf{w})]$. But from Lindquist and Picci (2015, Proposition 2.4.2), Assumption 2 implies that $\mathbf{E}[\mathbf{w}(t)|H_t^-(\mathbf{w}) \vee H_t^-(\mathbf{y})] = \mathbf{E}[\mathbf{w}(t)|H_t^-(\mathbf{w})]$, hence

$$\mathbf{e}_2(t) = \mathbf{w}(t) - \mathbf{E}[\mathbf{w}(t) \mid H_t^-(\mathbf{y}) \vee H_t^-(\mathbf{w})]. \tag{8}$$

Moreover, the transformed system (2) induces a relation between the output $\mathbf{y}$ and input $\mathbf{w}$. In detail, from (2b) we also have

$$\mathbf{e}_2(t) = \mathbf{w}(t) - C_{2,2}\bar{\mathbf{x}}_2(t). \tag{9}$$

Hence, substituting (9) in (2) yields the following realisation of $\mathbf{y}$

$$\begin{bmatrix}\bar{\mathbf{x}}_1(t+1)\\ \bar{\mathbf{x}}_2(t+1)\end{bmatrix} = \begin{bmatrix}A_{1,1} & A_{1,2}-K_{1,2}C_{2,2}\\ 0 & A_{2,2}-K_{2,2}C_{2,2}\end{bmatrix}\begin{bmatrix}\bar{\mathbf{x}}_1(t)\\ \bar{\mathbf{x}}_2(t)\end{bmatrix} \tag{10a}$$

$$+\begin{bmatrix}K_{1,2}\\ K_{2,2}\end{bmatrix}\mathbf{w}(t) + \begin{bmatrix}K_{1,1}\\ 0\end{bmatrix}\mathbf{e}_1(t) \tag{10b}$$

$$\mathbf{y}(t) = \begin{bmatrix}C_{1,1} & C_{1,2}\end{bmatrix}\begin{bmatrix}\bar{\mathbf{x}}_1(t)\\ \bar{\mathbf{x}}_2(t)\end{bmatrix} + \mathbf{e}_1(t) \tag{10c}$$

Note that $\mathbf{e}_1(t)$ is the innovation process of $\mathbf{y}$ (with respect to $\mathbf{w}$), i.e.,

$$\mathbf{e}_1(t) = \mathbf{y}(t) - \mathbf{E}[\mathbf{y}(t) \mid H_t^-(\mathbf{y}) \vee H_t^-(\mathbf{w})]. \tag{11}$$

Firstly, we claim that

$$\mathbf{e}_s(t) = \mathbf{e}_1(t) - \mathbf{E}[\mathbf{y}(t)|\mathbf{e}_2(t)] = \mathbf{e}_1(t) - D_0\mathbf{e}_2(t) \tag{12}$$

where[1] $D_0 = (\mathbf{E}[\mathbf{y}(t)\mathbf{e}_2^T(t)])^T(\mathbf{E}[\mathbf{e}_2(t)\mathbf{e}_2^T(t)])^{-1}$ is the minimum variance linear estimator of $\mathbf{y}(t)$ given $\mathbf{e}_2(t)$, see Lindquist and Picci (2015, Proposition 2.2.3.). In order to show (12), we first demonstrate that

$$H_t^-(\mathbf{y}) \vee H_{t+1}^-(\mathbf{w}) = (H_t^-(\mathbf{y}) \vee H_t^-(\mathbf{w})) \oplus H(\mathbf{e}_2(t)) \tag{13}$$

where $H(\mathbf{e}_2(t)) = \{\alpha^T\mathbf{e}_2(t) \mid \alpha \in \mathbb{R}^q\}$, is the space spanned by innovation process $\mathbf{e}_2(t)$, considered only at the time t. But from (8) it follows that the components of $\mathbf{e}_2(t)$ belong to $H_t^-(\mathbf{y}) \vee H_{t+1}^-(\mathbf{w})$. Hence,

$$(H_t^-(\mathbf{y}) \vee H_t^-(\mathbf{w})) \vee H(\mathbf{e}_2(t)) = H_t^-(\mathbf{y}) \vee H_{t+1}^-(\mathbf{w}).$$

Again from (8) it follows that $\mathbf{e}_2(t) \perp (H_t^-(\mathbf{w}) \vee H_t^-(\mathbf{y}))$, thus (13) holds. The relation (12) now follows since

$$\begin{aligned}&\mathbf{E}[\mathbf{y}(t) \mid H_t^-(\mathbf{y}) \vee H_{t+1}^-(\mathbf{w})]\\ &\quad= \mathbf{E}[\mathbf{y}(t) \mid H_t^-(\mathbf{y}) \vee H_t^-(\mathbf{w})] + \mathbf{E}[\mathbf{y}(t) \mid \mathbf{e}_2(t)]\\ &\quad= \mathbf{E}[\mathbf{y}(t) \mid H_t^-(\mathbf{y}) \vee H_t^-(\mathbf{w})] + D_0\mathbf{e}_2(t),\end{aligned} \tag{14}$$

and therefore, using (11) we can see that

$$\mathbf{e}_s(t) = \mathbf{e}_1(t) - D_0\mathbf{e}_2(t)$$

Now from (12) and (9) we get

$$\mathbf{e}_1(t) = \mathbf{e}_s(t) + D_0\mathbf{w}(t) - D_0C_{2,2}\bar{\mathbf{x}}_2(t),$$

which can be applied to (10) to obtain the realisation of $\mathbf{y}$ described in (3). Finally we are in a position to derive a formula for the minimum error variance estimate $\mathbf{E}[\mathbf{y}(t) \mid H_{t+1}^-(\mathbf{w})]$. That is, a formula for the orthogonal projection of $\mathbf{y}(t)$ given past and present values of $\mathbf{w}$. First define $\hat{\mathbf{x}}_g(t) = \mathbf{E}[\bar{\mathbf{x}}(t) \mid H_{t+1}^-(\mathbf{w})]$, then from (3) we get

$$\begin{aligned}&\mathbf{E}[\mathbf{y}(t) \mid H_{t+1}^-(\mathbf{w})]\\ &\quad= \tilde{C}\hat{x}_g(t) + D_0\mathbf{w}(t) + \mathbf{E}[\mathbf{e}_s(t)|H_{t+1}^-(\mathbf{w})] &(15)\\ &\quad= \tilde{C}\hat{x}_g(t) + D_0\mathbf{w}(t) &(16)\end{aligned}$$

where (16) follows from (5). Now (3) can be used to derive a dynamical expression for $\hat{\mathbf{x}}_g$ as follows

$$\begin{aligned}&\mathbf{E}[\bar{\mathbf{x}}(t+1) \mid H_{t+2}^-(\mathbf{w})]\\ &= \mathbf{E}\left[\tilde{A}\bar{\mathbf{x}}(t) + \tilde{K}\mathbf{w}(t) + \tilde{K}_e\mathbf{e}_s(t)\middle|H_{t+2}^-(\mathbf{w})\right]\end{aligned} \tag{17}$$

Clearly $\mathbf{E}[\mathbf{w}(t)|H_{t+2}^-(\mathbf{w})] = \mathbf{w}(t)$. For the state projection in (17) we have $\mathbf{E}[\bar{\mathbf{x}}(t)|H_{t+2}^-(\mathbf{w})] = \mathbf{E}[\bar{\mathbf{x}}(t)|H_{t+1}^-(\mathbf{w})]$. Indeed, the state vector $\bar{\mathbf{x}}(t)$ can be expressed as an infinite sum

$$\bar{\mathbf{x}}(t) = \sum_{i=1}^{\infty}\tilde{C}\tilde{A}^{i-1}\tilde{K}\mathbf{w}(t-i) + \sum_{i=1}^{\infty}\tilde{C}\tilde{A}^{i-1}\tilde{K}_e\mathbf{e}_s(t-i)$$

and hence for $r = 1, 2$

$$\begin{aligned}\mathbf{E}[\bar{\mathbf{x}}(t)|H_{t+r}^-(\mathbf{w})] &= \sum_{i=1}^{\infty}\tilde{C}\tilde{A}^{i-1}\tilde{K}\mathbf{E}[\mathbf{w}(t-i)|H_{t+r}^-(\mathbf{w})]\\ &\quad+\sum_{i=1}^{\infty}\tilde{C}\tilde{A}^{i-1}\tilde{K}_e\mathbf{E}[\mathbf{e}_s(t-i)|H_{t+r}^-(\mathbf{w})]\end{aligned}$$

Note that $\mathbf{E}[\mathbf{w}(t-i)|H_{t+r}^-(\mathbf{w})] = \mathbf{w}(t-i)$, $r = 1, 2$. Moreover, from (6a) we observe that $\mathbf{E}[\mathbf{e}_s(t-i)|H_{t+r}^-(\mathbf{w})] = 0$, $i \geq 1$, $r = 1, 2$. Hence,

$$\begin{aligned}\mathbf{E}[\bar{\mathbf{x}}(t)|H_{t+1}^-(\mathbf{w})] &= \sum_{i=1}^{\infty}\tilde{C}\tilde{A}^{i-1}\tilde{K}\mathbf{w}(t-i)\\ &= \mathbf{E}[\bar{\mathbf{x}}(t)|H_{t+2}^-(\mathbf{w})]\end{aligned}$$

Finally we have obtained (4) the formula for the minimum prediction error variance estimate of $\mathbf{y}(t)$ based on $H_{t+1}^-(\mathbf{w})$ (present and past of process $\mathbf{w}$) by using Lindquist and Picci (2015, (17.13a), Chapter 17).

5. Potential applications and examples

As it was mentioned before, the contribution of the paper was motivated by its role in computing parametrisations of LTI-ss predictors of $\mathbf{y}$ using $\mathbf{w}$. In turn, these parametrisations can be useful for developing system identification algorithms, or for applying PAC-Bayesian methods to LTI-ss representations. In order to elaborate on these applications, we review the basic formulation of the system identification problem.

Assume that the data used for system identification will be a sample $\{y(t) = \mathbf{y}(t)(\omega), w(t) = \mathbf{w}(t)(\omega)\}_{t=0}^N$, $\omega \in \Omega$ of $(\mathbf{y}, \mathbf{w})$. Consider a parameterised set of LTI systems $\{\Sigma(\theta) = (\hat{A}(\theta), \hat{B}(\theta), \hat{C}(\theta), \hat{D}(\theta), \mathbf{w})\}_{\theta\in\Theta}$ which serve as predictors, i.e., $\Sigma(\theta)$ generates a prediction $\hat{\mathbf{y}}_\theta(t)$ of output $\mathbf{y}(t)$ based on past values of $\mathbf{w}$ as follows

$$\begin{aligned}\hat{\mathbf{x}}_\theta(t+1) &= \hat{A}(\theta)\hat{\mathbf{x}}_\theta(t) + \hat{B}(\theta)\mathbf{w}(t)\\ \hat{\mathbf{y}}_\theta(t) &= \hat{C}(\theta)\hat{\mathbf{x}}_\theta(t) + \hat{D}(\theta)\mathbf{w}(t)\end{aligned}$$

Let us define the *prediction error* $V(\theta, \mathbf{y}, \mathbf{w}) = \mathbf{E}[\|\hat{\mathbf{y}}_\theta(t) - \mathbf{y}(t)\|_2^2]$.

Informally, system identification algorithms aim at finding a parameter value $\hat{\theta}_N$ so that the prediction error $V(\hat{\theta}_N)$ is small enough. Often we have a prior knowledge only on the class of systems which generated $(\mathbf{y}, \mathbf{w})$, but not on the predictors. Hence,

[1] In order to numerically compute D_0, we can use (11) to substitute $\mathbf{y}$, and since $\mathbf{e}_2 \perp H_t^-(\mathbf{w}) \vee H_t^-(\mathbf{y})$, we get $\mathbf{E}[\mathbf{E}[\mathbf{y}(t) \mid H_t^-(\mathbf{y}) \vee H_t^-(\mathbf{w})]\mathbf{e}_2^T] = 0$. Therefore $\mathbf{E}[\mathbf{y}(t)\mathbf{e}_2^T(t)] = \mathbf{E}[\mathbf{e}_1(t)\mathbf{e}_2^T(t)]$. Thus, one can compute D_0 from the covariance of innovation noise, i.e. $D_0 = Q_{1,2}Q_{2,2}^{-1}$.

in order to choose a parametrisation of predictors, we need a transformation from data generators to predictors.

This paper provides such a transformation. More precisely, assume that we have a parameterised set of models $\{\Sigma_{gen}(\theta) = (A_g(\theta), K_g(\theta), C_g(\theta), I, \mathbf{e}_{g,\theta})\}_{\theta\in\Theta}$ which could potentially generate the signal $(\mathbf{y}, \mathbf{w})$ as outputs, i.e. there exists $\theta_0 \in \Theta$, such that $\Sigma_{gen}(\theta_0)$ is a realisation of $(\mathbf{y}, \mathbf{w})$. We can then use the results of this paper to transform each system $\Sigma_{gen}(\theta)$ whose output is $(\mathbf{y}_\theta, \mathbf{w}_\theta)$ to an optimal predictor $\Sigma(\theta)$ which generates the best prediction $\hat{\mathbf{y}}_\theta(t)$ of $\mathbf{y}_\theta(t)$ based on$\{\mathbf{w}_\theta(s)\}_{s<t}$ for every $t \in \mathbb{Z}$. In particular, the predictor with the smallest prediction error is Σ_{θ_0}, i.e., it is the one which arises from the data generating system.

Moreover, if the dependence of the generators on the parameters is simple the resulting parameterised predictors will also have a relatively simple dependence on parameters. To illustrate this point, let us consider the following parametrisation of generating systems in forward innovation form $\{\Sigma_{gen}(\theta) = (A_g(\theta), K_g(\theta), C_g(\theta), I, \mathbf{e}_{g,\theta})\}_{\theta\in\Theta}$, where $\Theta \subseteq \mathbb{R}^{20}$, $Q(\theta) = \mathbf{E}[\mathbf{e}_{g,\theta}(t)\mathbf{e}_{g,\theta}^T(t)]$, and

$$A_g(\theta) = \begin{bmatrix} a_1(\theta) & a_3(\theta) & \theta_5 \\ a_2(\theta) & a_4(\theta) & \theta_6 \\ a_5(\theta) & a_6(\theta) & a_7(\theta) \end{bmatrix},$$

$$K_g(\theta) = \begin{bmatrix} \theta_7 & \theta_9 & \theta_{11} \\ \theta_8 & \theta_{10} & \theta_{12} \\ k_1(\theta) & k_2(\theta) & k_3(\theta) \end{bmatrix}$$

$$C_g(\theta) = \begin{bmatrix} \theta_{13}+\theta_{17} & \theta_{15}+\theta_{17} & \theta_{17} \\ \theta_{14}+\theta_{18} & \theta_{16}+\theta_{18} & \theta_{18} \\ c_7 & c_7 & c_7 \end{bmatrix}, Q(\theta) = \begin{bmatrix} q_1 & q_4 & \theta_{19} \\ q_2 & q_5 & \theta_{20} \\ \theta_{19} & \theta_{20} & q_9 \end{bmatrix}$$

where $a_i(\theta) = \theta_i + \theta_5$, $i \in \{1, 3\}$, $a_i(\theta) = \theta_i + \theta_6$, $i \in \{2, 4\}$ $a_5 = a7 - \theta_1 - \theta_2 - \theta_5 - \theta_6$, $a_6 = a7 - \theta_3 - \theta_4 - \theta_5 - \theta_6$, $a_7 = a7 - \theta_5 - \theta_6$, $k_1(\theta) = \theta_7 - \theta_8$, $k_2(\theta) = \theta_9 - \theta_{10}$, $k_3(\theta) = k_7 - \theta_{11} - \theta_{12}$, and a_7, k_7, c_7, q_9 are some constants. We assume that $\mathbf{y}(t)$ takes values in $\mathbb{R}^2$ and $\mathbf{w}(t)$ in $\mathbb{R}$. The construction of the paper leads then to the following parametrisation of predictors $\{\Sigma(\theta) = (\tilde{A}(\theta), \tilde{K}(\theta), \tilde{C}(\theta), D_0(\theta), \mathbf{w})\}_{\theta\in\Theta}$, where

$$\tilde{A}(\theta) = \begin{bmatrix} \theta_1 & \theta_3 & \theta_5 - c_7\,\theta_{11} - \frac{c_7\,\theta_7\,\theta_{19}}{q_9} - \frac{c_7\,\theta_9\,\theta_{20}}{q_9} \\ \theta_2 & \theta_4 & \theta_6 - c_7\,\theta_{12} - \frac{c_7\,\theta_8\,\theta_{19}}{q_9} - \frac{c_7\,\theta_{10}\,\theta_{20}}{q_9} \\ 0 & 0 & a_7 - c_7\,k_7 \end{bmatrix},$$

$$\tilde{K}(\theta) = \begin{bmatrix} \theta_{11} + \frac{\theta_7\,\theta_{19}}{q_9} + \frac{\theta_9\,\theta_{20}}{q_9} \\ \theta_{12} + \frac{\theta_8\,\theta_{19}}{q_9} + \frac{\theta_{10}\,\theta_{20}}{q_9} \\ k_7 \end{bmatrix},$$

$$\tilde{C}(\theta) = \begin{bmatrix} \theta_{13} & \theta_{15} & \theta_{17} - \frac{c_7\,\theta_{19}}{q_9} \\ \theta_{14} & \theta_{16} & \theta_{18} - \frac{c_7\,\theta_{20}}{q_9} \end{bmatrix}, \quad D_0(\theta) = \begin{bmatrix} \frac{\theta_{19}}{q_9} \\ \frac{\theta_{20}}{q_9} \end{bmatrix}$$

The dependence of $\Sigma(\theta)$ on θ is simple. If we use the existing literature (Gevers & Anderson, 1982), then we have to compute the transfer function $W_\theta(z)$ of the state-space representation $\Sigma_g(\theta)$, use it to compute a transfer function from $\mathbf{w}$ to $\mathbf{y}$ and then transform the latter transfer function to a state-space representation. The entries of $W_\theta(z)$ are already polynomials of θ, degree of which is at least 3. The subsequent transformations increase the complexity of the dependence on the parameter θ even further.

PAC-Bayesian methods The relative simplicity of the arising parametrisation of predictors is especially useful for applying *PAC-Bayesian bounds (Alquier, 2021; Eringis, Leth, Tan, Wisniewski, Esfahan et al., 2021)*, i.e., probabilistic inequalities of the form

$$\mathbf{P}\left(E_{f\sim\hat{\rho}}V(\theta) \le E_{\theta\sim\hat{\rho}}V_N(\theta) + \Psi_N(\delta)\right) > 1 - \delta$$

where $\hat{\rho}$ is a probability density on the parameter set Θ which depends on $\mathbf{y}$ and $\mathbf{w}$, $E_{\theta\sim\hat{\rho}}$ is the expectation operator wrt., $\hat{\rho}$, and $V_N(\theta) = \frac{1}{N}\sum_{t=0}^{N} \|\tilde{\mathbf{y}}_\theta(t) - \mathbf{y}(t)\|^2$ is the so called *empirical loss*. The term $\Psi_N(\delta)$ is the subject of research efforts on PAC-Bayesian bounds (Alquier, 2021; Eringis, Leth, Tan, Wisniewski, Esfahan et al., 2021; Eringis et al., 2022), its precise form is quite involved. Informally, $\hat{\rho}$ is the posterior density on the parameter values, which has been obtained from some prior density using the observed data and a form of Bayesian inference, i.e., $\hat{\rho}$ depends on $\{\mathbf{y}(t), \mathbf{w}(t)\}_{t=0}^N$ and it has been chosen so that the average empirical loss $E_{\theta\sim\hat{\rho}}V_N(\theta)$ is small enough. A system identification algorithm then corresponds to sampling a random element from $\hat{\rho}$, or by taking the maximal likelihood $\text{argmax}_{\theta\in\Theta}\hat{\rho}(\theta)$, see Alquier (2021) for a detailed discussion. PAC-Bayesian bounds provide guarantees for the prediction error for the thus obtained system identification algorithms.

The usefulness of PAC-Bayesian bounds depends on computing non-conservative estimates of the term $\Psi_N(\delta)$. The latter involves computing averages with respect to various densities on Θ of the matrices of the predictors (Alquier, 2021; Eringis, Leth, Tan, Wisniewski, Esfahan et al., 2021). Hence, in order to compute the estimates of $\Psi_N(\delta)$, the predictors should be functions of the parameters which are easy to compute. As the example above shows, the construction of this paper helps to obtain such parametrisations. The construction of this paper was in fact used to compute $\Psi_N(\delta)$ in Eringis et al. (2022).

System identification algorithm The transformation presented in this paper leads to alternative identification algorithms. Namely, one can use any of-the-shelf identification algorithm for identifying an autonomous stochastic LTI system realising $(\mathbf{y}, \mathbf{w})$ and then apply the transformation of the paper to the identified model in order to obtain a predictor. The description above defines a class of system identification algorithms, as every choice of the of-the-shelf identification algorithm results in a different predictor. We expect the consistency analysis of such algorithms could be easier: while consistency of autonomous stochastic LTI state-space representations is relatively well-understood (Lindquist & Picci, 2015), the case of systems with inputs is more involved (Chiuso & Picci, 2004; Katayama, 2005; Lindquist & Picci, 2015).

We evaluated the proposed system identification algorithm numerically: we generated data using generating systems of dimension 10, with output $\mathbf{y}(t)$ taking values in $\mathbb{R}^3$ and $\mathbf{w}(t)$ taking values in $\mathbb{R}^2$. Moreover, we considered two scenarios. In the first scenario, both the generating system and the predictors are complete black-boxes. In the second scenario, we know the subsystem of the generating system which generates $\mathbf{w}$. Hence, in the parametrisation of the generators the subsystem which generates $\mathbf{w}$ is known. For the sake of fairness, we constructed the parametrisation of the predictors from that of the generators, using the method of this paper. Numerical simulations reveal that the system identification algorithm which uses the construction of this paper performs similarly to the standard method, and it may outperform the standard one when the system generating $\mathbf{w}$ is known. The detailed description of the experiments can be found in Eringis, Leth, Tan, Wisniewski, Petreczky (2021).

6. Conclusion

We have proposed an explicit transformation from a joint LTI-ss representation of $(\mathbf{y}, \mathbf{w})^T$ to a stochastic LTI-ss representation of $\mathbf{y}$ in innovation form, driven by $\mathbf{w}$. We also discussed the potential applications of such a transformation to system identification. Note that the basic ingredients of the obtained results were realisation theory and the notion of Granger-causality (weak feedback free). Hence, the results could potentially be extended to other system classes for which these ingredients exist, such as bilinear systems (Jozsa, Petreczky, & Camlibel, 2020).

References

Alquier, Pierre (2021). User-friendly introduction to PAC-Bayes bounds. arXiv: 2110.11216.

Alquier, P., Ridgway, J., & Chopin, N. (2016). On the properties of variational approximations of gibbs posteriors. *Journal of Machine Learning Research*, *17*(239), 1–41.

Caines, P. (1976). Weak and strong feedback free processes. *IEEE Transactions on Automatic Control*, *21*(5), 737–739.

Caines, P. E. (1988). *Linear stochastic systems*. John Wiley and Sons.

Caines, P., & Chan, C. (1975). Feedback between stationary stochastic processes. *IEEE Transactions on Automatic Control*, *20*(4), 498–508.

Chiuso, Alessandro, & Picci, Giorgio (2004). On the ill-conditioning of subspace identification with inputs. *Automatica*, *40*(4), 575–589.

Eringis, D., Leth, J., Tan, Z.-H., Wisniewski, R., Esfahan, A. F., & Petreczky, M. (2021). PAC-Bayesian theory for stochastic LTI systems. In *2021 60th IEEE conference on decision and control* (pp. 6626–6633).

Eringis, D., Leth, J., Tan, Z.-H., Wisniewski, R., & Petreczky, M. (2021). Explicit construction of the minimum error variance estimator for stochastic LTI state-space systems. http://dx.doi.org/10.48550/ARXIV.2109.02384v3, arXiv: 2109.02384v3.

Eringis, Deividas, Leth, John, Tan, Zheng-Hua, Wisniewski, Rafal, & Petreczky, Mihaly (2022). PAC-Bayesian-like error bound for a class of linear time-invariant stochastic state-space models. http://dx.doi.org/10.48550/ARXIV.2212.14838.

Gevers, Michel, & Anderson, Brian D. O. (1982). On jointly stationary feedback-free stochastic processes. *IEEE Transactions on Automatic Control*, *27*, 431–436.

Granger, C. W. J. (1963). Economic processes involving feedback. *Information and Control*, *6*(1), 28–48.

Jozsa, Monika, Petreczky, Mihály, & Camlibel, M. Kanat (2018). Relationship between granger noncausality and network graph of state-space representations. *IEEE Transactions on Automatic Control*, *64*(3), 912–927.

Jozsa, Monika, Petreczky, Mihaly, & Camlibel, M. Kanat (2020). Causality and network graph in general bilinear state-space representations. *IEEE Transactions on Automatic Control*, *65*(8), 3623–3630.

Katayama, T. (2005). *Subspace methods for system identification*. Springer-Verlag.

Lindquist, A., & Picci, G. (2015). *Linear stochastic systems: a geometric approach to modeling, estimation and identification*. Springer.

van Schuppen, J. H. (2021). *Communications and control engineering, Control and system theory of discrete-time stochastic systems*. Springer International Publishing.