Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 55-30 (2022) 371–376

Data-driven approximation and reduction from noisy data in matrix pencils frameworks

Pauline Kergus * **Ion Victor Gosea** **

* CNRS, LAPLACE, Toulouse, France (e-mail: pauline.kergus@laplace.univ-tlse.fr).
** Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany (e-mail: gosea@mpi-magdeburg.mpg.de)

Abstract: This work aims at tackling the problem of learning surrogate models from noisy time-domain data by means of matrix pencil-based techniques, namely the Hankel and Loewner frameworks. A data-driven approach to obtain reduced order state-space models from time-domain input-output measurements for linear time-invariant (LTI) systems is proposed. This is accomplished by combining the aforementioned model order reduction (MOR) techniques with the signal matrix model (SMM) approach. The proposed method is illustrated by a numerical example consisting of a high-order building model.

Keywords: model order reduction, system identification, Loewner interpolation

1. INTRODUCTION

Numerous complex dynamical systems used in practical applications cannot be accurately described by physical models that are simple enough to be simulated or to be used for control design purposes. Model order reduction techniques then play a crucial role in obtaining a suitable complexity-accuracy trade-off. As recalled in Antoulas (2005), MOR is usually based on full knowledge of a complex and high-fidelity system description, derived from physics laws.

However, the increasing availability of data and the rise of data-driven applications require the incorporation of measurements when modeling or controlling a system. To that extent, data-driven reduction techniques, such as the Loewner Framework (LF) Mayo and Antoulas (2007), Vector Fitting (VF) Gustavsen and Semlyen (1999) or Adapative Antoulas Anderson (AAA) Nakatsukasa et al. (2018), are particularly appealing.

This paper focuses on the LF, which was mostly applied (with some exceptions) for noise-free data, obtained by simulating a high-fidelity model of the dynamical system under investigation. Indeed, as pointed out in Lefteriu et al. (2010), LF is quite sensitive to noisy (perturbed) data. Physical modes of the system may only be included in the model at the expense of overmodeling, which generally leads to high variances and overfitting. As a result, noisy data complicates the selection of the system's order and may lead to high approximation errors. To tackle this issue, in Lefteriu et al. (2010) the poles were selected according to their residue norm to make the Loewner framework more robust to noise. This approach has also been used in the context of data-driven control based on the LF in Kergus et al. (2018). In Ionita (2013), it is suggested that the choice of the frequencies (interpolation points), as well as the partition of the corresponding data points, impacts the robustness with respect to noise. This idea was also explored in Gosea et al. (2021) and Palitta and Lefteriu (2022), where different partitioning were studied for various numerical experiments. In Embree and Ioniţă (2022), the influence of the location and partition of the data points was studied through the pseudospectrum of the Loewner pencil. In Drmač and Peherstorfer (2019), it was shown that for Gaussian noise, the resulting Loewner model error grows at most linearly with the standard deviation of noise.

This work primarily aims at proposing a way to obtain reduced-order models (ROMs) through matrix pencils techniques, namely the LF and Hankel Framework (HF), that is more robust to noisy data. The objective is hence to enable the use of such techniques to obtain a ROM from measurements. In what we propose, the order is a tunable parameter, without considering available access to a high fidelity representation. To that extent, this work is at the crossroads of MOR and system identification (SI). The proposed method is based on matrix pencils approaches (LF and HF). The HF is rooted in realization theory Schutter (2000), in since it constructs a minimal LTI realization from Markov parameters, i.e. impulse response of discrete-time systems. Therefore, HF can be seen as a time-domain counterpart of LF Ionita and Antoulas (2012). In practice, the impulse response often has to be estimated from available input-output data. This is usually done through least-squares-based linear regression. In this work, we propose strategies for making this approach more robust to noisy data by using the SMM method introduced in Yin et al. (2020) (which allows estimating the impulse response of a system from noisy data).

To sum up, the proposed approach brings together aspects from MOR, realization theory and SI in an unified framework, which constitutes the main contribution of this work. Time-domain data, consisting of noisy input-output measurements, is used to estimate the finite impulse response of the system as in Yin et al. (2020), which constitutes a non parametric characterization of the underlying LTI system. The finite impulse response is then used to obtain a reduced-order, explicit model through the HF/LF.

10.1016/j.ifacol.2022.11.081

The rest of the paper is organized as follows. In Section 2, the problem under investigation is formulated. Here, the Loewner and Hankel frameworks, which constitute the basis of this work, are recalled. The proposed approach is then introduced in Section 3 and a detailed specification on tuning its hyperparameters is also provided. This method is then illustrated by a numerical example in Section 4, which is the Los Angeles Hospital building benchmark from the COMP$l_e ib$ library Leibfritz (2004), described by a 48th-order state-space model. Finally, the conclusion and outlook are discussed in Section 5.

2. PRELIMINARIES

2.1 Problem formulation

We consider an LTI discrete-time system with n_u inputs, n_y outputs and of order n_x, characterized by the following state-space realization:

$$\mathbf{\Sigma} : \begin{cases} \mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}u_t \\ y_t = \mathbf{C}\mathbf{x}_t + \mathbf{D}u_t \end{cases}, \tag{1}$$

with $\mathbf{x} \in \mathbb{R}^{n_x}$ the state vector, $\mathbf{u} \in \mathbb{R}^{n_u}$ the input vector, $\mathbf{y} \in \mathbb{R}^{n_y}$ the output vector, $\mathbf{A} \in \mathbb{R}^{n_x \times n_x}$, $\mathbf{B} \in \mathbb{R}^{n_x \times n_u}$, $\mathbf{C} \in \mathbb{R}^{n_y \times n_x}$ and $\mathbf{D} \in \mathbb{R}^{n_y \times n_u}$. The value of a vector v at the time step t is denoted $\mathbf{v}_t$.

The transfer function of $\mathbf{\Sigma}$ (1) is given by:

$$\mathbf{H}(z) = \mathbf{D} + \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}. \tag{2}$$

The Loewner Framework (LF) Mayo and Antoulas (2007), recalled in Section 2.2, can identify the underlying system from noise-free frequency-domain samples $\mathbf{H}(e^{\imath\omega_i})$ in (2). The Hankel Framework (HF) Schutter (2000), summarized in Section 2.3, relies on the impulse response $\{\mathbf{h}_k\}$ that connects the input and output samples as follows:

$$y_t = \sum_{k=-\infty}^{\infty} \mathbf{h}_k u_{t-k}. \tag{3}$$

The HF could be considered as a time-domain counterpart of the LF, as the frequency-domain representation (2) and the time-domain one (1) are connected: the impulse response coefficients, also known as Markov parameters, are defined as follows:

$$\mathbf{h}_k = \begin{cases} \mathbf{D}, & \text{if } k = 0, \\ \mathbf{C}\mathbf{A}^{k-1}\mathbf{B}, & \text{if } k > 0. \end{cases} \tag{4}$$

Note that $\mathbf{h}_k = 0$ for $k < 0$ as the systems under consideration are causal. The transfer function (2) can then be written as an Infinite Impulse Response (IIR) filter:

$$\mathbf{H}(z) = \sum_{k=0}^{\infty} \mathbf{h}_k z^{-k}. \tag{5}$$

These matrix pencils techniques (LF and HF) also allow to reduce the order of the obtained models in a straightforward manner.

Remark 1. (Descriptor/state-space forms). Both the LF and HF are inherently leading to a descriptor model formulation $(\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$:

$$\hat{\mathbf{\Sigma}} : \begin{cases} \hat{\mathbf{E}}\hat{\mathbf{x}}_{t+1} = \hat{\mathbf{A}}\hat{\mathbf{x}}_t + \hat{\mathbf{B}}u_t \\ y_t = \hat{\mathbf{C}}\hat{\mathbf{x}}_t + \hat{\mathbf{D}}u_t \end{cases}. \tag{6}$$

In practice, due to the reduction process, the $\hat{\mathbf{E}}$ matrix in (6) is full rank and therefore invertible. It is then possible to rewrite (6) in a standard state-space form as in (1), with the matrices $\big(\mathbf{E}^{-1}\mathbf{A}, \mathbf{E}^{-1}\mathbf{B}, \mathbf{C}, \mathbf{D}\big)$. The same considerations hold for HF.

While the LF and HF techniques have proven to be fairly successful when applied to MOR of given (large-scale) complex systems, they are indeed known to be quite sensitive to noisy data Lefteriu et al. (2010). The problem under consideration that is tackled in this paper is formulated below:

Problem 1. Given noisy data, how to obtain a linear reduced-order approximation of the underlying dynamical system through the LF or the HF frameworks?

In a sense, by using noisy data through these techniques, we more generally aim at bridging the gap between MOR, in which the underlying system is known but of complex or large-scale structure, and SI, which aims at building models from (noisy) measurements. As the proposed approach relies on the LF and HF frameworks, we will briefly summarize them in the two next subsections.

Remark 2. (Time-domain LF). Another time-domain counterpart of LF has been proposed in Peherstorfer et al. (2017): based on noise-free time-domain data $\{u_k, y_k\}_k$ and on the knowledge of a high-fidelity model, frequency-domain data is inferred to use the LF. In comparison, the present work proposes to obtain a non-parametric characterization of the system by estimating its Markov parameters, from which frequency-domain data can be inferred to be used in the LF. Contrary to Peherstorfer et al. (2017), the proposed approach does not require any description of the system and is more robust to noise.

2.2 The Loewner framework

Here, we briefly review the LF approach, see Antoulas et al. (2017) for more details. The LF is based on frequency-domain measurements $\{\mathbf{H}(z_k)\}_{k=1}^{N}$ corresponding to the transfer function (2), and finds a state-space model $\hat{\mathbf{H}}$ such that the following interpolation conditions are (approximately) fulfilled:

$$\hat{\mathbf{H}}(z_k) = \mathbf{H}(z_k)\ \forall k = 1 \dots N. \tag{7}$$

The available data is partitioned into two disjoint subsets, $\{\mathbf{H}(z_i)\}_{i=1}^{\frac{N}{2}}$ and $\{\mathbf{H}(z_j)\}_{j=1}^{\frac{N}{2}}$. The Loewner pencil $(\mathbb{L}, \mathbb{L}_s)$ is defined as follows

$$\mathbb{L}_{(i,j)} = \frac{\mathbf{H}(z_i) - \mathbf{H}(z_j)}{z_i - z_j},\ \mathbb{L}_{s\,(i,j)} = \frac{z_i\mathbf{H}(z_i) - z_j\mathbf{H}(z_j)}{z_i - z_j}, \tag{8}$$

while the data vectors $\mathbb{V}, \mathbb{W}^T \in \mathbb{R}^k$ are introduced as

$$\mathbb{V}_{(i)} = \mathbf{H}(z_i),\ \ \mathbb{W}_{(j)} = \mathbf{H}(z_j),\ \text{for } i, j = 1, \dots, \frac{N}{2}. \tag{9}$$

By assuming that the data is not redundant, a minimal realization is then given by:

$$\hat{\mathbf{E}} = -\mathbb{L},\ \ \hat{\mathbf{A}} = -\mathbb{L}_s,\ \ \hat{\mathbf{B}} = \mathbb{V},\ \ \hat{\mathbf{C}} = \mathbb{W},\ \ \hat{\mathbf{D}} = 0.$$

However, in practical applications, the Loewner pencil $(\mathbb{L}_s, \mathbb{L})$ is often singular and a ROM needs to be computed. In such cases, a singular value decomposition (SVD) of the Loewner matrices is typically performed in order to determine a suitable truncation index r and the corresponding projection matrices denoted with $\mathbf{X}_r$ and $\mathbf{Y}_r$. The projection matrices are computed based on the SVD of the Loewner matrix $\mathbb{L}$, with $\mathbf{X}_r$ chosen as the first r columns of $\mathbf{X}$ and $\mathbf{Y}_r$ as the first r columns of $\mathbf{Y}$:

$$\mathbb{L} = \mathbf{X}\mathbf{S}\mathbf{Y}^* \approx \mathbf{X}_r\mathbf{S}_r\mathbf{Y}_r^*. \tag{10}$$

Then, the reduced-order Loewner model of dimension r is given by the following matrices:

$$\hat{\mathbf{E}} = -\mathbf{X}_r^*\mathbb{L}\mathbf{Y}_r,\ \hat{\mathbf{A}} = -\mathbf{X}_r^*\mathbb{L}_s\mathbf{Y}_r,\ \hat{\mathbf{B}} = \mathbf{X}_r^*\mathbb{V},\ \hat{\mathbf{C}} = \mathbb{W}\mathbf{Y}_r,\ \hat{\mathbf{D}} = 0. \tag{11}$$

Remark 3. (Data partitioning). How to effectively separate the available data into two subsets still remains an open question. It is shown in Ionita (2013) that this partition impacts the robustness to noise. In Karachalios et al. (2021); Gosea et al. (2021), two different partitioning were numerically analyzed:

- "*alternate*" (the most recurrent way of separating data):
$$\{z_k\}_{k=1}^N = \{z_1, z_2, \dots z_{N-1}, z_N\}. \tag{12}$$
- "*half-half*" (an intuitive way of separating data):
$$\{z_k\}_{k=1}^N = \left\{z_1, \dots, z_{N/2}, z_{N/2+1}, \dots, z_N.\right\} \tag{13}$$

Fig. 1. Splitting schemes commonly used in the LF

As previously reported in Gosea et al. (2021), the effect of half-half partitioning is that the decay of the singular values of the Loewner matrix is clearer (more revealing) than for the alternate splitting (when dealing with noisy frequency-domain data). As a result, half-half LF seems to ease the order selection and hence avoids overfitting due to noise. Both types of partitioning are used jointly in this work, as explained in the next section.

2.3 The Hankel framework

While the LF interpolates the frequency response, the HF provides a model that interpolates the impulse response, similarly to the Ho-Kalman algorithm Ho and Kalman (1966) or Silverman realization Silverman (1971). Given the truncated impulse response $\mathbf{h} = [\mathbf{h}_0, \mathbf{h}_1, \cdots, \mathbf{h}_{N-1}]$, the resulting Hankel model is given in descriptor form by the following matrices:

$$\begin{gathered}\hat{\mathbf{E}} = \mathcal{H},\quad \hat{\mathbf{A}} = \mathcal{H}_s,\\ \hat{\mathbf{C}} = \begin{bmatrix}\mathbf{h}_1, & \mathbf{h}_2, \cdots, & \mathbf{h}_N\end{bmatrix},\quad \hat{\mathbf{B}} = \hat{\mathbf{C}}^T,\quad \hat{\mathbf{D}} = \mathbf{h}_0.\end{gathered} \tag{14}$$

with the Hankel pencil $(\mathcal{H}, \mathcal{H}_s)$ defined as follows:

$$\mathcal{H}_{(i,j)} = \mathbf{h}_{i+j-1},\ \mathcal{H}_{s(i,j)} = \mathbf{h}_{i+j}, \tag{15}$$

As in the LF, the dimension of the Hankel model (14) can be reduced by means of projection, using orthogonal matrices computed by means of applying an **SVD** for the Hankel matrix $\mathcal{H}$. In this case, we enforce approximation, i.e. by fitting a model which approximately explains the data. Additional insights on the HF were given in Ionita and Antoulas (2012).

3. FROM NOISY DATA TO REDUCED-ORDER MODELS

3.1 Overview of the proposed approach

To the best of our knowledge, most of the attempts to make the LF and HF matrix pencils identification techniques more robust to noisy data have consisted in changing the way the model is obtained as in Lefteriu et al. (2010).In this work, it is proposed to pre-process the noisy data instead. First, an estimation of the truncated impulse response $\{\tilde{\mathbf{h}}_k\}_{k=0}^{N-1}$ of the system is obtained from the available noisy measurements through the SMM approach, as proposed in Yin et al. (2020). This estimation forms a non-parametric model of the system, which is then parameterized and reduced through the HF or the LF.

While the estimated values $\{\tilde{\mathbf{h}}_k\}_{k=0}^{N-1}$ can be used directly in the HF, another possibility consists in applying a fast Fourier transform (FFT) to the impulse response to estimate frequency-response samples as follows:

$$\tilde{\mathbf{H}}_N(e^{\imath\omega_i}) = \sum_{k=0}^{N-1}\mathbf{h}_k e^{-\imath\omega_i k},\ \omega = \frac{2\pi i}{N},\ i = 0\dots N-1, \tag{16}$$

which is a truncated version of (5). The frequency-domain data estimated from (16) can then be used in the LF.

In this section, the SMM approach from Yin et al. (2020) is recalled in Section 3.2. The tuning knobs of the proposed approach, that combines SMM and matrix pencils approaches, are then detailed in Section 3.3. A synthesized algorithm that brings these different aspects together is then provided in Section 3.4.

3.2 Impulse response estimation: the SMM approach

Traditionally, Markov parameters $\mathbf{h}_k$'s can be obtained from input-output measurements $\{u_k, y_k\}_{k=0}^{N_s}$ by solving a linear system of equations based on (3): it is the least squares (LS) approach. It consists in identifying a N-th order FIR filter from the available data, rather than obtaining the true value of the first N-th Markov parameters. In practice, a very long impulse response sequence may be needed to reach a negligible truncation error with respect to (5), even for a low-order system. In Markovsky et al. (2005), a data-driven simulation approach, based on Willem's fundamental lemma, was proposed when noise-free input-output data are available. It allows to estimate the impulse response even when the truncation error is not negligible. The following assumptions are enforced:

(1) The LTI system is finite-dimensional and controllable;
(2) The input $\{u_k\}_{k=0}^{N_s}$ is persistently exciting of order $L = N + L_0$, with $L_0 > n_x$, meaning that the Hankel matrix $\mathcal{U} \in \mathbb{R}^{Ln_u \times M}$ (with $M = N_s - L + 1$)
$$\mathcal{U}_{(i,j)} = u_{i+j-2}. \tag{17}$$
has full row rank Willems et al. (2005).

Under these assumptions, the output trajectory of the system for an input $\mathbf{u} \in \mathbb{R}^{Nn_u}$, starting from initial conditions uniquely determined by the past input-output trajectory $\mathbf{u}_{ini} \in \mathbb{R}^{L_0 n_u}$ and $\mathbf{y}_{ini} \in \mathbb{R}^{L_0 n_y}$ for $L_0 \geq n_x$, is $\mathbf{y} = Y_f g$. Here, $g \in \mathbb{R}^{M'}$ is the solution of the linear system of equations:

$$\begin{bmatrix} U_p \\ Y_p \\ U_f \end{bmatrix} g = \begin{bmatrix} \mathbf{u}_{ini} \\ \mathbf{y}_{ini} \\ \mathbf{u} \end{bmatrix}, \tag{18}$$

where $U_p \in \mathbb{R}^{L_0 n_u \times M}$, $U_f \in \mathbb{R}^{Nn_u \times M}$ are matrices computed by using the available data such that:

$$\mathcal{U} = \begin{bmatrix} U_p \\ U_f \end{bmatrix}, \tag{19}$$

and similarly for $Y_p \in \mathbb{R}^{L_0 n_y \times M}$ and $Y_f \in \mathbb{R}^{Nn_y \times M}$ with the Hankel matrix $\mathcal{Y}$ of the output samples defined as $\mathcal{U}$ in (17).

In order to handle the case for which only noisy input-output measurements are available, the SMM approach in Yin et al. (2020) builds on Markovsky et al. (2005) and represents a maximum likelihood framework to obtain a statistically optimal implicit model. Additive i.i.d Gaussian output noise is considered:

$$\tilde{\mathbf{y}} = \mathbf{y} + w,\ w \sim \mathcal{N}(0, \sigma^2 \mathbb{I}). \tag{20}$$

As in Yin et al. (2020), the SMM approach is used to estimate the impulse response with $\mathbf{u}_{ini} = 0, \mathbf{y}_{ini} = 0$ and $\mathbf{u} = [1\, 0 \ldots 0]$. The estimate of the first N Markov parameters denoted with $\hat{\mathbf{h}}$ is explicitly given by $\hat{\mathbf{h}} = Y_f g_h$, where

$$\begin{aligned} g_h =&\ (F^{-1} - F^{-1}\mathcal{U}^T(\mathcal{U}F^{-1}\mathcal{U}^T)^{-1}\mathcal{U}F^{-1})Y_p^T \mathbf{y}_{ini} \\ &+ F^{-1}\mathcal{U}^T(\mathcal{U}F^{-1}\mathcal{U}^T)^{-1}\begin{bmatrix}\mathbf{u}_{ini}\\ \mathbf{u}\end{bmatrix}, \\ F =&\ Y_p^T Y_p + L\sigma^2\mathbb{I}. \end{aligned} \tag{21}$$

The result is unbiased for an arbitrary length N.

3.3 Tuning the hyper-parameters

a) Persistency of excitation and the choice of L_0: Persistency of excitation is the key assumption of Willem's fundamental lemma Willems et al. (2005) as it allows to characterize all possible trajectories of length N from the available data. However, it implies that the order of the underlying system is known. In the ideal case, $L_0 = n_x$ should be used in order to exploit the available data to the fullest extent. Nonetheless, the most important condition to be imposed is $L_0 \geq n_x$ so that $\mathbf{u}_{ini}$ and $\mathbf{y}_{ini}$ uniquely define the initial conditions. In practice, when n_x is unknown, a good choice for L_0 can be found by computing the cross-correlation R_{yu} of the measured output and the input signal and L_0 is then chosen as the minimal positive lag such that:

$$\forall \tau > L_0,\ |R_{yu}(\tau)| \leq \epsilon. \tag{22}$$

As the system is causal, the cross-correlation for negative lags is merely a numerical artifact and does not represent any real input-output relationship. For this reason, the threshold ϵ is fixed in this work as:

$$\epsilon = (1 + \alpha) \times \max\{|R_{yu}(\tau)| \text{ for } \tau < 0\}, \tag{23}$$

where the scalar $0 \leq \alpha \leq 1$ allows introducing an additional margin to avoid choosing a too large value for L_0.

b) The number N of estimated Markov parameters: A necessary condition for the matrix U to be of full row rank (the so-called persistency of excitation assumption) is $M \geq Ln_u$, which gives an upper bound N_{max} for the number of Markov parameters that can be estimated when a number of N_s input-output measurements are available:

$$N_{max} = \frac{N_s + 1}{n_u + 1} - L_0. \tag{24}$$

On the other hand, as the Hankel and shifted Hankel matrices are of size N, then N Markov parameters allow to obtain a model of order at most N through HF. In addition, when using the LF, the more Markov parameters are used, the lower the truncation error between (5) and (16) is. Consequently, after having chosen L_0 as previously explained, it is recommended to choose $N = N_{max}/2$ to enforce $M \ll Ln_u$. Then, one could decrease it if necessary until the matrix U is of full row rank.

c) Noise variance σ^2: The noise variance σ^2 is used in the SMM approach, see (21). In practice, this information might not be available. An approximation can be obtained through the LS approach. As recalled in Niu and Fisher (1995), for zero-mean white noise, an unbiased estimate of the variance σ^2 is given by:

$$\sigma^2 = \lim_{N_s \to \infty} \frac{\|\mathbf{h} - \mathbf{h}_{LS}\|_2^2}{N_s - N}, \tag{25}$$

where $\mathbf{h}$ are the first N true Markov parameters of the system and $\mathbf{h}_{LS}$ is the estimate by the classical LS approach, obtained by using N_s input-output samples. An approximation $\hat{\sigma}^2$ is then chosen as follows:

$$\hat{\sigma}^2 = \frac{\|\mathbf{h} - \mathbf{h}_{LS}\|_2^2}{N_s - N}. \tag{26}$$

d) Order of the reduced-order model: The order of the reduced-order model is a tunable parameter for both HF and LF. An adequate value is supposed to be chosen based on a rank-revealing decomposition of the Hankel or Loewner matrices. As detailed in Lefteriu et al. (2010), measurement noise complicates the choice of the reduced order r. In that case, it is possible to change the data partitioning in the LF in order to obtain a clearer SVD decay, as suggested in Gosea et al. (2021) and recalled in Remark 3. However, while half-half partitioning (13) reveals the system's order in a clear way and is robust to noise, it leads to less accurate models because the Loewner matrices tend to be ill-conditioned for this choice. At the same time, alternate partitioning (12) leads Loewner pencils that are diagonally dominant. Therefore, for a fixed order, this approach will result in more accurate models. This behavior has been pointed out in Ionita (2013), and more recently in Palitta and Lefteriu (2022) based on analyzing Cauchy matrices, which explicitly appear in the definition of Loewner matrices. For such reasons, we propose here to combine both types of data-partitioning in the LF to benefit from their respective advantages, see Algorithm 2.

3.4 Summary

Given noisy data $\{u_k, \tilde{y}_k\},\ k = 0 \ldots N_s - 1$, the proposed approach consists in tuning some hyper-parameters as explained in Section 3.3, before using the SMM approach from Yin et al. (2020) as recalled in Section 3.2. The resulting estimated Markov parameters $\mathbf{h}_{SMM}$, which constitute a non-parameterized model of the system, are then used in matrix pencil approaches, the HF (Algorithm 1) or the LF (Algorithm 2), allowing to obtain a linear reduced-order approximation $\left(\hat{\mathbf{E}}_r, \hat{\mathbf{A}}_r, \hat{\mathbf{B}}_r, \hat{\mathbf{C}}_r, \hat{\mathbf{D}}_r\right)$ of the underlying dynamical system (1).

These two techniques are referred to as SMM-HF and SMM-LF respectively. SMM-LF allows using different data-partitioning techniques in order to reveal the order of the system despite measurement noise, it is affected by the truncation of the Markov series when the truncation error is not negligible, while the SMM-HF is not sensitive to it. For this reason, it might be more interesting to use the HF once the order r has been determined from the SVD of the Loewner matrix built with half-half partitionning, which combines SMM-HF and SMM-LF.

Algorithm 1 SMM-HF

```
Inputs: Input-output time-domain data $\{u_k, \tilde{y}_k\},\ k = 0 \ldots N_s - 1$.
(1) Step 1: Tune the hyper-parameters $L_0$ from (22), $N = N_{max}/2$ based on (24) and decrease it until $U$ is full rank, and estimate the noise variance $\hat{\sigma}^2$ (26).
(2) Step 2: Using $L_0$, $N$ and $\hat{\sigma}^2$, estimate the Markov parameters $\mathbf{h}_{SMM}$ through SMM (21) as in Yin et al. (2020).
(3) Step 3: Apply the HF based on $\mathbf{h}_{SMM}$
```

```
Algorithm 2 SMM-LF
Inputs: Input-output time-domain data $\{u_k, \tilde{y}_k\}$, $k = 0 \dots N_s - 1$.
 (1) Steps 1 and 2: same as in Algorithm 1.
 (2) Step 3: Apply the LF on the FFT of $\mathbf{h}_{SMM}$
     (a) Use half-half partitioning (13) to determine the order $r$.
     (b) Use alternate partitioning (12) to build the Loewner model (11).
```

4. NUMERICAL EXAMPLE

The proposed approach is illustrated on the Los Angeles Hospital building benchmark from the COMP$l_e ib$ library Leibfritz (2004), described by a 48th-order state-space model.

To collect data, the high-order model is simulated using a normally distributed random input signal. The sampling period is $T_s = 15$ms and $N_s = 1000$ output samples y_k are collected. Additive output Gaussian noise of variance $\sigma^2 = 1 \cdot 10^{-7}$ is then considered, as in (20). 50 different noisy data sets are generated like this. Algorithms 1 and 2 are derived hereafter.

4.1 Step 1: Choice of the hyper parameters

The cross-correlation is computed for every noise realization and averaged. The threshold value is chosen as in (23) with $\alpha = 0.4$ and, according to (22), $L_0 = 66$ is taken, which slightly overestimates the order n_x of the system. The number of estimated Markov parameters is taken equal to $N = N_{max}/2 = 217$, and the corresponding matrix $\mathcal{U}$ is full row rank, which means that the input u is persistently exciting of order $L_0 + N$. The LS approach is applied and the resulting variance estimate is $\hat{\sigma}^2 = 1.27 \cdot 10^{-7}$.

4.2 Step 2: Impulse response estimation

Based on each noisy data set $\{u_k, \tilde{y}_k\}_{k=0}^{N_s-1}$, the SMM approach is used to estimate the first N Markov coefficients of the system, i.e. the first N samples of its impulse response. As in Yin et al. (2020), the fitting of the estimated impulse response $\hat{\mathbf{h}}$ to the true system impulse response $\mathbf{h}$ is defined by:

$$W = 100\left(1 - \sqrt{\frac{\sum_{i=1}^{N}(\mathbf{h}_i - \hat{\mathbf{h}}_i)^2}{\sum_{i=1}^{N}(\mathbf{h}_i - \overline{\mathbf{h}})^2}}\right), \quad (27)$$

with $\overline{\mathbf{h}}$ the average of the true Markov parameters $\mathbf{h}$. The results correspond to the level of performance presented in Yin et al. (2020): the SMM approach ($W = 54.2\%$) outperforms the LS ($W = 47\%$) one by obtaining a better median fit.

4.3 Step 3: Model approximation and reduction

The estimated impulse responses, denoted $\mathbf{h}_{LS}$ and $\mathbf{h}_{SMM}$ for the LS and SMM approach respectively, obtained in Step 2, one fore each noisy data set, are now used to obtain a parameterized model of the system through LF and HF.

a) Loewner framework: Frequency-domain data is inferred by performing a FFT as in (16) of the SMM estimated impulse response. For comparison purposes, frequency-domain data is also estimated as the ratio between the cross power spectral density of u and y, and the power spectral density of u, without taking noise into account. This last approach is referred to as *noisy LF* in this paragraph.

Once frequency-domain data is obtained, the Loewner pencil from (8) is then built using the two different data partitioning techniques (12) and (13) from Gosea et al. (2021). A SVD is performed on the Loewner matrix $\mathbb{L}$ to reveal the order of the underlying system. The average decay of the normalized singular values is visible on Figure 2: while alternate partitioning gives almost full-rank Loewner matrices with both the noisy LF and SMM-LF approaches, half-half partitioning leads to a Loewner matrix of order 48 for the SMM-LF approach and 60 for the noisy-LF approach (in average over the 50 noisy data sets). If allowing to approximate the order of the underlying system, half-half partitioning leads to less precise models, as highlighted in Gosea et al. (2021). Descriptor models are then obtained as in (11), based on alternate partitioning (12) as suggested in Algorithm 2. The order is chosen as $r = n_x = 48$. The proposed approach allows obtaining a better fit of the frequency response, on average, see Figure 3 which represents the average frequency response of the resulting models.

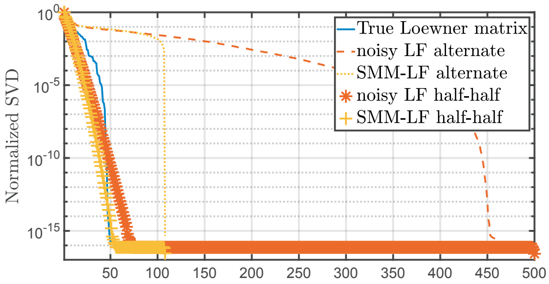

Fig. 2. Normalized SVD of the Loewner matrices built with the frequency-domain data inferred from the SMM approach or directly from the noisy time-domain data (*noisy LF*), and compared with the SVD decay of the Loewner matrix obtained with noise free frequency-domain data. Two types of data partitioning are used as in Gosea et al. (2021) to evaluate the order of the underlying system despite measurement noise.

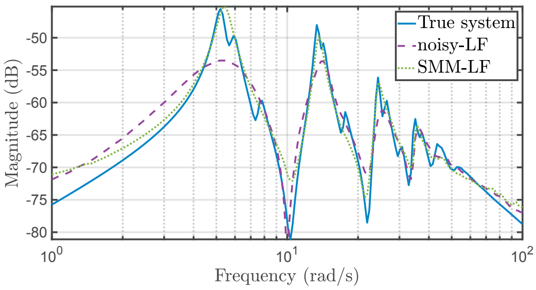

Fig. 3. Average frequency-response obtained when applying the SMM-LF and noisy LF procedures.

b) Hankel framework: The Hankel pencil from (15) is built and a SVD is performed on the Hankel matrix $\mathcal{H}$ to reveal the order of the underlying system. The average decay of the normalized singular values is visible on Figure 4 for the true Markov parameters $\mathbf{h}$ of the system and the estimated ones $\mathbf{h}_{LS}$ and $\mathbf{h}_{SMM}$. The same orders than for LF are chosen. The average impulse responses are visible on Figure 5, showing that the SMM-HF approach from Algorithm 1 outperforms the regular LS + HF approach.

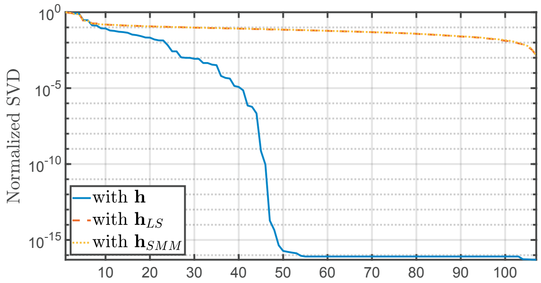

Fig. 4. Normalized SVD of the Hankel matrices built with Markov parameters estimated through the LS and SMM approaches and with the true Markov parameters of the system.

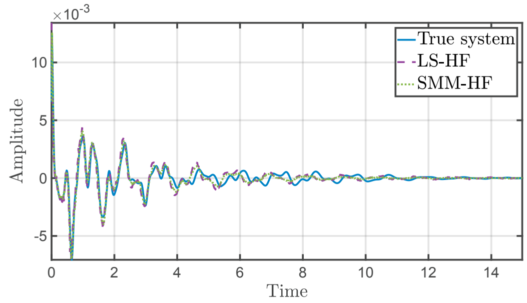

Fig. 5. Average impulse response obtained when applying the SMM-HF and LS-HF procedures.

5. CONCLUSIONS AND OUTLOOKS

In this work, a method to handle noisy data in matrix pencils frameworks, namely HF and LF, has been proposed. It relies on the SMM approach from Yin et al. (2020) to estimate the impulse response of the system from a noisy data set. The impulse response constitutes a non-parameterized model of the system, which is then used in the HF or LF to obtain a parameterized model and to reduce it. As in Gosea et al. (2021) Palitta and Lefteriu (2022), different data partitioning can be used to reveal the order of the system. As opposed to existing works such as Lefteriu et al. (2010), the new method proposes a preliminary step on the available data (the SMM approach), rather than modifying the way of obtaining the model. A thorough comparison between these methods and the proposed approach is left for future work (both in terms of computational complexity and also of accuracy of computed models). Connections to newly-proposed work in Wilber et al. (2021) could also be investigated (this work combines the classical Prony algorithm with the recently-proposed AAA algorithm mentioned in Gosea et al. (2021)).

Future work will also investigate the impact of noise level on the accuracy of the resulting models and it would be interesting to include pseudospectra analysis Embree and Ioniţă (2022) in the proposed approach. In addition, this work should be illustrated on real-world datasets. The proposed approach could also be used to improve the robustness to noise in the Loewner Data-Driven Control (L-DDC) framework Kergus et al. (2018), and to introduce a counterpart based on time-domain data relying on the HF the same way L-DDC relies on LF.

REFERENCES

Antoulas, A.C. (2005). *Approximation of large-scale dynamical systems*. SIAM, Philadelphia.

Antoulas, A.C., Lefteriu, S., and Ionita, A.C. (2017). A tutorial introduction to the Loewner framework for model reduction. In *Model Reduction and Approximation*, chapter 8. SIAM.

Drmač, Z. and Peherstorfer, B. (2019). Learning low-dimensional dynamical-system models from noisy frequency-response data with Loewner rational interpolation. *arXiv:1910.00110*.

Embree, M. and Ioniţă, A.C. (2022). Pseudospectra of Loewner matrix pencils. In *Realization and Model Reduction of Dynamical Systems*. Springer.

Gosea, I., Zhang, Q., and Antoulas, A. (2021). Data-driven modeling from noisy measurements. *Proceedings in Applied Mathematics and Mechanics*.

Gustavsen, B. and Semlyen, A. (1999). Rational approximation of frequency domain responses by vector fitting. *IEEE Transactions on power delivery*.

Ho, B.L. and Kalman, R.E. (1966). Effective construction of linear state variable models from input-output functions. *Regelungstechnik*, 14.

Ionita, A.C. and Antoulas, A.C. (2012). Matrix pencils in time and frequency domain system identification. *Control, Robotics and Sensors. Institution of Engineering and Technology*.

Ionita, A. (2013). *Lagrange rational interpolation and its applications to approximation of large-scale dynamical systems*. Ph.D. thesis.

Karachalios, D., Gosea, I.V., and Antoulas, A.C. (2021). The Loewner framework for system identification and reduction. In *Model Order Reduction: Volume I: System-and Data-Driven Methods and Algorithms*, 181–228. De Gruyter.

Kergus, P., Formentin, S., Poussot-Vassal, C., and Demourant, F. (2018). Data-driven control design in the Loewner framework: Dealing with stability and noise. In *European Control Conference*. IEEE.

Lefteriu, S., Ionita, A., and Antoulas, A. (2010). Modeling systems based on noisy frequency and time domain measurements. *Perspectives in Mathematical System Theory, Control, and Signal Processing*.

Leibfritz, F. (2004). Compleib, constraint matrix-optimization problem library-a collection of test examples for nonlinear semidefinite programs, control system design and related problems. *Dept. Math., Univ. Trier, Trier, Germany, Tech. Rep.*

Markovsky, I., Willems, J.C., Rapisarda, P., and De Moor, B. (2005). Data driven simulation with applications to system identification. *IFAC Proceedings Volumes*.

Mayo, A.J. and Antoulas, A.C. (2007). A framework for the solution of the generalized realization problem. *Linear algebra and its applications*.

Nakatsukasa, Y., Sète, O., and Trefethen, L. (2018). The AAA algorithm for rational approximation. *SIAM Journal on Scientific Computing*.

Niu, S. and Fisher, D.G. (1995). Simultaneous estimation of process parameters, noise variance, and signal-to-noise ratio. *IEEE transactions on signal processing*.

Palitta, D. and Lefteriu, S. (2022). An efficient, memory-saving approach for the Loewner framework. *Journal of Scientific Computing*.

Peherstorfer, B., Gugercin, S., and Willcox, K. (2017). Data-driven reduced model construction with time-domain Loewner models. *SIAM Journal on Scientific Computing*.

Schutter, B.D. (2000). Minimal state-space realization in linear system theory: an overview. *Journal of Computational and Applied Mathematics*.

Silverman, L. (1971). Realization of linear dynamical systems. *IEEE Transactions on Automatic Control*, 16(6), 554–567.

Wilber, H., Damle, A., and Townsend, A. (2021). Data-driven algorithms for signal processing with rational functions. *arXiv preprint arXiv:2105.07324*.

Willems, J.C., Rapisarda, P., Markovsky, I., and De Moor, B. (2005). A note on persistency of excitation. *Systems & Control Letters*.

Yin, M., Iannelli, A., and Smith, R. (2020). Maximum likelihood estimation in data-driven modeling and control. *arXiv:2011.00925*.