

# Stable Linear Subspace Identification: A Machine Learning Approach

Loris Di Natale,<sup>†</sup> Muhammad Zakwan,<sup>†</sup> Bratislav Svetožarevic,  
Philipp Heer, Giancarlo Ferrari-Trecate, Colin N. Jones

**Abstract**—Machine Learning (ML) and linear System Identification (SI) have been historically developed independently. In this paper, we leverage well-established ML tools — especially the automatic differentiation framework — to introduce SIMBa, a family of discrete linear multi-step-ahead state-space SI methods using backpropagation. SIMBa relies on a novel Linear-Matrix-Inequality-based free parametrization of Schur matrices to ensure the stability of the identified model.

We show how SIMBa generally outperforms traditional linear state-space SI methods, and sometimes significantly, although at the price of a higher computational burden. This performance gap is particularly remarkable compared to other SI methods with stability guarantees, where the gain is frequently above 25% in our investigations, hinting at SIMBa’s ability to simultaneously achieve state-of-the-art fitting performance and enforce stability. Interestingly, these observations hold for a wide variety of input-output systems and on both simulated and real-world data, showcasing the flexibility of the proposed approach. We postulate that this new SI paradigm presents a great extension potential to identify structured nonlinear models from data, and we hence open-source SIMBa on <https://github.com/Cemempamoi/simba>.

## I. INTRODUCTION

While linear System Identification (SI) matured decades ago [1], Machine Learning (ML) only rose to prominence in recent years, especially following the explosion of data collection and thanks to unprecedented computational power. In particular, large Neural Networks (NNs) have shown impressive performance on a wide variety of tasks [2], [3], leading to the recent boom of Deep Learning (DL) applications [4]. A key factor behind these successes has been the availability of efficient open-source libraries greatly accelerating the deployment of NNs, such as PyTorch and TensorFlow in Python. In particular, Automatic Differentiation (AD), at the core of the backpropagation algorithm [5], the backbone of NN training, nowadays benefits from extremely efficient implementations.

Given their effectiveness at grasping complex nonlinear patterns from data, NNs have recently been used for nonlinear system identification, where traditional SI methods struggle to compete [6]–[8]. NNs can be leveraged to create deep state-space models [9], deep subspace encoders [10], or

deep autoencoders [11], for example. While applying NNs to identify nonlinear systems can achieve good performance, it can underperform for linear systems, where methods assuming model linearity might achieve better accuracy [9].

Although nonlinear SI has attracted a lot of attention in the last years, the identification of Linear Time Invariant (LTI) models is, however, still of paramount importance to many applications. Indeed, linear models come with extensive theoretical properties [12] and lead to convex optimization problems when combined with convex cost functions in a Model Predictive Controller (MPC) [13], for example. Moreover, to date, numerous industrial applications still rely on the availability of linear models to conduct simulations, perform perturbation analysis, or design robust controllers following classical model-based techniques, such as  $\mathcal{H}_2$ ,  $\mathcal{H}_\infty$ , and  $\mu$ -synthesis [14].

In this work, we show how one can leverage ML tools — backpropagation and unconstrained Gradient Descent (GD) — for the identification of stable linear models, presenting a novel toolbox of System Identification Methods leveraging Backpropagation (SIMBa). Our research is related to the efforts in [15]–[17], where backpropagation was also used to identify LTI state-space models but only as part of specific frameworks and without considering stability constraints.

### A. Subspace identification for linear systems

State-of-the-art implementations of linear state-space SI often rely on subspace identification, such as the acclaimed MATLAB system identification toolbox [18] or the SIPPY Python package [19]. Both of them provide the three *traditional* Subspace Identification Methods (SIMs), namely N4SID, MOESP, and CVA [20].

Next to these traditional methods, SIPPY also proposes an implementation of *parsimonious* SIMs (PARSIMs), namely PARSIM-S [21], PARSIM-P [22], and PARSIM-K [23], which enforce causal models by removing non-causal terms. While the former two methods do not work with closed-loop data since they assume no correlation between the output noise and the input, PARSIM-K was specifically designed to alleviate this assumption.

### B. Enforcing stability

In practice, when the true system is known to be stable, one usually requires the identified model to also be stable [24]. In the case of traditional methods, a stability test can be done *a posteriori* and a stable model can always be recovered from the extended observability matrix [25], for example, but at the cost of a sometimes significant

This research was supported by the Swiss National Science Foundation under NCCR Automation, grant agreement 51NF40\_180545.

L. Di Natale, B. Svetožarevic, and P. Heer are with the Urban Energy Systems Laboratory, Swiss Federal Laboratories for Materials Science and Technology (Empa), 8600 Dübendorf, Switzerland. L. Di Natale, M. Zakwan, G. Ferrari-Trecate, and C.N. Jones are with the Laboratoire d’Automatique, Swiss Federal Institute of Technology Lausanne (EPFL), 1015 Lausanne, Switzerland.

<sup>†</sup> L. Di Natale and M. Zakwan contributed equally to this work. Corresponding author: L. Di Natale, [loris.dinatale@alumni.epfl.ch](mailto:loris.dinatale@alumni.epfl.ch).

performance loss [19]. Note that such a correction cannot be implemented for PARSIM methods.

Apart from such *post-hoc* modifications to ensure stability, one can also modify the Least Squares (LS) estimation at the heart of many identification procedures, either introducing custom weighting factors [26] or rewriting it as a constrained optimization problem [27], [28].

Alternatively, one can leverage parametrizations of stable matrices, such as the ones proposed in [29], [30], and then use projected gradients to approximate the LS solution while ensuring the resulting model remains stable at each step, for example [31]. A similar idea was utilized in [32], where the Perron-Frobenius theorem was leveraged to bound the eigenvalues of  $A$  and hence ensure the system remains stable *at all times*, even during the learning phase.

Finally, instead of directly constraining the state-space matrices, one can also simultaneously learn a model and a corresponding Lyapunov function for it, typically NN-based, thereby ensuring its stability by design [33]. This approach presents the advantage of naturally extending to nonlinear SI, contrary to all the others, but comes with a significant computational burden. Note that while SIMBa does not explicitly learn a Lyapunov function, it implicitly defines one to guarantee stability (see Section III). However, instead of learning it with an NN, we leverage Linear Matrix Inequalities (LMIs) to parametrize Schur matrices, inspired from [34].

### C. Contribution

This paper introduces SIMBa, a linear SI toolbox that guarantees the stability of the identified model through a novel free parametrization of Schur matrices. Leveraging ML tools, it is able to minimize the multi-step-ahead prediction error to improve upon the performance of traditional SI methods. Our main contributions can be summarized as follows:

- We propose SIMBa, a framework leveraging backpropagation and unconstrained GD for stable linear SI.
- We present a ready-to-use open-source Python implementation of SIMBa with a MATLAB interface. SIMBa is system-agnostic: it can seamlessly identify multi-input-multi-output systems, optimize different multi-steps-ahead performance metrics, deal with missing data and multiple trajectories, and it comes with smooth Graphical Processing Unit (GPU)-integration.
- We conduct an extensive empirical investigation of SIMBa's potential, exemplifying its flexibility and ability to attain state-of-the-art performance in various contexts, both on simulated and real-world data. Overall, we often observe an improvement of over 25% compared to the benchmark methods.

Altogether, we propose a new paradigm for the identification of stable linear models. While it comes with an additional computational burden, SIMBa very often outperforms traditional methods in our experiments. Furthermore, it presents interesting extension potential, for example, to integrate prior knowledge about the system, include tailored nonlinearities,

similarly to what was done in [35], or facilitate stable Koopman-based approaches like [16], [36].

*Organization:* After a few preliminaries in Section II, Section III presents the novel free parametrization of stable matrices used in SIMBa. Section IV then introduces the SIMBa toolbox and Section V provides empirical performance analyses. Finally, Section VI concludes the paper.

## II. PRELIMINARIES

Throughout this work, we are interested in the identification of discrete-time linear systems of the form

$$x_{k+1} = Ax_k + Bu_k + w_k \quad (1a)$$

$$y_k = Cx_k + Du_k + v_k, \quad (1b)$$

where  $x \in \mathbb{R}^n$  represents the state of the system,  $u \in \mathbb{R}^m$  its input,  $y \in \mathbb{R}^p$  its output, and  $w_k \in \mathbb{R}^n$  and  $v_k \in \mathbb{R}^p$  process and measurement noise, respectively. In general, state-space SI methods identify the unknown matrices  $A \in \mathbb{R}^{n \times n}$ ,  $B \in \mathbb{R}^{n \times m}$ ,  $C \in \mathbb{R}^{p \times n}$ , and  $D \in \mathbb{R}^{p \times m}$  from a data set  $\mathcal{D} = \{(u(0), y(0)), \dots, (u(l_s), y(l_s))\}_{s=1}^N$  of  $N$  input-output measurement trajectories  $s$  of length  $l_s$ .

In our experiments, we split the data into a training, a validation, and a test set of trajectories respectively denoted  $\mathcal{D}_{train}$ ,  $\mathcal{D}_{val}$ , and  $\mathcal{D}_{test}$ . The first set is used to fit the model, the second one to assess when to stop the algorithm so as to avoid overfitting the training data, and the third one to measure the performance of the identified model on unseen data, as classically done in ML pipelines [37].

Since the stability of identified models is often crucial, we are particularly interested in stable-by-design identification procedures. For systems of the form (1), stability is equivalent to the matrix  $A$  being Schur, i.e., ensuring that all its eigenvalues  $\lambda_i(A)$  be of magnitude smaller than one:

$$|\lambda_i(A)| < 1, \forall i = 1, \dots, n.$$

*Notation:* Given a symmetric matrix  $F \in \mathbb{R}^{q \times q}$ ,  $F \succ 0$  means  $F$  is positive definite. For a matrix  $V \in \mathbb{R}^{2q \times 2q}$ , we define its block components  $V_{11}, V_{12}, V_{21}, V_{22} \in \mathbb{R}^{q \times q}$  as:

$$\begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} := V. \quad (2)$$

## III. FREE PARAMETRIZATION OF SCHUR MATRICES

To run unconstrained GD in the search space and hence take full advantage of PyTorch without jeopardizing stability, we need to construct  $A$  matrices that are Schur *by design*. The following proposition provides such an LMI-based free parametrization of Schur matrices, inspired by [34].

**Proposition 1.** *For any  $W \in \mathbb{R}^{2n \times 2n}$ ,  $V \in \mathbb{R}^{n \times n}$ ,  $\tilde{\epsilon} \in \mathbb{R}$ , and given  $0 < \gamma \leq 1$ , let*

$$S := W^\top W + \epsilon \mathbb{I}_{2n}, \quad (3)$$

for  $\epsilon = \exp(\tilde{\epsilon})$ . Then

$$A = S_{12} \left[ \frac{1}{2} \left( \frac{S_{11}}{\gamma^2} + S_{22} \right) + V - V^\top \right]^{-1} \quad (4)$$

is Schur stable with  $|\lambda_i(A)| < \gamma, \forall i = 1, \dots, n$ .

*Proof.* We want to ensure that the magnitude of all the eigenvalues of  $A$  is bounded by  $\gamma$ . From [38, Theorem 2.2], we know this is the case if and only if the following LMI holds for some symmetric  $Q = Q^\top \succ 0$ :

$$\begin{bmatrix} \gamma Q & AQ \\ QA^\top & \gamma Q \end{bmatrix} \succ 0.$$

Taking its Schur complement, this is equivalent to

$$\gamma Q - A \frac{Q}{\gamma} A^\top \succ 0. \quad (5)$$

Defining the transformation  $T = [I, -A]$  and introducing a free parameter  $G \in \mathbb{R}^{n \times n}$ , this can be rewritten as

$$\begin{aligned} & \gamma Q - AGA^\top - A^\top G^\top A \\ & + AGA^\top + A^\top G^\top A - A \frac{Q}{\gamma} A^\top \succ 0 \\ \iff & T \begin{bmatrix} \gamma Q & AG \\ G^\top A^\top & G^\top + G - \frac{Q}{\gamma} \end{bmatrix} T^\top \succ 0 \\ \iff & \begin{bmatrix} \gamma Q & AG \\ G^\top A^\top & G^\top + G - \frac{Q}{\gamma} \end{bmatrix} \succ 0. \end{aligned} \quad (6)$$

In words,  $A$  is Schur stable with eigenvalues bounded by  $\gamma$  if and only if there exist  $Q \succ 0$  and  $G$  such that (6) holds.

Let us now parametrize the left-hand side of the above LMI by the matrix  $S$  in (3). Remarkably, since  $S$  is positive definite by construction for any choice of  $W$ , (6) will always be satisfied, ensuring the stability of  $A$ .

Finally, define

$$\begin{aligned} S_{11} &:= \gamma Q, & S_{12} &:= AG, \\ S_{21} &:= G^\top A^\top, & S_{22} &:= G^\top + G - \frac{Q}{\gamma}. \end{aligned}$$

This allows us to recover  $Q = \frac{S_{11}}{\gamma}$ , which is positive definite and symmetric by construction. We then note that

$$G^\top + G = \frac{Q}{\gamma} + S_{22} = \frac{S_{11}}{\gamma^2} + S_{22} \quad (7)$$

needs to hold. Since  $S_{11}$  and  $S_{22}$  are symmetric, (7) holds for any  $V \in \mathbb{R}^{n \times n}$  if we set

$$G = \frac{1}{2} \left( \frac{S_{11}}{\gamma^2} + S_{22} \right) + V - V^\top.$$

Remembering that  $A = S_{12}G^{-1}$  concludes the proof.  $\square$

Proposition 1 implies that we can run unconstrained gradient descent on two matrices  $W$  and  $V$  and *always* construct a Schur matrix  $A$  from them as in (4). Furthermore,  $\gamma$  allows us to tune the magnitude of the largest eigenvalue of  $A$ .

**Remark 1.** *Let  $A$  be a Schur matrix. Setting  $\gamma = 1$  in the above proof yields that  $A$  satisfies (6) for some  $Q \succ 0$  and any  $G \in \mathbb{R}^{n \times n}$ . By definition, there exists  $\epsilon > 0$  such that*

$$\begin{bmatrix} \gamma Q & AG \\ G^\top A^\top & G^\top + G - \frac{Q}{\gamma} \end{bmatrix} - \epsilon \mathbb{I}_{2n} =: \Gamma \succ 0.$$

Define  $V = 0$ ,  $W = \Gamma^{\frac{1}{2}}$ , and  $\tilde{\epsilon} = \log \epsilon$ . Then  $A$  can be constructed as in (4), with  $S$  as in (3), showing that the proposed parametrization captures all Schur matrices.

## IV. A SHORT INTRODUCTION TO SIMBA

This section briefly describes SIMBa, discussing how to use some of its most critical parameters. The core implementations rely on the Python `PyTorch` library, but we also provide a `MATLAB` interface, and the whole package can be found on <https://github.com/Cemempamoi/simba>. The default value of each parameter has been empirically tuned to achieve robust performance, and more details on how to use SIMBa can be found in [39].

### A. Optimization problem

SIMBa minimizes the multi-step-ahead prediction error:

$$\min_{A, B, C, D, x_0^{(s)}} \frac{1}{|Z|} \sum_{s \in Z} \left[ \frac{1}{l_s} \sum_{k=0}^{l_s} m_k^{(s)} \mathcal{L}_{train} \left( y^{(s)}(k), y_k^{(s)} \right) \right] \quad (8)$$

$$\text{s.t. } y_k^{(s)} = Cx_k^{(s)} + Du^{(s)}(k) \quad (9)$$

$$x_{k+1}^{(s)} = Ax_k^{(s)} + Bu^{(s)}(k), \quad (10)$$

where  $Z \in \mathcal{D}_{train}$  is a randomly sampled batch of trajectories, naturally handling the case where multiple training trajectories are provided, and  $\mathcal{L}_{train}$  is the training loss. Note that, if  $x_0$  is known, `learn_x0` can be disabled. Finally,  $m_k^{(s)}$  is a binary variable defaulting to one, where the superscript  $(s)$  stands for the sampled trajectory and the subscript  $k$  for the time step, that can be used in two ways. First,  $m_k^{(s)} = 0$  if  $y^{(s)}(k)$  does not exist in the data, giving SIMBa a natural means to deal with missing values. Second  $m_k^{(s)} = 0$  with probability  $0 \leq p < 1$ , where  $p$  is controlled through the `dropout` parameter, which is used to regularize the training procedure. This avoids overfitting the training data by randomly dropping points along the trajectories and might also provide empirical robustness to outliers present in the data.

Throughout our experiments, we use the Mean Square Error (MSE) loss, i.e.,  $\mathcal{L}_{train}(y, \hat{y}) = (y - \hat{y})^2$ , but it could seamlessly be substituted by any other metric of interest through `train_loss`. For instance, one could use the Mean Absolute Error (MAE), which is more robust against outliers.

SIMBa iteratively runs GD on batches of training data in (8) until `max_epochs` epochs have passed, where an *epoch* represents the fact that every training trajectory has been seen once. After each epoch, we evaluate the performance of the current model on the validation set. To that end, in the experiments in Section V, we set  $Z = \mathcal{D}_{val}$  and  $\mathcal{L}_{val} = \mathcal{L}_{train}$  in (8), but custom losses can be passed through `val_loss`. At the end of the training, we keep the model performing best on the validation data and test it on  $\mathcal{D}_{test}$ . Note that due to the nonconvexity of (8), we cannot expect SIMBa to converge to a global minimum but only a local one, rendering it typically sensitive to initializations and some hyperparameters.

### B. Initialization

To accelerate the convergence of SIMBa, it is possible to initialize the state-space matrices from the best

solution found either by MATLAB or SIPPY — hereafter referred to as the *initialization method* — with `init_from_matlab_or_ls=True`.<sup>1</sup>

Since  $A$  is usually required to be stable, relying on the construction introduced in Proposition 1, it cannot directly be initialized with the desired matrix  $A^*$  found by the initialization method. Instead, we use GD on the corresponding free parameters, i.e., we use PyTorch to solve

$$\min_{W,V} \mathcal{L}_{init}(A, A^*) \quad (11)$$

$$\text{s.t. } A \text{ as in (4)}, \quad (12)$$

and initialize  $A$  close to  $A^*$ .  $\mathcal{L}_{init}$  is the desired loss function, also defined as the MSE throughout this work. As mentioned before, GD on such a nonconvex objective is not guaranteed to converge, and the local optimum found through this procedure might hence induce a performance drop compared to the initialization method.

## V. NUMERICAL EXPERIMENTS

To showcase the ability of SIMBa to handle different systems and both simulated and real-world data, this Section presents an extensive performance analysis, benchmarking SIMBa with existing state-of-the-art linear state-space SI methods. Our results demonstrate the efficiency of the proposed multi-step SI approach leveraging Proposition 1, showing that it can enforce stability without sacrificing accuracy.

Throughout our analyses, we compare SIMBa to standard state-space methods implemented either in the Python-based SIPPY library [19] or the MATLAB SI toolbox [18]. Note that SIPPY assumes that  $x_0 = 0$ , so that results between both implementations are comparable in this setting, which is the case in Section V-A. When dealing with real-world input-output data in Section V-B, however, enforcing  $x_0 = 0$  leads to suboptimal performance, and we compare SIMBa with MATLAB’s performance when  $x_0$  is estimated.<sup>2</sup> Note that we set `‘Focus’ = ‘simulation’` when using the MATLAB SI toolbox for a fair comparison with SIMBa’s multi-step-ahead identification framework.

### A. Comparison using random stable models

To assess the performance of SIMBa on standard SI problems, we started by generating 50 random stable discrete-time state-space models, from which we simulated one trajectory of 300 steps, starting from  $x_0 = 0$ , for the training, validation, and testing data, respectively. For the three trajectories, each dimension of  $u \in \mathbb{R}^m$  was generated as a Generalised Binary Noise (GBN) signal with a switching probability of 0.1 [19]. Before fitting the methods, we then added white noise  $v \sim \mathcal{N}(0, 0.25)$  to the output training data. For this experiment, we arbitrarily chose  $n = 5$ ,  $m = 3$ ,  $p = 3$ , and the default parameters for SIMBa, except for the number of epochs, increased to 50,000 to ensure convergence.

<sup>1</sup>Since several benchmark methods are available, SIMBa uses the one achieving the best performance on the validation data.

<sup>2</sup>The code and data used for this paper can be found on <https://gitlab.nccr-automation.ch/loris.dinatale/simba-ecc>.

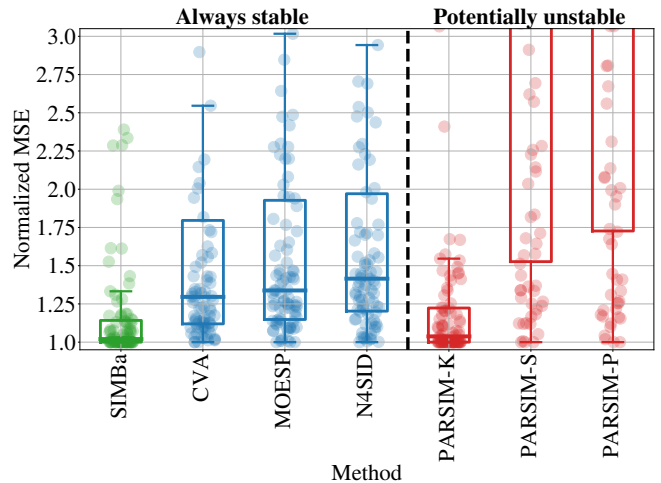


Fig. 1. Performance of input-output state-space SI methods on 50 randomly generated systems, where the MSEs have been normalized by the best-obtained error for each system. The performance of SIMBa (ours) is plotted in green, other stable SI methods in blue, while red indicates methods without stability guarantees. Key metrics are reported in Table I for clarity.

TABLE I  
NORMALIZED MSE OF EACH METHOD COMPARED TO THE BEST ONE,  
REPORTED FROM FIG. 1.

Method	0.25-quantile	Median	0.75-quantile
SIMBa	<b>1.00</b>	<b>1.02</b>	<b>1.14</b>
CVA	1.12	1.30	1.80
MOESP	1.15	1.34	1.93
N4SID	1.20	1.42	1.97
PARSIM-K	<b>1.00</b>	1.04	1.22
PARSIM-S	1.53	3.65	37.5
PARSIM-P	1.73	5.37	84.2

The performance of each SI method on the testing trajectory is plotted in Fig. 1, where green indicates SIMBa, blue other stable SI methods, and red PARSIMs, which cannot enforce stability. For each system, the MSE of each method was normalized with respect to the best-attained performance by any approach to generate the box plots, and the corresponding key metrics are reported in Table I. For a better visual representation, we overlaid the corresponding clouds of points, where we added random noise on the x-axis to distinguish them better. Note that this zoomed-in plot does not show one instance where SIMBa did not converge and attained poor performance, while it discards three such instances for PARSIM-K and many points with a normalized MSE between 3 and 7 for the methods in blue.

Overall, SIMBa shows the most robust performance, with 75% of its instances achieving an error within 14% of the best performance and half of them being near-optimal (see Table I). The only method coming close is PARSIM-K, but its performance is slightly more spread out and it cannot guarantee stability. If we only look at other stable SI methods, their median accuracy is at least 30% worse than the best one half of the time. In fact, their performance drop is more than 12% three times out of four, compared to approximately one-fourth of the time for SIMBa, and their

accuracy on the various systems is significantly more spread out. To summarize, SIMBa takes the best out of both worlds, simultaneously achieving state-of-the-art performance and stability guarantees.

### B. Performance on real-world input-output data

In this Section, we leverage DAISY, a database for SI [40], to test our framework on real-world data. In particular, we analyze the performance of SIMBa in detail on the data collected in a 120 MW power plant in Pont-sur-Sambre, France, where  $m = 5$  and  $p = 3$ . It gathers 200 data points with a sampling time of  $\delta = 1228.8$  seconds. Here, we first standardized the input and output data so that each dimension is zero-mean and has a standard deviation of one.<sup>3</sup> We use the first 100 and 150 samples for training and validation, respectively, and hold out the last 50 ones for testing the final performance of the models.<sup>4</sup>

We investigate four variations of SIMBa, encoded in their names: an “*i*” indicates instances with `init_from_matlab_or_ls=True`, and an “*L*” that SIMBa was run for more epochs to ensure convergence. Specifically, the number of epochs with “*L*” is pushed from 10,000 to 20,000 for *SIMBa<sub>i</sub>* and from 25,000 to 50,000 otherwise. We set `dropout=0`, `learn_x0=True` — since it is unknown —, and leave the other parameters at their default values. Since the true order of the system is unknown, one could leverage MATLAB’s SI toolbox to first find the most appropriate  $n$  and then run SIMBa to gain time. Here, we instead show that SIMBa dominates all the other methods from the MATLAB SI toolbox for *any* choice of  $n$ . The PARSIM methods are not analyzed here since they all diverged for at least one value  $n$  due to instability. Similarly, the stable methods from SIPPY achieved poor accuracy due to their assumption that  $x_0 = 0$ , which seems too restrictive for this data, and are thus omitted for clarity.

Fig. 2 reports the MSE of the different methods on the testing data normalized by the best performance obtained by MATLAB’s SI toolbox. This either corresponds to *N4SID*<sup>5</sup> (red crosses) or the Prediction Error Method (PEM) (orange triangles). The latter aims at improving the performance of the model found by N4SID and is thus expected to perform better on the training data. Both randomly initialized versions of SIMBa, reported in black and green, were run with 10 different seeds and the boxplot and clouds of points were generated as for Fig. 1. Since randomness has much less impact on initialized versions of SIMBa, we only report one instance of *SIMBa<sub>i</sub>* and *SIMBa<sub>iL</sub>* for clarity. Note that fitting a model with  $n \geq 7$  on 100 data samples is an

<sup>3</sup>Standardization generally has little impact on the performance, as analyzed in [39].

<sup>4</sup>Since  $x_0$  is estimated by SIMBa at training time, overlapping the training and validation data allows us to use the same initial state to validate its performance after each epoch. However, we let it run for 50 more steps to assess its extrapolation capability and avoid overfitting the training data. For testing, we rely on MATLAB’s `findstate` function to estimate  $x_0$ .

<sup>5</sup>Note that we set `N4Weight='auto'` — to automatically recover the best performance between the classical N4SID, CVA, and MOESP methods — and `Focus='simulation'` for a fair comparison with SIMBa, which is optimizing for the performance over the entire trajectory.

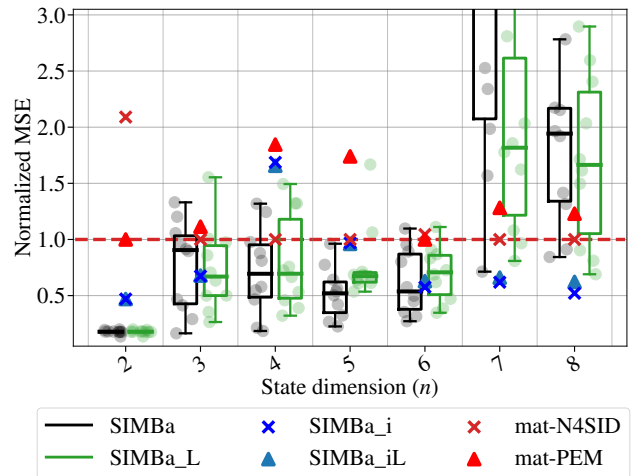


Fig. 2. MSE of the different methods on the power plant test data for different choices of state dimension  $n$ , normalized by the best performance obtained by MATLAB’s SI toolbox (red crosses and triangles). Black and green data show the performance of SIMBa over 10 runs with random initialization for shorter (*SIMBa*) and longer (*SIMBa<sub>L</sub>*) training times, respectively. Finally, blue crosses triangles represent the performance of one initialized version of SIMBa, *SIMBa<sub>i</sub>*, and a prolonged version *SIMBa<sub>iL</sub>*.

ill-posed problem, with more parameters than data points. Interestingly, however, SIMBa still manages to outperform MATLAB in some cases in this overparametrized setting, especially when it is initialized from MATLAB’s solution.

Impressively, SIMBa consistently attained the best performance for meaningful choices of  $n \leq 6$ . While the influence of randomness is non-negligible for randomly initialized versions, SIMBa always achieves the best accuracy, with improvements of up to 73–86% compared to MATLAB for different choices of  $n$ . Furthermore, half of the time, it outperforms the latter by more than 30–50% and 82% for  $n \geq 3$  and  $n = 2$ , respectively.

When being initialized with the solution of traditional SIMs, SIMBa always started from state-space matrices identified by MATLAB’s PEM method, which achieves the best performance amongst the baselines on the validation data.<sup>6</sup> Interestingly, SIMBa always improves PEM’s performance on the testing set — but sometimes not beyond N4SID, which attains the best accuracy amongst the baselines on this unseen data. In other words, PEM tends to overfit the training data and might start off SIMBa near a poor local minimum. While initializing with MATLAB’s solution allows one to converge faster, cutting the associated computational burden, it might hence not always improve the final performance.

### C. A note on the training time

*SIMBa<sub>L</sub>* was run for five times more epochs than *SIMBa<sub>i</sub>* in Section V-B, for example. Despite the small overhead required to fit  $A$  during the initialization procedure in (11), training *SIMBa<sub>i</sub>* still takes approximately only 20% of the time required to fit *SIMBa<sub>L</sub>*. In this experiment, training SIMBa ranged from 5 to 25 min — compared to a few

<sup>6</sup>Except for  $n = 2$ , where SIMBa was initialized from PARSIM-K.

seconds for MATLAB — on a MacBook Pro 2.6 GHz 6-Core Intel Core i7 laptop, irrespective of the choice of  $n$ . More details can be found in [39].

## VI. CONCLUSION

This paper presented SIMBa, a linear state-space SI toolbox leveraging well-established ML tools to minimize the multi-step-ahead prediction error and a novel LMI-based free parametrization of Schur matrices to ensure stability. SIMBa proved to be extremely flexible, worked with various types of systems on both simulated and real-world data, and achieved impressive performance despite enforcing stability, clearly outperforming traditional stable SI methods.

The significant average performance gains observed throughout our analyses — ranging from approximately 25–30% on simulated to 30–50% on real-world data — compared to other stable SI methods make up for the associated increased computational burden. While initializing SIMBa with the solution of traditional SI methods accelerates convergence toward meaningful solutions, it might be detrimental in the long run and get stuck in local minima.

For future works, it would be interesting to investigate the theoretical properties of SIMBa, such as the exact data requirements needed to attain good performance or its potential integration into Koopman-based approaches. We also plan to improve the open-source toolbox towards a general tool for knowledge-grounded structured nonlinear system identification.

## REFERENCES

- [1] L. Ljung, “System identification,” in *Signal analysis and prediction*. Springer, 1998, pp. 163–173.
- [2] Y. Goldberg, *Neural network methods for natural language processing*. Springer Nature, 2022.
- [3] J. Degraeve, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. de Las Casas *et al.*, “Magnetic control of tokamak plasmas through deep reinforcement learning,” *Nature*, vol. 602, no. 7897, pp. 414–419, 2022.
- [4] LeCun, Yann and Bengio, Yoshua and Hinton, Geoffrey, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [5] P. Werbos, “Beyond regression: New tools for prediction and analysis in the behavioral sciences,” *PhD thesis, Committee on Applied Mathematics, Harvard University, Cambridge, MA*, 1974.
- [6] L. Ljung, C. Andersson, K. Tiels, and T. B. Schön, “Deep learning and system identification,” *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 1175–1181, 2020.
- [7] M. Forgiione and D. Piga, “Continuous-time system identification with neural networks: Model structures and fitting criteria,” *European Journal of Control*, vol. 59, pp. 69–81, 2021.
- [8] S. L. Brunton and J. N. Kutz, *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press, 2022.
- [9] D. Gedon, N. Wahlström, T. B. Schön, and L. Ljung, “Deep state space models for nonlinear system identification,” *IFAC-PapersOnLine*, vol. 54, no. 7, pp. 481–486, 2021.
- [10] G. I. Beintema, M. Schoukens, and R. Tóth, “Deep subspace encoders for nonlinear system identification,” *Automatica*, vol. 156, p. 111210, 2023.
- [11] D. Masti and A. Bemporad, “Learning nonlinear state–space models using autoencoders,” *Automatica*, vol. 129, p. 109666, 2021.
- [12] W. J. Rugh, *Linear system theory*. Prentice-Hall, Inc., 1996.
- [13] K. R. Muske and J. B. Rawlings, “Model predictive control with linear models,” *AIChE Journal*, vol. 39, no. 2, pp. 262–287, 1993.
- [14] I. Khalil, J. Doyle, and K. Glover, *Robust and optimal control*. Prentice hall, 1996.
- [15] M. Schoukens, “Improved initialization of state-space artificial neural networks,” in *2021 European Control Conference (ECC)*. IEEE, 2021, pp. 1913–1918.
- [16] J. C. Schulze and A. Mitsos, “Data-Driven Nonlinear Model Reduction Using Koopman Theory: Integrated Control Form and NMPC Case Study,” *IEEE Control Systems Letters*, vol. 6, pp. 2978–2983, 2022.
- [17] J. Drgoňa, A. Tuor, E. Skomski, S. Vasisht, and D. Vrabie, “Deep learning explicit differentiable predictive control laws for buildings,” *IFAC-PapersOnLine*, vol. 54, no. 6, pp. 14–19, 2021.
- [18] L. Ljung, *System identification toolbox: User’s guide*. Citeseer, 1995.
- [19] G. Armenise, M. Vaccari, R. B. Di Capaci, and G. Pannocchia, “An open-source system identification package for multivariable processes,” in *2018 UKACC 12th International Conference on Control (CONTROL)*. IEEE, 2018, pp. 152–157.
- [20] S. J. Qin, “An overview of subspace identification,” *Computers & chemical engineering*, vol. 30, no. 10–12, pp. 1502–1513, 2006.
- [21] S. J. Qin, W. Lin, and L. Ljung, “A novel subspace identification approach with enforced causal models,” *Automatica*, vol. 41, no. 12, pp. 2043–2053, 2005.
- [22] S. J. Qin and L. Ljung, “Parallel QR implementation of subspace identification with parsimonious models,” *IFAC Proceedings Volumes*, vol. 36, no. 16, pp. 1591–1596, 2003.
- [23] G. Pannocchia and M. Calosi, “A predictor form PARSIMONIOUS algorithm for closed-loop subspace identification,” *Journal of Process Control*, vol. 20, no. 4, pp. 517–524, 2010.
- [24] I. R. Manchester, M. Revay, and R. Wang, “Contraction-based methods for stable identification and robust machine learning: a tutorial,” in *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 2955–2962.
- [25] J. M. Maciejowski, “Guaranteed stability with subspace methods,” *Systems & Control Letters*, vol. 26, no. 2, pp. 153–156, 1995.
- [26] W.-b. Huang, L. le Cao, F. Sun, D. Zhao, H. Liu, and S. Yu, “Learning Stable Linear Dynamical Systems with the Weighted Least Square Method,” in *IJCAI*, vol. 1599, 2016, p. 1605.
- [27] B. Boots, G. J. Gordon, and S. Siddiqi, “A constraint generation approach to learning stable linear dynamical systems,” *Advances in neural information processing systems*, vol. 20, 2007.
- [28] S. L. Lacy and D. S. Bernstein, “Subspace identification with guaranteed stability using constrained optimization,” *IEEE Transactions on automatic control*, vol. 48, no. 7, pp. 1259–1263, 2003.
- [29] N. Gillis, M. Karow, and P. Sharma, “A note on approximating the nearest stable discrete-time descriptor systems with fixed rank,” *Applied Numerical Mathematics*, vol. 148, pp. 131–139, 2020.
- [30] W. Jongeneel, T. Sutter, and D. Kuhn, “Efficient learning of a linear dynamical system with stability guarantees,” *IEEE Transactions on Automatic Control*, 2022.
- [31] G. Mamakoukas, I. Abraham, and T. D. Murphey, “Learning stable models for prediction and control,” *IEEE Transactions on Robotics*, 2023.
- [32] J. Drgoňa, A. R. Tuor, V. Chandan, and D. L. Vrabie, “Physics-constrained deep learning of multi-zone building thermal dynamics,” *Energy and Buildings*, vol. 243, p. 110992, 2021.
- [33] J. Z. Kolter and G. Manek, “Learning stable deep dynamics models,” *Advances in neural information processing systems*, vol. 32, 2019.
- [34] M. Revay, R. Wang, and I. R. Manchester, “Recurrent equilibrium networks: Flexible dynamic models with guaranteed stability and robustness,” *IEEE Transactions on Automatic Control*, 2023.
- [35] M. Zakwan, L. Di Natale, B. Svetozarevic, P. Heer, C. N. Jones, and G. F. Trecate, “Physically consistent neural ODEs for learning multi-physics systems,” *arXiv preprint arXiv:2211.06130*, 2022.
- [36] K. Loya, J. Buzhardt, and P. Tallapragada, “Koopman operator based predictive control with a data archive of observables,” *ASME Letters in Dynamic Systems and Control*, pp. 1–7, 2023.
- [37] M. A. Lones, “How to avoid machine learning pitfalls: a guide for academic researchers,” *arXiv preprint arXiv:2108.02497*, 2021.
- [38] M. Chilali and P. Gahinet, “H/sub/spl infin//design with pole placement constraints: an lmi approach,” *IEEE Transactions on automatic control*, vol. 41, no. 3, pp. 358–367, 1996.
- [39] L. Di Natale, M. Zakwan, P. Heer, G. Ferrari Trecate, and C. N. Jones, “SIMBa: System Identification Methods leveraging Backpropagation,” *arXiv preprint arXiv:2311.13889*, 2023.
- [40] B. De Moor, P. De Gerssem, B. De Schutter, and W. Favoreel, “DAISY: A database for identification of systems,” *JOURNAL A*, vol. 38, no. 4, p. 5, 1997.