A novel Data-Driven representation of Linear Systems based on Dynamic Mode Decomposition with Control to deal with noisy data

Francesco Giannini

Abstract—**In this paper, data generated by a plant that can be described as a linear time-invariant system subject to both process and measurement noise are considered. An original and structured organization of the collected input–output data is introduced, leading to an alternative representation of the system dynamics within an extended state-space framework. Based on this formulation, the Dynamic Mode Decomposition with Control algorithm is employed to identify a Data-Driven model in a lifted space. The resulting representation enhances robustness with respect to noise, allowing for a more accurate and reliable characterization of the underlying system dynamics in the presence of disturbances and measurement uncertainty. The proposed approach is assessed through numerical simulations, which illustrate its effectiveness in capturing the system behavior and mitigating the impact of noisy data, thereby highlighting its potential for data-driven modeling and control applications.**

I. Introduction

In control theory, the term *Data-Driven* refers to the use of available data to control a plant, more than a mathematical model of the system. The main reason behind this approach is that models are expensive and hard to obtain [1]. Moreover, identifying the dynamic equations governing the plan essentially means optimizing the model to the given data, and later optimizing the controller to the previously computed model. [2] Such a two-step procedure may be suboptimal because the real objective is seeking an optimal controller compatible with available data [3]. Nevertheless, in the *Data-Driven Control* paradigm, the two-step procedure is still used, and such methods are named *Indirect* [4] [5]. The aim of this paper is to propose an identification result useful to be integrated in *Direct Data-Driven Control* schemes.

All of the *Direct Data-Driven Control schemes* are based on a milestone theorem known as the *Fundamental Lemma* (FL) [6]. This theorem guarantees that the behavior [7] of a linear time-invariant (LTI) system coincides with the image of a Hankel matrix constructed from *informative* input–output data. In this context, the term *informative* means that the associated data matrix satisfies an appropriate rank condition. As a consequence, the system dynamics can be represented implicitly, without identifying a parametric model, directly from measured trajectories. As discussed in [8], this implicit representation of the underlying dynamics constitutes the theoretical foundation of *Direct Data-Driven Control* approaches, which exploit the FL to design control laws directly from data [9] [10]. Among these approaches, one of the most widely adopted is the data-driven formulation of Model Predictive Control (MPC), referred to as Data-enabled Predictive Control (DeePC) [11] [12].

A fundamentally different *Data-Driven Control* paradigm, which does not depend on the FL framework, is grounded in Koopman operator theory [13]. The linear nature of the Koopman operator, together with the forecasting simplicity offered by LTI models derived from its eigendecomposition, has contributed to its growing popularity in the analysis and learning of dynamical systems [14]. Within this framework, nonlinear system dynamics are described through a set of observable functions, and measured data are interpreted as samples of a finite dictionary of such observables. A finite-dimensional approximation of the underlying infinite-dimensional Koopman operator is then obtained from a limited dataset using data-driven techniques, most notably Dynamic Mode Decomposition (DMD) [15]. This approach enables the identification of linear representations of nonlinear dynamics, facilitating prediction and control using well-established linear system tools. The basic DMD algorithm is originally formulated for autonomous systems. Its extension that incorporates exogenous inputs is known as Dynamic Mode Decomposition with control (DMDc) [17].

The aim of this paper is to propose a novel system representation of an unknown liner plant subject to both process and measurement noise [16]. Such a representation is based on the FL and the idea is to exploit the results of the DMDc algorithm [17] to analyze the available data collected from the system. This integration of the FL with the DMDc algorithm enables the construction of a predictor that exhibits improved performance in the noisy case with respect to classic FL-based approach.

The paper outline is hereafter summarized. After a brief paragraph about notation and basic definitions, in Section II, the DMDc method is resumed. Section III introduces the *Data-Driven* representation of a linear plant in the noise-free setting based on the FL. Section IV is the core of this paper because it presents the novel plant's description and its original integration with the DMDc algorithm. All elements introduced in the paper are resumed in the algorithm in Section V. Section VI provides numerical results on the effectiveness of the presented methodology. Final considerations conclude the paper.

Francesco Giannini is with the Department of Environmental Engineering, University of Calabria, Rende (CS), Italy, francescogiannini97@libero.it
https://orcid.org/0000-0001-5892-2049

NOTATION AND DEFINITIONS

The symbol $\mathbb{N}$ represents the set of natural numbers, $\mathbb{R}$ real numbers, $\mathbb{R}^+$ real numbers greater than 0.

Given a vector $v \in \mathbb{R}^n$, its euclidean norm is denoted as $||v||$, while the notation $v(k_1 : k_2)$ denotes the vector extracted from v containing its elements from the index k_1 to k_2. The notation $v(0 : end)$ denotes the whole vector v.

Given a matrix $A \in \mathbb{R}^{n\times m}$ its transpose is denoted as A' while its pseudo-inverse as $A^\dagger$. In a matrix, if the element $A_{i,j}$ is substituted with the symbol $*$, that element is equal to the element $A'_{j,i}$.

The identity matrix of dimension n is denoted as I_n while 0_n denotes a matrix of zeros. $\mathbf{1_{n,m}}$ is a matrix composed of ones of dimension $n \times m$.

Definition 1: Given a signal $z(k)$ - with $k \in \mathbb{N}$ the time step -, the Hankel matrix [9] associated to z is defined as:

$$\mathcal{H}_L(z(k_0 : k_0 + N)) = \begin{bmatrix} z(k_0) & \dots & z(k_0+N-L) \\ z(k_0+1) & \dots & z(k_0+N-L+1) \\ \dots & & \\ z(k_0+L) & \dots & z(k_0+N) \end{bmatrix} \quad (1)$$

When the dimension of z is known, $\mathcal{H}_L(z) = \mathcal{H}_L(z(0 : end))$ □

II. BACKGROUND: KOOPMAN THEORY AND DMDC

The aim of Koopman Theory is to predict the evolution of an unknown system given the actual measurement $y(k)$:

$$\mathcal{K}y(k) := y(k + 1) \quad (2)$$

where $\mathcal{K}$ is the Koopman operator [13]. In general, obtaining an analytical representation of the Koopman operator from data is a challenging task [18]. For this reason, several approximation techniques have been developed and are widely adopted in practice. Among them, DMD provides a finite-dimensional linear approximation of the underlying system dynamics directly from snapshot data. This connection was originally established in [19] and has gained considerable interest in DMD, Koopman theory and Data-Driven Control. Nevertheless, standard DMD does not explicitly account for external inputs, which limits its applicability to controlled systems. Dynamic Mode Decomposition with Control (DMDc), originally proposed in [17], is the data-driven method that extends DMD including the effect of control inputs. The starting idea is to collect the data as:

$$\begin{cases} X^+ := \mathcal{H}_1(x(1 : end)) \\ X := \mathcal{H}_1(x(0 : end - 1)) \\ \Upsilon := \mathcal{H}_1(u(0 : end - 1)) \end{cases} \quad (3)$$

where $x \in \mathbb{R}^n$ is the state of the plant that generated the data, under the action of the input $u \in \mathbb{R}^m$.

Therefore, considering that the plant dynamic can be assumed as linear:

$$X^+ \approx AX + B\Upsilon \quad (4)$$

It is possible to simultaneously identify A and B as:

$$\begin{bmatrix} A & B \end{bmatrix} = X^+\Omega^\dagger \quad (5)$$

where

$$\Omega = \begin{bmatrix} X \\ \Upsilon \end{bmatrix} \quad (6)$$

The matrix Ω in (5) is generally a high-dimensional data matrix and can be approximated using the SVD:

$$\Omega = \tilde{U}\tilde{\Sigma}\tilde{V}^* \quad (7)$$

The matrix $\tilde{U}$ can be split into two matrices $\tilde{U} = \begin{bmatrix} \tilde{U}_1 \\ \tilde{U}_2 \end{bmatrix}$, to provide bases for X and Υ. See [20] for details.

III. THE DATA-DRIVEN REPRESENTATION OF A LINEAR PLANT IN THE NOISE-FREE SETTING

Consider the discrete-time systems described through *restricted behavior* $\mathcal{B}_L$ [7]:

$$\mathcal{B}_L = \{w \in \mathbb{R}^{qL} : w \text{ is a } L\text{-length trajectory of the plant}\} \quad (8)$$

A trajectory [21], is here considered as the vector composed of input $u(k) \in \mathbb{R}^m$ and measured output $y(k) \in \mathbb{R}^p$:

$$w(k) = \begin{bmatrix} u(k) \\ y(k) \end{bmatrix} \in \mathbb{R}^{q:=p+m}$$

In the *Data-Driven* framework, only one trajectory is supposed to be known. Such a trajectory must be informative, in the sense of the rank of the corresponding hankel matrix. In the noise-free setting, the trajectory must be *Persistently Exciting*:

Definition 2: Considering the system in (8), a trajectory w^d is called *Persistently Exciting* (PE) of order r if

$$rank(\mathcal{H}(w^d)) = r \quad (9)$$

□

This definition allows the introduction of the *Fundamental Lemma* [6] [9]:

Theorem 1: **Fundamental Lemma**
Given an unknown LTI system, the following holds.

1) If the input is persistently exciting of order $n + L$, any element of (8) can be expressed as

$$\begin{bmatrix} \mathcal{H}_L(u^d) \\ \mathcal{H}_L(y^d) \end{bmatrix} g \quad (10)$$

where g is a decision variable whose length depends on the size of the Hankel matrix of the data.

2) The *span* of the columns of the data matrix in (10) is contained into (8).

□

From this theorem, under the hypotheses of persistently exciting input data u^d, the following data-driven representation of a LTI system in the noise-free settings arises:

$$\begin{bmatrix} U_p \\ Y_p \\ U_f \\ Y_f \end{bmatrix} g = \begin{bmatrix} u_{ini} \\ y_{ini} \\ u \\ y \end{bmatrix} \tag{11}$$

where

$$\begin{bmatrix} U_p \\ U_f \end{bmatrix} := \mathcal{H}_{t_{ini}+N}(u^d) \tag{12}$$

$$\begin{bmatrix} Y_p \\ Y_f \end{bmatrix} := \mathcal{H}_{t_{ini}+N}(y^d) \tag{13}$$

Remark 1: Notice that the initial conditions (u_{ini}, y_{ini}) of length t_{ini} are supposed to be known. Therefore, equation (11) represents a data-driven prediction of the plant trajectory on an horizon N that can be used recursively. This predictor, in this paper, is indicated as *FL representation* of the dynamic system and used for comparison purposes. □

Remark 2: For $N = 1$, one obtains a data-driven representation with stable latent poles as resumed in [8]:

$$g^* = \begin{bmatrix} U_p \\ Y_p \\ U_f \end{bmatrix}^{\dagger} \begin{bmatrix} u_{ini} \\ y_{ini} \\ u(k) \end{bmatrix} \quad \rightarrow \quad y(k) = Y_f g^* \tag{14}$$

□

In the presence of noise, the *Direct Data-Driven Control* literature typically adopts a relaxed formulation of equation (11), as discussed in [3]. In such approaches, the exact consistency constraints imposed by the measured data are softened by introducing an additional term to the initial output vector y_{ini}. The norm of this additional term is then heavily penalized within the optimization problem used to compute the control action, thereby allowing feasibility in the presence of measurement noise and disturbances. While effective, this relaxation alters the original structure of the *Data-Driven* formulation of the cost and may introduce a trade-off between dealing with noise and optimizing control performance.

The aim of this paper is to apply the DMDc algorithm in an original manner to obtain an alternative representation of the underlying system dynamics in a lifted space. This representation is specifically designed to enhance robustness with respect to noise, without explicitly relaxing the data-consistency constraints.

IV. Towards an alternative representation

First of all, it is required to introduce the extended space:

$$\xi(k) := \begin{bmatrix} u(k - t_{ini}) \\ u(k - t_{ini} + 1) \\ \dots \\ u(k-1) \\ y(k - t_{ini}) \\ y(k - t_{ini} + 1) \\ \dots \\ y(k-1) \end{bmatrix} \tag{15}$$

W.r.t. this new variable, (14) becomes:

$$\xi(k+1) := \Phi \xi(k) + G u(k) \tag{16}$$

where

$$\Phi = \begin{bmatrix} 0 \quad I \quad 0 \quad 0 \\ 0 \quad 0 \quad 0 \quad 0 \\ 0 \quad 0 \quad 0 \quad I \\ Y_f \begin{bmatrix} U_p \\ Y_p \\ U_f \end{bmatrix}^{\dagger} (1 : mt_{ini} + pt_{ini}) \end{bmatrix} \tag{17}$$

and

$$G = \begin{bmatrix} 0 \\ Y_f \begin{bmatrix} U_p \\ Y_p \\ U_f \end{bmatrix}^{\dagger} (mt_{ini} + pt_{ini} + 1 : end) \end{bmatrix} \tag{18}$$

In the above equations, the dimensions of zero and identity matrices are omitted to improve readability. Moreover, the lower blocks in (17) and (18) illustrate the connection with the FL, as explained in *Remark* 2.

As in (3), data are organized as:

$$\begin{cases} \Xi = \mathcal{H}_1(\xi(0 : T-1)) \\ \Xi^+ = \mathcal{H}_1(\xi(1 : T)) \\ \Upsilon = \mathcal{H}_1(u(0 : T-1)) \end{cases} \tag{19}$$

where T is the total number of snapshots. Using these three data matrices, DMDc focuses on finding best-fit approximations to the mappings Φ and G in the lifted space:

$$\begin{cases} \tilde{\Phi} = \hat{U}^* \Xi^+ \tilde{V} \tilde{\Sigma}^{-1} \tilde{U}_1^* \hat{U} \\ \tilde{G} = \hat{U}^* \Xi^+ \tilde{V} \tilde{\Sigma}^{-1} \tilde{U}_2^* \end{cases} \tag{20}$$

where the matrices $\hat{U}, \tilde{V}, \tilde{U}$ come from the SVD decomposition of the data matrices:

$$\begin{cases} \begin{bmatrix} \Xi \\ \Upsilon \end{bmatrix} \approx \begin{bmatrix} \tilde{U}_1 \\ \tilde{U}_2 \end{bmatrix} \tilde{\Sigma} \tilde{V}^* \\ \\ \Xi^+ \approx \hat{U} \hat{\Sigma} \hat{V}^* \end{cases} \tag{21}$$

The reason behind the approximated equalities in (21) is that the dimension of the data matrices is reduced to eliminate the terms that have small singular values. Details in [17].

Remark 3: The use of the DMDc approach, because of the SVD decomposition of the data matrix reduces the impact of the noise [17] [23] [24].

A. Noisy data

Theorem 1 and the Data-Driven description of the Dynamics in (11) are based on an idealized scenario because implicitly assume that the data are not affected by measurement noise. As remarked, even a small additive noise component affecting y_{ini} could render the corresponding constraint infeasible. Moreover, the previous considerations rely on the additional assumption that the plant dynamics are strictly linear. In practice, however, physical systems are invariably affected by exogenous disturbances, environmental uncertainties, and unmodeled dynamics. To bridge this gap between theory and practice, it is crucial to explicitly account for process noise within the system description [16]. To rigorously characterize the stochastic nature of these uncertainties, this work models the disturbances as zero-mean Gaussian white noise with a covariance matrix S, defined as:

$$n(k) \sim \mathcal{N}(0, S) \tag{22}$$

In order to establish an even more realistic simulation scenario, measurement noise is explicitly injected into the simulated output data alongside process disturbances. Thus, control strategies validated under this comprehensive noise framework demonstrate superior reliability when transitioned to practical, inherently noisy operating environments.

V. The Proposed Algorithm

Based on the above developments, the following algorithm provides a trajectory predictor for the unknown system.

DMDc predictor in the lifted space

Collect the data w^d;
Stack the data as in (19);
Compute $\tilde{\Phi}$ and $\tilde{G}$ as in (20);
For *each sampling time* k :

1: **Project** the vector $\xi(k)$ into the lifted space:

$$z(k) = \hat{U}^* \xi(k) \tag{23}$$

2: **Evolve** the system in the lifted space:

$$z(k+1) = \tilde{\Phi} z(k) + \tilde{G} u(k)$$

3: **Lift back**:

$$E[\xi(k+1)] = \hat{U} z(k) \tag{24}$$

4: **Extract** the predicted measure:

$$E[y(k+1)] = E[\xi(k+1)(end-p:end)] \tag{25}$$

VI. Simulations

The following numerical simulations of the **DMDc predictor in the lifted space** algorithm have been performed in MATLAB R2024a on a *Lenovo IdeaPad L340* equipped with a *Intel-core i7 of the 9th generation* using the *Windows 11* operating system.

For simulation purposes, the following system has been considered:

$$\begin{cases} x(k+1) = \begin{bmatrix} 0.9 & -0.2 \\ 0.1 & 1.1 \end{bmatrix} x(k) + \begin{bmatrix} 0.2 \\ -0.2 \end{bmatrix} u(k) + n(k) \\ y(k) = \begin{bmatrix} 1 & 0 \end{bmatrix} x(k) + r(k) \end{cases} \tag{26}$$

where $n(k) \in \mathbb{R}^3$ is the process noise, $r(k) \in \mathbb{R}^2$ the measurement noise. For simulation purposes:

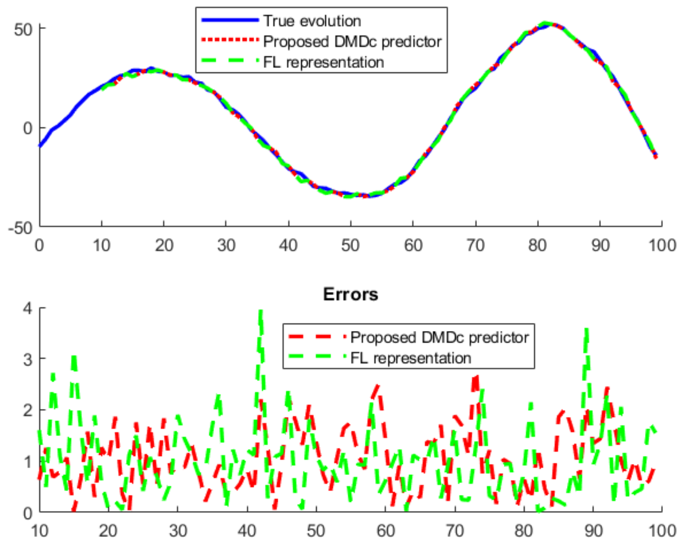

Fig. 1. Comparison between the true evolution of the system, the proposed DMDc predictor, and the predictor in (14).

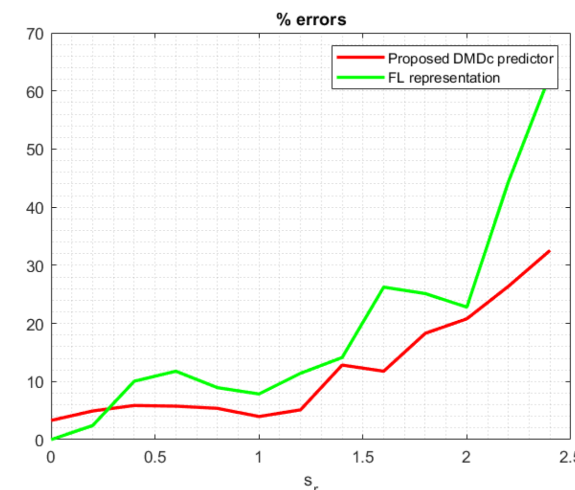

Fig. 2. Comparison between the proposed DMDc predictor, and the predictor in (14) w.r.t. the parameter s_r in (27).

$$\begin{cases} n(k) \sim \mathcal{N}(0, s_r\, I_2) \\ r(k) \sim \mathcal{N}(0, 10\, s_r) \end{cases} \tag{27}$$

Fig. 1 shows the evolution of the system under a random-generated sequence of control inputs, comparing the true evolution, the one based on the proposed DMDc predictor, and the one based on the FL in (14) with $s_r = 0.1$. The lower part of the figure assess that the proposal works better in terms of the following error:

$$e(k) = ||E[y(k)] - y(k)|| \tag{28}$$

where the expected value of the measurement is computed as in (25) and as in (14).

Fig. 2 presents a quantitative comparison between the proposed DMDc-based predictor and the predictor defined in

(14), evaluated as a function of the parameter s_r appearing in the noise model (27). The performance comparison is carried out using the mean value, computed over the entire trajectory, of the prediction error $e(k)$ defined in (28). As expected, in the noise-free case, the predictor in (14) achieves superior performance, since it provides an exact representation of the underlying linear time-invariant system. However, as the noise level increases and deviates from zero, a progressive degradation in the performance of this predictor is observed. In contrast, the proposed DMDc-based approach exhibits improved robustness to noise. The observed small oscillations in the performance curves can be attributed to the stochastic nature of the noise affecting the data generation process.

Conclusion

This paper has presented a data-driven identification approach for linear time-invariant systems affected by process and measurement noise, based on an original use of the Dynamic Mode Decomposition with Control algorithm. By organizing input–output data in a structured manner and constructing a linear representation of the system dynamics in a lifted space, the proposed method enhances robustness with respect to noisy measurements. Numerical simulations have demonstrated the ability of the identified model to accurately capture the system behavior in the noisy case better than the traditional solution based on the straightforward application of the Fundamental Lemma.

In future work, this paper will be the foundation for the development of a novel formulation of *Direct Data-Driven Control* schemes that does not require modifications to the objective function. to deal with noisy data In particular, this robust description of LTI systems will be used in the set-membership Data-Driven predictive strategy proposed in [25].

References

[1] M. Gevers, "Identification for control: From the early achievements to the revival of experiment design," *European Journal of Control*, vol. 11, no. 4–5, pp.335–352, 2005.

[2] G. Franzé, F. Giannini, V. Puig, and G. Fortino, "Embedding the state trajectories of nonlinear systems via multimodel linear descriptions: a data-driven-based algorithm," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024.

[3] I. Markovsky, L. Huang, and F. Dörfler, "Data-driven control based on the behavioral approach: From theory to applications in power systems," *IEEE Control Systems Magazine*, vol. 43, no. 5, pp. 28–68, 2023.

[4] G. Tang and K. Hauser, "A data-driven indirect method for nonlinear optimal control," *Astrodynamics*, vol. 3, no. 4, pp. 345–359, 2019.

[5] F. Giannini and D. Famularo, "A data-driven approach to set-theoretic model predictive control for nonlinear systems," *Information*, Vol. 15, no. 7, 2024.

[6] J. C. Willems, P. Rapisarda, I. Markovsky, and B. L. M. De Moor, "A note on persistency of excitation," *Systems & Control Letters*, vol. 54, no. 4, pp. 325–329, 2005.

[7] J. C. Willems, "The behavioral approach to open and interconnected systems," *IEEE Control Systems Magazine*, vol. 27, no. 6, pp. 46–99, 2007.

[8] J. Lee, N. H. Jo, H. Shim, F. Dörfler, and J. Kim, "Input-Output Data-Driven Representation: Non-Minimality and Stability," *arXiv preprint*, arXiv:2512.01238, 2025.

[9] C. De Persis and P. Tesi, "Formulas for data-driven control: Stabilization, optimality, and robustness," *IEEE Transactions on Automatic Control*, vol. 65, no. 3, pp. 909–924, 2019.

[10] M. Bajelani, W. Lucia, and K. Van Heusden, "Set-theoretic direct data-driven predictive control," in *2025 European Control Conference (ECC)*, pp. 2249–2254, 2025.

[11] J. Coulson, J. Lygeros, and F. Dörfler, "Data-enabled predictive control: In the shallows of the DeePC," in *18th European Control Conference (ECC)*, pp. 307–312, 2019.

[12] C. X. Zheng, R. Costa-Castelló and V. Puig, "Tube-based Robustification of Data Enabled Predictive Control with Guarantees," in *IEEE 64th Conference on Decision and Control (CDC)*, Rio de Janeiro, Pp. 1314-1319, 2025.

[13] B. O. Koopman, "Hamiltonian systems and transformation in Hilbert space," *Proceedings of the National Academy of Sciences*, Vol. 17, no. 5, pp. 315-318, 1931.

[14] P. Bevanda, M. Beier, A. Lederer, A. Capone, S. Sosnowski and S. Hirche, "Koopman-equivariant gaussian processes," *arXiv preprint* arXiv:2502.06645, 2025.

[15] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *Journal of fluid mechanics*, Vol. 656, Pp. 5-28, 2010.

[16] D. Simon, "Optimal state estimation: Kalman, H infinity, and nonlinear approaches," *John Wiley & Sons*, 2006.

[17] J. L. Proctor, S. L. Brunton, and J. N. Kutz, "Dynamic mode decomposition with control," *SIAM Journal on Applied Dynamical Systems*, Vol. 15, no. 1, pp. 142-161, 2016.

[18] E. Kaiser, J. N. Kutz, S. L. Brunton, "Data-driven discovery of koopman eigenfunctions for control," *Machine Learning: Science and Technology*, Vol. 2, 2021.

[19] C. W. Rowley, I. Mezić, S. Bagheri, P., Schlatter, and D. S. Henningson, "Spectral analysis of nonlinear flows," *Journal of fluid mechanics*, Vol. 641, Pp. 115-127, 2009.

[20] S. L. Brunton, J. N. Kutz, "Data-driven science and engineering: Machine learning, dynamical systems, and control," *Cambridge University Press*, 2022.

[21] I. Markovsky and F. Dörfler, "Identifiability in the behavioral setting," *IEEE Transactions on Automatic Control*, Vol. 68, no. 3, Pp. 1667-1677, 2022.

[22] A. Sasfi, I. Markovsky, A. Padoan, and F. Dörfler, "Gaussian behaviors: representations and data-driven control," *arXiv preprint arXiv:2504.15838*, 2025.

[23] J. Coulson, H. J. Van Waarde, J. Lygeros, and F. Dörfler, "A quantitative notion of persistency of excitation and the robust fundamental lemma," *IEEE Control Systems Letters*, vol. 7, pp. 1243–1248, 2022.

[24] D. Donoho and M. Gavish, "Minimax risk of matrix denoising by singular value thresholding," *Ann. Statist.*, Vol. 42, Pp. 2413-2440, 2014.

[25] F. Giannini, and F. Ranieri, "A Robust Direct Data-Driven Ellipsoidal Predictive Control scheme," *IEEE Control Systems Letters*, Vol. 10, pp. 163-168, 2026.