Data-driven dynamic interpolation and approximation

Ivan Markovsky [a] and Florian Dörfler [b]

[a] I. Markovsky is with the Department ELEC, Vrije Universiteit Brussel, 1050 Brussels, Belgium.

[b] F. Dörfler is with the Automatic Control Laboratory (IfA), ETH-Zürich, 8092 Zürich, Switzerland.

Abstract

The behavioral system theory and in particular a result that became known as the "fundamental lemma" give the theoretical foundation for nonparameteric representations of linear time-invariant systems based on Hankel matrices constructed from data. These "data-driven" representations led in turn to new system identification, signal processing, and control methods. This paper shows how the approach can be used further on for solving interpolation, extrapolation, and smoothing problems. The solution proposed and the resulting method are general—can deal simultaneously with missing, exact, and noisy data of multivariable systems—and simple—require only the solution of a linear system of equations. In the case of exact data, we provide conditions for existence and uniqueness of solution. In the case of noisy data, we propose an approximation procedure based on ℓ_1-norm regularization and validate its performance on real-life datasets. The results have application in missing data estimation and trajectory planning. They open a practical computational way of doing system theory and signal processing directly from data without identification of a transfer function or state-space system representation.

Key words: behavioral approach; system identification; data-driven methods; dynamic interpolation; missing data estimation; smoothing.

1 Introduction

Signal processing and control currently undergo a shift of paradigm from model-based to *data-driven*. One of the approaches for data-driven analysis and design is based on Hankel matrices constructed from observed data. A theoretical basis for this approach is given in the *behavioral setting* [23,19,16], where a dynamical system is defined as a set of trajectories—vector-valued signals without a priori separation of the variables into inputs and outputs. This makes the behavioral setting intrinsically data-driven. A setting that does not consider a dynamical system an input-output map is both general and useful, as illustrated by ample examples of modeling physical systems from first principles [24].

The key technical result that justifies the use of Hankel matrices constructed from observed data in data-driven analysis and design is the *fundamental lemma* [25]. It provides a nonparametric representation of a linear time-invariant system using observed data of the system, assuming that: 1. the data is exact (noise free), 2. the data generating system is controllable, and 3. an input component of the data is persistently exciting of a "sufficiently" high order. The fundamental lemma is originally derived for data consisting of a single trajectory. It is subsequently generalized to data consisting of multiple trajectories [22] and for uncontrollable systems [17]. The fundamental lemma and its generalizations give sufficient conditions and assume a given input/output partitioning of the data. An alternative result that gives necessary and sufficient conditions and does not assume a priori known input/output partitioning is [12, Theorem 15].

Both the fundamental lemma and [12, Theorem 15] lead to a data-driven nonparametric representation of linear time-invariant systems, which expresses the behavior of the system restricted to a finite-horizon as the image of a Hankel matrix constructed from data. The practical consequence of this fact is that any finite-horizon trajectory of the system can be expressed as a linear combination of the columns of the Hankel matrix. This led to algorithms based on the raw data rather than parametric representations, see [13].

This paper contributes to the line of research on data-driven signal processing using model representations based on Hankel matrices. We consider data-driven versions of the following well known system theory / signal processing problems for discrete-time linear time-invariant system:

- *simulation*—given initial conditions and input, find the corresponding output of the system;
- *smoothing*—given an inexact/noisy trajectory, find a trajectory of the system that is the optimal (in some specified

⋆ This paper was not presented at any IFAC meeting. Corresponding author I. Markovsky.

Email addresses: imarkovs@vub.be (Ivan Markovsky), dorfler@ethz.ch (Florian Dörfler).

sense) approximation of the given trajectory; and

- *prediction*—given inexact/noisy "past" trajectory, find a "future" trajectory of the system that is the optimal (in some specified sense) continuation of the past trajectory.

In the simulation problem the given data—initial conditions and input—is exact and the computed result—the output—is exact. In the smoothing and prediction problems the given data is a signal that is assumed to be an exact trajectory of the system corrupted by additive zero-mean Gaussian noise. This leads to maximum-likelihood estimation problems with the optimality criterion being a (weighted) 2-norm minimization of the fitting error. A recursive solution method is given by the *errors-in-variables Kalman smoother* [6,11]. The problem formulation and solution method of [6] is based on a transfer functions approach while the one of [11] on a state-space approach. In this paper, we use instead a nonparametric data-driven representation that requires a completely specified informative trajectory of the system. A related, more general, albeit less tractable solution approach is Hankel structured low-rank matrix completion [10].

A generalization of the simulation problem is:

- *interpolation*—given a subsample of values of a trajectory, find all trajectories of the system that fit the data.

The simulation problem is interpolation with the given values being the "past" inputs and outputs (initial conditions) and the "future" inputs. The extra flexibility of the interpolation problem, allowing for arbitrary combination of given variables at arbitrary moments of time, can be used for *missing data recovery*: the part of the trajectory that is not given is recovered from the given data and system. There are many situations when missing data occurs in practice, *e.g.*, partial observations due to failing sensors, packet losses in communication networks, and event-triggered estimation [7]. Another application of the interpolation problem is *trajectory planning* for control [5]. The design specification in trajectory planning is that the desired trajectory passes through some pre-specified points (the interpolation constraints).

Contrary to the simulation problem, which always has a unique solution, the interpolation problem may have a nonunique solution or no solution at all. We show that the set of solutions is affine and characterize it explicitly. The result can be used in optimal control for finding a trajectory that satisfies the interpolation conditions and is moreover optimal in some specified sense. In the case of no solution, *e.g.*, due to noisy data, we define an optimal approximation problem with optimality criterion being a weighted 2-norm minimization of the fitting error with respect to the given data. This approximation problem includes as special cases the smoothing and prediction problems. Thus, the general problem considered includes the simulation, interpolation, smoothing, and prediction problems.

In the data-driven versions of the problems, the data generating system is not given, and the aim is to find a direct map from the given (possibly noisy) data to the desired result (simulated, smoothed, predicted, or interpolated signal), without derivation of a parametric model representation. Instead of the data generating system one or more *complete* trajectories of the system that specify it implicitly are given.

The main contribution of the paper is posing and solving the data-driven dynamic interpolation and approximation problem. For exact data, we give existence and uniqueness conditions that are easily verifiable from the data. In the case of a nonunique solution, we show that the set of solutions is affine and explicitly characterize it. For inexact data, we propose an approximation procedure based on ℓ_1-norm regularization and validate its performance empirically. Empirical results on real-life datasets from the data-base for system identification DAISY [3] show that by tuning a hyper parameter, the ℓ_1-norm regularization method may outperform alternative model-based approaches.

The paper is organized as follows. Section 2 introduces the technical results used in the rest of the paper—specification of initial conditions by a prefix trajectory, trajectory-based representation, and its link to data-driven algorithms. The basic interpolation problem and its solution are presented in Sections 3. The assumptions needed for exact interpolation are relaxed in Section 4. The set of interpolants is characterized in case of a nonunique solution, an approximation problem is introduced when an exact interpolant does not exist, and a convex relaxation is proposed for dealing with noisy data. Section 5 shows the performance of the relaxation on real-life data sets.

2 Preliminaries

The set of q-variate discrete-time signals $w : \mathbb{N} \to \mathbb{R}^q$ is denoted by $(\mathbb{R}^q)^{\mathbb{N}}$. The *cut operator* $w|_L$ restricts w to the interval $[1,L]$ [1], *i.e.*, $w|_L := \big(w(1),\ldots,w(L)\big)$. With some abuse of notation, we view the finite L-samples long signal $w \in (\mathbb{R}^q)^L$ also as a qL-dimensional vector $w \in \mathbb{R}^{qL}$.

In the behavoral setting a discrete-time dynamical system $\mathscr{B}$ is defined as a set of trajectories, *i.e.*, $\mathscr{B} \subset (\mathbb{R}^q)^{\mathbb{N}}$. If the system $\mathscr{B}$ is linear, $\mathscr{B}$ is a subspace, and if it is time-invariant $\sigma\mathscr{B} = \mathscr{B}$, where $(\sigma w)(t) := w(t+1)$ is the *shift operator*. The class of linear time-invariant systems with q variables is denoted by $\mathscr{L}^q$.

The variables $w(t) \in \mathbb{R}^q$ of a system $\mathscr{B}$ can be partitioned into inputs $u(t) \in \mathbb{R}^m$ and outputs $y(t) \in \mathbb{R}^p$, *i.e.*, there is a permutation matrix $\Pi \in \mathbb{R}^{q\times q}$, such that $w = \Pi \left[\begin{smallmatrix} u \\ y \end{smallmatrix}\right]$. The partitioning of the variables into inputs and outputs is in general not unique. The number of inputs $\mathbf{m}(\mathscr{B})$, however, is invariant of the choice of the partitioning and is, therefore, a property of the system $\mathscr{B}$.

[1] $[1,L]$ denotes the interval of the integers from 1 to L.

Note 1 *Considering all variables on an equal footing,* i.e., *not introducing an a priori fixed input/output partitioning, is a more general modeling setting [24]. The difficulty of introducing an a priori fixed input/output partitioning is evident for example in modeling dynamic networks, such as the ones occurring in biological and complex man-made systems,* e.g., *circuits. In machine learning, the shift of paradigm from input/output maps to relations corresponds to the shift from supervised to unsupervised learning.*

For a linear time-invariant system $\mathscr{B} \in \mathscr{L}^q$, $\mathscr{B}|_L$ is the restriction of the behavior to the interval $[1,L]$, *i.e.*, $\mathscr{B}|_L := \{w|_L \mid w \in \mathscr{B}\}$. The restriction $\mathscr{B}|_L$ is a subspace of $\mathbb{R}^{qL}$ with dimension $\dim \mathscr{B}|_L = \mathbf{m}(\mathscr{B})L + \mathbf{n}(\mathscr{B})$, for $L \geq \mathbf{l}(\mathscr{B})$, where $\mathbf{m}(\mathscr{B})$ is the number of inputs, $\mathbf{l}(\mathscr{B})$ is the *lag*, and $\mathbf{n}(\mathscr{B})$ is the *order* of the system [12].

As shown in [14, Lemma 1], initial conditions for a trajectory $w \in \mathscr{B}$ can be specified by a prefix trajectory w_{ini} of length greater than or equal to the lag $\mathbf{l}(\mathscr{B})$ of the system.

Lemma 2 (Initial conditions specification [14]) *Let $\mathscr{B} \in \mathscr{L}^q$ admits an input/output partition $w = (u,y)$. Then, for any given $w_{\text{ini}} \in (\mathbb{R}^q)^{T_{\text{ini}}}$, with $T_{\text{ini}} \geq \mathbf{l}(\mathscr{B})$, and $u \in (\mathbb{R}^m)^L$, there is a unique $y \in (\mathbb{R}^p)^L$, such that $w_{\text{ini}} \wedge (u,y) \in \mathscr{B}|_{T_{\text{ini}}+L}$, where $w_{\text{p}} \wedge w_{\text{f}}$ denotes the concatenation of w_{p} and w_{f}.*

Consider a system $\mathscr{B} \in \mathscr{L}^q$. For any $w_{\text{d}} \in \mathscr{B}|_T$ and $L \in \mathbb{N}$, $1 \leq L \leq T$, we have that $\mathscr{B}|_L \subseteq \text{image}\, \mathscr{H}_L(w_{\text{d}})$, where $\mathscr{H}_L(w)$ the *Hankel matrix* with L block rows

$$\mathscr{H}_L(w) := \begin{bmatrix} w(1) & w(2) & \cdots & w(T-L+1) \\ w(2) & w(3) & \cdots & w(T-L+2) \\ \vdots & \vdots & & \vdots \\ w(L) & w(L+1) & \cdots & w(T) \end{bmatrix} \in \mathbb{R}^{qL \times T-L+1}.$$

If equality holds, *i.e.*,

$$\mathscr{B}|_L = \text{image}\, \mathscr{H}_L(w_{\text{d}}), \tag{1}$$

we have a data-driven representation of the system. Indeed, (1) characterizes all L-samples long trajectories of the system using directly the given data w_{d} without derivation of a parametric model for the system. However, for (1) to hold true additional conditions have to be satisfied. Such conditions are given in [25, Theorem 1] and [12, Theorem 15].

Theorem 3 ([Theorem 15 in [12]) *Consider $w_{\text{d}} \in \mathscr{B}|_T$, where $\mathscr{B} \in \mathscr{L}^q$ and $L \geq \mathbf{l}(\mathscr{B})$. Then, (1) holds if and only if*

$$\text{rank}\, \mathscr{H}_L(w_{\text{d}}) = \mathbf{m}(\mathscr{B})L + \mathbf{n}(\mathscr{B}). \tag{2}$$

Corollary 4 *Let $w_{\text{d}} \in \mathscr{B}|_T$, where $\mathscr{B} \in \mathscr{L}^q$, $L \geq \mathbf{l}(\mathscr{B})$, and (2) holds. Then, for any $w \in \mathscr{B}|_L$ there is a g, such that*

$$w = \mathscr{H}_L(w_{\text{d}})g. \tag{3}$$

Vice versa, any $g \in \mathbb{R}^{T-L+1}$ defines via (3) an L-samples long trajectory w of $\mathscr{B}$.

Note 5 (Interpretation of g) *In (3), the vector g "selects" a trajectory $w \in \mathscr{B}|_L$, in the sense that any $g \in \mathbb{R}^{T-L+1}$ specifies a trajectory $w \in \mathbb{R}^{qL}$. The vector g plays the role of the initial state $x(0)$ and input u in an input/state/output representation of the system. There is an important difference between g and $(x(0),u)$, however. The map $g \mapsto w$ is in general not injective,* i.e., *there may be many g's that map to the same w (see Lemma 6), while $(x(0),u)$ is in a one-to-one correspondence with w. Since each column of the Hankel matrix is a trajectory, from the perspective of motion primitives, dictionary learning, or basis functions [1], the Hankel matrix serves as a "trajectory library". A linear collection of elements (bases or motion primitives) from this library forms a new trajectory of $\mathscr{B}|_L$. Under the conditions of Theorem 3, this trajectory library is complete,* i.e., *it spans $\mathscr{B}|_L$.*

In general, a solution g to (3) is not unique.

Lemma 6 (Nonuniqueness of g) *Let $w_{\text{d}} \in \mathscr{B}|_T$, where $\mathscr{B} \in \mathscr{L}^q$, $L \geq \mathbf{l}(\mathscr{B})$, and (2) holds. Then, for any $w \in \mathscr{B}|_L$ the solution set $\mathscr{G}$ of (3) is nonempty and is given by $\mathscr{G} = g_{\text{p}} + \mathscr{N}_g$, where g_{p} is a particular solution and $\mathscr{N}_g$ is the null space of $\mathscr{H}_L(w_{\text{d}})$. Vice versa, for a given $g'_{\text{p}} \in \mathscr{G}$, any $g \in g'_{\text{p}} + \mathscr{N}_g$ define via (3) the same L-samples long trajectory $w = \mathscr{H}_L(w_{\text{d}})g_{\text{p}}$ of $\mathscr{B}$.*

Formulation (3) is used in [14] for solving data-driven prediction and control problems. Here, it is used for solving data-driven interpolation and approximation problems.

3 Problem formulation and basic solution method

A signal $w \in (\mathbb{R}^q)^L$ is also a qL-dimensional vector. For a vector of indices $I \in \mathbb{N}^K$, where $I_i \in \{1,\ldots,qL\}$, for $i = 1,\ldots,K$, we define $w|_I := [w_{I_1} \cdots w_{I_K}]^\top \in \mathbb{R}^K$, *i.e.*, $w|_I$ is the subvector of $w \in \mathbb{R}^{qL}$ with indices I.[2] Similarly, $\mathscr{H}_L(w_{\text{d}})|_I$ is the submatrix of $\mathscr{H}_L(w_{\text{d}})$ with row indices I. I_{given} denotes the indices of the *given elements* of the *to-be-interpolated trajectory* w. The set of indices I_{missing} of the missing elements of w is then the set difference of $\{1,\ldots,qL\}$ and I_{given}.

First we consider exact data $w_{\text{d}} \in \mathscr{B}|_T$. This assumption is relaxed in Section 4.3.

Problem 7 (Data-driven interpolation) *Given a trajectory $w_{\text{d}} \in \mathscr{B}|_T$ of a linear time-invariant system $\mathscr{B} \in \mathscr{L}^q$ and a partially specified trajectory $w|_{I_{given}}$, find a trajectory $\widehat{w} \in \mathscr{B}|_L$, such that $\widehat{w}|_{I_{given}} = w|_{I_{given}}$.*

Problem 7 defines a map $(w_{\text{d}}, w|_{I_{\text{given}}}) \mapsto \widehat{w} \in \mathscr{B}$, which recovers the missing values $w|_{I_{\text{missing}}}$, *i.e.*, $\widehat{w}|_{I_{\text{missing}}} = w|_{I_{\text{missing}}}$.

[2] The notation $w|_I$ for I being a set overloads the notation $w|_L$ defined earlier for $L \in \mathbb{N}$.

We call $\widehat{w}|_{I_{\text{missing}}}$ a *completion* of the given data $w|_{I_{\text{given}}}$. The constraint imposed in Problem 7 on the completion $\widehat{w}|_{I_{\text{missing}}}$ is that the complete trajectory $\widehat{w}$ is a valid trajectory of $\mathscr{B}$.

In the rest of the paper we refer to the following assumptions:

- A1. w_{d} satisfies condition (2),
- A2. $\operatorname{rank} \begin{bmatrix} w|_{I_{\text{given}}} & \mathscr{H}_L(w_{\text{d}})|_{I_{\text{given}}} \end{bmatrix} = \operatorname{rank} \mathscr{H}_L(w_{\text{d}})|_{I_{\text{given}}}$,
- A3. $\operatorname{rank} \mathscr{H}_L(w_{\text{d}})|_{I_{\text{given}}} = \operatorname{rank} \mathscr{H}_L(w_{\text{d}}) = \mathbf{m}(\mathscr{B})L + \mathbf{n}(\mathscr{B})$.

Assumption A1 is necessary and sufficient for the trajectory based representation (1). Assumption A2 is a consistency assumption: it guarantees that the given data $w|_{I_{\text{given}}}$ is consistent with the data generating system $\mathscr{B}$, *i.e.*, there is a completion $w|_{I_{\text{missing}}}$ of $w|_{I_{\text{given}}}$, such that $w \in \mathscr{B}|_L$. As shown next, Assumptions A1 and A2 guarantee existence of solution of Problem 7. For uniqueness, the stronger Assumption A3 is needed in place of Assumption A1. Verifying the assumptions requires, in addition to the data $(w_{\text{d}}, w|_{I_{\text{given}}})$, prior knowledge of the number of inputs $\mathbf{m}(\mathscr{B})$ and the order $\mathbf{n}(\mathscr{B})$ of the data generating system $\mathscr{B}$. However, the method proposed is based on the nonparameteric representation (1), and doesn't require prior knowledge of $\mathbf{m}(\mathscr{B})$ and $\mathbf{n}(\mathscr{B})$.

Selecting the equations in (3) corresponding to $w|_{I_{\text{given}}}$, gives us the interpolation condition:

$$\text{there is a } g, \text{ such that } w|_{I_{\text{given}}} = \mathscr{H}_L(w_{\text{d}})|_{I_{\text{given}}} g. \qquad (4)$$

Any solution g of (4) defines via $\widehat{w} := \mathscr{H}_L(w_{\text{d}})g$ an interpolant of $w|_{I_{\text{given}}}$. This leads to the basic algorithm for data-driven interpolation, given in Algorithm 1. Note that the algorithm does not require the number of inputs $\mathbf{m}(\mathscr{B})$ and the order $\mathbf{n}(\mathscr{B})$ of the data generating system $\mathscr{B}$.

Algorithm 1 Data-driven interpolation.

Input: w_{d}, I_{given}, and $w|_{I_{\text{given}}}$.
1: Solve $w|_{I_{\text{given}}} = \mathscr{H}_L(w_{\text{d}})|_{I_{\text{given}}} g$ for g.
2: Let $\widehat{w} := \mathscr{H}_L(w_{\text{d}})g$.
Output: $\widehat{w}$.

Proposition 8 (Existence of $\widehat{w}$) *Assuming A1 and A2, 1) Problem 7 has a solution and 2) Algorithm 1 computes a solution of Problem 7.*

PROOF. The proof of the proposition that Problem 7 has a solution is constructive. It follows from the proposition that Algorithm 1 computes a solution of Problem 7. Therefore, we need to prove only the latter. In order to do this, we need to show that: 1) $\widehat{w}$ is a trajectory of $\mathscr{B}$, *i.e.*, $\widehat{w} \in \mathscr{B}|_L$, and 2) $\widehat{w}$ interpolates the given data, *i.e.*, $\widehat{w}|_{I_{\text{given}}} = w|_{I_{\text{given}}}$.

Using Assumption A1, by Theorem 3, (1) holds true. By Assumption A2, there is a solution g to (4). Then, for any solution g to (4), $\widehat{w}$, defined via (3) interpolates the given data $w|_{I_{\text{given}}}$ and, by Corollary 4, $\widehat{w}$ is a trajectory of $\mathscr{B}$. □

Proposition 9 (Uniqueness of $\widehat{w}$) *Assuming A1–A3, Problem 7 has a unique solution, computed by Algorithm 1.*

PROOF. Under Assumptions A1 and A2, by Proposition 8, Algorithm 1 computes a trajectory $\widehat{w}$ of the system $\mathscr{B}$ that interpolates the given data $w|_{I_{\text{given}}}$. Algorithm 1 computes a particular solution g_{p} of (4). The solution set of (4) is $g_{\text{p}} + \mathscr{N}_g'$, where $\mathscr{N}_g'$ is the null space of $\mathscr{H}_L(w_{\text{d}})|_{I_{\text{given}}}$. Under Assumption A3, however, $\mathscr{N}_g' = \mathscr{N}$ — the null space of $\mathscr{H}_L(w_{\text{d}})$. Therefore, by Lemma 6, the interpolant $\widehat{w}$ is unique even though g_{p} is an arbitrary solution of (4). □

Note 10 (Representation of the map $w|_{I_{\text{given}}} \mapsto w|_{I_{\text{missing}}}$) *Under the assumptions of Proposition 9, the map $w|_{I_{given}} \mapsto w|_{I_{missing}}$ is well-defined. It is linear and allows the following explicit representation by w_{d}*

$$w|_{I_{missing}} = \mathscr{H}_L(w_{\text{d}})|_{I_{missing}} \left(\mathscr{H}_L(w_{\text{d}})|_{I_{given}}\right)^+ w|_{I_{given}},$$

where M^+ denotes the pseudo-inverse of a matrix M.

Note 11 (Minimal number of samples for exact recovery) *For A3 to hold true, the number of given samples K must be at least $K_{\min} := \mathbf{m}(\mathscr{B})L + \mathbf{n}(\mathscr{B})$.*

Note 12 (Computational complexity) *The computational complexity of Algorithm 1 is determined by the system of equations (4) on step 1. Equation (4) has K equations and $T - L + 1$ unknowns. With K fixed and T growing, the computational complexity is linear in T. With K growing linearly, due to Assumption A1, T must also grow at least linearly. This implies that both dimensions of the system are growing. As a result, the computational complexity of Algorithm 1 is cubic in K [21].*

4 Generalizations: nonunique solution, approximation of $w|_{I_{\text{given}}}$, and inexact w_{d}

The generalizations considered in this section relax the Assumptions A1–A3 for existence and uniqueness of an exact solution of the data-driven interpolation Problem 7. We start in Section 4.1 by characterizing the set of interpolants $\widehat{w}$ when Assumption A3 is not satisfied (and as a result the solution of Problem 7 is not unique). Then, in Section 4.2 as a result of dropping Assumption A2, we define an approximation problem, where $w|_{I_{\text{given}}}$ is approximated optimally in a weighted least squares sense by $\widehat{w}$. A special case of the approximation / missing data estimation problem is the errors-in-variables Kalman smoothing problem. The solution of the general data-driven approximation problem is given by a relatively minor modification of Algorithm 1: on step 1, g is an approximate weighted least squares / least norm solution of (4) (rather than an exact one). Section 4.3 presents a convex relaxation for the problem of interpolation with inexact/noisy data w_{d}.

4.1 Nonunique solution

As shown in Section 2, a consequence of dropping Assumption A3 is that Problem 7 has a nonunique solution. The solution set is affine and the following corollary of Proposition 8 gives it a data-driven characterization.

Corollary 13 (Nonuniqueness of $\widehat{w}$) *Assuming A1 and A2, the solution set to Problem 7 is $\widehat{\mathscr{W}} := \widehat{w}_{\rm p} + \mathscr{N}_w$, where $\widehat{w}_{\rm p} = \mathscr{H}_L(w_{\rm d})g_{\rm p}$ and $\mathscr{N}_w = \mathscr{H}_L(w_{\rm d})\mathscr{N}_g'$, with $g_{\rm p}$ a particular solution of (4) and $\mathscr{N}_g'$ the null space of $\mathscr{H}_L(w_{\rm d})|_{I_{given}}$.*

The characterization of the set of all solutions is needed in the trajectory planning control application. The set of interpolants defines the feasible set for further optimization, *i.e.*, a unique solution is selected from the set of all solutions using additional control objectives, typically a cost function. In addition to trajectory planning the characterization of all interpolants is used in subspace identification in two special cases: characterization of all free responses and all zero initial conditions responses of the system.

4.2 Approximation

When the consistency Assumption A2 is not satisfied, Problem 7 has no solution. In this section, we define an alternative problem to Problem 7, which aims for an approximate solution, *i.e.*, instead of requiring an exact interpolation of the given data $w|_{I_{\rm given}}$, we aim for an optimal in some sense approximation of $w|_{I_{\rm given}}$ by a trajectory $\widehat{w}$ of the system.

Assumption A2 may not be satisfied because $w|_{I_{\rm given}}$ is corrupted by noisy or because it is not generated by a linear time-invariant system of a bounded complexity. In either case the optimality criterion used is deterministic and minimizes the weighted 2-norm of the approximation error

$$\|w|_{I_{\rm given}} - \widehat{w}|_{I_{\rm given}}\|_W := \sqrt{(w|_{I_{\rm given}} - \widehat{w}|_{I_{\rm given}})^\top W (w|_{I_{\rm given}} - \widehat{w}|_{I_{\rm given}})}, \quad (5)$$

where $W \in \mathbb{R}^{K\times K}$ is a positive definite matrix. This leads to the following problem.

Problem 14 (Data-driven approximation) *Given a trajectory $w_{\rm d} \in \mathscr{B}|_T$ of a linear time-invariant system $\mathscr{B} \in \mathscr{L}^q$, a partially specified trajectory $w|_{I_{given}}$, and a positive definite weight matrix $W \in \mathbb{R}^{K\times K}$,*

$$\begin{aligned} &\underset{\widehat{w}}{\text{minimize}} \ \|w|_{I_{given}} - \widehat{w}|_{I_{given}}\|_W \\ &\textit{subject to} \quad \widehat{w} \in \mathscr{B}|_L. \end{aligned} \quad (6)$$

Proposition 15 *Assuming A1, Problem 14 has a solution*

$$\widehat{w} = \mathscr{H}_L(w_{\rm d})\big(W^{1/2}\mathscr{H}_L(w_{\rm d})|_{I_{given}}\big)^+ W^{1/2} w|_{I_{given}}, \quad (7)$$

where M^+ denotes the pseudo-inverse of a matrix M and $W^{1/2}$ denotes the matrix square root of W.

PROOF. Using Assumption A1, by Corollary 4, we have

$$\widehat{w} \in \mathscr{B}|_L \quad \iff \quad \text{there is a } g, \text{ such that } \widehat{w} = \mathscr{H}_L(w_{\rm d})g.$$

Therefore, (6) is equivalent to the weighted linear least squares problem

$$\min_g \ \|w|_{I_{\rm given}} - \mathscr{H}_L(w_{\rm d})|_{I_{\rm given}} g\|_W.$$

An least-norm optimal solution is given by

$$\widehat{g} = \big(W^{1/2}\mathscr{H}_L(w_{\rm d})|_{I_{\rm given}}\big)^+ W^{1/2} w|_{I_{\rm given}}.$$

Then, (6) follows from (3). □

For $K \geq K_{\rm min}$ given samples, the solution (7) of (6) is unique. Also, if exact solution exists, (7) is exact, *i.e.*, Problem 7 and Algorithm 1 are special cases of Problem 14 and (7).

Note 16 (Simultaneous approximation and interpolation) *A generalization of the data-driven approximation problem is to add equality constraints in (6) in order to achieve exact interpolation of certain specified data points. Since* $\operatorname{rank} \mathscr{H}_L(w_{\rm d})|_{I_{given}} = K_{\rm min}$*, up to* $K_{\rm min}$ *equality constraints can be added as exact interpolation conditions, while retaining feasibility. The resulting problem is equality constrained least-squares minimization and is still convex. We do not present the solution of this modification of Problem 14 here.*

4.3 Inexact/noisy data $w_{\rm d}$

In this section, we relax the basic assumption $w_{\rm d} \in \mathscr{B}|_T$, *i.e.*, $w_{\rm d}$ is no longer assumed to be exact. Instead, it is assumed that $w_{\rm d}$ is generated in the *errors-in-variables* setup $w_{\rm d} = \overline{w}_{\rm d} + \widetilde{w}_{\rm d}$, where $\overline{w}_{\rm d}$ is a trajectory of a bounded complexity linear time-invariant system (the "true" system) and $\widetilde{w}_{\rm d}$ is the measurement noise [20]. Define the complexity of $\mathscr{B}$ as the triple $c = \mathbf{c}(\mathscr{B}) := \big(\mathbf{m}(\mathscr{B}), \mathbf{l}(\mathscr{B}), \mathbf{n}(\mathscr{B})\big)$. With $\mathscr{L}_c^q$ denoting the set of linear time-invariant systems with complexity bounded by c, the assumption about $\overline{w}_{\rm d}$ is that there is a $\mathscr{B} \in \mathscr{L}_c^q$, such that $\overline{w}_{\rm d} \in \mathscr{B}|_T$. The assumption about the measurement noise $\widetilde{w}_{\rm d}$ is that it is zero mean Gaussian with covariance matrix $s^2 I$. The maximum-likelihood identification problem of the true system $\mathscr{B}$ from the data $w_{\rm d}$ is:

$$\begin{aligned} &\underset{\widehat{w}_{\rm d}, \widehat{\mathscr{B}}}{\text{minimize}} \ \|w_{\rm d} - \widehat{w}_{\rm d}\| \\ &\text{subject to} \quad \widehat{w}_{\rm d} \in \widehat{\mathscr{B}}|_T \text{ and } \widehat{\mathscr{B}} \in \mathscr{L}_c^q. \end{aligned} \quad (8)$$

Using the equivalence between trajectories of bounded complexity linear time-invariant systems and rank deficiency of

Hankel matrices (see [10, Lemma 3]), we restate (8) as Hankel structured low-rank approximation problems:

$$\begin{aligned} &\underset{\widehat{w}_\mathrm{d}}{\text{minimize}} \ \|w_\mathrm{d} - \widehat{w}_\mathrm{d}\| \\ &\text{subject to} \quad \operatorname{rank} \mathscr{H}_{\ell+1}(\widehat{w}_\mathrm{d}) \leq (\ell+1)m + n, \end{aligned}$$

where $m = \mathbf{m}(\mathscr{B})$, $\ell = \mathbf{l}(\mathscr{B})$, and $n = \mathbf{n}(\mathscr{B})$. Due to the rank constraint, the Hankel structured low-rank approximation problem is non-convex.

The data-driven interpolation problem can be defined as the bi-level problem:

$$\begin{aligned} &\underset{g}{\text{minimize}} \ \|w|_{I_\text{given}} - \mathscr{H}_L(\widehat{w}_\mathrm{d}^*)|_{I_\text{given}} g\|_W \\ &\text{subject to} \quad \widehat{w}_\mathrm{d}^* = \arg \min_{\widehat{w}_\mathrm{d}, \widehat{\mathscr{B}}} \ \|w_\mathrm{d} - \widehat{w}_\mathrm{d}\| \\ &\qquad\qquad\qquad \text{subject to} \quad \widehat{w}_\mathrm{d} \in \widehat{\mathscr{B}}|_T \text{ and } \widehat{\mathscr{B}} \in \mathscr{L}_c^q \end{aligned} \tag{9}$$

or equivalently

$$\begin{aligned} &\underset{g}{\text{minimize}} \ \|w|_{I_\text{given}} - \mathscr{H}_L(\widehat{w}_\mathrm{d}^*)|_{I_\text{given}} g\|_W \\ &\text{subject to} \quad \widehat{w}_\mathrm{d}^* = \arg \min_{\widehat{w}_\mathrm{d}} \ \|w_\mathrm{d} - \widehat{w}_\mathrm{d}\| \\ &\qquad\qquad\qquad \text{subject to} \quad \operatorname{rank} \mathscr{H}_{\ell+1}(\widehat{w}_\mathrm{d}) \leq (\ell+1)m + n, \end{aligned}$$

which is also non-convex due to the rank constraint in the inner optimization problem. Note that contrary to Algorithm 1, the maximum-likelihood data-driven interpolation problem (9) requires prior knowledge of the true system's complexity $\mathbf{c}(\mathscr{B})$.

The approach for solving (9) presented next is based on a convex relaxation similar to the one used in [4] for solving data-driven control problems. Since $\operatorname{rank} \mathscr{H}_L(\widehat{w}_\mathrm{d}^*)|_{I_\text{given}} \leq L\mathbf{m}(\mathscr{B}) + \mathbf{n}(\mathscr{B})$, adding the constraint

$$\|g\|_0 \leq L\mathbf{m}(\mathscr{B}) + \mathbf{n}(\mathscr{B}) \tag{10}$$

to the primary optimization problem of (9) does not change the problem. The relaxation proposed

(1) replaces the nonconvex constraint (10) by the convex one $\|g\|_1 \leq \alpha$, where $\alpha > 0$ is a hyper-parameter, and
(2) drops the nonconvex constraint $\widehat{w}_\mathrm{d} \in \widehat{\mathscr{B}} \in \mathscr{L}_c^q$, which leads to the trivial inner problem solution $\widehat{w}_\mathrm{d}^* = w_\mathrm{d}$.

The resulting convex optimization problem

$$\begin{aligned} &\underset{g}{\text{minimize}} \ \|w|_{I_\text{given}} - \mathscr{H}_L(w_\mathrm{d})|_{I_\text{given}} g\|_W \\ &\text{subject to} \quad \|g\|_1 \leq \alpha \end{aligned}$$

is by strong duality equivalent to the problem

$$\underset{g}{\text{minimize}} \ \tfrac{1}{2}\|w|_{I_\text{given}} - \mathscr{H}_L(w_\mathrm{d})|_{I_\text{given}} g\|_W^2 + \lambda \|g\|_1, \tag{11}$$

for some $\lambda > 0$. Problem (11) is known in statistics and machine learning as the Least Absolute Shrinkage and Selection Operator (LASSO) [1]. It is a convex optimization problem and there are readily available methods for its solution.

The hyper-parameter λ in (11) controls the accuracy–complexity trade off and is related to prior belief of the model complexity in the maximum-likelihood estimation problem. It has to be chosen "sufficiently large" in order to ensure the desired sparsity (10) of g, however, not "too large" because then $\widehat{w}$ tends towards the trivial zero solution. An optimal value λ^* for λ is the smallest value, for which the solution g of (11) has $L\mathbf{m}(\mathscr{B}) + \mathbf{n}(\mathscr{B})$ nonzero elements. λ^* can be found by bisection.

Note 17 (2-norm regularizer and $\lambda = 0$) *Although (7) is derived under the assumption of exact data w_d of a linear time-invariant system, it can be used as a heuristic for data-driven interpolation in the case of inexact data w_d. The use of the pseudo-inverse in (7) implies that the least-norm solution for g is selected. Minimization of $\|g\|$ is an alternative regularization to the $\|g\|_1$ in (11). Since (7) does not depend on hyper-parameters and (11) is undefined for $\lambda = 0$, we define the solution of (11) for $\lambda = 0$ to be (7).*

It is shown in [4] that for a related control problem the ℓ_1-norm regularization performs favorably in case of noisy data and nonlinear systems. Next, we test empirically the effectiveness of (11) for missing data estimation.

5 Simulation examples

The simulation results are made reproducible in the sense of [2] by providing the implementation of the method and the data generating scripts. The computational environment used is Matlab. The files reproducing the simulation results are available from `http://homepages.vub.ac.be/~imarkovs/software/ddint.tar`. The code is presented in a literate programming style [8] `http://homepages.vub.ac.be/~imarkovs/software/ddint.pdf`.

In this section, we show the performance of the ℓ_1-norm heuristic (11) on five datasets from the data-base for system identification DAISY [3]. Each dataset consists of two time series—measured input and measured output of a physical process. The datasets names, number of samples, number of inputs, and number of outputs are given in Table 1. For more details about the physical processes and the measurement experiments, we refer the reader to [3]. Note that these are measurements of real-life systems that do not necessarily satisfy the linearity and time-invariance assumptions.

The time series are split into two parts: the first 75% is used as the "data trajectory" w_d, the remaining 25% as the "to-be-interpolated trajectory" w. Randomly sampled variables $w|_{I_\text{missing}}$ of w are removed and estimated via data-driven interpolation using w_d and $w|_{I_\text{given}}$. The ratio $f = K/(qL)$

Table 1
Five datasets from the DAISY database are used for empirical evaluation of the (11) method. (T—number of samples, m—number of inputs, p—number of outputs) .

	data set name	T	m	p
1	Distillation column	90	5	3
2	pH process	2001	2	1
3	Hair dryer	1000	1	1
4	Heat flow density	1680	2	1
5	Heating system	801	1	1

of the number of missing elements to the total number of elements of w is selected as $f = 0.2$, *i.e.*, 20% missing values.

For the numerical solution of (11), we use the alternating direction method of multipliers [18]. The performance of the method is evaluated by the relative percentage error in the estimation of the missing values:

$$e_{\text{missing}} := \frac{\| w|_{I_{\text{missing}}} - \widehat{w}|_{I_{\text{missing}}} \|}{\| w|_{I_{\text{missing}}} \|} \, 100\%,$$

where $\widehat{w}$ is the computed solution. Since the performance of the (11) heuristic depends of the choice of the λ hyper-parameter, we evaluate e_{missing} over a grid of values for λ.

Table 2 shows the best value of e_{missing} obtained for the λ's in the grid. As a baseline for comparison, we use the performance of (7) (see Note 17) and the model-based method (9). A more detailed view of the performance of the ℓ_1-norm heuristic (11) is given in Figure 1. The left plot shows e_{missing} as a function of λ and the right plot shows the sparsity of the solution g, corresponding to the minimum value of e_{missing}. The theoretically optimal number of nonzero elements $L\mathbf{m}(\mathscr{B}) + \mathbf{n}(\mathscr{B})$ (the value of $\mathbf{n}(\mathscr{B})$ is taken from [15]) is shown in the figure as the vertical dotted line. In two of the examples (hair dryer and heating system) the theoretical and empirically obtained optimal sparsity levels coincide. In the other three examples the theoretical number of nonzero elements is much larger (not shown in the figures) than the empirical one. Hence, a "simpler" data-driven model explains better the observed trajectories than are more "complex" parametric model.

The results show improvement of the performance of the baseline method (7) by (11) for a suitably chosen value of λ. In other datasets from DAISY, however, such an improvement was not observed, *i.e.*, the best performance of (11) is achieved for $\lambda = 0$. Another caveat is the need of selecting λ. In practice, e_{missing} can not be evaluated, so that a surrogate of e_{missing} should be used instead. Cross-valuation methods offer such surrogates. For the model-based method, we used the value of $\mathbf{l}(\mathscr{B})$ suggested in [15]. The poor performance of (9), which we attribute to bias due to the non-linear nature of the examples, makes the results obtained by the data-driven methods even more remarkable.

Another caveat is the need of selecting λ. In practice, e_{missing} can not be evaluated, so that a surrogate of e_{missing} should be used instead. Cross-valuation methods offer such surrogates.

Table 2
Relative percentage error e_{missing} for (7), (9), and (11) (best result for a grid of values for λ) on the DAISY datasets.

	data set name	(7)	(9)	(11)
1	Distillation column	19.24	17.44	9.30
2	pH process	38.38	85.71	12.19
3	Hair dryer	12.35	8.96	7.06
4	Heat flow density	7.16	44.10	3.98
5	Heating system	0.92	1.35	0.36

6 Conclusions and outlook

Data-driven representations avoid parametric model identification. The methods presented in the paper allow us to solve nontrivial problems such as interpolation and approximation of trajectories of linear time-invariant systems using only basic linear algebra. The methods lead to general, simple, and practical algorithms. The generality and utility of the algorithms were illustrated on numerical examples. Dealing with inexact / noisy data w_{d} was approached by a convex relaxation leading to an ℓ_1-norm regularization problem. We showed the effectiveness of the relaxation on datasets from the DAISY database. Topics for future work are the computationally efficiency of the algorithms, recursive implementations of the methods that are suitable for real-time applications, alternative methods for dealing with inexact / noisy data, and statistical analysis of the resulting estimators.

Acknowledgements

The research leading to these results has received funding from: ERC Grant 258581, FWO projects G028015N and G090117N; and FNRS–FWO EOS Project 30468160.

References

[1] Steven L. Brunton and J. Nathan Kutz. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control.* Cambridge University Press, 2019.

[2] J. Buckheit and D. Donoho. *Wavelets and statistics*, chapter Wavelab and reproducible research. Springer-Verlag, 1995.

[3] B. De Moor, P. De Gersem, B. De Schutter, and W. Favoreel. DAISY: A database for identification of systems. *Journal A*, 38(3):4–5, 1997. Available from `http://homes.esat.kuleuven.be/~smc/daisy/`.

[4] F. Dörfler, J. Coulson, and I. Markovsky. Bridging direct & indirect data-driven control formulations via regularizations and relaxations. Technical report, arXiv:2101.01273, 2021. Available at `https://arxiv.org/abs/2101.01273`.

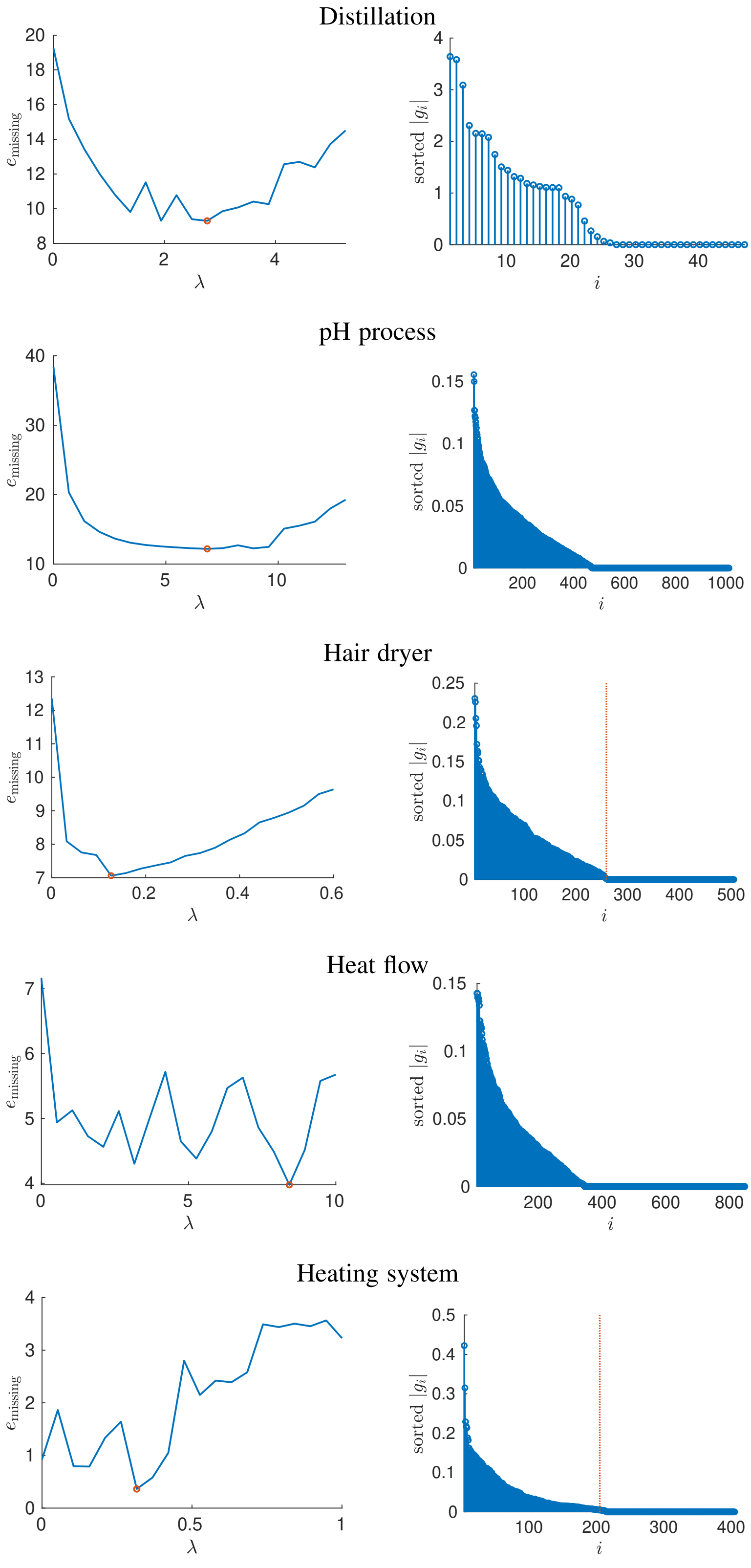

Fig. 1. Left: relative percentage error e_{missing} as a function of λ. Right: sparsity of g for the optimal value of λ, shown by plotting the sorted in decreasing value $|g_1|, \ldots, |g_{T-L+1}|$. The dotted vertical line shows the theoretically expected sparsity according to (10).

[5] E. Frazzoli, M. A. Dahleh, and E. Feron. Real-time motion planning for agile autonomous vehicles. In *Proc. American Control Conf.*, volume 1, pages 43–49, 2001.

[6] R. Guidorzi, R. Diversi, and U. Soverini. Optimal errors-in-variables filtering. *Automatica*, 39:281–289, 2003.

[7] J. P. Hespanha, P. Naghshtabrizi, and Y. Xu. A survey of recent results in networked control systems. *Proc. of the IEEE*, 95(1):138–162, 2007.

[8] D. Knuth. Literate programming. *Comput. J.*, 27(2):97–111, 1984.

[9] I. Markovsky. On the most powerful unfalsified model for data with missing values. *Systems & Control Lett.*, 95:53–61, 2016.

[10] I. Markovsky. A missing data approach to data-driven filtering and control. *IEEE Trans. Automat. Contr.*, 62:1972–1978, April 2017.

[11] I. Markovsky and B. De Moor. Linear dynamic filtering with noisy input and output. *Automatica*, 41(1):167–171, 2005.

[12] I. Markovsky and F. Dörfler. Identifiability in the behavioral setting, 2020. Available at `http://homepages.vub.ac.be/~imarkovs/publications/identifiability.pdf`.

[13] I. Markovsky and F. Dörfler. Behavioral systems theory in data-driven analysis, signal processing, and control. Technical report, Vrije Universiteit Brussel, 2021. Available at `http://homepages.vub.ac.be/~imarkovs/publications/overview-ddctr.pdf`.

[14] I. Markovsky and P. Rapisarda. Data-driven simulation and control. *Int. J. Control*, 81(12):1946–1959, 2008.

[15] I. Markovsky and S. Van Huffel. High-performance numerical algorithms and software for structured total least squares. *J. Comput. Appl. Math.*, 180(2):311–331, 2005.

[16] I. Markovsky, J. C. Willems, S. Van Huffel, and B. De Moor. *Exact and Approximate Modeling of Linear Systems: A Behavioral Approach*. SIAM, 2006.

[17] V. Mishra, I. Markovsky, and B. Grossmann. Data-driven tests for controllability. *Control Systems Letters*, 5:517–522, 2020.

[18] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1:123–231, 2014.

[19] J. Polderman and J. C. Willems. *Introduction to Mathematical Systems Theory*. Springer-Verlag, 1998.

[20] T. Söderström. Errors-in-variables methods in system identification. *Automatica*, 43:939–958, 2007.

[21] L. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, 1997.

[22] H. J. van Waarde, C. De Persis, M. K. Camlibel, and P. Tesi. Willems' fundamental lemma for state-space systems and its extension to multiple datasets. *IEEE Control Systems Letters*, 4:602–607, 2020.

[23] J. C. Willems. From time series to linear system—Part I. Finite dimensional linear time invariant systems, Part II. Exact modelling, Part III. Approximate modelling. *Automatica*, 22, 23:561–580, 675–694, 87–115, 1986, 1987.

[24] J. C. Willems. The behavioral approach to open and interconnected systems: Modeling by tearing, zooming, and linking. *Control Systems Magazine*, 27:46–99, 2007.

[25] J. C. Willems, P. Rapisarda, I. Markovsky, and B. De Moor. A note on persistency of excitation. *Systems & Control Lett.*, 54(4):325–329, 2005.