Identifiability in the Behavioral Setting

Ivan Markovsky and Florian Dörfler

***Abstract*—The behavioral approach to system identification starts from a given time series without a priori separation of the variables into inputs and outputs. The available identifiability conditions however require persistency of excitation of an input component of the time series which implicitly assumes that an input/output partitioning of the variables is given. In addition, a standard identifiability assumption is that the true data generating system is controllable. The conditions of controllability and persistency of excitation are sufficient but not necessary.**

Motivated by the need to infer linear-time invariant models from rank deficient Hankel matrices and to use such matrices as data-driven predictors in signal processing and control, we derive necessary and sufficient identifiability conditions that do not require a priori given input/output partitioning of the variables nor controllability of the true system. The prior knowledge needed for identifiability is the number of inputs, lag, and order of the true system. The results are based on a modification of the notion of a most powerful unfalsified model for finite data and a novel algorithm for its computation.

The results in the paper are derived assuming exact data, however, low-rank approximation allows their application in the case of noisy data. We compare empirically low-rank approximation of the Hankel, Page, and trajectory matrices in the errors-in-variables setting. Although the Page and trajectory matrices are unstructured, the parameter estimates obtained from them are less accurate than the one obtained from the Hankel matrix.

***Index Terms*—Behavioral system theory, System identification, Most powerful unfalsified model, Hankel matrix.**

EDICS: SSP-IDEN, SSP-PARE, SSP-SYSM

I. Introduction

A well known result in system identification is:

> (FL) If a signal w is a trajectory of a linear time-invariant system with m inputs, order n, and lag ℓ, then the rank of a Hankel matrix constructed of w with $L \geq \ell$ block rows is upper bounded by $mL+n$. Equality holds under the assumptions of controllability of the system and persistency of excitation of an input component of w of order $n+L$ [1].

We refer to the result of [1] as the *fundamental lemma*. Its importance is due to the fact that it provides *identifiability conditions*: under the assumptions of controllability and persistency of excitation of the input of order $n+\ell+1$, the data generating system can be inferred back from the data w.

Implicit in the assumptions of the fundamental lemma is an a priori known input/output partitioning of the variables $w=(u,y)$ and knowledge of the system's lag ℓ and order n. Moreover, controllability is required, which is an overly restrictive assumption as it excludes for example the class of autonomous systems. Finally, the result is originally derived for the case of a single given time series. The fundamental lemma is generalized separately for multiple time series [2], the Page matrix, and uncontrollable systems [3]. The generalizations, however, still assume a given input/output partitioning and persistency of excitation of an input component of w.

Apart from giving identifiability conditions, the fundamental lemma has important practical applications. The implication of the fundamental lemma that the image of a Hankel matrix constructed from the data coincides with the set of all finite trajectories of the system was effectively used in subspace-type algorithms [4]. This work led to new method for data-driven simulation and control [5], [6], [7], [8], [9], [10] as well as new identification methods [4], [11], [12].

An outstanding open problem for the application of the fundamental lemma in practice is dealing with approximation due to noisy data, disturbances, and model uncertainty. Indeed, the result of [1] is for exact data and nontrivial modifications are needed for its application in case of noisy data. One approach for dealing with noise is using structured low-rank approximation [13], [12]. This leads to statistically optimal (maximum-likelihood solution) in the errors-in-variables setting [14], however, the resulting optimization problems are nonconvex. Subspace methods and convex relaxations based on the nuclear norm heuristic offer suboptimal solutions that can be used as initial approximation for the nonconvex optimization methods. Recently, new approaches for dealing with noise are developed in the context of data-driven control. In [15], an approximation based on the distributionally robust theory is proposed and in [16], [17] a solution for a robust state feedback control based on the S-lemma is presented.

The fundamental lemma provides sufficient but not necessary identifiability conditions. The original motivation for this work is the converse of FL, which we state as a question.

> (CFL) Is a signal w, for which the associated Hankel matrix with L block rows is rank deficient, a trajectory of a linear time-invariant system with lag $\ell \leq L$? If so, how can this system be inferred from w?

The converse of the fundamental lemma is missing in the literature despite the fact that it is implicitly used in methods based on Hankel structured low-rank approximation [18], [12]. The CFL questions led us to:

1) revision of the notion of the *most powerful unfalsified model* [19] for finite time series,
2) a novel algorithm for its computation, and
3) new identifiability results.

These are the main contributions of the paper. Although, as in the original work [1] on the fundamental lemma as well as in the follow up publications [4], [5], [6], [9], [2], [3], the results

I. Markovsky is with the Department ELEC, Vrije Universiteit Brussel, 1050 Brussels, Belgium (e-mail: imarkovs@vub.be).

F. Dörfler is with the Automatic Control Laboratory (IfA), ETH-Zürich, 8092 Zürich, Switzerland (e-mail: dorfler@ethz.ch).

are derived assuming exact data, we relax and generalize the assumptions, demonstrating the practical

4) applicability of the theory for the case of noisy data.

We consider the errors-in-variables setting [14] and use as approximation methods low-rank approximation of the Hankel, Page, and trajectory matrices.

Outline

Section II defines the notation and collects basic results used in the paper. The focus is on *finite length trajectories*. An important result is formula ($\mathscr{B}|_L$, KER), which gives a matrix representation of the behavior restricted to a finite interval in terms of the parameters of a kernel representation. This allows us to use linear algebra (instead of polynomial algebra) as the main tool for the analysis of linear time-invariant systems.

Section III reviews the *integer invariants* of a linear time-invariant system. The integer invariants are used in Lemma 4 to give an explicit formula for the dimension of the behavior restricted to an interval of any length. This is the core technical result of the paper on which the other results are based. In Section III-B, using Corollary 5 of Lemma 4 we define the notion of model complexity.

Section IV introduces the notion of most powerful unfalsified model. First, in Section IV-A we recall the definition in the case of infinite time series: minimization of model complexity over the set of exact models. Lemma 10 in Section IV-B shows that in case of finite time series this definition leads to an autonomous model irrespective of the data. In Section IV-C, we present a modification of the most powerful unfalsified model for finite time series. The modification leads to a constructive procedure (Algorithm 1) for the computation of a minimal kernel representation of the most powerful unfalsified model in case of finite time series.

The main results of the paper are in Section V. Theorem 15 gives necessary and sufficient identifiability conditions that do not require a priori known input/output partitioning and controllability of the true system. The answers to the questions in CFL are given in Section VI-A. They are based on the modified notion of most powerful unfalsified model for finite time series. Generically, rank deficiency of the Hankel matrix implies existence of an exact bounded complexity linear time-invariant model, however, in nongeneric cases this is not true. Other applications of the theory developed in the paper, discussed in Sections VI-B–VI-D, are identifiability by filtering spurious annihilators and using low-rank approximation of the unstructured Page and trajectory matrices for approximation in the presence of noise. Section VII shows simulation results of identification with noisy data in the errors-in-variables setting. Although the Page and trajectory matrices are unstructured, the parameter estimates obtained from them are surprisingly less accurate than the one obtained from the Hankel matrix.

II. Preliminaries

The goals of this section are to give the necessary background for the technical results in the rest of the paper and to set the notation. We focus on the case of finite length, *i.e.*, the restriction of the behavior to a finite interval. In Section II-A, we define the Hankel matrix as obtained by consecutive application of the shift and cut operators on the time series. Section II-B introduces the notions of a shift-invariant subspace and linear time-invariant system and defines the kernel, input/output, and input/state/output representations. Again, we focus on the finite length case. An important result of this section is the matrix characterization ($\mathscr{B}|_L$, KER) of the restricted behavior in terms of a kernel representation.

A. Hankel matrices

We use interchangeably the terms *time series, sequence, and (discrete-time) trajectory*. The set of finite length q-variate, T-samples long, real-valued time series $w = \big(w(1),\ldots,w(T)\big)$, where $w(t) \in \mathbb{R}^q$ is denoted by $(\mathbb{R}^q)^T$. The set of infinite length q-variate, real-valued time series $w = \big(w(1),w(2),\ldots\big)$, where $w(t) \in \mathbb{R}^q$ is denoted by $(\mathbb{R}^q)^{\mathbb{N}}$.

Selection of a part of a time series, defined over a subinterval of the time interval, is called *restriction*. The corresponding operator is called the *cut*. For a time series $w \in (\mathbb{R}^q)^T$ or $w \in (\mathbb{R}^q)^{\mathbb{N}}$ and an integer L, $1 \leq L \leq T$, we define the *cut operator*

$$w|_L := \big(w(1),\ldots,w(L)\big).$$

Applied on a set of time series $\mathscr{W} \subset (\mathbb{R}^q)^T$ or $\mathscr{W} \subset (\mathbb{R}^q)^{\mathbb{N}}$, the cut operator acts on all time series in the set, *i.e.*,

$$\mathscr{W}|_L := \{\, w|_L \mid w \in \mathscr{W} \,\}.$$

The derivative operator plays a key role for continuous-time dynamical systems. Its discrete-time equivalent is the *unit shift operator* $\sigma : (\mathbb{R}^q)^{\mathbb{N}} \mapsto (\mathbb{R}^q)^{\mathbb{N}}$ defined by $(\sigma w)(t) := w(t+1)$. Applied on a set of time series $\mathscr{W} \subset (\mathbb{R}^q)^{\mathbb{N}}$, σ acts on all elements of $\mathscr{W}$, *i.e.*, $\sigma\mathscr{W} := \{\, \sigma w \mid w \in \mathscr{W} \,\}$. For a finite length time series $w \in (\mathbb{R}^q)^T$ and an integer $0 \leq \tau \leq T-1$,

$$\sigma^\tau w := \big(w(\tau+1),\ldots,w(T)\big),$$

so that σ can be applied to $w \in (\mathbb{R}^q)^T$ at most $T-1$ times. The resulting time series put next to each other form the two dimensional array of numbers shown in Table I. Question

TABLE I
Consecutive application of the shift operator on a finite time series results in a Hankel matrix with missing values.

w	σw	$\cdots$	$\sigma^{T-1}w$
$w(1)$	$w(2)$	$\cdots$	$w(T)$
$w(2)$	$\vdots$	$\iddots$	?
$\vdots$	$w(T)$	$\iddots$	$\vdots$
$w(T)$	?	$\cdots$	?

marks indicate unspecified elements due to the finiteness of the time series.

Restriction of $w, \sigma w, \ldots, \sigma^{T-L}w$ to the interval $[1,L]$, for some $1 \leq L \leq T$, is equivalent to consecutive application of the shift and cut operators. The *shift-and-cut* operator is used in [20] for state construction. Here we use it for defining the *Hankel matrix* of depth L

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

which is a fully defined upper-left block in Table I.

Although $\mathscr{H}_L(w)$ is defined for all integers $L \in [1,T]$, we will require that $\mathscr{H}_L(w)$ has at least as many columns as rows. This limits the range of values for L to $[1, L_{\max}]$, where

$$L_{\max} := \left\lfloor \frac{T+1}{q+1} \right\rfloor \tag{$L_{\max}$}$$

is the largest integer smaller than or equal to $(T+1)/(q+1)$.

In the case of an infinite sequence w, $\mathscr{H}_L(w)$ is the one-side infinite Hankel matrix

$$\mathscr{H}_L(w) = \begin{bmatrix} w|_L & (\sigma w)|_L & (\sigma^2 w)|_L & \cdots \end{bmatrix} \in \mathbb{R}^{qL\times\infty}.$$

We define also the two-side infinite Hankel matrix

$$\mathscr{H}(w) = \begin{bmatrix} w & \sigma w & \sigma^2 w & \cdots \end{bmatrix} \in \mathbb{R}^{\infty\times\infty}.$$

The *Page matrix* [21] $\mathscr{P}_L(w)$ of w with L block rows is obtained from the Hankel matrix $\mathscr{H}_L(w)$ by column selection:

$$\begin{aligned}\mathscr{P}_L(w) &:= \begin{bmatrix} w|_L & (\sigma_L w)|_L & \cdots & (\sigma_{(T'-1)L} w)|_L \end{bmatrix} \in \mathbb{R}^{qL\times T'} \\ &= \begin{bmatrix} w_1 & w_{L+1} & \cdots & w_{(T'-1)L+1} \\ \vdots & \vdots & & \vdots \\ w_L & w_{2L} & \cdots & w_{T'L} \end{bmatrix},\end{aligned} \tag{$\mathscr{P}_L(w)$}$$

where $T' := \lfloor T/L \rfloor$. Like the Hankel matrix $\mathscr{H}_L(w)$, the Page matrix $\mathscr{P}_L(w)$ also consists of L-samples long trajectories, however, unlike the Hankel matrix, the Page matrix has no repeated elements on the anti-diagonals.

The operator $\mathscr{P}_L$ constructing the Page matrix is called the *lifting operator*. It is used for identification of linear periodically time-varying systems [22] as well as an alternative to the Hankel matrix $\mathscr{H}_L(w)$ for data-driven control [15], approximate realization of linear time-invariant systems [21], and time series analysis [23].

A generalization of the Hankel matrix for a set

$$\mathscr{W} := \{ w^1, \dots, w^N \}, \quad w^i \in (\mathbb{R}^q)^{T_i}$$

of N time series is the *mosaic-Hankel matrix* [24], [25], [2]

$$\mathscr{H}_L(\mathscr{W}) := \begin{bmatrix} \mathscr{H}_L(w^1) & \cdots & \mathscr{H}_L(w^N) \end{bmatrix} \in \mathbb{R}^{qL\times\sum_{i=1}^N (T_i - L)}. \tag{$\mathscr{H}_L(\mathscr{W})$}$$

The range of values for L is $[1, L_{\max}]$, where $L_{\max}$ is defined as in ($L_{\max}$) with $T := \max\{T_1, \dots, T_N\}$. Blocks $\mathscr{H}_L(w^i)$ in ($\mathscr{H}_L(\mathscr{W})$) for which $T_i < L$ are missing.

A special case of the mosaic-Hankel matrix when all time series w^i have length $T_1 = \cdots = T_N = L$, *i.e.*, $w^i \in (\mathbb{R}^q)^L$ for all $i \in \{1, \dots, N\}$, is the *trajectory matrix*

$$\mathscr{T}(\mathscr{W}) := \begin{bmatrix} w_1^1 & w_1^2 & \cdots & w_1^N \\ \vdots & \vdots & & \vdots \\ w_L^1 & w_L^2 & \cdots & w_L^N \end{bmatrix} \in \mathbb{R}^{qL\times N}. \tag{$\mathscr{T}(\mathscr{W})$}$$

The trajectory matrix is used for dictionary learning [26]. The Page matrix $\mathscr{P}_L(w)$ is a special trajectory matrix obtained by taking $w^i = (\sigma^{(i-1)L} w)|_L$, for $i \in \{1, \dots, T'\}$.

B. Shift-invariant subspaces and linear time-invariant systems

In the behavioral setting, *dynamical systems* are sets of trajectories. The trajectories of a discrete-time system are time series, so that a discrete-time system $\mathscr{B}$ is a subset of $(\mathbb{R}^q)^{\mathbb{N}}$. The system $\mathscr{B} \subset (\mathbb{R}^q)^{\mathbb{N}}$ is linear if $\mathscr{B}$ is a subspace and time-invariant if $\mathscr{B}$ is shift-invariant.

Definition 1 (Shift-invariant spaces)**.** The set $\mathscr{W} \subset (\mathbb{R}^q)^{\mathbb{N}}$ is *shift-invariant* if for any $w \in \mathscr{W}$ and $\tau \in \mathbb{N}$, $\sigma^\tau w \in \mathscr{W}$. The set $\mathscr{W} \subset (\mathbb{R}^q)^T$ is shift-invariant if for any $w \in \mathscr{W}$ and $\tau \in \{0, 1, \dots, T-1\}$, there is $v \in \mathscr{W}$, such that $\sigma^\tau w = v|_{T-\tau}$.

The set of linear time-invariant systems with q variables is denoted by $\mathscr{L}^q$. A finite-dimensional linear time-invariant system $\mathscr{B}$ admits a *kernel representation*

$$\mathscr{B} = \ker R(\sigma) := \{ w \mid R(\sigma) w = 0 \}, \tag{KER}$$

where the operator $R(\sigma)$ is defined by the polynomial matrix

$$\begin{aligned} R(z) &= R_0 + R_1 z + \cdots + R_\ell z^\ell \\ &= \begin{bmatrix} R^1(z) \\ \vdots \\ R^g(z) \end{bmatrix} = \begin{bmatrix} R_0^1 + R_1^1 z + \cdots + R_{\ell_1}^1 z^{\ell_1} \\ \vdots \\ R_0^g + R_1^g z + \cdots + R_{\ell_g}^g z^{\ell_g} \end{bmatrix} \in \mathbb{R}^{g\times q}[z]. \end{aligned} \tag{R}$$

(KER) is called *minimal* if the number of equations g is as small as possible over all kernel representations of $\mathscr{B}$. As shown in Proposition 3, in a minimal kernel representation $g = p$—the number of outputs of $\mathscr{B}$—and $\ell := \deg R$ is also minimized over all kernel representations of $\mathscr{B}$.

Definition 2 (Annihilator)**.** An operator $r(\sigma)$, $r(z) \in \mathbb{R}^{1\times q}[z]$, such that $r(\sigma)\mathscr{B} = 0$ is called an *annihilator* of $\mathscr{B}$.

The rows $R^1, \dots, R^g$ of R are annihilators of $\mathscr{B} = \ker R(\sigma)$. Moreover, the rows of R form a basis for all annihilators of $\mathscr{B}$.

The *multiplication matrix* of

$$r = r_0 + r_1 z + \cdots + r_\ell z^\ell \in \mathbb{R}^{1\times q}[z]$$

with $L > \ell := \deg r$ block columns is the $(L-\ell) \times qL$ matrix

$$\mathscr{M}_L(r) := \begin{bmatrix} r_0 & r_1 & \cdots & r_\ell & & \\ & r_0 & r_1 & \cdots & r_\ell & \\ & & \ddots & \ddots & & \ddots \\ & & & r_0 & r_1 & \cdots & r_\ell \end{bmatrix}.$$

For $L \leq \ell$, we define $\mathscr{M}_L(r)$ to be a $0 \times qL$ empty matrix, $\operatorname{rank} \mathscr{M}_L(r) = 0$, and $\ker \mathscr{M}_L(r) = \mathbb{R}^{qL}$.

Consider the polynomial operator $R(\sigma)$ in a kernel representation (KER). We have the following characterization of the restricted behavior of the system $\mathscr{B} = \ker R(\sigma)$ to the interval $[1, L]$ in terms of (KER)

$$\mathscr{B}|_L = \ker \begin{bmatrix} \mathscr{M}_L(R^1) \\ \vdots \\ \mathscr{M}_L(R^g) \end{bmatrix} =: \ker \mathscr{M}_L(R). \tag{$\mathscr{B}|_L$, KER}$$

For a permutation matrix $\Pi \in \mathbb{R}^{q\times q}$ and an integer $0 < m < q$ define via

$$\begin{bmatrix} u \\ y \end{bmatrix} := \Pi^{-1} w \tag{$w \mapsto (u,y)$}$$

a *partitioning* of the variables $w(t) \in \mathbb{R}^q$ into $u(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}^{q-m}$. Let Π_u be the projection of w on the u variable, *i.e.*, $\Pi_u w := u$. Acting on a set, Π_u projects all elements in the set, which results in a new set. The partitioning $(w \mapsto (u,y))$ is an *input/output partitioning* of $\mathscr{B}$ if $\Pi_u \mathscr{B} = (\mathbb{R}^m)^{\mathbb{N}}$, *i.e.*, u is a free variable.

Let $\mathscr{B} = \ker R(\sigma)$ be a minimal kernel representation of $\mathscr{B}$. The partitioning $(w \mapsto (u,y))$ is an *input/output partitioning* of $\mathscr{B}$ if and only if with $\begin{bmatrix} Q & -P \end{bmatrix} := R\Pi^{-1}$, with $P \in \mathbb{R}^{p\times p}$ non-singular [27]. The resulting *input/outut representation* is

$$\mathscr{B}_{\text{i/o}}(P,Q,\Pi) = \{\, \Pi \begin{bmatrix} u \\ y \end{bmatrix} \mid Q(\sigma)u = P(\sigma)y \,\}. \tag{I/O}$$

A finite dimensional linear time-invariant system $\mathscr{B}$ admits an *input/state/output representation*

$$\begin{aligned} \mathscr{B} = \mathscr{B}_{\text{ss}}(A,B,C,D,\Pi) := \{\, \Pi \begin{bmatrix} u \\ y \end{bmatrix} \mid \text{there is } x \in (\mathbb{R}^n)^{\mathbb{N}}, \\ \text{such that } \sigma x = Ax + Bu,\ y = Cx + Du \,\}, \end{aligned} \tag{I/S/O}$$

where $\Pi \in \mathbb{R}^{q\times q}$ is a permutation and $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] \in \mathbb{R}^{(n+p)\times(n+m)}$. (I/S/O) is called *minimal* if $n := \dim A$ is as small as possible over all input/state/output representations of $\mathscr{B}$. Associated to $\mathscr{B} = \mathscr{B}_{\text{ss}}(A,B,C,D,\Pi)$, is the *autonomous sub-behavior*

$$\begin{aligned} \mathscr{B}_{\text{ss}}(A,C) := \{\, y \mid \text{there is } x \in (\mathbb{R}^n)^{\mathbb{N}}, \\ \text{such that } \sigma x = Ax,\ y = Cx \,\}. \end{aligned}$$

III. Integer invariants and model complexity

The concept of integer invariants is a key one in the theory of linear time-invariant systems. The integer invariants are used in Lemma 4 to characterize the dimension of the restricted behavior over an interval of any length. A related result, stated as Corollary 5, gives the dimension of the restriction of the behavior over intervals of length larger than the lag. Based on Corollary 5, in Section III-B we define the model complexity as the triple: (number of inputs, lag, order).

A. Integer invariants

For a linear time-invariant system $\mathscr{B}$ we define the following integers, called *invariants* of $\mathscr{B}$:

- *input cardinality* $\mathbf{m}(\mathscr{B}) :=$ number of inputs in (I/O),
- *output cardinality* $\mathbf{p}(\mathscr{B}) :=$ number of outputs in (I/O),
- *order* $\mathbf{n}(\mathscr{B}) :=$ minimal number of states in (I/S/O),
- *lag* $\mathbf{l}(\mathscr{B}) :=$ minimal degree of R in (KER), and
- *structure indices* $\big(\boldsymbol{\ell}_1(\mathscr{B}),\dots,\boldsymbol{\ell}_p(\mathscr{B})\big)$, $\boldsymbol{\ell}_i(\mathscr{B}) := \deg R^i$ in a minimal kernel representation (KER) of $\mathscr{B}$. Without loss of generality, we assume $\boldsymbol{\ell}_1(\mathscr{B}) \le \cdots \le \boldsymbol{\ell}_p(\mathscr{B})$ and define $\boldsymbol{\ell}_0(\mathscr{B}) := 0$, $\boldsymbol{\ell}_{p+1}(\mathscr{B}) := \infty$ for any $\mathscr{B}$.

In order to be well defined, the invariants of $\mathscr{B}$ should not depend on the choice of the representation. This non-obvious fact is proven in [27].

Proposition 3 (Wellposedness of the invariants [27])**.** *Let $\mathscr{B} = \ker R(\sigma)$ be a minimal kernel representation and $\mathscr{B} = \mathscr{B}_{ss}(A,B,C,D,\Pi)$ be a minimal input/state/output representation of a linear time-invariant system $\mathscr{B}$. Then, $\mathbf{m}(\mathscr{B})$, $\mathbf{p}(\mathscr{B})$, $\mathbf{n}(\mathscr{B})$, $\mathbf{l}(\mathscr{B})$, and $\big(\boldsymbol{\ell}_1(\mathscr{B}),\dots,\boldsymbol{\ell}_p(\mathscr{B})\big)$ are invariant of the representations. Moreover,*

- $\mathbf{p}(\mathscr{B}) = \operatorname{row\,dim} C = \operatorname{row\,dim} R$,
- $\mathbf{m}(\mathscr{B}) = \operatorname{col\,dim} B = q - \operatorname{row\,dim} R$,
- $\mathbf{l}(\mathscr{B}) = \max\{\, \boldsymbol{\ell}_1(\mathscr{B}),\dots,\boldsymbol{\ell}_p(\mathscr{B}) \,\}$, *and*
- $\mathbf{n}(\mathscr{B}) = \boldsymbol{\ell}_1(\mathscr{B}) + \cdots + \boldsymbol{\ell}_p(\mathscr{B})$.

Note that

$$(\boldsymbol{\ell}_0(\mathscr{B}), \boldsymbol{\ell}_1(\mathscr{B})] \cup \cdots \cup (\boldsymbol{\ell}_p(\mathscr{B}), \boldsymbol{\ell}_{p+1}(\mathscr{B}))$$

define a partitioning of $\mathbb{N}$, so that for any $L \in \mathbb{N}$, there is $k(L) \in \{\,0,1,\dots,p\,\}$, such that $L \in (\boldsymbol{\ell}_{k(L)}(\mathscr{B}), \boldsymbol{\ell}_{k(L)+1}(\mathscr{B})]$.

The following core technical lemma shows that $\mathscr{B}|_L$ is a piecewise linear function of L with kinks at the structure indices. Initially, $\mathscr{B}|_L$ increases at the rate of q. At each kink point the rate of increase drops by 1. For $L > \ell$, the rate reaches the constant $\mathbf{m}(\mathscr{B}) = q - \mathbf{p}(\mathscr{B})$.

Lemma 4 (Dimension of $\mathscr{B}|_L$)**.** *Let $\mathscr{B}$ be a linear time-invariant system. Then,*

$$\dim \mathscr{B}|_L = \big(q - k(L)\big)L + \boldsymbol{\ell}_1(\mathscr{B}) + \cdots + \boldsymbol{\ell}_{k(L)}(\mathscr{B}). \tag{$\dim \mathscr{B}|_L$}$$

Proof. Consider a minimal kernel representation $\mathscr{B} = \ker R(\sigma)$. From ($\mathscr{B}|_L$, KER), we have that

$$\dim \mathscr{B}|_L = qL - \operatorname{rank} \mathscr{M}_L(R).$$

Since $\mathscr{M}_L(R)$ is full row rank, we can find its rank from the row dimension

$$\begin{aligned} \operatorname{rank} \mathscr{M}_L(R) &= \operatorname{row\,dim} \mathscr{M}_L(R) \\ &= \sum_{i=1}^{k(L)} (L - \ell_i) = Lk(L) - \sum_{i=1}^{k(L)} \ell_i. \end{aligned} \tag{$*$}$$

For $L < \ell_1$, $k(L) = 0$, $\mathscr{M}_L(R)$ is an empty matrix, and $\operatorname{rank} \mathscr{M}_L(R) = 0$. Then, in ($*$), $\sum_{i=1}^{0} \ell_i = 0$. □

Lemma 4 shows that after an irregular increase of the dimension of $\mathscr{B}|_L$ in the interval $[1,\ell]$, in the final stage $L \in [\ell,\infty)$, $\mathscr{B}|_L$ increases linearly at a rate $\mathbf{m}(\mathscr{B})$.

Corollary 5. *For $L \ge \mathbf{l}(\mathscr{B})$, $\dim \mathscr{B}|_L$ is an affine function of L with slope determined by $\mathbf{m}(\mathscr{B})$ and offset $\mathbf{n}(\mathscr{B})$,*

$$\dim \mathscr{B}|_L = \mathbf{m}(\mathscr{B})L + \mathbf{n}(\mathscr{B}),\ \textit{for all } L \ge \mathbf{l}(\mathscr{B}). \tag{$\dim \mathscr{B}|'_L$}$$

In the special case of an autonomous system, after an increase of the dimension of $\mathscr{B}|_L$ in the interval $[1,\ell]$, $\mathscr{B}|_L$ becomes constant and is equal to the order of $\mathscr{B}$. The fact that $\mathscr{B}|_L = \mathbf{n}(\mathscr{B})$, for an linear time-invariant autonomous system and $L \ge \mathbf{l}(\mathscr{B})$ is well known from realization theory.

B. Model complexity

Linear systems are subspaces. The system's complexity is related to its dimension—the higher dimensional the subspace, the more complex the corresponding system. For a linear time-invariant system $\mathscr{B}$ with $\mathbf{m}(\mathscr{B}) > 0$ inputs, however, $\dim \mathscr{B} = \infty$, despite the fact that $\mathscr{B}$ admits representations (*e.g.*, (KER) and (I/S/O)) with finite dimensional parameter vectors. Therefore, we characterize the complexity of $\mathscr{B}$ by the increase of the dimension of $\mathscr{B}|_L$ as L grows. As shown in Lemma 4, this increase is fully characterized by the structure

indices $\big(\boldsymbol{\ell}_1(\mathscr{B}),\dots,\boldsymbol{\ell}_p(\mathscr{B})\big)$. By Corollary 5 however, for $L>\ell$, the increase of $\mathscr{B}|_L$ depends only on $\mathbf{m}(\mathscr{B})$ and $\mathbf{n}(\mathscr{B})$.

Definition 6 (Model complexity)**.** The complexity $\mathbf{c}(\mathscr{B})$ of a linear time-invariant system $\mathscr{B}$ is the triplet

$$\mathbf{c}(\mathscr{B}) := \big(\mathbf{m}(\mathscr{B}),\mathbf{l}(\mathscr{B}),\mathbf{n}(\mathscr{B})\big).$$

Complexities are compared by the lexicographic ordering:

> a model with more inputs is more complex than a model with fewer inputs irrespective of their lags and orders.

The class of bounded complexity linear time-invariant system $\mathscr{L}^{q,n}_{m,\ell}$ is defined as

$$\mathscr{L}^{q,n}_{m,\ell} := \{\,\mathscr{B}\in\mathscr{L}^q \mid \mathbf{c}(\mathscr{B})\le(m,\ell,n)\,\}.$$

In addition to systems with order n, lag ℓ, and m inputs, $\mathscr{L}^{q,n}_{m,\ell}$ includes all *lower complexity* systems—systems with fewer than m inputs, and/or lag smaller than ℓ, and/or order smaller than n. When we need to specify precisely the complexity of a system we use the notation

$$\partial\mathscr{L}^{q,n}_{m,\ell} := \{\,\mathscr{B}\in\mathscr{L}^q \mid \mathbf{c}(\mathscr{B})=(m,\ell,n)\,\}.$$

Note 7 (Missing letters in $\mathscr{L}^{q,n}_{m,\ell}$ / $\partial\mathscr{L}^{q,n}_{m,\ell}$). Missing letters n, ℓ, or m in the notation $\mathscr{L}^{q,n}_{m,\ell}$ mean that there are no bounds imposed on the corresponding system's invariants—order, lag, and input cardinality, respectively. "$\bullet$" can be used as a place holder for unrestricted quantities. For example, $\mathscr{L}^{q}_{m,\ell}$ is the set of linear time-invariant systems with at most m inputs and lag upper bounded by ℓ, while $\mathscr{L}^{q}_{\bullet,\ell}$ is the set of systems with an upper bound on the lag but no bound on the number of inputs.

Note that the order $\mathbf{n}(\mathscr{B})$ of a system $\mathscr{B}\in\mathscr{L}^q$ must satisfy the implicit constraints

$$\mathbf{l}(\mathscr{B})\le\mathbf{n}(\mathscr{B})\le\mathbf{p}(\mathscr{B})\mathbf{l}(\mathscr{B}).$$

Therefore, $\mathscr{L}^{q,n}_{m,\ell}\subseteq\mathscr{L}^{q}_{m,\ell}$, *i.e.*, the model class $\mathscr{L}^{q,n}_{m,\ell}$ specifies a more refined structure than $\mathscr{L}^{q}_{m,\ell}$.

Similarly, missing letters in the notation $\partial\mathscr{L}^{q,n}_{m,\ell}$ imply that the corresponding invariants are not restricted.

IV. The most powerful unfalsified model

The "clear and rational foundation under the problem of obtaining models from time series", developed by Jan C. Willems in [28], culminates in the notion of the *most powerful unfalsified model*. The definition of the most powerful unfalsified model $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_\mathrm{d})$ for the data $\mathscr{W}_\mathrm{d}$ in the model class $\mathscr{L}^q$, recalled in Section IV-A, is based on the notion of model complexity defined in Section III-B. The complexity (m,ℓ,n) is minimized in the lexicographic order over the constraint that the model is exact. In [28] infinite length data is considered.

In the case of finite length data, in Section IV-B we show that independent of the data $\mathscr{W}_\mathrm{d}$, $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_\mathrm{d})$ is a trivial autonomous system (Lemma 10). This critical issue is overlooked in the literature. In order to avoid the trivial solution a bound on the model complexity should be imposed.

Section IV-C presents a modification of the most powerful unfalsified model for finite data and an algorithm for computing a minimal kernel representation for it. As a side result, we show that if two systems have behaviors that coincide over an interval of length L, where L is larger than the lags of the systems, then the systems coincide.

A. Infinite length case

Definition 8 (Most powerful unfalsified model, infinite length data)**.** The most powerful unfalsified model $\mathscr{B}=\mathscr{B}_{\mathrm{MPUM}_{,}\mathscr{M}}(\mathscr{W}_\mathrm{d})$ for the data $\mathscr{W}_\mathrm{d}\subset(\mathbb{R}^q)^{\mathbb{N}}$ in the model class $\mathscr{M}$ is defined by the following properties:

1) $\mathscr{B}$ is *in the model class, i.e.*, $\mathscr{B}\in\mathscr{M}$,
2) $\mathscr{B}$ is *unfalsified, i.e.*, $\mathscr{W}_\mathrm{d}\subseteq\mathscr{B}$, and
3) $\mathscr{B}$ is *most powerful, i.e.*, any other model $\mathscr{B}'$ for which properties 1 and 2 hold is no more powerful than $\mathscr{B}$, *i.e.*,

$$\mathbf{c}(\mathscr{B})\le\mathbf{c}(\mathscr{B}'). \qquad (\mathscr{B}\subseteq\mathscr{B}')$$

Definition 8 leads to the multi-objective optimization

$$\min_{\mathscr{W}_\mathrm{d}\subseteq\mathscr{B}\in\mathscr{M}} \mathbf{c}(\mathscr{B}). \qquad \text{(OPT)}$$

In this paper, the model classes considered are $\mathscr{L}^q$, $\mathscr{L}^{q,n}_{m,\ell}$, and $\partial\mathscr{L}^{q,n}_{m,\ell}$. The complexity can be upper bounded or fixed by the *hyper-parameters* m, ℓ, n. If the model class is the unrestricted $\mathscr{L}^q$ (*i.e.*, hyper-parameters are not specified), we use the notation $\mathscr{B}_{\mathrm{MPUM}}$, *i.e.*, $\mathscr{L}^q$ is skipped. The dataset $\mathscr{W}_\mathrm{d}$ may consist of a single trajectory $w_\mathrm{d}\in(\mathbb{R}^q)^{\mathbb{N}}$ or of multiple trajectories,

$$\mathscr{W}_\mathrm{d}=\{\,w^1_\mathrm{d},\dots,w^N_\mathrm{d}\,\},\ w^i_\mathrm{d}\in(\mathbb{R}^q)^{\mathbb{N}}.$$

The most powerful unfalsified model in $\mathscr{L}^q$ always exists and is unique. Indeed,

$$\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_\mathrm{d})=\operatorname{image}\{\,\mathscr{W}_\mathrm{d}\cup\sigma\mathscr{W}_\mathrm{d}\cup\sigma^2\mathscr{W}_\mathrm{d}\cup\cdots\,\}. \qquad (\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_\mathrm{d}))$$

An insightful way of expressing $(\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_\mathrm{d}))$ is by the image of the Hankel matrix constructed from the data. For a data set consisting of a single trajectory $\mathscr{W}_\mathrm{d}=\{\,w_\mathrm{d}\,\}$, $w_\mathrm{d}\in(\mathbb{R}^q)^{\mathbb{N}}$,

$$\mathscr{B}_{\mathrm{MPUM}}(w_\mathrm{d})=\operatorname{image}\mathscr{H}(w_\mathrm{d}).$$

For multiple trajectories $\mathscr{W}_\mathrm{d}=\{\,w^1_\mathrm{d},\dots,w^N_\mathrm{d}\,\}$, $\mathscr{B}_{\mathrm{MPUM}}(w_\mathrm{d})$ is given by the image of the mosaic-Hankel matrix:

$$\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_\mathrm{d})=\operatorname{image}\underbrace{\begin{bmatrix}\mathscr{H}(w^1_\mathrm{d}) & \cdots & \mathscr{H}(w^N_\mathrm{d})\end{bmatrix}}_{\mathscr{H}(\mathscr{W}_\mathrm{d})}.$$

All subsequent results in the paper (for finite as well as infinite data) hold true for multiple trajectories by using mosaic-Hankel matrices in place of Hankel matrices.

Although, by definition $\mathscr{B}_{\mathrm{MPUM}}$ imposes a priori no bound on the complexity, depending on the data $\mathscr{W}_\mathrm{d}$, the actual model $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_\mathrm{d})$ may have bounded complexity.

Lemma 9 (Complexity of $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_\mathrm{d})$)**.** *Given a set of time series $\mathscr{W}_d\subset(\mathbb{R}^q)^{\mathbb{N}}$, let ℓ be the smallest natural number, such that there are m and n satisfying the equations*

$$\dim\operatorname{image}\mathscr{H}_L(\mathscr{W}_d)=mL+n,\ \textit{for all } L\in\{\,\ell,\ell+1,\dots\,\}. \quad (*)$$

Then, $\mathscr{B}_{MPUM}(\mathscr{W}_d) \in \partial\mathscr{L}^{q,n}_{m,\ell}$.

Proof. By definition, $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_\mathrm{d}) = \text{image}\,\mathscr{H}(\mathscr{W}_\mathrm{d})$. Since

$$\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_\mathrm{d})|_L = \text{image}\,\mathscr{H}_L(\mathscr{W}_\mathrm{d}),$$

by $(*)$ and $(\dim \mathscr{B}|_L)$, we have that $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_\mathrm{d}) \in \partial\mathscr{L}^{q,n}_{m,\ell}$. □

Contrary to $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_\mathrm{d})$, which always exists, the most powerful unfalsified model $\mathscr{B}_{\mathrm{MPUM},\mathscr{L}^{q,n}_{m,\ell}}(\mathscr{W}_\mathrm{d})$ in a model class of bounded complexity $\mathscr{L}^{q,n}_{m,\ell}$ may not exist.

B. Finite length case

In the finite length case without complexity specification, Definition 8 leads to an autonomous model.

Lemma 10. *For finite length data,* $\mathscr{B}_{MPUM}(\mathscr{W}_d)$ *is autonomous,* i.e., $\mathbf{m}\big(\mathscr{B}_{MPUM}(\mathscr{W}_d)\big) = 0$.

Proof. By the lexicographic ordering of the complexity, any autonomous system is less complex than any system with inputs. Therefore, for the proof it suffices to show that there exists an exact autonomous model. Consider first a single finite trajectory $w_\mathrm{d} \in (\mathbb{R}^q)^T$. The claim is that there is an autonomous linear time-invariant system $\mathscr{B}$ that fits w_d exactly. Let $\mathscr{B} = \mathscr{B}_{\mathrm{ss}}(A,C)$ be a minimal state space representation of the model. Then, $w_\mathrm{d} \in \mathscr{B}|_T$ if and only if the system of equations $w_\mathrm{d} = \mathscr{O}x_{\mathrm{ini}}$, where $\mathscr{O}$ is the extended observability matrix $\mathscr{O} := \mathrm{col}(C, CA, \ldots, CA^{T-1})$, has a solution $x_{\mathrm{ini}} \in \mathbb{R}^{\mathbf{n}(\mathscr{B})}$. There is a solution x_{ini}, for *any* $w_\mathrm{d} \in (\mathbb{R}^q)^T$ when $\mathscr{O}$ is full row rank. This can be guaranteed by choosing the order n sufficiently large. For example, take $n = qT$ and

$$C = \begin{bmatrix} I_q & 0 & \cdots & 0 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & I_q & 0 & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & I_q \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Then, $\mathscr{O} = I_{qT}$ and the solution is $w_\mathrm{d} = x_{\mathrm{ini}}$.

In case of a finite number of finite length trajectories $\{w^1_\mathrm{d}, \ldots, w^N_\mathrm{d}\}$, the argument above can be applied repeatedly for each trajectory w^i_d, resulting in an exact autonomous model $\mathscr{B}^i = \mathscr{B}_{\mathrm{ss}}(A^i, C^i)$. Define then $\mathscr{B}$ as

$$\mathscr{B} = \mathscr{B}_{\mathrm{ss}}\big(\mathrm{diag}(A^1, \ldots, A^N), \begin{bmatrix} C^1 & \cdots & C^N \end{bmatrix}\big),$$

which is an exact autonomous model for $\mathscr{W}_\mathrm{d}$. □

Lemma 10 shows that without a suitable bound on the complexity, in the finite length case, the most powerful unfalsified model always leads to a trivial model. This fact is observed in [29], where as a possible solution it is proposed to fix the number of inputs, *i.e.*, (in our notation) the model $\mathscr{B}_{\mathrm{MPUM},\partial\mathscr{L}^q_m}(\mathscr{W}_\mathrm{d})$ is considered. Next, we modify the definition of the most powerful unfalsified model for the finite data case in order to avoid the trivial model when prior knowledge about the number of inputs, lag, or order is not given.

C. Modification and algorithm

From the results in Section IV-B it seems that an upper bound on the complexity of the model should be a priori given in order to avoid the trivial model in the finite time case. In fact, prior knowledge about the complexity of the model need not be given by the hyper parameters (m, ℓ, n) because there is a universal upper bound on the lag in case of finite data

$$\mathscr{W}_\mathrm{d} = \{w^1_\mathrm{d}, \ldots, w^N_\mathrm{d}\}, \quad w^i_\mathrm{d} \in (\mathbb{R}^q)^{T_i}. \tag{$\mathscr{W}_\mathrm{d}$}$$

In order to see this, let $T := \max\{T_1, \ldots, T_N\}$, define

$$L_{\max} := \left\lfloor \tfrac{T+1}{q+1} \right\rfloor$$

and recall Definition 2 of an annihilator. Based on the finite data $\mathscr{W}_\mathrm{d}$, only annihilators of degree up to $L_{\max}-1$ can be determined from $\mathscr{W}_\mathrm{d}$. Indeed, an annihilator of degree $L-1$ is computed from the left kernel of the Hankel matrix $\mathscr{H}_L(\mathscr{W}_\mathrm{d})$, however, for $L > L_{\max}$ the left kernel of $\mathscr{H}_L(\mathscr{W}_\mathrm{d})$ is trivial due to the fact that $\mathscr{H}_L(\mathscr{W}_\mathrm{d})$ has more rows than columns. Therefore, we redefine the notion of most powerful unfalsified model in the finite length case by incorporating the bound $L_{\max}-1$ on the lag in the definition of $\mathscr{B}_{\mathrm{MPUM}}$.

Definition 11 (Most powerful unfalsified model, finite length data)**.** The most powerful unfalsified model $\mathscr{B} = \mathscr{B}_{\mathrm{MPUM},\mathscr{M}}(\mathscr{W}_\mathrm{d})$ for the data $(\mathscr{W}_\mathrm{d})$ in the model class $\mathscr{M}$ is defined by the following properties:

1) $\mathscr{B}$ is in the model class and has lag $\mathbf{l}(\mathscr{B}) < L_{\max}$, *i.e.*,
$$\mathscr{B} \in \mathscr{M} \cap \mathscr{L}^q_{\bullet, L_{\max}-1},$$
2) $\mathscr{B}$ is *unfalsified, i.e.*,
$$w^i_\mathrm{d} \in \mathscr{B}|_{T_i}, \quad \text{for all } i \in \{1, \ldots, N\}, \qquad (\mathscr{W}_\mathrm{d} \subset \mathscr{B})$$
3) $\mathscr{B}$ is *most powerful, i.e.*, any other model $\mathscr{B}'$ for which properties 1 and 2 hold is no more powerful than $\mathscr{B}$, *i.e.*, $(\mathscr{B} \subseteq \mathscr{B}')$.

Next, we present an algorithm for computing $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_\mathrm{d})$. The algorithm constructs sequentially the annihilators of degrees $0, 1, \ldots, L_{\max}-1$ from the left kernels of

$$\boxed{\mathscr{H}_1(\mathscr{W}_\mathrm{d})}, \quad \boxed{\mathscr{H}_2(\mathscr{W}_\mathrm{d})}, \ldots, \boxed{\mathscr{H}_{L_{\max}}(\mathscr{W}_\mathrm{d})}.$$

As a result, the algorithm delivers a minimal kernel representation (KER) of $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_\mathrm{d})$. By stopping prematurely when annihilators of degree up to $\ell_{\max} < L_{\max}-1$ are computed, the resulting set of annihilators define $\mathscr{B}_{\mathrm{MPUM},\mathscr{L}^q_{\bullet,\ell_{\max}}}(\mathscr{W}_\mathrm{d})$. This allows us to easily include a given bound $\ell_{\max}$ on the lag.

The proposed computational method is given in Algorithm 1. Step 4 computes the difference Δr between the predicted rank $Lm + \sum_{i=0}^{p} \ell_i$, based on the currently computed annihilators, and the actual rank of the Hankel matrix $\mathscr{H}_L(\mathscr{W}_\mathrm{d})$. If $\Delta r > 0$, then there are potential new annihilators in the left kernel of $\mathscr{H}_L(\mathscr{W}_\mathrm{d})$. Step 6 computes a basis R_{new} for the subspace spanned by the potential new annihilators. Steps 7 and 8 construct the potential new annihilators $R^1_{\mathrm{new}}(z), \ldots, R^{\Delta r}_{\mathrm{new}}(z)$ from R_{new}. For a potential new annihilator $R^i_{\mathrm{new}}(z)$ to be an actual annihilator and therefore be part of the kernel representation of $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_\mathrm{d})$, its degree must be $L-1$ (or else, it should have been detected on a previous step of the algorithm). Therefore, if degree $R^i_{\mathrm{new}}(z) = L-1$, Steps 11 and 12 adapt the model including $R^i_{\mathrm{new}}(z)$ as an annihilator.

Algorithm 1 Computation of a minimal kernel representation of the most powerful unfalsified model $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}})$.

Input: Data $(\mathscr{W}_{\mathrm{d}})$.

Let

$$m := q, \quad p := 0, \quad R^{(0)}(z) := [], \quad \ell_0 = 0.$$

{start with the trivial model $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}}) = (\mathbb{R}^q)^{\mathbb{N}}$}

Let $T := \max\{T_1, \ldots, T_N\}$ and define $(L_{\max})$.

for $L = 1 : L_{\max}$ **do**

  Let

$$\Delta r := Lm + \sum_{i=0}^{p} \ell_i - \operatorname{rank} \mathscr{H}_L(\mathscr{W}_{\mathrm{d}}).$$

  {$\Delta r =$ predicted rank minus actual rank}

  if $\Delta r > 0$ **then**

    Find R_{new}, such that

$$\ker \begin{bmatrix} \mathscr{M}_L(R^{(L-1)}) \\ R_{\mathrm{new}} \end{bmatrix} = \operatorname{left\,ker} \mathscr{H}_L(\mathscr{W}_{\mathrm{d}}).$$

    {R_{new} contains potential new annihilators}

    Let

$$R_{\mathrm{new}} =: \begin{bmatrix} R_{\mathrm{new},0} & R_{\mathrm{new},1} & \cdots & R_{\mathrm{new},L-1} \end{bmatrix}.$$

    Let

$$R_{\mathrm{new}}(z) := R_{\mathrm{new},0} + R_{\mathrm{new},1} z^1 + \cdots + R_{\mathrm{new},L-1} z^{L-1} =: \begin{bmatrix} R^1_{\mathrm{new}}(z) \\ \vdots \\ R^{\Delta r}_{\mathrm{new}}(z) \end{bmatrix}.$$

    {potential new annihilators}

    for $i = 1 : \Delta r$ **do**

      if $R^i_{\mathrm{new},L-1} \neq 0$ {*i.e.*, degree $R^i_{\mathrm{new}}(z) = L-1$} **then**

        Let

$$m := m - 1, \quad p := p + 1, \quad \ell_p := L - 1.$$

        {add a new annihilator to the model}

        Let

$$R^{(L)}(z) := \begin{bmatrix} R^{(L-1)}(z) \\ R^i_{\mathrm{new}}(z) \end{bmatrix}.$$

        {update the kernel representation with $R^i_{\mathrm{new}}(z)$}

      end if

    end for

  end if

end for

Output: Minimal kernel representation

$$\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}}) = \ker R(\sigma).$$

The following lemma is a simple but important consequence of Algorithm 1. In particular, its Corollary 13 is used in the proof of the identifiability result Theorem 15 in Section V.

Lemma 12 (The map $\mathscr{B}|_{\ell+1} \mapsto \mathscr{B}$ is well defined)**.** *The linear time-invariant system $\mathscr{B} \in \mathscr{L}^{q,n}_{m,\ell}$ is completely specified by its restriction $\mathscr{B}|_{\ell+1}$ to the interval $[1, \ell+1]$,* i.e.*, there is a unique extension of $\mathscr{B}|_{\ell+1} \subset (\mathbb{R}^q)^{\ell+1}$ to $\mathscr{B} \in \mathscr{L}^{q,n}_{m,\ell}$.*

Proof. The proof is constructive. 1) Compute a basis $(\mathscr{W}_{\mathrm{d}})$ with $T_1 = \cdots = T_N = \ell + 1$ for $\mathscr{B}|_{\ell+1}$, *i.e.*, $(\mathscr{W}_{\mathrm{d}})$ such that $\operatorname{span} \mathscr{W}_{\mathrm{d}} = \mathscr{B}|_{\ell+1}$ and $N = \dim \mathscr{B}|_{\ell+1} = m(\ell+1) + n$. 2) For $(\mathscr{W}_{\mathrm{d}})$, computed on step 1, Algorithm 1 produces a kernel representation of $\mathscr{B}$. This kernel representation can be used then to reconstruct $\mathscr{B}$. □

Corollary 13 ($\mathscr{B}|_{\ell+1} = \mathscr{B}'|_{\ell+1} \iff \mathscr{B} = \mathscr{B}'$)**.** *If two linear time-invariant systems $\mathscr{B}, \mathscr{B}' \in \mathscr{L}^q$ have the same behaviors over the interval $[1, L]$, where $L \geq \max\{\mathbf{l}(\mathscr{B}), \mathbf{l}(\mathscr{B}')\} + 1$, then they coincide.*

V. Identifiability

Let the data $(\mathscr{W}_{\mathrm{d}})$ be generated by a system $\bar{\mathscr{B}} \in \mathscr{M}$, *i.e.*,

$$w^i_{\mathrm{d}} \in \bar{\mathscr{B}}|_{T_i}, \quad \text{for all } i \in \{1, \ldots, N\}. \qquad (\mathscr{W}_{\mathrm{d}} \subset \bar{\mathscr{B}})$$

We refer to $\bar{\mathscr{B}}$ as the *true data generating system* and pose the questions:

1) Can we recover $\bar{\mathscr{B}}$ from $\mathscr{W}_{\mathrm{d}}$?
2) If so, how can we obtain $\bar{\mathscr{B}}$?

Question 1 is called an *identifiability question*. Question 2 is an *exact identification problem*. Identifiability ensures well-posedness of the exact identification problem $\mathscr{W}_{\mathrm{d}} \mapsto \bar{\mathscr{B}}$.

The identification principle is the Occam's razor—minimization of the complexity over all exact models in a predefined model class $\mathscr{M}$, see (OPT). Using the Occam's razor principle with the prior information that the true model is linear time-invariant, the identifiability question becomes the question of guaranteeing that the most powerful unfalsified model of $\mathscr{W}_{\mathrm{d}}$ in $\mathscr{L}^q$ coincides with $\bar{\mathscr{B}}$. The following theorems gives conditions on $\mathscr{W}_{\mathrm{d}}$, under which $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}}) = \bar{\mathscr{B}}$, so that $\bar{\mathscr{B}}$ is identifiable from $\mathscr{W}_{\mathrm{d}}$ in $\mathscr{L}^q$.

Theorem 14. *Let the data $(\mathscr{W}_d)$ be generated by a system $\bar{\mathscr{B}} \in \partial \mathscr{L}^{q,n}_{m,\ell}$,* i.e.*, $(\mathscr{W}_d \subset \bar{\mathscr{B}})$. Then, $\bar{\mathscr{B}}$ is identifiable from $\mathscr{W}_d$,* i.e.*, $\mathscr{B}_{MPUM}(\mathscr{W}_d) = \bar{\mathscr{B}}$, if*

$$\ell < L_{\max} := \left\lfloor \tfrac{T+1}{q+1} \right\rfloor \quad \textit{and} \quad \mathbf{p}\big(\mathscr{B}_{MPUM}(\mathscr{W}_d)\big) = q - m. \quad \text{(ID1)}$$

Proof. The condition $\ell < L_{\max}$ is necessary. Otherwise, there is an annihilator of $\bar{\mathscr{B}}$ with lag larger than or equal to $L_{\max}$ that can not be detected from the data. Under the assumption $\ell < L_{\max}$, all annihilators of $\bar{\mathscr{B}}$ are detected by Algorithm 1, *i.e.*, they are among the annihilators of $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}})$. However, $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}})$ may include additional annihilators that are not annihilators of $\bar{\mathscr{B}}$. This possibility is ruled out by the condition $\mathbf{p}\big(\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}})\big) = p = q - m$. Then, the annihilators of $\bar{\mathscr{B}}$ and $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}})$ coincide and therefore $\bar{\mathscr{B}} = \mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}})$. □

In order to check the identifiability conditions (ID1) of Theorem 14, one has to know the lag ℓ and the number of inputs m

of the true system $\bar{\mathscr{B}}$. The conditions (ID1) are sufficient. The following theorem gives a necessary and sufficient condition that requires prior knowledge of the complexity (m,ℓ,n) of $\bar{\mathscr{B}}$.

Theorem 15. *Let the data $(\mathscr{W}_d)$ be generated by a system $\bar{\mathscr{B}} \in \partial\mathscr{L}^{q,n}_{m,\ell}$,* i.e.*, $(\mathscr{W}_d \subset \bar{\mathscr{B}})$. Then, $\bar{\mathscr{B}}$ is identifiable from $\mathscr{W}_d$,* i.e.*, $\mathscr{B}_{MPUM}(\mathscr{W}_d) = \bar{\mathscr{B}}$ if and only if*

$$\mathit{rank}\,\mathscr{H}_{\ell+1}(\mathscr{W}_d) = m(\ell+1)+n. \tag{ID2}$$

Proof. Since the data is exact, $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}}) \subseteq \bar{\mathscr{B}}$, so that, in particular, $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}})|_{\ell+1} \subseteq \bar{\mathscr{B}}|_{\ell+1}$. By the rank condition, we have that $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}})|_{\ell+1} = m(\ell+1)+n$. On the other hand, $\bar{\mathscr{B}}|_{\ell+1} = m(\ell+1)+n$. Therefore, $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}})|_{\ell+1} = \bar{\mathscr{B}}|_{\ell+1}$. Then, by Corollary 13, $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}}) = \bar{\mathscr{B}}$. □

Theorems 14 and 15 apply to uncontrollable as well as controllable systems and do not require a priori known input/output partitioning of the variables nor controllability. The rank condition (ID2) can be viewed as a generalized persistency of excitation condition for the data $\mathscr{W}_{\mathrm{d}}$.

Note 16 (Comparison with the fundamental lemma). The fundamental lemma gives alternative identifiability conditions:

1) $\bar{\mathscr{B}}$ is controllable,
2) $\bar{\mathscr{B}}$ admits an input/output partitioning $w = \left[\begin{smallmatrix} u \\ y \end{smallmatrix}\right]$,
3) u_{d} is persistently exciting of order $\ell+n+1$, *i.e.*,

$$\operatorname{rank}\mathscr{H}_{\ell+n+1}(u_{\mathrm{d}}) = m(\ell+n+1). \tag{PE}$$

Condition 3 is similar to (ID2): it is a rank condition for a Hankel matrix that requires knowledge of the true model's complexity (m,ℓ,n). The additional assumptions 1 and 2, however, impose a loss of generality. Although there is always a subset of m variables that are inputs, in general, not *all* subsets of m variables are inputs of $\bar{\mathscr{B}}$, so that the choice of the input/output partition cannot be made without loss of generality. Finally, uncontrollable systems are excluded from consideration. In particular, the fundamental lemma does not include autonomous systems. In contrast, in the case of an autonomous system, Theorem 15 yields the well known (from realization theory) identifiability condition $\operatorname{rank}\mathscr{H}_{\ell+1}(w_{\mathrm{d}}) = n$.

A noteworthy generalization of the new identifiability results over the fundamental lemma is that they apply to multiple time series. The corresponding generalization of the data structure is from a Hankel matrix to a mosaic-Hankel matrix. Using data from multiple time series has underappreciated potential for system. For example, as illustrated in Section VII, the following corollary has relevance for approximation in the errors-in-variables setting.

Corollary 17. *Let the data $(\mathscr{W}_d)$ be generated by a system $\bar{\mathscr{B}} \in \partial\mathscr{L}^{q,n}_{m,\ell}$,* i.e.*, $(\mathscr{W}_d \subset \bar{\mathscr{B}})$ and let $T_i = \ell+1$ for all $i \in \{1,\ldots,N\}$. Then, $\bar{\mathscr{B}}$ is identifiable from $\mathscr{W}_d$ if and only if*

$$\mathit{rank}\,\mathscr{T}(\mathscr{W}_d) = m(\ell+1)+n.$$

(The trajectory matrix $\mathscr{T}(\mathscr{W}_{\mathrm{d}})$ is defined in $(\mathscr{T}(\mathscr{W}))$.)

Note 18. A generalization of the fundamental lemma for multiple time series, involving a mosaic-Hankel matrix, is presented in [2]. The persistency of excitation assumption in [2], however, does not allow for a result similar to Corollary 17 involving the trajectory matrix $\mathscr{T}(\mathscr{W}_{\mathrm{d}})$. An analogous corollary holds for the Page matrix. In [30], an analog to Corollary 17 involving the Page matrix $(\mathscr{P}_L(w))$ is presented under more stringent persistency of excitation assumptions.

Finally, the following corollary stating that the image of the Hankel matrix $\mathscr{H}_L(w_{\mathrm{d}})$ coincides with the true system's behavior $\bar{\mathscr{B}}|_L$ (*i.e.*, it spans all L-samples long trajectories of $\bar{\mathscr{B}}$). The result is used for data-driven simulation and control.

Corollary 19. *Let the data $(\mathscr{W}_d)$ be generated by a system $\bar{\mathscr{B}} \in \partial\mathscr{L}^{q,n}_{m,\ell}$,* i.e.*, $(\mathscr{W}_d \subset \bar{\mathscr{B}})$. Then,* image $\mathscr{H}_L(\mathscr{W}) = \bar{\mathscr{B}}_L$*, for $L > \ell$ if and only if $\mathit{rank}\,\mathscr{H}_L(\mathscr{W}_d) = mL+n$.*

Trivial modifications of Corollary 19 allow us to use Page and trajectory matrices in place of the mosaic-Hankel matrix.

VI. Applications

A. An answer to the questions in CFL

In this section, we come back to the questions CFL posed in the introduction. The signal w_{d} is an exact trajectory of a linear time-invariant system with lag $\ell \leq L$ if and only if the most powerful unfalsified model $\mathscr{B}_{\mathrm{MPUM},\mathscr{L}^q_{\bullet,L-1}}(w_{\mathrm{d}})$ is nontrivial. Rank deficiency of $\mathscr{H}_L(w)$ is a necessary condition for existence of a nontrivial model, however, as shown in the following example it is not sufficient. The issue is that a potential annihilator obtained from the left kernel of the Hankel matrix $\mathscr{H}_L(w)$ may not have the required degree $L-1$.

Example 20. An example of w_{d} for which the Hankel matrix $\mathscr{H}_L(w_{\mathrm{d}})$ is rank deficient but there is no exact model with lag $\ell < L$ is $w_{\mathrm{d}} = (0,\ldots,0,1) \in \mathbb{R}^T$. For all L, $\operatorname{rank}\mathscr{H}_L(w_{\mathrm{d}}) = 1$. In particular, $\mathscr{H}_2(w_{\mathrm{d}})$ is rank deficient. However, there is no linear time-invariant system with lag $\ell = 1$ that fits w_{d}. Indeed the left kernel of $\mathscr{H}_2(w_{\mathrm{d}})$ is spanned by the vector $\begin{bmatrix} 1 & 0 \end{bmatrix}$, which gives rise to a polynomial operator $r(\sigma) = 1$. This operator is of degree 0 instead of 1, so that it does not define an linear time-invariant system with lag 1. This nongeneric situation is detected in Algorithm 1 by the condition on step 5. For the data in the example, Algorithm 1 returns the trivial model $\mathscr{B}_{\mathrm{MPUM}}(w_{\mathrm{d}}) = \mathbb{R}^{\mathbb{N}}$. Note, however, that if exactness of the signal w_{d} is checked by restricting to the interval $[1, T-L]$, the rank deficiency of $\mathscr{H}_L(w)$ is equivalent to existence of a nontrivial model with lag $\ell < L$.

Algorithm 1 gives a constructive procedure for computing the exact model for w_{d} with lag $\ell < L$, provided that it exists.

B. Spurious annihilators

Algorithm 1 detects the annihilators of degree up to $\ell_{\max} := L_{\max} - 1$. Assume as in the identifiability problem that the data is generated by a linear time-invariant system $\bar{\mathscr{B}} = \ker R(\sigma) \in \partial\mathscr{L}^{q,n}_{m,\ell}$. Then, the identified model $\widehat{\mathscr{B}} := \mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}})$ contains the annihilators $R^1,\ldots,R^g$ of $\bar{\mathscr{B}}$ of degree up to $\ell_{\max}$. Depending on the data $\mathscr{W}_{\mathrm{d}}$, however, $\widehat{\mathscr{B}}$ may contain additional annihilators of degree up to $\ell_{\max}$ that are not annihilators of $\bar{\mathscr{B}}$. We call these annihilators *spurious*.

Definition 21 (Spurious annihilators)**.** Let the data $(\mathscr{W}_{\mathrm{d}})$ be generated by a system $\bar{\mathscr{B}} \in \partial\mathscr{L}^{q,n}_{m,\ell}$, *i.e.*, $(\mathscr{W}_{\mathrm{d}} \subset \bar{\mathscr{B}})$, $\bar{\mathscr{R}}$ be the

set of annihilators of $\bar{\mathscr{B}}$, and $\widehat{\mathscr{R}}$ be the set of annihilators of $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}})$. The elements of the set difference $\widehat{\mathscr{R}} \setminus \bar{\mathscr{R}}$ are spurious annihilators of $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}})$ (with respect to $\bar{\mathscr{B}}$).

Existence of spurious annihilators prevents identifiability. Indeed, by definition, $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}}) = \bar{\mathscr{B}}$ if and only if $\mathscr{B}_{\mathrm{MPUM}}(\mathscr{W}_{\mathrm{d}})$ has no spurious annihilators. Consequently, all identifiability results can be understood as giving conditions ensuring that the model $\widehat{\mathscr{B}}$ does not include spurious annihilators. A prototypical example of a spurious annihilator is an annihilator for a set of input variables. In the fundamental lemma such annihilators are avoided by the persistency of excitation assumption (PE) for the inputs. Indeed, persistency of excitation of u_{d} of order L implies that there are no annihilators of order $L-1$. The fundamental lemma uses a separation of the annihilators into spurious and non-spurious based on degree: any annihilator of degree larger than an a priori known threshold degree $\ell_{\max}$ is spurious.

We envisage a range of other identifiability conditions that filter the spurious annihilators based on other types of prior knowledge about the true system $\bar{\mathscr{B}}$.

C. Why is persistency of excitation of order $L+n$ needed?

The fundamental lemma ensures that there are no spurious annihilators by the persistency of excitation assumption. As stated in [1] however a surprising fact is that persistency of excitation of order more than $\ell+1$, namely $\ell+n+1$, is required. Indeed, persistency of excitation of order $\ell+1$ is sufficient to ensure that the inputs alone have no annihilators of order up to ℓ. The need of extra persistency of excitation did not become elucidated in the follow-up publications [9], [3], [2], giving alternative proofs and generalizations of the fundamental lemma. From Theorem 14, it is clear that the extra persistency of excitation is needed in order to ensure that the outputs are also sufficiently excited, so that the full Hankel matrix $\mathscr{H}_{\ell+1}(w_{\mathrm{d}})$ achieves its maximum rank $m(\ell+1)+n$.

The following example shows a single-input single-output controllable system that is not identifiable from data $w_{\mathrm{d}} = (u_{\mathrm{d}}, y_{\mathrm{d}})$ with persistently exciting input of order less than $2n+1$.

Example 22. Consider a controllable system $\bar{\mathscr{B}} \in \mathscr{L}^{2,n}_{1,n}$ defined by (I/O) with $\deg Q = n$. Define the *input nulling behavior*

$$\mathscr{B}_{u,0} := \{\, u \mid Q(\sigma)u = 0 \,\}.$$

Since $\mathscr{B}_u \in \mathscr{L}^{1,n}_{0,n}$, there are persistently exciting signals $u_{\mathrm{d}} \in \mathscr{B}_u$ of order n. For any such input signal u_{d}, there is a corresponding output $y_{\mathrm{d}} = 0$ of $\bar{\mathscr{B}}$. The resulting trajectory $(u_{\mathrm{d}}, 0) \in \bar{\mathscr{B}}$, however, does not allow identifiability of $\mathscr{B}$.

Assuming an extra persistency of excitation of order $2n+1$ for u_{d} *and* controllability of $\bar{\mathscr{B}}$, the output y_{d} is guaranteed to be exited independently of the initial conditions, so that the rank condition (ID2) holds.

D. Approximation using Page/trajectory matrix

The Page matrix $\mathscr{P}_L(w)$ is introduced in [21] as an alternative to the Hankel matrix $\mathscr{H}_L$ for system realization. Since $\mathscr{P}_L(w_{\mathrm{d}})$ is a submatrix of $\mathscr{H}_L(w_{\mathrm{d}})$,

$$\operatorname{image} \mathscr{P}_L(w_{\mathrm{d}}) \subseteq \operatorname{image} \mathscr{H}_L(w_{\mathrm{d}}).$$

Equality does not hold in general. Assuming that it holds, a minimal order exact model can be obtained either from $\mathscr{P}_L(w_{\mathrm{d}})$ or from $\mathscr{H}_L(w_{\mathrm{d}})$. The models obtained from $\mathscr{P}_L(w_{\mathrm{d}})$ and $\mathscr{H}_L(w_{\mathrm{d}})$ are identical, however, stronger assumptions are needed in order to use $\mathscr{P}_L(w_{\mathrm{d}})$.[1] Therefore, from the point of view of exact realization there is no advantage of using the Page matrix.

It [21], it is suggested that the Page matrix has advantage when an approximation is needed in case of noisy data. Based on empirical evidence, it is reported that approximate model realization using Kung's method, *i.e.*, unstructured low-rank approximation based on truncation of the singular value decomposition followed by Ho-Kalman realization using the factors of the low-rank approximation gives better results when using the Page matrix in place of the Hankel matrix. The rational for this is that the Page matrix $\mathscr{P}_L(w_{\mathrm{d}})$ has no repeated elements and the low-rank approximation $\widehat{D}$ of $\mathscr{P}_L(w_{\mathrm{d}})$, based on truncation of the singular value decomposition, is optimal in the sense of minimizing the Frobenius norm of the approximation error $\mathscr{P}_L(w_{\mathrm{d}}) - \widehat{D}$. Another argument in favor of using the Page matrix in case of noisy data is given in [15], [23]. In [15], it is shown that the Page matrix is a strictly better distributionally robust predictor of the behavior.

Assuming that w_{d} is an exact trajectory of a linear time-invariant system $\bar{\mathscr{B}} \in \delta\mathscr{L}^{q,n}_{m,\ell}$, both the Page matrix $\mathscr{P}_L(w_{\mathrm{d}})$, for $L > \ell+1$ and the Hankel matrix $\mathscr{H}_L(w_{\mathrm{d}})$ have *shift-invariant structure* apart from the rank deficiency

$$\operatorname{rank} \mathscr{P}_L(w_{\mathrm{d}}) \leq \operatorname{rank} \mathscr{H}_L(w_{\mathrm{d}}) \leq Lm+n.$$

For $\mathscr{H}_L(w_{\mathrm{d}})$, the shift-invariant structure manifests itself in the pattern of repeated elements. For $\mathscr{P}_L(w_{\mathrm{d}})$, however, the shift-invariant structure is not evident in a pattern of the matrix elements. The fact that low-rank approximation does not impose shift-invariance of the approximation renders Kung's method using both the Page matrix with $L > \ell+1$ as well as the Hankel matrix heuristics.

The Page matrix can be viewed alternatively as a trajectory matrix constructed from short segments of a long trajectory. More generally, using multiple experiments instead of a single long experiment the appropriate data matrix is the mosaic Hankel matrix. Using data from multiple experiments has advantages for identification of unstable and uncontrollable systems. For example, multiple experiments allow consistent estimation of an autonomous system [31]. In the special case of using data $\mathscr{W}_{\mathrm{d}}$ consisting of L-samples long trajectories, the mosaic-Hankel matrix $\mathscr{H}_L(\mathscr{W}_{\mathrm{d}})$ becomes the trajectory matrix

$$\mathscr{H}_L(\mathscr{W}_{\mathrm{d}}) = \mathscr{T}(\mathscr{W}_{\mathrm{d}}) = \begin{bmatrix} w_{\mathrm{d}}^1 & \cdots & w_{\mathrm{d}}^N \end{bmatrix} \in \mathbb{R}^{qL \times N},$$

i.e., each column of $\mathscr{H}_L(\mathscr{W}_{\mathrm{d}})$ is a different trajectory. Like the Page matrix $\mathscr{P}_L(w_{\mathrm{d}})$, the trajectory matrix $\mathscr{T}(\mathscr{W}_{\mathrm{d}})$ has no repeated elements. In what follows, we treat the Page matrix and the trajectory matrix together although they use different data (a single long trajectory vs multiple short trajectories).

[1] A simple necessary conditions is that the number of columns of $\mathscr{H}_L(w_{\mathrm{d}})$ and $\mathscr{P}_L(w_{\mathrm{d}})$ are greater than $\dim \bar{\mathscr{B}}|_L = mL+n$, for $L > \ell$. This leads to, $T_{\min} := (m+1)L+n-1$ samples, for the Hankel matrix, and $T'_{\min} := mL^2+nL$ samples for the Page matrix: more data is needed for using the Page matrix.

In the special case $L = \ell + 1$, both $\mathscr{P}_{\ell+1}(w_\mathrm{d})$ and $\mathscr{T}(\mathscr{W}_\mathrm{d})$ loose the shift-invariance structure. This renders unstructured low-rank approximation of $\mathscr{P}_{\ell+1}(w_\mathrm{d})$ and $\mathscr{T}(\mathscr{W}_\mathrm{d})$ statistically optimal (maximum likelihood) for model identification from the corresponding data: multiple short trajectories obtained as segments from w_d or independent experiments $\mathscr{W}_\mathrm{d}$.

Consider two noisy data sets: 1) $w_\mathrm{d} \in (\mathbb{R}^q)^T$ and 2) $(\mathscr{W}_\mathrm{d})$ with $T_1 = \cdots = T_N = \ell + 1$ that have the same total number of samples, *i.e.*, $T = N(\ell+1)$. Although low-rank approximation of $\mathscr{T}(\mathscr{W}_\mathrm{d})$ yields a maximum likelihood estimator using the data $\mathscr{W}_\mathrm{d}$, it is not evident that this estimator has superior performance than low-rank approximation of the Hankel matrix $\mathscr{H}_{\ell+1}(w_\mathrm{d})$. Indeed, avoiding the shift-invariant structure by using multiple $(\ell+1)$-long trajectories results in an ℓ-times reduction in the number of columns, which also affects the approximation accuracy. Empirical comparison of the three methods—low-rank approximation using the Hankel, Page, and trajectory matrices—in the errors-in-variance setting is presented in the following section.

VII. Simulation examples with noisy data

In this section, we do empirical comparison of approximate system identification methods based on low-rank approximation of the Hankel matrix, the Page matrix, and the trajectory matrix. The simulation results are made reproducible in the sense of [32] by providing the implementation of the methods 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/identifiability-code.tar and the code is presented in a literate programming style[33], [34] here: http://homepages.vub.ac.be/~imarkovs/software/identifiability-code.pdf.

A. Simulation setup

We use the benchmark example of [35], which is a $n =$ 4th order single-input single-output system $\bar{\mathscr{B}}$ defined by an input/output representation (I/O) with Equivalently, the system is defined by a kernel representation (KER) with a parameter

$$\bar{R}(z) = \underbrace{\begin{bmatrix} -\bar{Q}_0 & \bar{P}_0 \end{bmatrix}}_{\bar{R}_0} + \underbrace{\begin{bmatrix} -\bar{Q}_1 & \bar{P}_1 \end{bmatrix}}_{\bar{R}_1} z + \cdots + \underbrace{\begin{bmatrix} -\bar{Q}_4 & \bar{P}_4 \end{bmatrix}}_{\bar{R}_4} z^4.$$

The normalization $\bar{P}_4 = 1$ is used in order to make the parameter vector $\bar{R} := \begin{bmatrix} \bar{R}_0 & \bar{R}_1 & \bar{R}_2 & \bar{R}_3 & \bar{R}_4 \end{bmatrix}$ unique.

Two different data sets are generated in the errors-in-variables setup:

- a single T-samples long trajectory w_d,
- N, $(\ell+1) = 5$-samples long trajectories $(\mathscr{W}_\mathrm{d})$.

In the case of a single trajectory, $w_\mathrm{d} = \bar{w} + \tilde{w}$, where $\bar{w} \in \bar{\mathscr{B}}|_T$ is a random trajectory of the system $\bar{\mathscr{B}}$ and $\tilde{w}$ is a zero mean white Gaussian noise with standard deviation s that is varied from 0 (exact data) to 0.1. In the case of multiple trajectories, the noise parameters are identical. The total number of samples in the data sets is the same, *i.e.*, $T = N(\ell+1)$.

B. Identification methods

The identification methods compared compute an estimate $\widehat{R}$ of the model parameter $\bar{R}$ from the approximate left kernel of a data matrix:

- the Hankel matrix $D = \mathscr{H}_5(w_\mathrm{d})$,
- the Page matrix $D = \mathscr{P}_5(w_\mathrm{d})$, and
- the trajectory matrix $D = \mathscr{T}(\mathscr{W})$.

The approximation of the left kernel is obtained from the optimal in the Frobenius norm unstructured rank-4 approximation of D, computed by truncation of the singular value decomposition of D. The computed parameter estimate $\widehat{R}$ is normalized by redefining it as $\widehat{R} := \widehat{R}/\widehat{R}_{10}$, *i.e.*, making $\widehat{P}_4 = 1$.

The methods based on the singular value decomposition of the Hankel matrix is heuristic. A statistically optimal (maximum likelihood) estimator in the errors-in-variables setup is obtained by structured low-rank approximation of the Hankel matrix $\mathscr{H}_5(w_\mathrm{d})$ [13]. For comparison in the simulation example, we include also the results of the maximum likelihood estimator, computed by the method of [25], [18].

C. Validation criteria

As a validation criterion used is the relative parameter error:

$$e_{\bar{R}} := \|\bar{R} - \widehat{R}\| / \|\bar{R}\|, \qquad (e_{\bar{R}})$$

averaged over $N = 500$ Monte-Carlo repetitions of the estimation with different noise realizations. In addition, we show the error of approximation of the data $\mathscr{W}_\mathrm{d}$ by a given model $\mathscr{B}$:

$$e_{\mathscr{W}_\mathrm{d}} := \sqrt{\sum_{i=1}^{N} \min_{\widehat{w}^i \in \mathscr{B}|_{T_i}} \|w_\mathrm{d}^i - \widehat{w}^i\|_2^2}.$$

D. Results and discussion

Figure 1 shows the parameter error $e_{\bar{R}}$ and the error of approximation of the data $\mathscr{W}_\mathrm{d}$ averaged over $N = 500$ Monte-Carlo repetitions. Low-rank approximation of the trajectory matrix is optimal in terms of the $e_{\mathscr{W}_\mathrm{d}}$ criterion. Indeed, $e_{\mathscr{W}_\mathrm{d}}$ is the criterion that this method minimizes. With respect to the parameter error $e_{\bar{R}}$, however, low-rank approximation of the Hankel matrix $\mathscr{H}_5(w_\mathrm{d})$ gives better results than low-rank approximation of the Page matrix and the trajectory matrix. This apparent contradiction can be explained by the fact that $\operatorname{col\,dim} \mathscr{T}(\mathscr{W}_\mathrm{d}) = \operatorname{col\,dim} \mathscr{P}_{\ell+1}(w_\mathrm{d}) = 100$, while $\operatorname{col\,dim} \mathscr{H}_5(w_\mathrm{d}) = 494$. Consistency of the estimators corresponding to the methods compared imply that the more columns the data matrix has, the smaller the estimation error $e_{\bar{R}}$ is. Figure 2 illustrates this.

The empirical results suggest that although all methods yield consistent estimators, the one based on low-rank approximation of the Hankel matrix is more efficeint than the ones based on the Page and trajectory matrices. This is due to the fact that although w_d and $\mathscr{W}_\mathrm{d}$ have the same number of samples, w_d is more "informative" for identification of the true system than $\mathscr{W}_\mathrm{d}$ and the Hankel matrix exploits this information more effectively than the Page matrix.

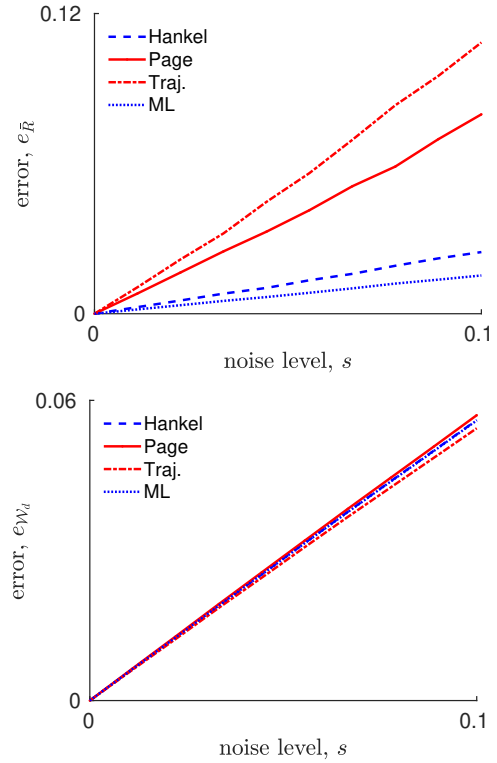

Fig. 1. Low-rank approximation of the trajectory matrix $\mathscr{T}(\mathscr{W}_\mathrm{d})$ (dashed-dotted line) is optimal in the data approximation criterion $e_{\mathscr{W}_\mathrm{d}}$, however, it is suboptimal with respect to the parameter estimation error $e_{\bar{R}}$ and yields worse results than low-rank approximation of the Hankel matrix $\mathscr{H}_5(w_\mathrm{d})$ (dashed line). The statistically optimal maximum likelihood (ML) estimate (dotted line) is computed by structured low-rank approximation of $\mathscr{H}_5(w_\mathrm{d})$.

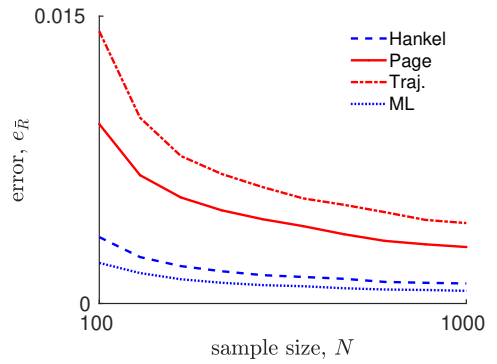

Fig. 2. The estimation errors of the methods converge to zero as the total number of samples goes to infinity (consistency), however, at different rates (efficiency). The efficiency of the low-rank approximation of the Hankel matrix $\mathscr{H}_5(w_\mathrm{d})$ (dashed line) is better than the low-rank approximation of the trajectory matrix $\mathscr{T}(\mathscr{W}_\mathrm{d})$ (dashed-dotted line). Optimal efficiency has the maximum likelihood (ML) estimator (dotted line), computed by structured low-rank approximation of $\mathscr{H}_5(w_\mathrm{d})$.

VIII. Conclusions and outlook

Using formula ($\mathscr{B}|_L$, KER), which is a finite dimensional matrix representation of the restriction of the behavior, we derived an explicit formula for the dimension of the behavior restricted to an interval of any length (Lemma 4). This result is the basis for the other results in the paper—definition of model complexity, revision of the notion of most powerful unfalsified model for finite data, and identifiability conditions.

We showed that the classical notion of the most powerful unfalsified model proposed for infinite time series is inadequate in case of finite time series. Irrespective of the data it results in a trivial autonomous model (Lemma 10). We proposed a natural modification of the most powerful unfalsified model for the case of finite time series (Definition 11) that avoids the trivial model without using priori knowledge about the model's complexity. This led us to a constructive algorithm (Algorithm 1) for computation of the most powerful unfalsified model. The algorithm computes recursively a minimal kernel representation of the model. The key computational step is detecting rank deficiency of generalized Hankel matrices with increasing depths constructed from the data.

Assuming that the data is generated by a linear time-invariant system of bounded complexity, Theorem 15 gives a necessary and sufficient condition for the most powerful unfalsified model to coincide with the data generating system. This condition does not require a priori known input/output partitioning and controllability of the true system. It uses only the prior knowledge about the complexity of the true system.

Some practical implications and directions for future work suggested by the results in the paper are given next.

- Using low-rank Hankel matrices in system theory requires a caveat: in nongeneric cases, the tail of the time series may not be consistent with the model.
- In classical identifiability results such, as the fundamental lemma, the spurious annihilators are distinguished from the true system's annihilators based on a degree separation. New identifiability conditions can be derived using more general separation criteria.
- We did not delve into numerical computations issues related to the implementation of Algorithm 1. This topic has connections with work on numerical linear algebra methods for Hankel structured matrices, see for example [36], [37], [38]. In particular incorporating approximation in Algorithm 1 is a interesting topic for further work.
- Using a trajectory matrix with $\ell+1$ block-rows leads to an unstructured data matrix, so that in this case approximation by truncation of the singular value decomposition yields optimal approximation. This observation allows us to avoid the nonconvex optimization of the structured low-rank approximation problem, however, data from multiple short experiments is less informative than data from one long experiment with the same total number of samples. Empirical results show that overall low-rank approximation of the Hankel matrix gives more accurate model parameter estimates than low-rank approximation of the trajectory matrix. This contradicts empirical results reported in [15], [23] on using the trajectory matrix in data-driven MPC control so that a further research in this direction is needed.
- Other applications of the results in the paper that will be explored elsewhere are interpolation of trajectories and data-driven errors-in-variables smoothing.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007–2013) / ERC Grant agreement number 258581 "Structured low-rank approximation: Theory, algorithms, and applications" and Fund for Scientific Research Vlaanderen (FWO) projects G028015N "Decoupling multivariate polynomials in nonlinear system identification" and G090117N "Block-oriented nonlinear identification using Volterra series"; and Fonds de la Recherche Scientifique (FNRS) – FWO Vlaanderen under Excellence of Science (EOS) Project no 30468160 "Structured low-rank matrix / tensor approximation: numerical optimization-based algorithms and applications".

REFERENCES

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

[2] 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," *Control Systems Letters*, vol. 4, no. 3, pp. 602–607, 2020.

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

[4] I. Markovsky, J. C. Willems, P. Rapisarda, and B. De Moor, "Algorithms for deterministic balanced subspace identification," *Automatica*, vol. 41, no. 5, pp. 755–766, 2005.

[5] I. Markovsky and P. Rapisarda, "Data-driven simulation and control," *Int. J. Contr.*, vol. 81, no. 12, pp. 1946–1959, 2008.

[6] T. Maupong and P. Rapisarda, "Data-driven control: A behavioral approach," *Syst. Control Lett.*, vol. 101, pp. 37–43, 2017.

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

[8] C. De Persis and P. Tesi, "Formulas for data-driven control: Stabilization, optimality, and robustness," *IEEE Trans. Automat. Contr.*, vol. 65, pp. 909–924, 2020.

[9] H. van Waarde, J. Eising, H. Trentelman, and K. Camlibel, "Data informativity: A new perspective on data-driven analysis and control," *IEEE Trans. Automat. Contr.*, 2019.

[10] J. Berberich and F. Allgower, "A trajectory-based framework for data-driven system analysis and control," in *Proc. European Control Conference (ECC)*, 2020.

[11] T. Katayama, *Subspace Methods for System Identification*. Springer, 2005.

[12] I. Markovsky, *Low-Rank Approximation: Algorithms, Implementation, Applications*, 2nd ed. Springer, 2019.

[13] ——, "Structured low-rank approximation and its applications," *Automatica*, vol. 44, no. 4, pp. 891–909, 2008.

[14] T. Söderström, "Errors-in-variables methods in system identification," *Automatica*, vol. 43, pp. 939–958, 2007.

[15] J. Coulson, J. Lygeros, and F. Dörfler, "Distributionally robust chance constrained data-enabled predictive control," https://arxiv.org/abs/2006.01702, May 2020.

[16] C. D. Persis and P. Tesi, "Low-complexity learning of linear quadratic regulators from noisy data," arXiv:2005.01082, 2020.

[17] H. van Waarde, K. Camlibel, and M. Mesbahi, "From noisy data to feedback controllers: Non-conservative design via a matrix S-lemma," arXiv:2006.00870, 2020.

[18] I. Markovsky, "A software package for system identification in the behavioral setting," *Control Eng. Practice*, vol. 21, no. 10, pp. 1422–1436, 2013.

[19] J. C. Willems, "From time series to linear system—Part II. Exact modelling," *Automatica*, vol. 22, no. 6, pp. 675–694, 1986.

[20] P. Rapisarda and J. C. Willems, "State maps for linear systems," *SIAM J. Control Optim.*, vol. 35, no. 3, pp. 1053–1091, 1997.

[21] A. Damen, P. Van den Hof, and A. Hajdasinski, "Approximate realization based upon an alternative to the hankel matrix: the page matrix," *Control Lett.*, vol. 2, pp. 202–208, 1982.

[22] I. Markovsky, J. Goos, K. Usevich, and R. Pintelon, "Realization and identification of autonomous linear periodically time-varying systems," *Automatica*, vol. 50, pp. 1632–1640, 2014.

[23] A. Agarwal, J. Amjad, D. Shah, and D. Shen, "Model agnostic time series analysis via matrix estimation," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 2, no. 3, pp. 1–39, 2018.

[24] G. Heinig, "Generalized inverses of Hankel and Toeplitz mosaic matrices," *Linear Algebra Appl.*, vol. 216, no. 0, pp. 43–59, Feb. 1995.

[25] I. Markovsky and K. Usevich, "Software for weighted structured low-rank approximation," *J. Comput. Appl. Math.*, vol. 256, pp. 278–292, 2014.

[26] S. Brunton, J. Proctor, and N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.

[27] J. C. Willems, "Paradigms and puzzles in the theory of dynamical systems," *IEEE Trans. Automat. Contr.*, vol. 36, no. 3, pp. 259–294, 1991.

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

[29] V. Mishra and I. Markovsky, "The set of linear time-invariant unfalsified models with bounded complexity is affine," Dept. ELEC, Vrije Universiteit Brussel, Tech. Rep., 2019.

[30] J. Coulson, J. Lygeros, and F. Dörfler, "Regularized and distributionally robust data-enabled predictive control," in *Proc. of the IEEE Conf. on Decision and Control*, Nice, France, December 2019, pp. 7165–7170.

[31] I. Markovsky and R. Pintelon, "Identification of linear time-invariant systems from multiple experiments," *IEEE Trans. Signal Process.*, vol. 63, no. 13, pp. 3549–3554, 2015.

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

[33] D. Knuth, "Literate programming," *Comput. J.*, vol. 27, no. 2, pp. 97–111, 1984.

[34] N. Ramsey, "Literate programming simplified," *IEEE Software*, vol. 11, pp. 97–105, 1994.

[35] I. Landau, D. Rey, A. Karimi, A. Voda, and A. Franco, "A flexible transmission system as a benchmark for robust digital control," *European Journal of Control*, vol. 1, no. 2, pp. 77–96, 1995.

[36] G. Heinig and P. Jankowski, "Kernel structure of block Hankel and Toeplitz matrices and partial realization," *Linear Algebra Appl.*, vol. 175, pp. 1–30, 1992.

[37] G. Heinig and K. Rost, *Algebraic methods for Toeplitz-like matricies and operators*. Birkhauser, 1984.

[38] T. Laudadio, N. Mastronardi, and P. Van Dooren, "The generalized schur algorithm and some applications," *Axioms*, vol. 7, p. 81, 2018.

[39] M. Galrinho, R. Prota, M. Ferizbegovic, and H. Hjalmarsson, "Weighted null-space fitting for identification of cascade networks," in *18th IFAC Symposium on System Identification*, vol. 51, no. 15, 2018, pp. 856–861.