Automatica 169 (2024) 111869

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Technical communique

Identification from data with periodically missing output samples ☆

Ivan Markovsky [a,*], Mohammad Alsalti [b], Victor G. Lopez [b], Matthias A. Müller [b]

[a] Catalan Institution for Research and Advanced Studies (ICREA), Pg. Lluis Companys 23, Barcelona, Spain and the International Centre for Numerical Methods in Engineering (CIMNE), Gran Capitàn, Barcelona, Spain
[b] Leibniz University Hannover, Institute of Automatic Control, 30167 Hannover, Germany

ARTICLE INFO

Article history:
Received 29 October 2023
Received in revised form 10 June 2024
Accepted 17 June 2024
Available online 22 August 2024

Keywords:
System identification
Missing data
Behavioral approach
Lifting

ABSTRACT

The identification problem in case of data with missing values is challenging and currently not fully understood. For example, there are no general nonconservative identifiability results, nor provably correct data efficient methods. In this paper, we consider a special case of periodically missing output samples, where all but one output sample per period may be missing. The novel idea is to use a lifting operation that converts the original problem with missing data into an equivalent standard identification problem. The key step is the inverse transformation from the lifted to the original system, which requires computation of a matrix root. The well-posedness of the inverse transformation depends on the eigenvalues of the system. Under an assumption on the eigenvalues, which is not verifiable from the data, and a persistency of excitation-type assumption on the data, the method based on lifting recovers the data-generating system.

© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

1. Introduction

The standard setup in system identification is that the data consists of equidistantly sampled input/output variables of the to-be-identified system. If even a single variable is missing at one moment of time, the whole dataset cannot be used by standard identification methods. Due to the time-series nature of the data, unless the missing variable appears in the first or the last sample, it cannot be skipped. It should be estimated together with the system. This led to the development of specialized methods for identification of linear time-invariant systems from data with missing values, see, *e.g.*, Broersen, De Waele, and Bos (2004), Hansson and Wallin (2012), Isaksson (1993), Liu, Hansson, and Vandenberghe (2013), Markovsky (2016), Markovsky and Usevich (2013), Wallin and Hansson (2014), Wang, Li, and Stoica (2006). Alternatively, the data can be split into two datasets—one before the missing sample and one after it. The two data sets should be fitted then by one model. For multiple missing variables, scattered over different moments of time, the data becomes fragmented into multiple short data sets. The worst distribution of the missing samples is periodic and the best is lumped missing data in the beginning or at the end. In the former case, the data is maximally fragmented into short sub-trajectories, while in the latter case the missing samples can be discarded, leaving one maximally long complete trajectory. Efficient methods that fit simultaneously multiple data sets are proposed in Markovsky and Pintelon (2015). Applied to the problem of identification from data with missing values, however, they are not data efficient because of unexploited relations among the data sets.

The approach based on simultaneous estimation of the missing data and identification of the underlying data-generating system is data efficient, however, it leads to a nonconvex optimization problem. There are different heuristics in the literature that can be used for solving it. We mention the local optimization method based on the variables projections (Markovsky, 2013; Markovsky & Usevich, 2013), methods using convex relaxations based on the nuclear norm (Fazel, 2002; Liu et al., 2013), and the subspace method of Markovsky (2016). Particularly interesting is the subspace method of Markovsky (2016) as it is noniterative and has theoretical guarantees to recover the data-generating system in case of exact data satisfying suitable assumptions. These assumptions, however, cannot be checked a priori as they are not explicitly given as in the classical identification problem (Willems, Rapisarda, Markovsky, & De Moor, 2005). A modification of the method of Markovsky (2016) and a priori verifiable sufficient conditions on the input for identifiability are derived in Alsalti, Markovsky, Lopez, and Müller (2024). The method of Markovsky (2016) and its modification in Alsalti et al. (2024), however, are not data efficient, *i.e.*, there are cases when the system is identifiable but the methods cannot identify it.

☆ The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Andrew Lamperski under the direction of Editor André Tits.

* Corresponding author.
E-mail addresses: imarkovsky@cimne.upc.edu (I. Markovsky), alsalti@irt.uni-hannover.de (M. Alsalti), lopez@irt.uni-hannover.de (V.G. Lopez), mueller@irt.uni-hannover.de (M.A. Müller).

https://doi.org/10.1016/j.automatica.2024.111869
0005-1098/© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

In this paper, we consider identification of discrete-time linear time-invariant systems from data with periodically missing output samples and fully specified input samples. The problem considered is related to, but different from, identification from non-uniformly sampled data, where the inputs as well as the outputs are missing (Zhang, Han, Ge, Ning, & Zhang, 2023). Thus, although existing methods for identification from non-uniformly sampled data, *e.g.*, the ones of Tsang and Billings (1995), Xie, Liu, Yang, and Ding (2010), Zhu, Telkamp, Wang, and Fu (2009), can be applied to the problem considered in this paper, they do not take into account the given inputs and are therefore less efficient than the specialized method proposed in the paper.

As explained above, the periodic distribution of the missing values is the worst case in the sense that it leads to a maximal fragmentation of the data. Indeed, if P is the period of the missing values, the longest complete sub-trajectory of the data trajectory has length $P-1$. Note that classical system identification methods, such as the N4SID and MOESP methods (Van Overschee & De Moor, 1996; Verhaegen & Dewilde, 1992), require trajectories with length at least equal to the lag ℓ of the system plus one in order to form a data matrix of sufficient size. Therefore, for $P \leq \ell+1$ classical identification methods cannot recover the system using the splitting method. In this setup, the methods of Markovsky (2016) and Alsalti et al. (2024) can recover the data-generating system when $P = \ell+1$ and the order of the system is larger than the number of outputs, but not when $P < \ell+1$. This made us hypothesize that the system is not identifiable when $P < \ell+1$. The present paper rejects the hypothesis, providing a method that recovers the system for any periodicity P of the missing output samples as well as any number of outputs. Moreover, as few as one output sample per period may be given. Thus, the method proposed in the paper is more data efficient than the one of Markovsky (2016). In order to show its full potential, we make the most severe assumption about the missing data and without loss of generality assume that the given output sample is the first one of the period.

Section 2 introduces the lifting operation by its action on a trajectory. Applied on a system, it results in a system that preserves linearity and time-invariance. Theorem 1 shows how the lifting operation changes the complexity of the system, which is defined as the triple: number of inputs, lag, and order. Theorem 2 shows the effect of the lifting operation on an input/state/output representation of the system. Based on the results in Section 2, Section 3 presents the proposed identification method for identification with periodically missing output samples. The key step of the method is the inverse transformation of the identified lifted system to the original system. The operation requires computing the Pth root of a matrix. The main result is stated in Theorem 4, which gives conditions under which the method recovers the true data-generating system, *i.e.*, an identifiability result for identification with periodically missing output samples. Section 4 empirically validates and illustrates the method.

The novel idea at the core of the proposed method is to use a *lifting operation* which subsamples the data at a rate P, then concatenates the P input/output samples over one period into a single sample. Provided that the data-generating system is linear time-invariant and the data is exact, the lifted data is also a trajectory of a linear time-invariant system, called the *lifted system*. Selecting the given elements from the outputs of the lifted system results in a complete trajectory from which the lifted system's dynamics can be identified by standard system identification methods. The inverse transformation from the lifted system to the original system is well-posed provided that the eigenvalues of the original system (*i.e.*, the eigenvalues of the A matrix in a minimal state–space representation) are in the sector

$$\mathscr{C}_P := \{\, \lambda \in \mathbb{C} \mid -\pi/P < \angle\lambda < \pi/P \,\} \qquad (\mathscr{C}_P)$$

where $\angle\lambda$ is the angle of λ and the lifted system has no real negative or zero eigenvalues (Bini, Higham, & Meini, 2005). The condition on the eigenvalues of the original system is not verifiable from the data and restricts the applicability of the method.

2. Methodology: the lifting operation

We use the behavioral approach, where a dynamical system $\mathscr{B}$ is defined as a set of trajectories (Willems, 1986,1987). Thus, $w \in \mathscr{B}$ means that w is a trajectory of the system $\mathscr{B}$. The restriction of the trajectory w and the system $\mathscr{B}$ to the finite interval $[1, T]$ is denoted by $w|_T$ and $\mathscr{B}|_T$, respectively. The class of linear time-invariant systems with q variables is denoted by $\mathscr{L}^q$.

The complexity of $\mathscr{B} \in \mathscr{L}^q$ determines the dimension of the set $\mathscr{B}|_T$ of finite-horizon trajectories. More specifically, as shown in Markovsky and Dörfler (2023, Corollary 5), there are integers $\mathbf{m}(\mathscr{B})$, $\mathbf{l}(\mathscr{B})$, and $\mathbf{n}(\mathscr{B})$ (that are properties of the system $\mathscr{B}$), such that

$$\dim \mathscr{B}|_T = \mathbf{m}(\mathscr{B})T + \mathbf{n}(\mathscr{B}), \text{ for all } T \geq \mathbf{l}(\mathscr{B}). \qquad (\text{DIM})$$

It can be shown that $\mathbf{m}(\mathscr{B})$ is the number of inputs in an input/output partitioning of the variables, $\mathbf{l}(\mathscr{B})$ is the observability index also called the *lag* of $\mathscr{B}$, and $\mathbf{n}(\mathscr{B})$ is the order of $\mathscr{B}$ (for details see Markovsky and Dörfler (2021, Section 2.2)). The class of linear time-invariant systems with fixed complexity is denoted by

$$\partial\mathscr{L}^q_{m,\ell,n} = \{\, \mathscr{B} \in \mathscr{L}^q \mid \mathbf{m}(\mathscr{B}) = m,\ \mathbf{l}(\mathscr{B}) = \ell,\ \mathbf{n}(\mathscr{B}) = n \,\}$$

and the class of linear time-invariant systems with bounded complexity is denoted by $\mathscr{L}^q_{m,\ell,n}$, where m, ℓ, n are upper bounds of the number of inputs, lag, and order, respectively.

A bounded complexity linear time-invariant system $\mathscr{B} \in \mathscr{L}^q_{m,\ell,n}$ admits an *input/state/output representation*

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

where σ is the shift operator, defined by $(\sigma w)(t) := w(t+1)$, and $\Pi \in \mathbb{R}^{q\times q}$ is a permutation matrix.

For $w = \big(w(1), \ldots, w(T)\big) \in (\mathbb{R}^q)^T$, the *Hankel matrix* with L block rows, where $1 \leq L \leq T$, constructed from w, is

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

For a trajectory $w \in \mathscr{B}|_T$ and a natural number $P \leq T$, we define the *lifted trajectory* $w' \in (\mathbb{R}^{qP})^{T'}$ as

$$w' = \big(w'(1), w'(2), \ldots, w'(T')\big) = \underbrace{\left(\begin{bmatrix} w(1) \\ \vdots \\ w(P) \end{bmatrix}, \begin{bmatrix} w(P+1) \\ \vdots \\ w(2P) \end{bmatrix}, \ldots, \begin{bmatrix} w\big((T'-1)P+1\big) \\ \vdots \\ w(T'P) \end{bmatrix} \right)}_{\text{lift}_P(w)},$$

where $T' := \lfloor T/P \rfloor$ is the largest integer smaller than or equal to T/P. The lifting operator lift_P is used for identification of periodically time-varying systems (Markovsky, Goos, Usevich, & Pintelon, 2014). In this paper, we use it for identification from data with periodically missing output samples. The matrix $\text{lift}_P(w)$ coincides with the Page matrix of w with depth P (Damen, Van den Hof, & Hajdasinski, 1982).

Acting on a system $\mathscr{B}$, the lifting operator creates a system

$$\mathscr{B}' = \text{lift}_P(\mathscr{B}) := \{\, \text{lift}_P(w) \mid w \in \mathscr{B} \,\}.$$

The following theorem asserts that if $\mathscr{B}$ is linear time-invariant, $\mathrm{lift}_P(\mathscr{B})$ is also linear time-invariant. The theorem characterizes the complexity of $\mathrm{lift}_P(\mathscr{B})$ in terms of the complexity of $\mathscr{B}$.

Theorem 1 (*Complexity of $\mathrm{lift}_P(\mathscr{B})$*)**.** *Let $\mathscr{B} \in \partial\mathscr{L}^{q}_{m,\ell,n}$ and $P \in \mathbb{N}$. Then, $\mathrm{lift}_P(\mathscr{B}) \in \partial\mathscr{L}^{Pq}_{m',\ell',n'}$, where*

$$m' = mP, \quad \ell' = \lceil \ell/P \rceil, \quad \text{and} \quad n' = n$$

($\lceil \ell/P \rceil$ is the smallest integer larger than or equal to ℓ/P).

Proof. By construction $\mathscr{B}|_{LP} = \mathscr{B}'|_L$ for any $L, P \in \mathbb{N}$. Using this fact and (DIM) applied to $\mathscr{B} \in \partial\mathscr{L}^{q}_{m,\ell,n}$, we have

$$\dim \mathscr{B}|_{LP} = \dim \mathscr{B}'|_L = mPL + n, \quad \text{for } PL \geq \ell. \tag{$*$}$$

Using Markovsky and Dörfler (2023, Corollary 5) applied to $\mathscr{B}' \in \partial\mathscr{L}^{qP}_{m',\ell',n'}$, we have

$$\dim \mathscr{B}'|_L = m'L + n', \quad \text{for } L \geq \ell'. \tag{$**$}$$

The complexity of $\mathscr{B}'$ follows by comparison of ($*$) and ($**$).

In what follows we assume that the system $\mathscr{B}$ admits an input/output partitioning $w = \left[\begin{smallmatrix} u \\ y \end{smallmatrix}\right]$, where $u(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}^p$, with $p := q - m$. Equivalently, in the input/state/output representation (I/S/O) we assume $\Pi = I$ (and skip it from the notation). The following result shows a particular input/state/output representation of the lifted system $\mathrm{lift}_P(\mathscr{B})$ that manifest a link to the parameters of an input/state/output representation of $\mathscr{B}$.

Theorem 2 (*State–Space Representation of $\mathrm{lift}_P(\mathscr{B})$*)**.** *Let $\mathscr{B} = \mathscr{B}_{\mathrm{ss}}(A, B, C, D)$ and $P \in \mathbb{N}$. Then, the lifted system has a state-space representation*

$$\mathrm{lift}_P(\mathscr{B}) = \mathscr{B}_{\mathrm{ss}}(A', B', C', D'), \quad \text{with}$$
$$A' = A^P, \quad B' = \begin{bmatrix} A^{P-1}B & \cdots & AB & B \end{bmatrix},$$
$$C' = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{P-1} \end{bmatrix}, \quad D' = \begin{bmatrix} D & & & \\ CB & D & & \\ \vdots & \ddots & \ddots & \\ CA^{P-1}B & \cdots & CB & D \end{bmatrix}. \tag{SS}$$

Moreover, if $\mathscr{B}_{\mathrm{ss}}(A, B, C, D)$ is minimal, then $\mathscr{B}_{\mathrm{ss}}(A', B', C', D')$ is also minimal.

Proof. For any trajectory $w' = \left[\begin{smallmatrix} u' \\ y' \end{smallmatrix}\right] \in \mathscr{B}'$ of the lifted system and for any $t \in \mathbb{N}$, we have that $w'(t) \in \mathscr{B}_{\mathrm{ss}}(A, B, C, D)|_P$. Therefore,

$$x\big(tP + 1\big) = A^P x\big((t-1)P + 1\big) + \begin{bmatrix} A^{P-1}B & \cdots & AB & B \end{bmatrix} u'(t)$$

and

$$y'(t) = C'x\big((t-1)P + 1\big) + D'u'(t),$$

where C' and D' are defined in (SS). The result follows by comparing the above equations and

$$\begin{aligned} \sigma^P x &= A'x + B'u' \\ y' &= C'x + D'u'. \end{aligned}$$

If $\mathscr{B}_{\mathrm{ss}}(A, B, C, D)$ is minimal, $\dim(A) = \mathbf{n}(\mathscr{B})$ and by Theorem 1, $\mathbf{n}(\mathscr{B}') = \mathbf{n}(\mathscr{B})$. The minimality of $\mathscr{B}_{\mathrm{ss}}(A', B', C', D')$ follows from $\dim(A') = \dim(A) = \mathbf{n}(\mathscr{B}')$.

3. Identification with periodically missing outputs

In this section, we consider the problem of identification from data with periodically missing output samples. Let $\mathscr{B} \in \mathscr{L}^{q}_{m,\ell,n}$ be the data-generating system and $w_{\mathrm{d}} \in \mathscr{B}|_{T_{\mathrm{d}}}$ be the data trajectory with missing output samples with period P. As explained in the introduction, the problem is challenging when $P \leq \mathbf{l}(\mathscr{B}) + 1$. Moreover, for the case $P < \mathbf{l}(\mathscr{B}) + 1$ there are no alternative methods that solve this problem. Next, we show how the problem can be solved using the lifting operation and a back transformation derived from (SS). Since the method converts the problem of identification with missing data to an identification with complete data, any method can be used for the latter. The method proposed for the missing data estimation problem is provably correct under given conditions and can deal with the situation of $\mathbf{l}(\mathscr{B}) = 1$ as well as up to $P-1$ missing output samples per period.

For simplicity assume that only one sample is given per period and it is the first one. Let $w'_{\mathrm{d}} := \mathrm{lift}_P(w_{\mathrm{d}})$ and let w''_{d} be the trajectory obtained from w'_{d} by removing the missing elements. It follows from Theorems 1 and 2 that w''_{d} is a trajectory of a system $\mathscr{B}'' \in \mathscr{L}^{mP+p}_{mP,\ell'',n}$ for some ℓ'' with a state–space representation

$$\begin{aligned} \mathscr{B}'' &= \mathscr{B}_{\mathrm{ss}}(A'', B'', C'', D'') \\ &:= \mathscr{B}_{\mathrm{ss}}\Big(A^P, \begin{bmatrix} A^{P-1}B & \cdots & AB & B \end{bmatrix}, C, \begin{bmatrix} D & 0 & \cdots & 0 \end{bmatrix}\Big). \end{aligned}$$

Under the generalized persistency of excitation (Markovsky & Dörfler, 2023)

$$\mathrm{rank}\, \mathscr{H}_{\ell''+1}(w''_{\mathrm{d}}) = mP(\ell'' + 1) + \mathbf{n}(\mathscr{B}), \tag{GPE}$$

the system $\mathscr{B}''$ can be recovered back from w''_{d} by computing the *most powerful unfalsified model* $\mathscr{B}_{\mathrm{mpum}}(w''_{\mathrm{d}})$, *i.e.*, the least complex model that fits the data exactly (Willems, 1986).

Remark 3 (*Upper Bound on the Lag of $\mathscr{B}''$*)**.** An upper bound $\ell'' \leq P\ell$ can be used in (GPE) as well as in the following results: Theorems 4 and 6 and ($T_{\min}$). Moreover, under certain assumptions (Krauss, Lopez, & Müller, 2022), it can be guaranteed that $\ell'' = \ell$.

A representation of $\mathscr{B}_{\mathrm{mpum}}(w''_{\mathrm{d}})$ can be computed in practice using standard identification methods. This leads to the following procedure for system identification from data with periodically missing output samples:

(1) find $\mathscr{B}_{\mathrm{mpum}}(w''_{\mathrm{d}}) := \mathscr{B}_{\mathrm{ss}}(A'', B'', C'', D'')$,
(2) compute the Pth matrix root $\widehat{A}$ of A'', and
(3) let $\widehat{B}$ be the last block-element of B'', $\widehat{C} = C''$, and $\widehat{D}$ the first block-element of D''.

The identified model $\widehat{\mathscr{B}}$ is then given by $\widehat{\mathscr{B}} := \mathscr{B}_{\mathrm{ss}}(\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D})$.

Our main result is the following theorem that gives conditions under which the model identified by the method above coincides with the true data-generating system.

Theorem 4 (*Identifiability Conditions*)**.** *Assuming that*

A.1 $\mathscr{B}''$ is identifiable from w''_{d}, i.e., *(GPE) holds,*
A.2 $\mathscr{B}''$ has no real negative or zero eigenvalues, and
A.3 the eigenvalues of $\mathscr{B}$ are in the sector $\mathscr{C}_P$ (see ($\mathscr{C}_P$))

the identified model is exact, i.e., *$\mathscr{B}_{\mathrm{ss}}(\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D}) = \mathscr{B}$.*

Proof. By Assumption A.1 and Markovsky and Dörfler (2023, Theorem 17), $\mathscr{B}''$ is identifiable, *i.e.*, $\mathscr{B}_{\mathrm{mpum}}(w''_{\mathrm{d}}) = \mathscr{B}''$. By Assumption A.2, A'' has a unique Pth root $\widehat{A}$ with eigenvalues in $\mathscr{C}_P$, see Gallier (2008, Theorem 4.8) and Bini et al. (2005). By Assumption A.3, $\widehat{A} = A$ up to a similarity transformation.

The conclusion $\mathscr{B}_{\mathrm{ss}}(\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D}) = \mathscr{B}$ follows from Theorem 2, *i.e.*, there is a representation $\mathscr{B}_{\mathrm{ss}}(\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D})$ of $\mathscr{B}$, such that the

corresponding lifted state–space representation matches the computed model parameters A'', B'', C'', D'' from the data w_{d}''. If $\mathscr{B}_{\mathrm{ss}}(A'', B'', C'', D'')$ is minimal, then $\mathscr{B}_{\mathrm{ss}}(\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D})$ is also minimal.

Remark 5 (*Minimal Number of Samples for Identifiability*)**.** For Assumption A.1 to be satisfied, the number of samples T should be at least

$$T_{\min} := P\big((mP+1)(\ell''+1)+\mathbf{n}(\mathscr{B})-1\big). \qquad (T_{\min})$$

The alternative identifiability result stated in the following theorem is equivalent to the fundamental lemma (Willems et al., 2005, Lemma 1) for the problem of system identification with periodically missing output samples. The order of persistency of excitation $\mathrm{PE}(u)$ of the signal $u \in (\mathbb{R}^m)^T$ is the maximal $L \in \mathbb{N}$, for which the Hankel matrix $\mathscr{H}_L(u)$ is full row-rank.

Theorem 6. *Assuming that*

A.1' the data-generating system $\mathscr{B}$ is controllable,
A.2' $\mathrm{PE}(u') = \ell'' + \mathbf{n}(\mathscr{B}) + 1$, and
A.3' Assumptions A.2 and 3 of Theorem 4 hold,

the identified model is exact, i.e., $\mathscr{B}_{\mathrm{ss}}(\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D}) = \mathscr{B}$.

Proof. The result follows by replacing the identifiability assumption (Assumption A.1) in Theorem 4 by the sufficient identifiability conditions of Willems et al. (2005).

4. Numerical examples

This section illustrates the effectiveness of the proposed method, implemented in the function `sysid_pmo`, on numerical examples and compares its performance with the performance of the following alternative methods:

`misdata` the method of Ljung (1999, Section 14.2), implemented in the `misdata` function of the System Identification Toolbox of Matlab,

`ident` the structured low-rank approximation method of Markovsky and Usevich (2013), implemented in the `ident` function of Markovsky (2013), initialized with the true system,

`hmc_nn` the nuclear norm method (Fazel, 2002; Liu et al., 2013), implemented using the CVX package (Grant & Boyd, 2017),

`mpum_md` the subspace method of Alsalti et al. (2024), and

`irdata` identification from periodically missing data (inputs and outputs), implemented in the System Identification Toolbox of Matlab (Ljung, 2023, Page 2-35).

The proposed method `sysid_pmo` is the most specialized of the compared methods. The method `irdata` is also for periodically missing data, but ignores some of the given input samples in our setup, while `sysid_pmo` fully exploits the special pattern of periodically missing outputs.

The first example, presented in Section 4.1, uses exact data from randomly generated systems satisfying the assumptions of Theorem 4. The aim of this example is to confirm numerically the correctness of Theorem 4. The second example, presented in Section 4.2, uses noisy data from a linear time-invariant system that models the effect of fluid infusion on the change of blood volume and serves as a simplified physiological model of cardiovascular hemodynamics (Alsalti, Tivay, Jin, Kramer, & Hahn, 2022; Bighamian, Reisner, & Hahn, 2016).

The numerical examples are reproducible with code available from https://imarkovs.github.io/mpum-md.tgz.

Table 1
In an experiment with a random 5th order single-input single-output system $\mathscr{B}$ satisfying the assumptions of Theorem 4, `sysid_pmo` recovers exactly the true system $\mathscr{B}$ from a trajectory w_{d} with 100 samples and alternating given-missing outputs.

Method	$\mathrm{dist}(\mathscr{B}, \widehat{\mathscr{B}})$	Execution time (sec)
`sysid_pmo`	10^{-6}	0.23
`misdata`	1.47	13.82
`ident`	10^{-5}	0.03
`hmc_nn`	1.43	16.03
`mpum_md`	1.53	0.05
`irdata`	1.52	0.07

4.1. Exact data from randomly generated systems satisfying the assumptions of Theorem 4

In this section, the data generating system $\mathscr{B}$ is random stable (obtained with the function `drss`). The simulation parameters are: the number of inputs m, the number of outputs p, the order n of $\mathscr{B}$, the period P of the missing output samples, and the number of samples T_{d}. If Assumption A.3 is not satisfied for $\mathscr{B}$, it is discarded and a new system is generated. This is repeated until a system $\mathscr{B}$ that satisfies Assumption A.3 is found. In the experiment, we chose $m = p = 1$, $n = 5$, $P = 2$, $T_{\mathrm{d}} = 100$ and average the results for 100 randomly generated systems.

The data w_{d} is a random trajectory of $\mathscr{B}$ with periodically missing outputs with one given sample per period. The methods compared are applied on the data w_{d}. The identified systems are validated by computing the distance $\mathrm{dist}(\mathscr{B}, \widehat{\mathscr{B}})$ between the true system $\mathscr{B}$ and the identified system $\widehat{\mathscr{B}}$, where $\mathrm{dist}(\mathscr{B}, \widehat{\mathscr{B}})$ is the principal angle between $\mathscr{B}|_{T_{\mathrm{d}}}$ and $\widehat{\mathscr{B}}|_{T_{\mathrm{d}}}$ (Fazzi & Markovsky, 2023; Padoan, van Coulson, Waarde, Lygeros, & Dörfler, 2022). The results, shown in Table 1, are averaged over 100 random systems, generated as explained above. The results indicate that among the alternative methods only `ident` recovers exactly (up to the numerical errors) the data-generating system. However, `ident` is initialized with the true system, which in practice is not available. The error $\mathrm{dist}(\mathscr{B}, \widehat{\mathscr{B}}) = 10^{-6}$ for the proposed method is an empirical confirmation of Theorem 4.

4.2. Case study: Cardiovascular hemodynamics

In this section we use a simplified linear time-invariant model of the cardiovascular hemodynamics that describes the effect of fluid infusion on the change of blood volume (Alsalti et al., 2022; Bighamian et al., 2016). The rationale for considering exact inputs but missing outputs in this application is that the input is administered by a clinician and is therefore exactly known while the output is difficult and time consuming to measure.

Denoting the change in blood volume with ΔV_B, the rate of fluid infusion with u, and the total infused volume of fluid with $\breve{u}(t)$, the cardiovascular hemodynamics model is

$$\frac{d}{dt}\begin{bmatrix}\Delta V_B \\ \breve{u}\end{bmatrix} = \begin{bmatrix}-K & \frac{K}{1+\alpha} \\ 0 & 0\end{bmatrix}\begin{bmatrix}\Delta V_B \\ \breve{u}\end{bmatrix} + \begin{bmatrix}1 \\ \frac{1}{1+\alpha}\end{bmatrix} u$$
$$y = \begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}\Delta V_B \\ \breve{u}\end{bmatrix}. \qquad (\text{CT-SYS})$$

Here α and K are patient-specific parameters specifying the steady-state split of fluid gain between the intravascular and extravascular compartments and the inter-compartmental rate of fluid shift, respectively. Being patient-specific α and K may differ from patient to patient and are in practice unknown. Therefore, data-driven methods that do not assume a given model and can deal with missing outputs are perfectly suited for the application.

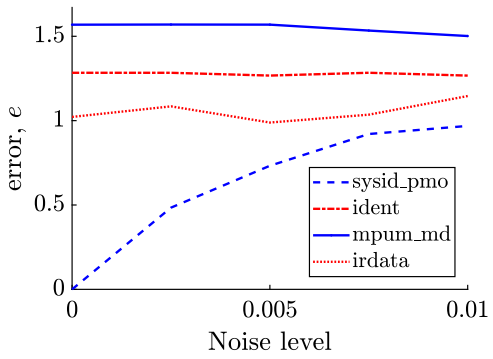


Fig. 1. In a Monte-Carlo experiment with the cardiovascular hemodynamics model (CT-SYS) discretized with period 1 min and noisy output measurements collected with a period 3 min, the proposed method `sysid_pmo` gives the most accurate estimate.

The continuous-time system (CT-SYS) with parameters $K = 0.5$ and $\alpha = 1.3$ is discretized with a sampling period 1 minute with the zero-order hold method. The data w_{d} is then simulated using the discretized system with fluid (*i.e.*, the input) administered in boluses for $T_{\mathrm{d}} = 150$ minutes. The change in the blood volume (*i.e.*, the output) is corrupted by zero mean white Gaussian process noise. We show results when the blood volume is measured with period 3 minutes, *i.e.*, 1 measurement per 3 samples.

The methods, except for `misdata` and `hmc_nn` (which are excluded from the comparison due to overly long computational times), are applied on the data and the distance measure $\mathrm{dist}(\mathscr{B}, \widehat{\mathscr{B}})$, introduced in Section 4.1, is computed. Fig. 1, shows the averaged over 100 Monte-Carlo repetitions of the experiment error $e := \frac{1}{N}\sum_{i=1}^{N} \mathrm{dist}(\mathscr{B}, \widehat{\mathscr{B}}^{(i)})$ as a function of the noise levels. The result indicates that the proposed method `sysid_pmo` gives the most accurate model among the methods compared.

5. Conclusions

We conjecture that the method proposed in the paper is data efficient for the problem of periodically missing output samples. Relaxing the assumptions and investigating the data efficiency of the method is left for future work.

Acknowledgments

The research leading to these results has received funding from the Catalan Institution for Research and Advanced Studies (ICREA), the Fond for Scientific Research Vlaanderen (FWO) project G033822N, and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant No 948679).

References

Alsalti, M., Markovsky, I., Lopez, V. G., & Müller, M. A. (2024). Data-based system representations from irregularly measured data. *IEEE Transactions on Automatic Control*.

Alsalti, M., Tivay, A., Jin, X., Kramer, G. C., & Hahn, J.-O. (2022). Design and in silico evaluation of a closed-loop hemorrhage resuscitation algorithm with blood pressure as controlled variable. *Journal of Dynamic Systems, Measurement, and Control*, *144*(2), Article 021001.

Bighamian, R., Reisner, A. T., & Hahn, J.-O. (2016). A lumped-parameter subject-specific model of blood volume response to fluid infusion. *Frontiers in Physiology*, *7*, 390.

Bini, D., Higham, N., & Meini, B. (2005). Algorithms for the matrix *p*th root. *Numerical Algorithms*, 349–378.

Broersen, P., De Waele, S., & Bos, R. (2004). Autoregressive spectral analysis when observations are missing. *Automatica*, *40*(9), 1495–1504.

Damen, A., Van den Hof, P., & Hajdasinski, A. (1982). Approximate realization based upon an alternative to the Hankel matrix: The page matrix. *Systems & Control Letters*, *2*, 202–208.

Fazel, M. (2002). *Matrix rank minimization with applications*. (PhD thesis), Elec. Eng. Dept., Stanford University.

Fazzi, A., & Markovsky, I. (2023). Distance problems in the behavioral setting. *European Journal of Control*, *74*, Article 100832.

Gallier, J. (2008). Logarithms and square roots of real matrices. arXiv:0805.0245.

Grant, M., & Boyd, S. (2017). CVX: Matlab software for disciplined convex programming. stanford.edu/boyd/cvx.

Hansson, A., & Wallin, R. (2012). Maximum likelihood estimation of Gaussian models with missing data—Eight equivalent formulations. *Automatica*, *48*(9), 1955–1962.

Isaksson, A. (1993). Identification of ARX-models subject to missing data. *IEEE Transactions on Automatic Control*, *38*(5), 813–819.

Krauss, I., Lopez, V. G., & Müller, M. A. (2022). Sample-based observability of linear discrete-time systems. In *IEEE 61st conf. on decision and control* (pp. 4199–4205).

Liu, Z., Hansson, A., & Vandenberghe, L. (2013). Nuclear norm system identification with missing inputs and outputs. *Systems & Control Letters*, *62*, 605–612.

Ljung, L. (1999). *System identification*. Prentice-Hall: Theory for the User.

Ljung, L. (2023). *System identification toolbox*. The MathWorks: User's Guide.

Markovsky, I. (2013). A software package for system identification in the behavioral setting. *Control Engineering Practice*, *21*, 1422–1436.

Markovsky, I. (2016). On the most powerful unfalsified model for data with missing values. *Systems & Control Letters*, 53–61.

Markovsky, I., & Dörfler, F. (2021). Behavioral systems theory in data-driven analysis, signal processing, and control. *Annual Reviews in Control*, *52*, 42–64.

Markovsky, I., & Dörfler, F. (2023). Identifiability in the behavioral setting. *IEEE Transactions on Automatic Control*, *68*(3), 1667–1677.

Markovsky, I., Goos, J., Usevich, K., & Pintelon, R. (2014). Realization and identification of autonomous linear periodically time-varying systems. *Automatica*, *50*, 1632–1640.

Markovsky, I., & Pintelon, R. (2015). Identification of linear time-invariant systems from multiple experiments. *IEEE Transactions on Signal Processing*, *63*(13), 3549–3554.

Markovsky, I., & Usevich, K. (2013). Structured low-rank approximation with missing data. *SIAM Journal on Matrix Analysis and Applications*, *34*(2), 814–830.

Padoan, A., van Coulson, J., Waarde, H. J., Lygeros, J., & Dörfler, F. (2022). Behavioral uncertainty quantification for data-driven control. arXiv:2204.02671.

Tsang, K. M., & Billings, S. A. (1995). Identification of systems from non-uniformly sampled data. *International Journal of Systems Science*, *26*(10), 1823–1837.

Van Overschee, P., & De Moor, B. (1996). Subspace identification for linear systems. *Theory, implementation, applications*. Boston: Kluwer.

Verhaegen, M., & Dewilde, P. (1992). Subspace model identification, part 1: The output-error state-space model identification class of algorithms. *International Journal of Control*, *56*, 1187–1210.

Wallin, R., & Hansson, A. (2014). Maximum likelihood estimation of linear SISO models subject to missing output data and missing input data. *International Journal of Control*, *87*(11), 2354–2364.

Wang, Y., Li, J., & Stoica, P. (2006). Spectral analysis of signals. *The missing data case*. Morgan Claypool.

Willems, J. C. (1986). From time series to linear system—Part II. Exact modelling. *Automatica*, *22*(6), 675–694.

Willems, J. C. (1986,1987). 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.

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

Xie, L., Liu, Y., Yang, H., & Ding, F. (2010). Modelling and identification for non-uniformly periodically sampled-data systems. *IET Control Theory & Applications*, *4*(5), 784–794.

Zhang, X.-M., Han, Q.-L., Ge, X., Ning, B., & Zhang, B.-L. (2023). Sampled-data control systems with non-uniform sampling: A survey of methods and trends. *Annual Reviews in Control*, *55*, 70–91.

Zhu, Y., Telkamp, H., Wang, J., & Fu, Q. (2009). System identification using slow and irregular output samples. *Journal of Process Control*, *19*(1), 58–67.