Systems & Control Letters 95 (2016) 53–61

Contents lists available at ScienceDirect

Systems & Control Letters

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

The most powerful unfalsified model for data with missing values

Ivan Markovsky [*]

Department ELEC, Vrije Universiteit Brussel (VUB), Pleinlaan 2, Building K, B-1050 Brussels, Belgium

ARTICLE INFO

Article history:
Available online 2 February 2016

Keywords:
Most powerful unfalsified model
Exact system identification
Subspace methods
Missing data
Low-rank matrix completion

ABSTRACT

The notion of the most powerful unfalsified model plays a key role in system identification. Since its introduction in the mid 80s, many methods have been developed for its numerical computation. All currently existing methods, however, assume that the given data is a *complete* trajectory of the system. Motivated by the practical issues of data corruption due to failing sensors, transmission lines, or storage devices, we study the problem of computing the most powerful unfalsified model from data with missing values. We do not make assumptions about the nature or pattern of the missing values apart from the basic one that they are a part of a trajectory of a linear time-invariant system. The identification problem with missing data is equivalent to a Hankel structured low-rank matrix completion problem. The method proposed selects rank deficient complete submatrices of the incomplete Hankel matrix. Under specified conditions the kernels of the submatrices form a nonminimal kernel representation of the data generating system. The final step of the algorithm is reduction of the nonminimal kernel representation to a minimal one. Apart from its practical relevance in identification, missing data is a useful concept in systems and control. Classic problems, such as simulation, filtering, and tracking control can be viewed as missing data estimation problems for a given system. The corresponding identification problems with missing data are "data-driven" equivalents of the classical simulation, filtering, and tracking control problems.

1. Introduction

1.1. Context and aim of the paper

The behavioral approach to systems and control was developed from "a need to put a clear and rational foundation under the problem of obtaining models from time series" [1, page 561]. One of the key ideas that came out from the original work [1–3] of Jan Willems is the notion of the "most powerful unfalsified model", or MPUM for short. The MPUM is "unfalsified" in the sense that it is an exact model for the given data and "most powerful" in the sense that it is the least complicated exact model. Thus, the MPUM is an optimal exact model for the data.

A candidate model $\widehat{\mathscr{B}}$ for the data w_{d} is unfalsified if w_{d} is a trajectory of $\widehat{\mathscr{B}}$. In the behavioral setting this fact is conveniently written as $w_{\mathrm{d}} \in \widehat{\mathscr{B}}$. (By definition the model is the set of all valid trajectories.) Restricting to the class of linear time-invariant models and assuming that the number of the input variables (which is a well defined quantity, see [1, Section 4]) is a priori fixed, the complexity of the model can be quantified by its order or by its lag. Let $\mathscr{L}_{\mathtt{m}}$ be the set of linear time-invariant systems with at most $\mathtt{m}$ inputs. The MPUM for the data w_{d} in the model class $\mathscr{L}_{\mathtt{m}}$ is defined as

$$\mathscr{B}_{\mathrm{mpum}}(w_{\mathrm{d}}) := \arg \underbrace{\min_{\widehat{\mathscr{B}} \in \mathscr{L}_{\mathtt{m}}} \operatorname{lag}(\widehat{\mathscr{B}})}_{\text{most powerful}} \quad \text{subject to} \quad \underbrace{w_{\mathrm{d}} \in \widehat{\mathscr{B}}}_{\text{unfalsified model}}. \tag{MPUM}$$

Apart from defining the notion of the MPUM, in [2], Jan Willems developed algorithms that implement the mapping $w_{\mathrm{d}} \mapsto \mathscr{B}_{\mathrm{mpum}}(w_{\mathrm{d}})$. These algorithms motivated the development of the subspace identification methods. The so-called "deterministic subspace identification" problem, see [4, Chapter 2] and [5, Chapter 7], is the problem of computing a state space representation of an exact model from data. Unlike the methods of [2], the subspace identification methods assume a priori given upper bound of the lag or the order of the model. If this bound is over specified, a nonminimal representation of $\mathscr{B}_{\mathrm{mpum}}(w_{\mathrm{d}})$ is computed. Subsequently, it is reduced to a minimal one. Thus instead of optimizing over the model complexity, the subspace methods use model reduction in order to find the MPUM.

The class of the subspace methods was generalized to approximate identification in the ARMAX [4,6] and errors-in-variables [7]

[*] Fax: +32 2 629 28 50.
E-mail address: ivan.markovsky@vub.ac.be.

http://dx.doi.org/10.1016/j.sysconle.2015.12.012
0167-6911/© 2015 Elsevier B.V. All rights reserved.

settings, identification in closed-loop [8], identification of dissipative and lossless systems [9,10], and other constrained identification problems. The subspace methods are practically successful and are still developed theoretically, generalized to new problems, and improvement computationally.

In this paper, we consider the exact (deterministic) identification problem in the case of data with missing values. Apart from the preliminary results [11] by the author, currently there are no subspace methods that address this problem. We do not make assumptions about the nature or pattern of the missing values apart from the basic one that they are a part of a valid trajectory of a linear time-invariant system with a given number of inputs and bounded lag. The missing elements can be both inputs and outputs of the system under some given input/output partitioning of the variables and they can appear in any pattern in time: periodically, randomly, or in blocks of consecutive time samples.

1.2. Literature review and contribution

Most of the currently existing literature on identification with missing data addresses special cases, such as specific patterns of occurrence of the missing data, or uses heuristics for estimation of the missing data, such as interpolation methods, followed by classical identification from the completed data. Three important special cases and three state-of-the-art methods that solve the general problem are reviewed next.

Special cases

The following special identification problems with missing data were considered in the literature:

- partial realization problem,
- missing input and output values in a single block, and
- missing values in the output only.

The partial realization problem is an exact identification problem from data consisting of the first few samples of the impulse response. This problem can be posed and solved as an extension of the given samples of the impulse response, *i.e.*, estimation of the missing output values, after the given ones. Kalman derived an analytical solution [12,13] for this problem. This solution, however, does not generalize to other patterns of missing data.

Another special identification problem with missing data considered in the literature [14] is the problem when missing are $w_\text{d}(t), w_\text{d}(t+1), \dots, w_\text{d}(t+\text{lag}(\mathscr{B}))$. In this case, the identification problem with missing data is equivalent to identification from two independent data sets: $w_\text{d}^1 = \big(w_\text{d}(1), \dots, w_\text{d}(t-1)\big)$ and $w_\text{d}^2 = \big(w_\text{d}(t+\text{lag}(\mathscr{B})+1), \dots, w_\text{d}(T)\big)$, where T is the number of samples of w_d. This result also does not generalize to other patterns of missing values.

The special case when the missing data is restricted to the output variables only can be handled by the classical prediction error identification methods [15,16]. The predictor is used to estimate the missing output values from the inputs, the current guess of the model and the initial conditions.

Optimization-based methods

The general identification problem with missing data can be approached by choosing a representation of the model and optimizing the complexity over the model parameters and the missing values, subject to the constraint that the completed data is a trajectory of the system. This leads to a nonconvex optimization problem. Three classes of methods that use this approach are:

- modification of the classical prediction error methods,
- methods developed in the structure low-rank approximation setting [17–19], and
- convex relaxation methods based on the nuclear norm heuristic.

All these methods are designed for estimation from noisy as well as missing data.

The approach using the prediction error methods for missing data estimation in the outputs was recently generalized in [20] to missing values of both inputs and outputs. Standard nonlinear local optimization methods are used. These methods require initial values for the optimization variables (model parameters and missing values) and the results depend on their closeness to a "good" locally optimal solution. Similar in spirit but different in implementation details [21–23] are the methods developed in the structure low-rank approximation setting.

An approach that gained popularity lately due to its success in compressive sensing is relaxation of the problem to a convex one by using the nuclear norm in lieu of the rank [24]. In [25], system identification with missing data is handled by (1) completion of the missing data using the nuclear norm heuristic (this step requires solution of a convex optimization problem), and (2) identification of a model parameter from the completed sequence using classical subspace identification methods. In the context of identification from noisy data, the optimization problem on step 1 involves a trade-off between the model complexity and the model accuracy. This trade-off is set by a user defined hyper-parameter. In the context of the exact identification problem considered in this paper, there is no trade-off parameter, see Section 5.1.

Contribution and organization of the paper

Our main contribution is a subspace type method for exact identification of a linear time-invariant system from data with missing values. Compared with the method based on the nuclear norm heuristic, the subspace method uses only linear algebra operations such as kernel computation and solution of linear systems of equations, which makes it computationally more efficient.

The subspace algorithm proposed in the paper selects complete submatrices of the incomplete Hankel matrix constructed from the data by (1) grouping together the columns of the Hankel matrix which have missing elements at the same positions and (2) skipping the missing elements. If the resulting submatrices have sufficiently many columns, their left kernels carry information about the most powerful unfalsified model.

Another contribution of the paper is using the missing data estimation methods for solving systems and control related problems, besides model identification. More specifically, we solve simulation and output tracking control problems by Algorithms 1 and 2, presented in Section 4. These examples show that missing data estimation is a unifying tool for systems and control related problems.

The paper is organized as follows. In Section 2 we define the notation being used. Section 3 defines formally the problems considered:

1. verification when a sequence w_d with missing data is a trajectory of a *given* linear time-invariant system $\mathscr{B}$, and
2. methods for computing $\mathscr{B}_\text{mpum}(w_\text{d})$ from a sequence w_d with missing values.

The solution to these problems is presented in Sections 4 and 5, respectively. The methods proposed are illustrated on examples and their advantages and disadvantages are compared. Conclusions and directions for future work are given in Section 6.

2. Preliminaries

Missing data values are denoted by the symbol NaN ("not a number"). The extended set of real numbers $\mathbb{R}_\text{e}$ is the union of the