Delft University of Technology

Fac. of Information Technology and Systems **Control Systems Engineering**

Technical report bds:99-07

Minimal State-Space Realization in Linear System Theory: An Overview*

B. De Schutter

To cite this work, please refer to the published version:

B. De Schutter, "Minimal state-space realization in linear system theory: An overview," *Journal of Computational and Applied Mathematics,* Special Issue on Numerical Analysis in the 20th Century – Vol. I: Approximation Theory, vol. 121, no. 1–2, pp. 331–354, Sept. 2000. doi:10.1016/S0377-0427(00)00341-1

Control Systems Engineering
Faculty of Information Technology and Systems
Delft University of Technology
Delft, The Netherlands
Current URL: https://www.dcsc.tudelft.nl

* This report can also be downloaded via https://dspub.eu/99-07

Minimal State-Space Realization in Linear System Theory: An Overview

B. De Schutter*

Keywords: minimal realization, linear system theory, state space models

Abstract

We give a survey of the results in connection with the minimal state space realization problem for linear time-invariant systems. We start with a brief historical overview and a short introduction to linear system theory. Next we present some of the basic algorithms for the reduction of non-minimal state space realizations and for the minimal state space realization of infinite or finite sequences of Markov parameters of linear time-invariant systems. Finally we discuss some extensions of this problem to other classes of systems and point out some related problems.

1 Introduction

1.1 Overview

In this paper we give an overview of the results in connection with the minimal state space realization problem for linear time-invariant (LTI) systems. The reason for focusing on LTI systems is that on the one hand they form a very simple class of systems that can be analyzed rather easily and for which many analytic and numerical results are available, but that on the other hand they have been used to solve many problems that appear in practice in a very satisfactory way. For sake of simplicity and conciseness, we will limit ourselves mainly to finite-dimensional discrete-time systems with real inputs and outputs in this paper. This choice is also motivated by the fact that most physical systems have real inputs and by the fact that some concepts (especially the Markov parameters) have a more natural physical interpretation for discrete-time systems than for continuous-time systems. Furthermore, most of the techniques for discrete-time systems with real-valued inputs and outputs are also valid for systems with complex inputs and outputs and for continuous-time systems.

In general the minimal state space realization problem for LTI systems can be formulated as follows: "Given some data about an LTI system, find a state space description of minimal size that explains the given data." The data are typically the impulse response of the system, the step response, input-output measurements, frequency response data, or more general frequency measurements. The minimal state space realization problem starting from impulse responses (or more general: sequences of Markov parameters) has been studied since the early 1960s and many algorithms have been developed to solve the problem. In this paper we will give an overview of some of these algorithms. At the end of the paper we will also

*Control Lab, Faculty of Information Technology and Systems, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands, email: b.deschutter@its.tudelft.nl

briefly discuss the minimal state space realization problem for some other classes of dynamical systems. Furthermore, we will also point out the relation between the minimal state space realization problem and more involved problems such as model reduction and identification.

This paper is organized as follows. In Sections 1.2 and 1.3 we give a brief overview of the history of linear system theory and we discuss the main differences between the state space representation and the transfer function representation of linear systems. In Section 2 we give a short and informal introduction to some of the basic concepts of linear system theory that are used in this paper. In Section 3 we treat various aspects of the minimal state space realization problem for LTI systems and discuss some algorithms for solving this problem. Finally, we consider some related problems and extensions of the basic minimal state space realization problem for LTI systems.

In order to limit the already large number of references in the bibliography of this paper we have selected a small subset of possible references, thereby aiming at historical papers, seminal papers, survey papers and reference works. Whenever we refer to a general book or paper, the reference is also intended to encompass the references included in that work.

1.2 Some historic notes on linear system theory and state space models[1]

Linear systems have already been studied for a long time and from many different points of view: in physics, mathematics, engineering, and so on. In an engineering context linear systems have been extensively studied since the 1930s. In those early days most researchers used frequency domain techniques (i.e. input-output or transfer function descriptions). Moreover, most of this work was done for single-input single-output (SISO) systems. At first sight the frequency domain techniques did not seem to extend satisfactorily to the multi-input multi-output (MIMO) systems that became increasingly important in aerospace, process control, and econometric applications in the late 1950s. This fact, and the importance of time-varying systems and time-domain characteristics in aerospace problems, led to a renewed interest in the state space description of linear systems, triggered by the work of Bellman and Kalman. The papers [18] and [32] give a good idea of the situation around 1960. The state space formulation led to many new ideas for systems design and feedback control. In the early 1970s Popov and Rosenbrock [43] have shown that many of the scalar transfer function concepts developed for SISO systems could also be extended to matrix transfer functions for MIMO systems. Now we could say that transfer functions descriptions (which are basically frequency domain methods) and state space descriptions (which are more oriented towards the time domain) are only two extremes of a whole spectrum of possible descriptions of finite-dimensional LTI systems. We can work exclusively with one description or the other, but we can also easily translate results from one framework to another, and it really depends on the application we have in mind which method best suits our needs.

In this paper we will only consider state space descriptions. The minimal realization problem for transfer functions is related to Padé approximation of rational functions, a topic that will be discussed in the contributions in this volume by Bultheel and De Moor, Guillaume, and Wuytack [9, 23, 60] (see also Section 4.1).

In the next section we will briefly discuss some differences between the state space description and the transfer function description of a linear system.

[1] This section is based on [31].

1.3 State space models versus transfer functions

The most important differences between the state space representation and the transfer function representation of a given system are [12, 48]:

- The transfer function of an LTI system describes the relation between the input and the output of the system under the assumption that the system is initially relaxed (i.e. the initial state is zero). Hence, if this assumption does not hold, the description is not applicable[2]. In contrast to the state space description, the transfer function representation does not reveal what will happen if the system is not initially relaxed (e.g. observable modes can be excited due to a nonzero initial state but may not appear in the transfer function due to pole-zero cancellation).

- The transfer function formulation does not reveal the behavior inside the system, such as unobservable unstable modes. Therefore, the transfer function matrix cannot always be used to study the stability properties of an LTI system. This problem of hidden pole-zero cancellation was not really understood prior to the work of Gilbert [18] and Kalman [32] who proved that the input-output description reveals only the controllable and observable part of a dynamical system.

- Although most results that are available for MIMO state space descriptions can now also be obtained in the transfer function approach, the state space formulation stays the most elegant way of dealing with generalizations like MIMO systems or nonlinear systems. Moreover, in practice the state space formulation is very important for numerical computations and controller design.

- The state space formulation can easily be extended to the time-varying case (see also Section 4.7). The extension of the transfer function to the time-varying case has not been very successful.

2 Linear system theory

In this section we give an informal introduction to some concepts of linear system theory that will be used in the subsequent sections. The notation used in this section and the following sections is mainly based on [31]. Unless explicitly indicated otherwise, the proofs of the theorems and properties given below can be found in [31]. Other introductions to linear system theory can be found in [12, 50].

2.1 Notation

The set of the real numbers is denoted by $\mathbb{R}$. All the vectors that appear in this paper are assumed to be column vectors, i.e. matrices with one column. If a is a vector then a_i represents the ith component of a. If A is a matrix then a_{ij} and $(A)_{ij}$ represent the entry on the ith row and the jth column of A. To select rows, columns and submatrices of a given matrix A we use the following Matlab-like notation. The ith row of A is denoted by $A(i,:)$, and the jth column by $A(:,j)$. Let i,j with $i<j$ be two row indices of A, and let k,l with $k<l$ be two column

[2]Note that this assumption does hold for the minimal state space realization problem based on the sequence of Markov parameters of an LTI system, which is the main topic of this paper.

indices of A. The submatrix of A consisting of the entries on rows $i, i+1, \ldots, j$ and columns $k, k+1, \ldots, l$ is denoted by $A(i{:}j, k{:}l)$. The submatrix of A consisting of rows $i, i+1, \ldots, j$ is denoted by $A(i{:}j, :)$. Similarly, the submatrix of A consisting of columns $k, k+1, \ldots, l$ is denoted by $A(:, k{:}l)$. The $n \times n$ identity matrix is denoted by I_n and the $m \times n$ zero matrix by $0_{m,n}$. If the dimensions of the identity matrix or the zero matrix are not indicated, they should be clear from the context.

2.2 Linear time-invariant systems

A system or model is said to be time-invariant if its response to any arbitrary input signal does not depend on absolute time. Consider a time-invariant system and let $\mathcal{S}(u)$ be the output of the system if the input signal u is applied to the system. Then we say that the system is linear if for every input signal u_1, u_2 and for every $c_1, c_2 \in \mathbb{R}$ we have $\mathcal{S}(c_1u_1 + c_2u_2) = c_1\mathcal{S}(u_1) + c_2\mathcal{S}(u_2)$. If we know and are interested in the inputs and outputs of the system at each time instant, then we will use a continuous-time model. On the other hand, in sampled or digital systems we often only know the signals of the system at certain discrete time instants (e.g. at integer multiples of the sampling period). This leads to discrete-time models.

The behavior of a continuous-time LTI system with m inputs and l outputs can be described by a model of the form

$$\begin{aligned} \frac{dx(t)}{dt} &= Ax(t) + Bu(t) && (1)\\ y(t) &= Cx(t) + Du(t) && (2) \end{aligned}$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{l\times n}$ and $D \in \mathbb{R}^{l\times m}$, and where u is the input of the system, y the output and x the state. Similarly, the evolution of a discrete-time LTI system can be described by a model of the form

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) && (3)\\ y(k) &= Cx(k) + Du(k)\ . && (4) \end{aligned}$$

The models (1) – (2) and (3) – (4) are called state space models. The number of components of the state vector x is called the order of the model. A state space model will be represented by the 4-tuple (A, B, C, D) of system matrices.

The Markov parameters G_k of an LTI system are defined by

$$G_0 = D \quad \text{and} \quad G_k = CA^{k-1}B \quad \text{for } k = 1, 2, \ldots \tag{5}$$

We say that (A, B, C, D) is a realization of the sequence $\{G_k\}_{k=0}^{\infty}$ if (5) holds. The realization is minimal if the model order is minimal. The model order of a minimal realization is called the minimal system order or sometimes also the McMillan degree of the system.

Consider a discrete-time LTI system and assume that $x(0) = 0$. If we apply a unit impulse $e(\cdot)$ defined by

$$e(k) = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{otherwise,} \end{cases}$$

to the ith input of the system and if we apply a zero signal to the other inputs, then the output of the system is given by

$$y(0) = D(:, i) \quad \text{and} \quad y(k) = CA^{k-1}B(:, i) \quad \text{for } k = 1, 2, \ldots$$

This output is called the impulse[3] response due to an impulse at the ith input. Note that $y(k)$ corresponds to the ith column of the matrix $CA^{k-1}B$ for $k = 1, 2, 3, \ldots$ Therefore, the sequence D, CB, CAB, CA^2B, $\ldots$ is called the *impulse response* of the system. Note that the terms of this sequence corresponds to the Markov parameters of the system. So for a discrete-time LTI system the sequence $\{G_k\}_{k=0}^{\infty}$ of Markov parameters corresponds to the impulse response of the system.

Remark 2.1 For a continuous-time LTI system the situation is a little bit more complicated: let $y^i(t)$ be the output of a continuous-time LTI system with model (1) – (2) if we apply a Dirac impulse to the ith input and a zero signal to the other inputs. The matrix-valued function $Y(\cdot) = [\, y^1(\cdot) \;\; y^2(\cdot) \;\; \ldots \;\; y^m(\cdot) \,]$ is called the impulse response of the system. It can be shown that the Taylor series expansion of $Y(\cdot)$ around the point $t = 0$ is given by

$$Y(t) = \sum_{k=0}^{\infty} G_k \frac{t^k}{k!} \;\; .$$

So for a continuous-time LTI system the relation between the Markov parameters and the impulse response is given by

$$G_k = \left. \frac{d^{k-1} Y(t)}{dt^{k-1}} \right|_{t=0} \;\; . \tag{6}$$

2.3 Controllability and observability

Consider a 4-tuple (A, B, C, D) of system matrices of an LTI system and let N be a positive integer. We define

$$\begin{aligned} \mathcal{O}_N(C, A) &= \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{N-1} \end{bmatrix} \\ \mathcal{C}_N(B, A) &= \begin{bmatrix} B & AB & \ldots & A^{N-1}B \end{bmatrix} \;\; . \end{aligned}$$

If n is the model order of the realization (A, B, C, D) then $\mathcal{O}_n(C, A)$ is called the observability matrix of the realization and $\mathcal{C}_n(A, B)$ is called the controllability matrix.

A realization (A, B, C, D) is called observable if the observability matrix $\mathcal{O}_n(C, A)$ has full rank. If a realization is observable, then we can always reconstruct the initial state $x(0)$ from observing the output evolution for $k \geq 0$ or $t \geq 0$ provided that we also know the input evolution for $k \geq 0$ or $t \geq 0$.

A realization is (A, B, C, D) is called controllable if the controllability matrix $\mathcal{C}_n(A, B)$ has full rank. If a realization is controllable, then for any initial state it is always possible to design an input sequence that steers the system to a desired final state.

The concepts observability and controllability are dual in the sense that a realization (A, B, C, D) is observable if and only if the dual realization (A^T, C^T, B^T, D) is controllable, and vice versa.

[3]Note that some authors prefer to use the term "pulse response" for the discrete-time case and reserve the term "impulse response" for the continuous-time case. However, in this paper we follow the terminology of [31] in which the term "impulse response" is used for both the discrete-time and the continuous-time case.

The following theorem which is due to Kalman gives a characterization of minimal state space realizations:

Theorem 2.2 *A realization (A, B, C, D) is minimal if and only if it is controllable and observable.*

In general, a state space realization of a given LTI system is not unique. Nevertheless, minimal state space representations are unique up to a change of basis of the state space, or equivalently, any two minimal state space realizations are connected by a unique similarity transformation [18, 32]:

Proposition 2.3 *If (A, B, C, D) and $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ are two minimal state space realizations of a given LTI system, there exists a unique invertible matrix T such that*

$$\tilde{A} = T^{-1}AT, \quad \tilde{B} = T^{-1}B, \quad \tilde{C} = CT \quad \text{and} \quad \tilde{D} = D \ . \tag{7}$$

Furthermore, the matrix T can be specified as $T = \mathcal{C}\tilde{\mathcal{C}}^T(\tilde{\mathcal{C}}\tilde{\mathcal{C}}^T)^{-1} = ((\tilde{\mathcal{O}}^T\tilde{\mathcal{O}})^{-1}\tilde{\mathcal{O}}^T\mathcal{O})^{-1}$ with $\mathcal{C} = \mathcal{C}_\rho(A, B)$, $\tilde{\mathcal{C}} = \mathcal{C}_\rho(\tilde{A}, \tilde{B})$, $\mathcal{O} = \mathcal{O}_\rho(C, A)$ and $\tilde{\mathcal{O}} = \mathcal{O}_\rho(\tilde{C}, \tilde{A})$ where ρ is the minimal system order.

The similarity transformation (7) corresponds to a transformation of the state $\tilde{x}(\cdot) = Tx(\cdot)$ where $x(\cdot)$ and $\tilde{x}(\cdot)$ are the state vectors of the realizations (A, B, C, D) and $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ respectively. Each choice of basis for the state space will lead to another state space representation (i.e. other system matrices). This results in several possible canonical forms such as the observer canonical form, the observability canonical form, the controller canonical form, etc. [31]. Different properties stand out more clearly in different realizations, and some canonical forms may have advantages in some applications. Note however that the input-output properties of the system such as the transfer function, the Markov parameters, the impulse response, and so on are not changed by similarity transformations.

In the next section we turn to the main topic of this paper: the minimal state space realization problem for LTI systems.

3 The minimal state space realization problem for LTI systems

3.1 Overview

The origins of the minimal state space realization problem lie in the early 1960s. The minimal state space realization problem for (continuous) LTI systems was first stated by Gilbert [18], who gave an algorithm for transforming a transfer function into a system of differential equations (i.e. a state space description). A second algorithm for the problem was given around the same time by Kalman [32]. The approach of Gilbert was based on partial-fraction expansions and worked under the assumption that each entry of the transfer function matrix has distinct poles. Kalman's algorithm was based on the theory of controllability and observability and reduced a non-minimal state space realization until it became minimal (cf. Theorem 2.2). Ho and Kalman [26, 27] approached the minimal realization problem from an entirely new point of view: they solved the problem starting from the sequence of Markov parameters of the system. Their algorithm will be discussed extensively below. All these algorithms assume that the entire sequence of Markov parameters is available. However, many times only a limited number of Markov parameters is available. The corresponding minimal *partial* state

space realization problem for MIMO systems was first explored by Kalman [34] and Tether [54]. Later, Rissanen [42] gave a recursive solution of the SISO version of this problem (which he claims can easily be extended to the MIMO case).

Most of the early work on the minimal state space realization problem dealt with the realization given the sequence of Markov parameters of the system. From a system-theoretical point of view this problem is often regarded as being somewhat academic. Nevertheless, there are several reasons why the minimal state space realization problem for LTI systems deserves to be studied :

- This problem is one of the most fundamental problems in system theory and can be considered as a simplified version of problems with noisy data, nonlinear models, etc. that occur frequently in practice. Before we deal with these more complex problems, it is useful to study the simplified version, which might lead to additional insight in the original problems. As such the solution of the minimal state space realization problem can also be seen as the first step towards problems such as model reduction and identification, which are of important practical interest.

- In order to analyze systems it is advantageous to have a compact description of the system. The aim of the minimal state space realization problem is to find a state space model of minimal size of the given system. Moreover, minimal realization techniques can also be used to reduce the order of existing state space models.

- Since the minimal realization is both controllable and observable, it is a good basis for designing an observer to estimate the states of the system from measurements of the outputs, and also for subsequently designing a state feedback controller (using e.g. pole placement).

- Furthermore, the minimal state space realization problem can be solved very elegantly using linear matrix algebra methods, that can be implemented in a numerically stable way.

The minimal state space realization problem has attracted much attention since the early 1960s, which has resulted in a wide variety of algorithms to solve the problem. In the next sections we will discuss some of these minimal state space realization algorithms.

In the remainder of the paper we will only consider discrete-time systems since for these systems the Markov parameters coincide with the terms of the impulse response, whereas for continuous-time systems the relation between the Markov parameters and the impulse response is more complicated (see Remark 2.1). Nevertheless, if we have in some way obtained the Markov parameters of a continuous-time LTI system then the techniques discussed below can also be used to obtain a minimal state space realization of that system. Note however that (6) implies that matching an increasing number of Markov parameters of a continuous-time system means placing increasing emphasis on the high-frequency behavior of the system, which is more susceptible to noise.

In general the basic minimal state space realization methods can be classified into two main groups:

- The first group consists of methods that start with a non-minimal realization which could be obtained fairly easily and then reduce it to get a realization that is both controllable and observable and therefore also minimal. These methods will be discussed in Section 3.2.

- The second group consists of those methods that start with the impulse response (or Markov parameters) of the system and obtain the minimal realization directly by suitable transformations of the resulting Hankel matrix. These methods are treated in Section 3.3.

Afterwards we will also consider the minimal partial realization problem in Section 3.4, and the realization or approximation of noisy measurements of the impulse response (in Section 3.5) and the step response (in Section 3.6).

3.2 Minimal realization based on reduction of non-minimal realizations

Suppose that we have a (not necessarily minimal) nth order state space realization (A, B, C, D) of a given LTI system. Rosenbrock [43] has developed a procedure to transform this realization into a minimal realization in two steps. In fact, this algorithm is merely a small modification to the standard algorithm for reducing matrices to echelon form [37]. Rosenbrock's method works as follows. The matrices A, B and C are put in a matrix

$$P = \left[\begin{array}{c|c} A & B \\ \hline C & 0 \end{array}\right] \quad .$$

By applying a similarity transformation on P that consists of a sequence of elementary row operations (such as interchanging two rows or adding the multiple of a row to another row) on the first n rows of P and the corresponding column operations on the first n columns of P, the matrix P can be transformed into a matrix of the form

$$\tilde{P} = \left[\begin{array}{cc|c} A_{11} & 0 & 0 \\ A_{21} & A_{22} & B_2 \\ \hline C_1 & C_2 & 0 \end{array}\right] \stackrel{\text{def}}{=} \left[\begin{array}{c|c} \tilde{A} & \tilde{B} \\ \hline \tilde{C} & 0 \end{array}\right]$$

where (A_{22}, B_2, C_2, D) is controllable. Since $(\tilde{A}, \tilde{B}, \tilde{C}, D)$ is connected to (A, B, C, D) by a similarity transformation, it is also a realization of the given system. Furthermore, since $\tilde{C}\tilde{A}^k\tilde{B} = C_2 A_{22}^k B_2$ for $k = 0, 1, 2, \ldots$, the 4-tuple (C_2, A_{22}, B_2, D) is a (controllable) state space realization of the given system. By an analogous procedure on the matrix

$$Q = \left[\begin{array}{c|c} A_{22}^T & C_2^T \\ \hline B_2^T & 0 \end{array}\right]$$

we obtain an observable realization. The resulting realization is then both controllable and observable and therefore also minimal (cf. Theorem 2.2).

A variant of Rosenbrock's method is implemented in the `minreal` command of Matlab. A stabilized version of Rosenbrock's algorithm is given in [56]. This algorithm is implemented in the freeware subroutine library SLICOT [7], which provides Fortran implementations of numerical algorithms for computations in systems and control theory.

3.3 Minimal realization of impulse responses

In this section we consider the problem of constructing a minimal realization starting from the impulse response $\{G_k\}_{k=0}^{\infty}$ of the system. Note that we always have $D = G_0$. Therefore, the problem of reconstructing D can be separated from the construction of A, B and C.

Many algorithms for minimal state space realization of impulse responses use the following block Hankel matrix:

$$H_{r,r'}(\mathcal{G}) = \begin{bmatrix} G_1 & G_2 & G_3 & \dots & G_{r'} \\ G_2 & G_3 & G_4 & \dots & G_{r'+1} \\ G_3 & G_4 & G_5 & \dots & G_{r'+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ G_r & G_{r+1} & G_{r+2} & \dots & G_{r+r'-1} \end{bmatrix} .$$

Note that if (A, B, C, D) is a realization of the impulse response $\mathcal{G}$ then we have

$$H_{r,r'}(\mathcal{G}) = \mathcal{O}_r(C, A)\mathcal{C}_{r'}(A, B) \ .$$

We also define the shifted block Hankel matrix $\overline{H}_N(\mathcal{G})$ as

$$\overline{H}_{r,r'}(\mathcal{G}) = \begin{bmatrix} G_2 & G_3 & G_4 & \dots & G_{r'+1} \\ G_3 & G_4 & G_5 & \dots & G_{r'+2} \\ G_4 & G_5 & G_6 & \dots & G_{r'+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ G_{r+1} & G_{r+2} & G_{r+3} & \dots & G_{r+r'} \end{bmatrix} .$$

The order of any minimal state space realization of the sequence $\mathcal{G} = \{G_k\}_{k=0}^{\infty}$ is given by

$$\rho = \operatorname{rank} H_{\infty,\infty}(\mathcal{G}) \ .$$

This result was discovered independently by Ho [25, 26, 27], Silverman [47], and Youla and Tissi [61].

Note that it is not always necessary to build the semi-infinite Hankel matrix $H_{\infty,\infty}(\mathcal{G})$ to determine the minimal system order. Indeed, if there is a linear relation between the Markov parameters of the form

$$G_{r+j} = \sum_{k=0}^{r-1} \alpha_k G_{k+j} \quad \text{for } j = 0, 1, 2, \dots \tag{8}$$

with $\alpha_0, \alpha_1, \dots, \alpha_{r-1} \in \mathbb{R}$, then we have $\rho = \operatorname{rank} H_{r,r}(\mathcal{G})$ [26, 27]. If the system matrix A of a (possible non-minimal) realization of the system is available, then a linear relation of the form (8) can easily be derived from the characteristic equation of the matrix A in combination with the Cayley-Hamilton theorem.

The use of Hankel matrices in realization theory was developed independently in the work of Ho and Kalman [25, 26, 27], Silverman [47], and Youla and Tissi [61]. These minimal realization algorithms can be divided into two groups:

- Some algorithms first determine the observable part of a system, and then the controllable part of the resulting system (or vice versa). Since the observability and controllability are dual concepts (see Section 2.3), the basic requirement is an algorithm for determining the controllable part. Most algorithms achieve this by selecting a largest set of linearly independent columns from the controllability matrix and use this set to construct a suitable transformation matrix (which removes the uncontrollable part). The resulting algorithms are quite complex. The algorithm of Silverman, which will be discussed more extensively below, belongs to this group.

- Another group of algorithms is based on a decomposition of the Hankel matrix. Both the algorithm of Ho and the algorithm of Youla and Tissi belong to this group.

3.3.1 Silverman's algorithm

The following theorem characterizes the sequences of Markov parameters that can be realized by an LTI system [48]:

Theorem 3.1 *An infinite sequence of Markov parameters $\mathcal{G} = \{G_k\}_{k=0}^{\infty}$ is realizable by an LTI state space model if and only if there exist positive integers r, r' and ρ such that*

$$\operatorname{rank} H_{r,r'}(\mathcal{G}) = \operatorname{rank} H_{r+1,r'+j}(\mathcal{G}) = \rho \tag{9}$$

for $j = 1, 2, \ldots$ The integer ρ then is the minimal system order.

In theory, the entire infinite sequence $\mathcal{G}$ is needed to determine realizability since in general it is not true that $\operatorname{rank} H_{r,r'+1} = \operatorname{rank} H_{r,r'}$ implies that (9) holds for all positive integers j [48]. However, for r large enough the rank of the Hankel matrix satisfies $\operatorname{rank} H_{r,r}(\mathcal{G}) = \rho$ where ρ is the minimal system order.

Let r, r' and ρ be determined as in Theorem 3.1. The method of Silverman [48, 49] is based on finding linearly independent rows in $H_{r,r'}(\mathcal{G})$. Let G be the submatrix of $H_{r,r'}(\mathcal{G})$ formed by the first ρ linearly independent rows of $H_{r,r'}(\mathcal{G})$, and let $\tilde{G}$ be the submatrix of $H_{r+1,r'}$ positioned l rows below G. Let F be the nonsingular $\rho \times \rho$ matrix formed by the first ρ linearly independent columns of G, and let $\tilde{F}$ be the $\rho \times \rho$ matrix occupying the same column positions in $\tilde{G}$ as does F in G. Let F_1 be the $l \times \rho$ matrix occupying the same column positions in $H_{1,r'}(\mathcal{G})$ as does F in G. If we define $A = \tilde{F}F^{-1}$, $B = G(:, 1{:}m)$, $C = F_1 F^{-1}$, and $D = G_0$ then (A, B, C, D) is a minimal state space realization of $\mathcal{G}$.

3.3.2 Ho's algorithm

The celebrated algorithm of Ho [26, 27] can be stated as follows:

1. Determine a linear relation of the form (8) or select r large enough (e.g. larger than or equal to the order of another — possibly non-minimal — realization if that is available) and define $\rho = \operatorname{rank} H_{r,r}(\mathcal{G})$.

2. Find nonsingular matrices P and Q such that[4]

$$PH_{r,r}(\mathcal{G})Q = \begin{bmatrix} I_\rho & 0 \\ 0 & 0 \end{bmatrix} \quad . \tag{10}$$

3. Now define

$$\begin{aligned} A &= E_{\rho,rl} P \overline{H}_{r,r}(\mathcal{G}) Q E_{\rho,rm}^T \\ B &= E_{\rho,rl} P H_{r,r}(\mathcal{G}) E_{m,rm}^T \\ C &= E_{l,rl} H_{r,r}(\mathcal{G}) Q E_{\rho,rm}^T \\ D &= G_0 \end{aligned}$$

   where $E_{p,q}$ is the $p \times q$ block matrix $[\, I_p \;\; 0_{p,q-p} \,]$.

[4] This is a standard problem in linear algebra. Apart from noting that P and Q may be taken to be lower and upper triangular, Ho and Kalman did not specify a particular matrix decomposition to be used in [26, 27].

This yields a minimal state space realization (A, B, C, D) of the sequence $\mathcal{G}$. Related algorithms using a reduced (i.e. smaller) Hankel matrix are described in [13, 44].

Note that (10) corresponds to a decomposition of the matrix $H_{r,r}(\mathcal{G})$ as $H_{r,r}(\mathcal{G}) = H_{\mathrm{o}} H_{\mathrm{c}}$ with $H_{\mathrm{o}} \in \mathbb{R}^{rl \times \rho}$ and $H_{\mathrm{c}} \in \mathbb{R}^{\rho \times rm}$ full rank matrices (with rank ρ). The algorithm of Youla and Tissi [61] is also based on such a decomposition of $H_{r,r}(\mathcal{G})$. It can be shown that for any full rank matrix decomposition $H_{r,r}(\mathcal{G}) = H_{\mathrm{o}} H_{\mathrm{c}}$ with $H_{\mathrm{o}} \in \mathbb{R}^{rl \times \rho}$ and $H_{\mathrm{c}}^{\rho \times rm}$ satisfying

$$\operatorname{rank} H_{\mathrm{o}} = \operatorname{rank} H_{\mathrm{c}} = \operatorname{rank} H_{r,r}(\mathcal{G}) = \rho \ ,$$

there exist matrices A, B, C from a ρ-dimensional state space model such that

$$H_{\mathrm{o}} = \mathcal{O}_r(C, A) \quad \text{and} \quad H_{\mathrm{c}} = \mathcal{C}_r(A, B) \ .$$

Furthermore, $\overline{H}_{r,r}(\mathcal{G}) = H_{\mathrm{o}} A H_{\mathrm{c}}$. The matrices A, B and C can then be constructed as follows: $A = H_{\mathrm{o}}^{+} \overline{H}_N(\mathcal{G}) H_{\mathrm{c}}^{+}$ where M^{+} is the pseudo-inverse of the matrix M, $B = H_{\mathrm{c}}(:, 1{:}m)$, and $C = H_{\mathrm{o}}(1{:}l, :)$.

A numerically very reliable procedure for both the full rank decomposition of $H_{r,r}(\mathcal{G})$ and for the construction of the pseudo-inverses H_{o}^{+} and H_{c}^{+} is the singular value decomposition (SVD) [22, 28]. The SVD also yields the most reliable numerical calculation of the rank of a matrix. The SVD of a matrix $M \in \mathbb{R}^{m \times n}$ is a decomposition of the form $M = U \Sigma V^T$ with $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ orthogonal matrices and $\Sigma \in \mathbb{R}^{m \times n}$ a diagonal matrix with $(\Sigma)_{11} \geq (\Sigma)_{22} \geq \ldots \geq 0$. The number of nonzero diagonal entries is equal to the rank of M.

The SVD can be used for the decomposition of the Hankel matrix $H_{r,r}(\mathcal{G})$ in the second step of Ho's algorithm as follows. Compute the SVD of $H_{r,r}(\mathcal{G})$: $H_{r,r}(\mathcal{G}) = U \Sigma V$ and define $H_{\mathrm{o}} = U \Sigma^{\frac{1}{2}}$ and $H_{\mathrm{c}} = \Sigma^{\frac{1}{2}} V^T$. This yields a decomposition that is equivalent to (10). The use of the SVD for the decomposition of the Hankel matrix was introduced by Zeiger and McEwen in their paper [62] in which they considered the problem of determining approximate state space realizations of noisy data (see also Section 3.5).

Remark 3.2 In general the system matrices A, B, C and D that result from the minimal realization algorithms discussed above do not exhibit a specific structure, i.e. all the system matrices are filled with nonzero coefficients. This implies that in general all $\rho(\rho + l + m) + lm$ entries have to be computed where ρ is the minimal system order. This has motivated work on algorithms that provide state space models with specific canonical structures such as e.g. the method of Ackerman and Bucy [1]. This method also consists in determining a set of linearly independent rows in the matrix $H_{r,r}(\mathcal{G})$. The resulting realization is in the canonical form of Bucy and has at most $\rho(l + m)$ parameters (the other entries are fixed at either 0 or 1).

3.4 The minimal partial realization problem

Now we assume that only a finite number of Markov parameters is available. So given a finite sequence $\mathcal{G}_N = \{G_k\}_{k=0}^{N}$ we want to find a 4-tuple (A, B, C, D) such that $D = G_0$ and $CA^{k-1}B = G_k$ for $k = 1, 2, \ldots, N$. In that case we say that (A, B, C, D) is a *partial realization* of $\mathcal{G}_N$. Note that trivially we have $D = G_0$. The 4-tuple (A, B, C, D) is said to be a minimal partial realization of $\mathcal{G}_N$ if and only if the size of A is minimal among all other partial realizations of $\mathcal{G}_N$.

Clearly, a minimal partial realization always exists. However, uniqueness (even up to a similarity transformation) is only guaranteed under certain conditions [54]:

Proposition 3.3 *Given a finite sequence* $\mathcal{G}_N = \{G_k\}_{k=0}^N$ *such that*

$$\operatorname{rank} H_{r,r'}(\mathcal{G}_N) = \operatorname{rank} H_{r+1,r'}(\mathcal{G}_N) = \operatorname{rank} H_{r,r'+1}(\mathcal{G}_N)$$

for some positive integers r, r' *with* $r + r' = N$*, then the extension of the sequence* $\mathcal{G}_N$ *to the infinite sequence* $\mathcal{G}_\infty = \{G_k\}_{k=N+1}^\infty$ *for which*

$$\operatorname{rank} H_{p',p}(\mathcal{G}_\infty) = \operatorname{rank} H_{r',r}(\mathcal{G}_\infty) = \operatorname{rank} H_{r',r}(\mathcal{G}_N)$$

with $p' + p = N + k$ *for* $k = 1, 2, \ldots$*, is unique.*

If the conditions of this proposition hold, we can still apply the algorithms that are developed for the full minimal realization problem [34, 54]:

Proposition 3.4 *The minimal partial realization problem of the sequence* $\mathcal{G}_N = \{G_k\}_{k=0}^N$ *may be solved by Ho's algorithm if and only if there exist positive integers* r *and* r' *with* $r + r' = N$ *such that*

$$\operatorname{rank} H_{r',r}(\mathcal{G}_N) = \operatorname{rank} H_{r',r}(\mathcal{G}_N) = \operatorname{rank} H_{r'+1,r}(\mathcal{G}_N) \tag{11}$$

The dimension of the minimal partial realization is equal to $\operatorname{rank} H_{r',r}$.

If the rank condition (11) is satisfied then any pair of two different minimal partial realizations of the sequence $\mathcal{G} = \{G_k\}_{k=0}^N$ are connected by a similarity transformation.

Note that if the rank condition (11) is satisfied and if we have a partial realization of $\{G_k\}_{k=0}^N$, then we cannot be sure that this minimal partial realization is also a realization of the entire sequence $\mathcal{G}_\infty = \{G_k\}_{k=0}^\infty$ since $\operatorname{rank} H_{r',r}(\mathcal{G}_\infty)$ may increase if we increase r or r'.

If we have a finite sequence $\mathcal{G}_N = \{G_k\}_{k=0}^N$ for which the rank condition (11) does not hold for any positive integer r, then the only possibility for utilizing Proposition 3.4 is to try to extend $\mathcal{G}_N$ to a longer sequence until (11) is satisfied. There could exist many extensions that satisfy the rank condition and each extension might yield a different minimal system order. Therefore, we now look for the extension that yields that smallest minimal system order among all possible extensions of $\mathcal{G}_N$ that satisfy the rank condition (11). A characterization of the resulting minimal system order is too complex to state here, but can be found in [34]. A similar result was discovered simultaneously and independently by Tether [54].

The procedure of reduction of a non-minimal state space representation of a finite sequence $\mathcal{G}_N$ of Markov parameters to a controllable and observable one does not necessarily lead to a minimal realization of $\mathcal{G}_N$. A compression algorithm to reduce an arbitrary finite realization of $\mathcal{G}_N$ to a minimal realization is given in [21]. This paper also provides a criterion for the minimality of a partial realization and an expression for the minimal system order.

Rissanen [42] has developed a recursive algorithm for the minimal partial state space realization problem. His algorithm is based on a decomposition of $p \times q$ submatrices $H_{p,q}$ of the Hankel matrix $H_{\infty,\infty}(\mathcal{G})$ as PQ with $P \in \mathbb{R}^{p\times\rho}$ a lower triangular matrix with 1s on the diagonal and with certain entries of the matrix $Q \in \mathbb{R}^{\rho\times q}$ set to 0 so that entries in the lower triangular part of P can be computed recursively one by one and such that the numbers already calculated do not change if extra rows or columns are added to $H_{p,q}$. This yields an efficient algorithm for subsequently computing minimal partial state space realizations of the finite sequences $\{G_k\}_{k=0}^N$, $\{G_k\}_{k=0}^{N+1}$, ... where more data are taken into account as new measurements become available. In contrast to the other minimal realization algorithms discussed above, which require a complete recalculation of all parameters each time a new measurement becomes available, this algorithm has the advantage that only a few new parameters need to be calculated to extend a partial realization.

3.5 Minimal realization of noisy measurements of the impulse response

In practice, we will never have the exact Markov parameters of an LTI system at our disposition, but we will have measured data which are disturbed by noise. Furthermore, in practice we will also only have a finite number of terms. Now we ask ourselves how we can extract the underlying LTI state space model from these noisy measurements.

If the minimal system order of the underlying "real" LTI system is ρ, then the measured sequence $\mathcal{G}_N = \{G_k\}_{k=0}^N$ can in general not be generated exactly by a ρth order state space model. Furthermore, the Hankel matrix $H_{r,r}(\mathcal{G})$ will generically be of full rank, which implies that it is not possible to construct a low-order state space realization that exactly matches the given sequence $\mathcal{G}_N$. Therefore, it may be better to make a good low-order approximation of the sequence $\mathcal{G}_N$ rather than to try to match it exactly. Here we already enter the field of identification which will be discussed more extensively in Section 4.3. However, since Ho's algorithm can easily be extended to the special case of this section, we already treat it here. The method presented here is due to Kung [35] and is based on the SVD:

1. Given the sequence $\mathcal{G}_N = \{G_k\}_{k=0}^N$, construct a Hankel matrix $H_{r,r'}(\mathcal{G})$ with $r + r' = N$.
2. Compute the SVD of $H_{r,r'}(\mathcal{G})$: $H_{r,r'}(\mathcal{G}) = U\Sigma V^T$. Look how the singular values $(\Sigma)_{ii}$ decrease as a function of the index i, and decide how many singular values are significant. The remaining singular values will be neglected. Let ρ be the number of singular values that are retained.
3. Construct $U_\rho = U(:, 1{:}\rho)$, $V_\rho = V(:, 1{:}\rho)$ and $\Sigma_\rho = \Sigma(1{:}\rho, 1{:}\rho)$.
4. Now apply Ho's algorithm to the matrix $H_{\text{red}}(\mathcal{G}) = U_\rho \Sigma_\rho V_\rho^T$. Since $H_{\text{red}}(\mathcal{G})$ has rank ρ, the order of the resulting minimal state space realization will be equal to ρ.

A related algorithm is given in [62] in which the SVD was also used, but no method for determining the resulting system order was specified.

Since in general the matrix $H_{\text{red}}(\mathcal{G})$ will not have a block Hankel structure, the Markov parameters of the resulting realization (A, B, C, D) will not exactly match the blocks of $H_{\text{red}}(\mathcal{G})$.

3.6 Minimal realization based on step response data

In many industrial processes we have step response measurements available instead of impulse response data. A straightforward way to do the realization then is to construct impulse response data by differencing or differentiating the step response data. However, this operation is not attractive since it will introduce an amplification of high-frequency noise in the data. As an alternative approach for discrete-time LTI systems, it is possible to use the step response data directly in a realization method that is a modified version of the Kung method. This modification is due to Van Helmont, Van der Weiden and Anneveld [57], and consists in applying similar operations as the Kung algorithm of Section 3.5 but this time on the matrix

$$T_{r,r'} = \begin{bmatrix} S_1 & S_2 & S_3 & \dots & S_{r'} \\ S_2 & S_3 & S_4 & \dots & S_{r'+1} \\ S_3 & S_4 & S_5 & \dots & S_{r'+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_r & S_{r+1} & S_{r+2} & \dots & S_{r+r'-1} \end{bmatrix} - \begin{bmatrix} S_0 & S_0 & S_0 & \dots & S_0 \\ S_1 & S_1 & S_1 & \dots & S_1 \\ S_2 & S_2 & S_2 & \dots & S_2 \\ \vdots & \vdots & \ddots & \vdots & \\ S_{r-1} & S_{r-1} & S_{r-1} & \dots & S_{r-1} \end{bmatrix}$$

with $r + r' = N + 1$ where $\{S_k\}_{k=0}^{N}$ is the measured step response.

In practice, the measurements that are available will not necessarily be impulse response or step response data, but general input-output data. Since these data will in general always contain noise, an exact realization of the data by an LTI model (of low order) will not be possible. This brings us to the topic of identification, which will be discussed in the next section together with other related problems and extensions of the minimal state space realization problem for LTI systems.

4 Related problems and extensions

4.1 Rational approximation

If we apply the z-transform to the discrete-time LTI state space model (3) – (4) and if we assume that the initial condition of the system is $x(0) = 0$, then we obtain the following relation between the input and the output of the system:

$$Y(z) = H(z)U(z)$$

with the transfer function $H(\cdot)$ of the system given by

$$H(z) = C(zI - A)^{-1}B + D = \sum_{k=0}^{\infty} G_k z^{-k} \ . \tag{12}$$

Since

$$H(z) = \frac{1}{\det(zI - A)} C \operatorname{adj}(zI - A)B + D$$

where $\operatorname{adj}(M)$ represents the adjoint matrix of M, the transfer function will always be a rational (matrix) function.

If we have a state space representation of a system, then the transfer function can be computed using (12). On the other hand, if we have a SISO transfer function

$$H(z) = \frac{\sum_{i=0}^{n} a_{n-i} z^i}{\sum_{i=0}^{n} b_{n-i} z^i}$$

of a discrete-time LTI system with b_0 normalized to 1, then a possible state space representation is given by the 4-tuple (A, B, C, D) with

$$A = \begin{bmatrix} -b_1 & -b_2 & \dots & -b_{n-1} & -b_n \\ 1 & 0 & \dots & 0 & \\ 0 & 1 & \dots & 0 & \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$C = \begin{bmatrix} a_1 - b_1 a_0 & a_2 - b_2 a_0 & \dots & a_n - b_n a_0 \end{bmatrix} \quad \text{and} \quad D = a_0 \ .$$

A similar result holds for SISO continuous-time LTI models. For the MIMO case, the SISO state space models that correspond to the individual transfer functions from each input to each output, could be stacked into one large MIMO state space model. However, the resulting state space models will in general certainly not be minimal. Therefore, several authors have developed methods to transform transfer function matrices into a minimal state space realization (see e.g. [33, 41]).

Since the state space representation can be converted into a transfer function and vice versa, we can also rephrase the minimal realization problem of Section 3.3 as follows: "Given the sequence of Markov parameters of an LTI system, determine the transfer function of the system with minimal McMillan degree." Since this transfer function is a rational function, this leads to the problem of approximation a given series by a rational function. This problem is related to the Padé approximation problem. For more information on this topic the reader is referred to [10, 11] and the contributions of Bultheel and De Moor, Guillaume, and Wuytack in this volume [9, 23, 60].

4.2 Model reduction

In many practical applications high-order LTI state space models are obtained (e.g. by combining models of separate components to build the model of a large plant, as the result of a filter or controller design, and so on). It is often desirable to replace them by lower-order models without introducing too much errors. Consequently, a wide variety of model reduction methods have been proposed. We shall concentrate on one method since it is connected to Ho's algorithm. It can be shown that the state space model obtained using Ho's algorithm with SVD will be "balanced". The idea of balanced realizations of systems has first been introduced to the control area by Moore [39] and uses similarity transformations to put the system in a form from which reduced models can be obtained. Loosely speaking, in a balanced realization every state is as controllable as it is observable. As a consequence, the states can be ordered in terms of their contribution to the input-output properties of the system. In order to model reduction the states with the least contribution can be removed.

More information on this topic can be found in [19].

4.3 Identification

In practice the input-output measurements of a system will be disturbed by sensor and process noise. Furthermore, there will be nonlinear effects, modeling errors and so on, which makes that the given data can almost never be explained by a linear model. This brings us to the topic of identification, where we want to determine a linear model that explains the given data as well as possible (and that has also good generalization properties).

There are several approaches to generate a linear model of a system. We could e.g. start from first principles and write down the basic physical laws that govern the behavior of the system. If the resulting model is nonlinear, we could linearize it in the operating point of the system in order to obtain a linear model. This "white-box" approach works for simple examples, but its complexity increases rapidly for real-world systems. An alternative approach is system identification, which is also called the "black-box" approach[5]. In system identification we first collect measurements of the input-output behavior of the system and afterwards we

[5]Note that there also exists a "grey-box" approach that is used when the state space equations of the system are known up to some unknown parameters, which are estimated using a parameter estimation method.

compute a model that explains the measured data. The field of identification has developed rapidly during the past decades. We can now distinguish two main groups of algorithms to identify linear LTI models on the basis of measured data: prediction error methods and subspace methods. Let us now briefly discuss these two main groups of techniques.

The prediction error methods were developed by Ljung and his co-workers [36]. In prediction error methods the model of the system is first parameterized in some canonical way, and then the model parameters are determined such that the measurements are explained as accurately as possible by the model. This is done by formulating a constrained optimization problem with the unknown parameters of the model as variables, with a measure of the deviation between the measured data and the predictions obtained from the model as the objective function, and the model equations as the constraints.

In the beginning of the 1990s a new type of linear system identification algorithms, called subspace methods, emerged. Subspace identification algorithms yield state space models and consist of two steps [14]. Most subspace methods first estimate the states of the system explicitly or implicitly using a projection of certain subspaces generated from the data. Next, they determine the state space model by a linear least squares method.

So in subspace methods the identification problem is reduced to a simple least squares problem, whereas in prediction error methods generally nonlinear optimization problems have to be solved. Since subspace identification methods do not involve nonlinear optimization techniques (which are in general iterative), they are faster than prediction error methods. Another advantage is that subspace methods — provided they are implemented correctly — have better numerical properties than prediction error methods. Furthermore, they do not suffer from problems with local minima. The price to be paid is that subspace methods are suboptimal.

Since giving an overview of this domain is beyond a scope of this paper, we refer the interested reader to the following papers and books for more information on this topic. An excellent recent overview of subspace identification methods can be found in [14]. Prediction error methods are described in [36]. Some other key references for the field are [5, 6, 8, 51].

In the next sections we will discuss the minimal realization problem for state space models that are not linear time-invariant. Although for most of these cases there exist theoretical characterizations of the minimal state space realization, for almost all of the cases there are currently no efficient algorithms to compute minimal realizations (except for the linear time-varying case).

4.4 Positive linear systems

Positive linear systems are LTI systems for which the state and the output are always nonnegative for any nonnegative input signal. Positive linear models appear when we have a system in which the variables must take nonnegative value due to nature of the underlying physical system. Typical examples of positive linear systems are networks of reservoirs, industrial processes involving chemical reactors, heat exchangers and distillation columns, age-structure population models, compartmental systems (which are frequently used for modeling transport and accumulation phenomena of substances in human body), water and atmospheric pollution models, stochastic models with probabilities as state variables, and many other models commonly used in economy and sociology.

So a discrete-time positive LTI system (or positive linear system for short) is a system

that can be described by a model of the form

$$x(k+1) = Ax(k) + Bu(k) \tag{13}$$
$$y(k) = Cx(k) + Du(k) \tag{14}$$

in which the components of the input, the state and the output are always nonnegative. This implies that the entries of the system matrices A, B, C and D are also nonnegative [55].

Now we consider the minimal state space realization problem for positive linear systems: "Given the impulse response $\mathcal{G} = \{G_k\}_{k=0}^{\infty}$ of a positive linear system, determine a positive state space realization (A, B, C, D) of $\mathcal{G}$ with the dimension of A as small as possible." Although the problem of finding a finite-dimensional positive state space realization for positive systems has been solved, the minimal positive state space realization problem has not been solved completely yet [2]. If $\mathcal{G} = \{G_k\}_{k=0}^{\infty}$ is the impulse response of the system, then in contrast to general discrete-time LTI systems, the rank of the Hankel matrix $H_{\infty,\infty}(\mathcal{G})$ is only a lower bound for the minimal positive system order, and there are systems for which the actual minimal positive system order is larger than the rank of the Hankel matrix. In general the minimal positive system order can be characterized as follows [55]:

Proposition 4.1 *Given the impulse response $\mathcal{G} = \{G_k\}_{k=0}^{\infty}$ of a positive linear system with l inputs, the minimal positive system order is equal to the smallest integer ρ for which there exist matrices $H_{\mathrm{o}} \in \mathbb{R}^{\infty\times\rho}$, $H_{\mathrm{c}} \in \mathbb{R}^{\rho\times\infty}$ and $A \in \mathbb{R}^{\rho\times\rho}$ such that*

$$H_{\infty,\infty}(\mathcal{G}) = H_{\mathrm{o}}H_{\mathrm{c}} \tag{15}$$
$$H_{\mathrm{o}}A = \hat{H}_{\mathrm{o}} \tag{16}$$

where $\hat{H}_{\mathrm{o}}$ is the matrix obtained by removing the first l rows of H_{o}.

However, there exist no efficient algorithms to compute a minimal decomposition of the form (15) – (16). It is easy to verify that if we have a minimal decomposition of the form (15) – (16) of $H_{\infty}(\mathcal{G})$ then the 4-tuple $(A, H_{\mathrm{c}}(:, 1{:}m), H_{\mathrm{o}}(1{:}l, :), G_0)$ is a minimal state space realization of the given impulse response. More information on this problem can be found in [17, 55].

4.5 Max-plus-algebraic models

In this section we focus on state space models for a class of discrete-event systems. Typical examples of discrete-event systems are manufacturing systems, telecommunication networks, railway traffic networks, and multi-processor computers. One of the characteristic features of discrete-event systems, as opposed to the continuous-variable systems[6] considered above, is that their dynamics are *event-driven* as opposed to time-driven. An event corresponds to the start or the end of an activity. For a manufacturing system possible events are: the completion of a part on a machine, a machine breakdown, or a buffer becoming empty.

In general, models that describe the behavior of discrete-event systems are nonlinear, but there exists a class of discrete-event systems for which the model becomes "linear" when formulated in the max-plus algebra, which has maximization (represented by $\oplus$) and addition (represented as $\otimes$) as its basic operations. Loosely speaking, this class of discrete-event systems can be characterized as the class of deterministic time-invariant discrete-event systems in which only synchronization and no concurrency occurs. If we write down a model for the

[6] I.e. systems the behavior of which can be described by difference or differential equations.

behavior of such a system, then the operations maximization and addition arise as follows. Synchronization corresponds to maximization (a new activity can only start when all the preceding activities have been finished, i.e. after the maximum of the finishing times of the preceding activities), whereas the duration of activities corresponds to addition (the finishing time of an activity is the starting time plus the duration of the activity). This leads to a model of the following form[7]:

$$x(k+1) = A \otimes x(k) \oplus B \otimes u(k) \quad (17)$$
$$y(k) = C \otimes x(k) \ . \quad (18)$$

For a manufacturing system, $u(k)$ would typically represent the time instants at which raw material is fed to the system for the $(k+1)$th time, $x(k)$ the time instants at which the machines start processing the kth batch of intermediate products, and $y(k)$ the time instants at which the kth batch of finished products leaves the system.

Note that the description (17) – (18) closely resembles the state space description (3) – (4) for discrete-time LTI systems, but with + replaced by $\oplus$ and $\times$ by $\otimes$. Therefore, we say that (17) – (18) is a max-plus-linear model, i.e. a model that is linear in the max-plus algebra.

The reason for using the symbols $\oplus$ and $\otimes$ to denote maximization and addition is that there is a remarkable analogy between $\oplus$ and addition, and between $\otimes$ and multiplication: many concepts and properties from conventional linear algebra and linear system theory (such as the Cayley-Hamilton theorem, eigenvectors and eigenvalues, Cramer's rule, ...) can be translated to the max-plus algebra and max-plus-algebraic system theory by replacing + by $\oplus$ and $\times$ by $\otimes$. However, since there does not exist a max-plus-algebraic equivalent of the minus operator, we cannot straightforwardly transfer all the techniques from linear system theory to the max-plus-algebraic system theory.

We can also define the minimal state space realization problem for max-plus-linear time-invariant discrete-event systems. This problem is strongly related to the minimal realization problem for positive linear systems that was considered in the previous section (e.g. with the proper change of notation, Proposition 4.1 also holds for max-plus-linear time-invariant systems). Just as for positive linear systems, there are currently no efficient, i.e. polynomial-time, algorithms to solve the general max-plus-algebraic minimal state space realization problem, and there are strong indications that the problem is at least NP-hard. Nevertheless, there are also some special cases for which efficient algorithms exist. An recent overview of the current status of research and the open questions in connection with this problem is given in [15, 40].

4.6 Multi-dimensional minimal state space realization

In recent years there has been an increasing interest in the study of multi-dimensional systems, due to a wide range of applications in image processing, seismological data, geophysics, computer tomography, control of multi-pass processes, and so on. An n-dimensional state space model has the following form:

$$\overset{\star}{x} = Ax + Bu(i_1, i_2, \ldots, i_n)$$
$$y(i_1, i_2, \ldots, i_n) = Cx + Du(i_1, i_2, \ldots, i_n)$$

[7] The max-plus-algebraic matrix sum and product are defined in the same way as in linear algebra but with + replaced by $\oplus$ and $\times$ by $\otimes$. So $(A \oplus B)_{ij} = a_{ij} \oplus b_{ij} = \max(a_{ij}, b_{ij})$ and $(A \otimes B)_{ij} = \bigoplus_k a_{ik} \otimes b_{kj} = \max_k (a_{ik} + b_{kj})$.

with

$$\overset{\star}{x} = \begin{bmatrix} x_{11}(i_1+1,i_2,\ldots,i_n) \\ x_{12}(i_1,i_2+1,\ldots,i_n) \\ \vdots \\ x_{1n}(i_1,i_2,\ldots,i_n+1) \\ \hline x_{21}(i_1+1,i_2,\ldots,i_n) \\ x_{22}(i_1,i_2+1,\ldots,i_n) \\ \vdots \\ x_{2n}(i_1,i_2,\ldots,i_n+1) \\ \hline \vdots \\ x_{mn}(i_1,i_2,\ldots,i_n+1) \end{bmatrix} \quad \text{and} \quad x = \begin{bmatrix} x_{11}(i_1,i_2,\ldots,i_n) \\ x_{12}(i_1,i_2,\ldots,i_n) \\ \vdots \\ x_{1n}(i_1,i_2,\ldots,i_n+1) \\ \hline x_{21}(i_1,i_2,\ldots,i_n) \\ x_{22}(i_1,i_2,\ldots,i_n) \\ \vdots \\ x_{2n}(i_1,i_2,\ldots,i_n) \\ \hline \vdots \\ x_{mn}(i_1,i_2,\ldots,i_n) \end{bmatrix} .$$

The minimal state space realization problem and the model reduction problem play an important role in the analysis and design of multi-dimensional systems because of the large amount of data involved in multi-dimensional signal processing. However, the general problem of minimal state space realization of multidimensional systems has not been solved even for 2-dimensional systems. Nevertheless, for some special cases minimal state space realization methods have been derived. For more information the interested reader is referred to [3, 38] and the references therein.

4.7 Linear time-varying models

The system matrices in the state space models of the previous sections were constant over time. However, we can also consider time-varying linear systems in which the system matrices also depend on time:

$$\begin{aligned} x_{k+1} &= A_k x_k + B_k u_k \\ y_k &= C_k x_k + D_k \ . \end{aligned}$$

Some authors even consider models in which the dimensions of the system matrices may change over time. Minimal state space realizations for linear time-varying systems can also be characterized as being both controllable and observable [16]. Furthermore, the algorithm of Youla and Tissi can be extended to yield minimal state space realizations for time-varying linear systems. We refer to [4, 16, 20, 45] for more information on this topic.

4.8 Nonlinear models

When we use linear models to model physical systems, we are making some assumptions that correspond to an idealization of the real world, which is in fact nonlinear. Although LTI models turn out to be able to approximate many real-world systems and processes very well in practice, sometimes nonlinear models are required. In general a discrete-time nonlinear time-invariant model has the following form:

$$\begin{aligned} x_{k+1} &= f(x_k, u_k) \\ y_k &= g(x_k, u_k) \ . \end{aligned}$$

We can also define a state space realization and a minimal state space realization for nonlinear systems. In analogy with linear systems, some authors define a minimal realization of a

nonlinear system as a realization that is both controllable and observable [53]. However, where for a linear systems the dimension of the minimal realization can easily be determined from the impulse response or input-output data of the system, the situation is far more complicated for nonlinear systems. For more information in this context, the reader is referred to [24, 29, 30, 46, 53].

There are many other classes of linear and nonlinear time-invariant or time-varying systems (such as linear systems that operate on finite fields or integers (instead of real numbers), descriptor systems, periodic systems, ...) for which minimal state space realization results exist, but it would be beyond the scope of this paper to discuss them all. More information on this topic can be found in [31, 52, 58, 59] and the references therein.

5 Conclusion

In this paper we have given an overview of the minimal state space realization problem for linear time-invariant systems and discussed some related problems and extensions. The basic problem has been solved satisfactorily since the mid 1960s and has led to a renewed research in various fields such as model reduction, approximation and identification. Especially for general nonlinear systems and special classes of nonlinear systems there still is much active research going on.

References

[1] J.E. Ackermann and R.S Bucy, "Canonical minimal realization of a matrix of impulse response sequences," *Information and Control*, no. 3, pp. 224–231, Oct. 1971.

[2] B.D.O. Anderson, "Positive system realizations," in *Open Problems in Mathematical Systems and Control Theory* (V.D. Blondel, E.D. Sontag, M. Vidyasagar, and J.C. Willems, eds.), ch. 2, London: Springer-Verlag, 1999.

[3] G.E. Antoniou, P.N. Paraskevopoulous, and S.J. Varoufakis, "Minimal state-space realization of factorable 2-D transfer functions," *IEEE Transactions on Circuits and Systems*, vol. 35, no. 8, Aug. 1988.

[4] D.Z. Arov, M.A. Kaashoek, and D.R. Pik, "Optimal time-variant systems and factorization of operators, I: Minimal and optimal systems," *Integral Equations and Operator Theory*, vol. 31, no. 4, pp. 389–420, Aug. 1998.

[5] K.J. Åström and P. Eykhoff, "System identification — A survey," *Automatica*, vol. 7, no. 2, pp. 123–162, Mar. 1971.

[6] K.J. Åström and T. Söderström, eds., *Automatica*, vol. 31, no. 12, Dec. 1995. Special Issue on Trends in System Identification.

[7] P. Benner, V. Sima, V. Mehrmann, S. Van Huffel, and A. Varga, "SLICOT – A subroutine library in systems and control theory," in *Applied and Computational Control, Signals, and Circuits, Volume 1* (B. Datta, ed.), ch. 10, pp. 504–546, Birkhäuser Boston, 1999. See also `http://www.win.tue.nl/niconet/NIC2/slicot.html`.

[8] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel, *Time series analysis: Forecasting and Control*. Englewood Cliffs, New Jersey: Prentice-Hall, 3rd ed., 1994.

[9] A. Bultheel and B. De Moor, "Rational approximation in linear systems and control." Invited paper for volume 1 on Approximation Theory of the NA20 project of *Journal of Computational and Applied Mathematics*, 1999.

[10] A. Bultheel and M. Van Barel, "Padé techniques for model reduction in linear system theory: A survey," *Journal of Computational and Applied Mathematics*, vol. 14, no. 3, pp. 401–438, Mar. 1986.

[11] A. Bultheel and M. Van Barel, *Linear Algebra, Rational Approximation and Orthogonal Polynomials*, vol. 6 of *Studies in Computational Mathematics*. Amsterdam, The Netherlands: Elsevier, 1997.

[12] C.T. Chen, *Linear System Theory and Design*. New York: Holt, Rinehart and Winston, 1984.

[13] C.T. Chen and D.P. Mital, "A simplified irreducible realization algorithm," *IEEE Transactions on Automatic Control*, vol. 17, pp. 535–537, Aug. 1972.

[14] B. De Moor, P. Van Overschee, and W. Favoreel, "Algorithms for subspace state-space system identification: An overview," in *Applied and Computational Control, Signals, and Circuits, Volume 1* (B. Datta, ed.), ch. 6, pp. 271–335, Birkhäuser Boston, 1999.

[15] B. De Schutter and G.J. Olsder, "The minimal state space realization problem in the max-plus algebra: An overview," Tech. rep., Control Laboratory, Fac. of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands, May 1999. Submitted for publication.

[16] P. Dewilde and A.J. van der Veen, *Time-Varying Systems and Computations*. Boston: Kluwer Academic Publishers, 1998.

[17] L. Farina and L. Benvenuti, "Positive realizations of linear systems," *Systems & Control Letters*, vol. 26, pp. 1–9, 1995.

[18] E.G. Gilbert, "Controllability and observability in multi-variable control systems," *SIAM Journal on Control*, vol. 1, no. 2, pp. 128–151, 1963.

[19] K. Glover, "All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds," *International Journal of Control*, vol. 39, no. 6, pp. 1115–1193, June 1984.

[20] I. Gohberg and M.A. Kaashoek, "On minimality and stable minimality of time-varying linear systems with well-posed boundary conditions," *International Journal of Control*, vol. 43, no. 5, pp. 1401–1411, May 1986.

[21] I. Gohberg, M.A. Kaashoek, and L. Lerer, "On minimality in the partial realization problem," *Systems & Control Letters*, vol. 9, pp. 97–104, 1987.

[22] G.H. Golub and C.F. Van Loan, *Matrix Computations*. Baltimore, Maryland: The John Hopkins University Press, 1989.

[23] P. Guillaume, "Multivariate Padé approximation." Invited paper for volume 1 on Approximation Theory of the NA20 project of *Journal of Computational and Applied Mathematics*, 1999.

[24] R. Hermann and A. Krener, "Nonlinear controllability and observability," *IEEE Transactions on Automatic Control*, vol. 22, pp. 728–740, 1977.

[25] B.L. Ho, *An effective construction of realizations from input/output descriptions*. PhD thesis, Stanford University, Stanford, California, 1966.

[26] B.L. Ho and R.E. Kalman, "Effective construction of linear state-variable models from input/output functions," in *Proceedings of the 3rd Annual Allerton Conference on Circuit and System Theory* (Monticello, Illinois, Oct. 1965) (M.E. Van Valkenburg, ed.), pp. 449–459, 1965. See also [27].

[27] B.L. Ho and R.E. Kalman, "Effective construction of linear, state-variable models from input/output functions," *Regelungstechnik*, vol. 14, no. 12, pp. 545–548, 1966. This paper is more or less a reprint of [26].

[28] R.A. Horn and C.R. Johnson, *Matrix Analysis*. Cambridge, United Kingdom: Cambridge University Press, 1985.

[29] A. Isidori, *Nonlinear Control Systems*. Springer, 3rd ed., 1995.

[30] B. Jakubczyk, "Existence and uniqueness of realizations of nonlinear systems," *SIAM Journal on Control and Optimization*, vol. 18, no. 4, pp. 455–471, July 1980.

[31] T. Kailath, *Linear Systems*. Englewood Cliffs, New Jersey: Prentice-Hall, 1980.

[32] R.E. Kalman, "Mathematical description of linear dynamical systems," *SIAM Journal on Control*, vol. 1, no. 2, pp. 152–192, 1963.

[33] R.E. Kalman, "Irreducible realizations and the degree of a rational matrix," *Journal of the Society for Industrial and Applied Mathematics*, vol. 13, no. 2, pp. 520–544, June 1965.

[34] R.E. Kalman, "On minimal partial realizations of a linear input/output map," in *Aspects of Network and System Theory* (R.E. Kalman and N. DeClaris, eds.), pp. 385–407, New York: Holt, Rinehart and Winston, 1971.

[35] S.Y. Kung, "A new identification and model reduction algorithm via singular value decomposition," in *Proceedings of the 12th Asilomar Conference on Circuits, Systems and Computers*, Pacific Grove, California, pp. 705–714, 1978.

[36] L. Ljung, *System Identification: Theory for the User*. Upper Saddle River, New Jersey: Prentice-Hall, 2nd ed., 1999.

[37] D.Q. Mayne, "An elementary derivation of Rosenbrock's minimal realization algorithm," *IEEE Transactions on Automatic Control*, pp. 306–307, June 1973.

[38] S.H. Mentzelopoulou and N.J. Theodorou, "n-dimensional minimal state-space realization," *IEEE Transactions on Circuits and Systems*, vol. 38, no. 3, pp. 340–343, Mar. 1991.

[39] B.C. Moore, "Principal component analysis in linear systems: controllability, observability and model reduction," *IEEE Transactions on Automatic Control*, vol. 26, no. 1, pp. 17–31, Feb. 1981.

[40] G.J. Olsder and B. De Schutter, "The minimal realization problem in the max-plus algebra," in *Open Problems in Mathematical Systems and Control Theory* (V.D. Blondel, E.D. Sontag, M. Vidyasagar, and J.C. Willems, eds.), ch. 32, pp. 157–162, London: Springer-Verlag, 1999.

[41] I.S. Pace and S. Barnett, "Efficient algorithms for linear system calculations — Part II Minimal realization," *International Journal of Systems Science*, vol. 5, no. 5, pp. 413–424, 1974.

[42] J. Rissanen, "Recursive identification of linear systems," *SIAM Journal on Control and Optimization*, vol. 9, no. 3, pp. 420–430, Aug. 1971.

[43] H.H. Rosenbrock, *State-Space and Multivariable Theory*. London: Thomas Nelson and Sons, 1970.

[44] P. Rózsa and N.K. Sinha, "Efficient algorithm for irreducible realization of a rational matrix," *International Journal of Control*, vol. 20, no. 5, pp. 739–751, Nov. 1974.

[45] W.J. Rugh, *Linear System Theory*. Upper Saddle River, New Jersey: Prentice-Hall, 2nd ed., 1996.

[46] J.M.A. Scherpen and W.S. Gray, "Minimality and similarity invariants of a nonlinear state space realization." Accepted for publication in *IEEE Transactions on Automatic Control*, Apr. 1999.

[47] L.M. Silverman, *Structural Properties of Time-Variable Linear Systems*. PhD thesis, Dept. of Electrical Engineering, Columbia University, New York, 1966.

[48] L.M. Silverman, "Realization of linear dynamical systems," *IEEE Transactions on Automatic Control*, vol. 16, no. 6, Dec. 1971.

[49] L.M. Silverman and H.E. Meadows, "Equivalence and synthesis of time-variable linear systems," in *Proceedings of the 4th Annual Allerton Conference on Circuit and System Theory* (Monticello, Illinois, Oct. 1966) (W.R. Perkins and J.B. Cruz, jr., eds.), pp. 776–784, 1965.

[50] R.E. Skelton, *Dynamic Systems Control*. New York: John Wiley & Sons, 1988.

[51] T. Söderström and P. Stoica, *System Identification*. London: Prentice-Hall, 1989.

[52] J. Sreedhar, P. Van Dooren, and P. Misra, "Minimal order time invariant representation of periodic descriptor systems," in *Proceedings of the 1999 American Control Conference*, San Diego, California, pp. 1309–1313, June 1999.

[53] H.J. Sussmann, "Existence and uniqueness of minimal realizations of nonlinear systems," *Mathematical Systems Theory*, vol. 10, no. 3, pp. 263–284, 1977.

[54] A.J. Tether, "Construction of minimal linear state-variable models from finite input-output data," *IEEE Transactions on Automatic Control*, vol. 17, no. 4, pp. 427–436, Aug. 1970.

[55] J.M. van den Hof, *System Theory and System Identification of Compartmental Systems*. PhD thesis, Faculty of Mathematics and Natural Sciences, University of Groningen, Groningen, The Netherlands, Nov. 1996.

[56] P.M. Van Dooren, "The generalized eigenstructure on linear system theory," *IEEE Transactions on Automatic Control*, vol. 26, no. 1, pp. 111–129, Feb. 1981.

[57] J.B. van Helmont, A.J.J. van der Weiden, and H. Anneveld, "Design of optimal controllers for a coal fired Benson boiler based on a modified approximate realization algorithm," in *Application of Multivariable System Techniques* (R. Whalley, ed.), London, pp. 313–320, Elsevier, 1990.

[58] A. Varga, "Computation of minimal realizations of periodic systems," in *Proceedings of the 37th IEEE Conference on Decision and Control*, Tampa, Florida, pp. 3825–3830, Dec. 1998.

[59] Y. Wang and E.D. Sontag, "Orders of input/output differential equations and state space dimensions," *SIAM Journal on Control and Optimization*, vol. 33, no. 4, pp. 1102–1127, July 1995.

[60] L. Wuytack, "Padé approximation." Invited paper for volume 1 on Approximation Theory of the NA20 project of *Journal of Computational and Applied Mathematics*, 1999.

[61] D.C. Youla and P. Tissi, "n-port synthesis via reactance extraction – Part I," *IEEE International Convention Record*, vol. 14, pt. 7, pp. 183–205, 1966.

[62] H.P. Zeiger and A.J. McEwen, "Approximate linear realizations of given dimension via Ho's algorithm," *IEEE Transactions on Automatic Control*, vol. 19, no. 2, p. 153, Apr. 1974.