Thoughts on System Identification

Jan C. Willems

SISTA-ESAT, K.U. Leuven, B-3001 Leuven, Belgium
Jan.Willems@esat.kuleuven.be

1 Introduction

It is a pleasure to dedicate this Festschrift article to Keith Glover on the occasion of his 60-th birthday. Keith came to MIT in 1969, armed with a degree from Imperial College, a couple of years of industrial experience, and a Kennedy scholarship, ready for some serious research. I, at that time a young assistant professor at MIT, had the good fortune to become Keith's M.Sc. and Ph.D. supervisor, who he could choose freely because his scholarship made him, so to speak, 'self-supporting'.

In his M.Sc. thesis Keith applied circle-criterion based stability conditions to numerical integration routines [9]. His Ph.D. dissertation [10] was about System Identification (SYSID), more specifically, about Identifiability. The main results of his dissertation were published in [11]. It was one of the first articles that dealt with the parametrization of linear systems, an intricate subject that was studied very deeply later on. This research reached its apogee in Hannan and Deistler's book [15] where the relevance for SYSID algorithms of parametrizations and canonical forms of multivariable linear systems is pursued in great depth.

Keith's dissertation showed a penetrating understanding of linear systems. It was a forebode of his later work in this area, culminating in his classic papers on model reduction [12] and on the double Riccati equation solution of the $\mathscr{H}_\infty$ feedback control problem [7], both among the most cited in the field of systems and control.

A Festschrift is a welcome occasion to write an article with a personal and historical flavor. Because of the occasion, I chose the subject of Keith Glover's Ph.D. dissertation, SYSID. My interest in this area remained originally limited to the implications of the structure of linear systems. This situation changed with the *Automatica* papers [35]. These contain, in addition to the first comprehensive exposition of the behavioral approach to systems theory, a number of new ideas and subspace type-algorithms for SYSID. The aim of the present article is to explain in a somewhat informal style my own personal point of view on SYSID. Among other things, I will describe in some detail the many representations of linear time-invariant systems, leading up to some exact deterministic SYSID algorithms based on the notion of the

most powerful unfalsified model. I will then explain the idea behind subspace algorithms, where the state trajectory is constructed directly from the observations and a system model in state form is deduced from there. Subsequently, I will discuss the role of latent variables in SYSID. This leads in a natural way to stochastic models. I will finish with some remarks on the rationale, or lack of it, of viewing SYSID in a stochastic framework.

I view SYSID in terms of the following picture from classical statistics:

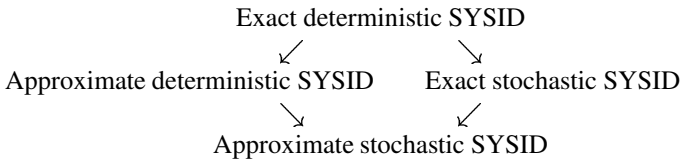
$$\text{ID}: \ \mathbb{D} \in \mathscr{D} \ \ \mapsto \ \ \mathbb{M} \in \mathscr{M}. \tag{ID}$$

Here $\mathbb{D}$ is the *data*, $\mathscr{D}$ the *data class*, $\mathscr{M}$ the *model class*, ID the *identification procedure*, and $\mathbb{M} = \text{ID}(\mathbb{D})$ the model chosen from $\mathscr{M}$ on the basis of the data $\mathbb{D}$ through the identification procedure ID. Our interest is in the case that the data consists of a time-series (throughout, time-series means *vector* time-series)

$$\tilde{w}_{\mathtt{T}} \ = \ \big(\tilde{w}(1), \tilde{w}(2), \ldots, \tilde{w}(\mathtt{T})\big) \qquad \tilde{w}(\mathtt{t}) \in \mathbb{R}^{\mathtt{w}} \tag{$\tilde{w}_{\mathtt{T}}$}$$

i.e. $\mathbb{D} \in (\mathbb{R}^{\mathtt{w}})^{\mathtt{T}}$. However, while our primary aim is at *finite* time-series $\mathtt{T} < \infty$, we will occasionally consider $\mathtt{T} \to \infty$ and even $\mathtt{T} = \infty$. So, we may as well take for the data class $\mathscr{D} = \cup_{\mathtt{T} \in \mathbb{N} \cup \{\infty\}} (\mathbb{R}^{\mathtt{w}})^{\mathtt{T}}$. The model class $\mathscr{M}$ consists of linear time-invariant, deterministic or stochastic, dynamical systems. The intent of this essay is to explain some of the issues involved in defining and choosing $\mathscr{M}$, and setting up an identification procedure (ID).

I prefer to deal with SYSID by going from the simple to the complex:



In view of space limitations, I will not discuss the approximate stochastic case.

2 Deterministic models

Accordingly, I will start with deterministic systems.

The first question which we need to confront is: What are we after? What does an identification procedure aim at? *When we accept a mathematical model as a description of a phenomenon, what do we really do?* Mind you, I am not asking a philosophical question. I am not inviting an erudite discourse on the convoluted relation between a real world manifestation of a physical or an economic phenomenon and a mathematical description of it. The question is meant to be a mathematical one.

The answer to this question is simple, evident, but startlingly enlightening (in German, a *'Aha Erlebnis'*). A model is simply a subset, $\mathscr{B}$, of a universe, $\mathscr{U}$, of a priori possibilities. This subset $\mathscr{B} \subseteq \mathscr{U}$ is called the *behavior* of the model. Thus, before the phenomenon was captured in a model, all outcomes from $\mathscr{U}$ were in principle possible. But after we accept $\mathscr{B}$ as the model, we declare that only outcomes

from $\mathscr{B}$ are possible. For example, the ideal gas law states that the temperature T, pressure P, volume V, and quantity (number of moles) N of an ideal gas satisfy $PV/NT = R$, with R a universal constant. So, before Boyle, Charles, and Avogadro got into the act, T, P, V, and N may have seemed unrelated, yielding $\mathscr{U} = \mathbb{R}_+^4$. The ideal gas law restricts the possibilities to $\mathscr{B} = \{(T,P,V,N) \in \mathbb{R}_+^4 \mid PV/NT = R\}$.

In most applications, a model is given in terms of a set of equations (think of the basic laws of physics, e.g. Maxwell's equations). The set of solutions to these equations is then the behavior. In the context of modeling, equations simply serve as a representation of their solution set. This view comes in very useful, since it gives a clear and unambiguous answer to the question when two sets of equations represent the same model. Often, it is trivial that a transformation of equations (as changing the order in which Maxwell's equations are written down) does not change a model, but for some transformations this may be much more difficult to establish (think of expressing Maxwell's equations in terms of potential functions).

In the case of *dynamical* systems, the phenomenon which is being modelled produces functions that map the set of time instances relevant to the model to the *signal space*. This is the space in which these functions take on their values. For the sake of concreteness, and because of its relevance to SYSID, I will assume in this article that the set of relevant time instances is $\mathbb{N}$ (the theory is analogous for $\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{R}_+$). This choice means that we are studying discrete time systems and that we postulate that the model is valid for all time (of course, this does not mean that observations need to extend over all of $\mathbb{N}$). I also assume throughout that the signal space is a finite dimensional real vector space, typically $\mathbb{R}^{\mathtt{w}}$.

Following our idea of a model, the behavior for the dynamical systems which we consider is therefore a collection $\mathscr{B}$ of functions mapping the time set $\mathbb{N}$ into the signal space $\mathbb{R}^{\mathtt{w}}$. A dynamical model can hence be identified with its *behavior* $\mathscr{B} \subseteq (\mathbb{R}^{\mathtt{w}})^{\mathbb{N}}$. The behavior is hence a family of maps from $\mathbb{N}$ to $\mathbb{R}^{\mathtt{w}}$. Of course, also for dynamical systems the behavior $\mathscr{B}$ is usually specified as the set of solutions of equations, for the case at hand typically difference equations. As dynamical models, difference equations thus merely serve as a representation of their solution set. Note that this immediately leads to a notion of equivalence and to canonical forms for difference equations. These are particularly relevant in the context of dynamical systems, because of the multitude of, usually over-parameterized, representations of the behavior of a dynamical system.

3 Deterministic linear dynamical systems

SYSID usually employs as the model class dynamical systems that are (i) linear, (ii) time-invariant, and (iii) that satisfy a third property, related to the finite dimensionality of the underlying state space, or to the rationality of a transfer function. It is, however, clearer and advantageous to approach this situation in a more intrinsic way, by imposing this third property directly on the behavior, and not on a representation of it. In this section, I discuss this model class in considerable detail.

A behavior $\mathscr{B} \subseteq (\mathbb{R}^{\mathtt{w}})^{\mathbb{N}}$ is said to be *linear* if $w \in \mathscr{B}, w' \in \mathscr{B}$, and $\alpha \in \mathbb{R}$ imply $w + w' \in \mathscr{B}$ and $\alpha w \in \mathscr{B}$, and *time-invariant* if $\sigma\mathscr{B} \subseteq \mathscr{B}$. The *shift*, σ, plays a central role in SYSID. It is defined by $(\sigma f)(\mathtt{t}) := f(\mathtt{t}+1)$, and often called the *backwards* shift, since it shifts the time function f backwards. The third property that enters into the specification of the model class is completeness. $\mathscr{B}$ is called *complete* if it has the following property:

$$[\![w : \mathbb{N} \to \mathbb{R}^{\mathtt{w}} \text{ belongs to } \mathscr{B}]\!] \Leftrightarrow [\![w|_{[1,\mathtt{t}]} \in \mathscr{B}|_{[1,\mathtt{t}]} \text{ for all } \mathtt{t} \in \mathbb{N}]\!].$$

In words: $\mathscr{B}$ is complete if we can decide that $w : \mathbb{N} \to \mathbb{R}^{\mathtt{w}}$ is 'legal' (i.e. belongs to $\mathscr{B}$) by verifying that everyone of its 'prefixes' $\big(w(1), w(2), \ldots, w(\mathtt{t})\big)$ is 'legal' (i.e. belongs to $\mathscr{B}|_{[1,\mathtt{t}]}$). So, roughly speaking, $\mathscr{B}$ is complete iff the laws of $\mathscr{B}$ do not involve what happens at ∞. Requirements as $w \in \ell_2(\mathbb{N}, \mathbb{R}^{\mathtt{w}})$, w has compact support, or $\lim_{\mathtt{t}\to\infty} w(\mathtt{t})$ exists, risk at obstructing completeness. However, often crucial information about a complete $\mathscr{B}$ can be obtained by considering its intersection with $\ell_2(\mathbb{N}, \mathbb{R}^{\mathtt{w}})$, or its compact support elements, etc.

Recall the following standard notation. $\mathbb{R}[\xi]$ denotes the polynomials with real coefficients in the indeterminate ξ, $\mathbb{R}(\xi)$ the real rational functions, and $\mathbb{R}^{\mathtt{n}_1 \times \mathtt{n}_2}[\xi]$ the polynomial matrices with real $\mathtt{n}_1 \times \mathtt{n}_2$ matrices as coefficients. When the number of rows is irrelevant and the number of columns is $\mathtt{n}$, the notation $\mathbb{R}^{\bullet \times \mathtt{n}}[\xi]$ is used. So, in effect, $\mathbb{R}^{\bullet \times \mathtt{n}}[\xi] = \cup_{\mathtt{k} \in \mathbb{N}} \mathbb{R}^{\mathtt{k} \times \mathtt{n}}[\xi]$. A similar notation is used for polynomial vectors, or when the number of rows and/or columns is irrelevant. The degree of $P \in \mathbb{R}^{\bullet \times \bullet}[\xi]$ equals the largest degree of its entries, and is denoted by $\deg(P)$. Given a time-series $w : \mathbb{N} \to \mathbb{R}^{\mathtt{w}}$ and a polynomial matrix $R \in \mathbb{R}^{\mathtt{v} \times \mathtt{w}}[\xi]$, say $R(\xi) = R_0 + R_1\xi + \cdots + R_{\mathtt{L}}\xi^{\mathtt{L}}$, we can form the new $\mathtt{v}$-dimensional time-series

$$R(\sigma)w = R_0 w + R_1 \sigma w + \cdots + R_{\mathtt{L}}\sigma^{\mathtt{L}} w.$$

Hence $R(\sigma) : (\mathbb{R}^{\mathtt{w}})^{\mathbb{N}} \to (\mathbb{R}^{\mathtt{v}})^{\mathbb{N}}$, with $R(\sigma)w : \mathtt{t} \in \mathbb{N} \mapsto R_0 w(\mathtt{t}) + R_1 w(\mathtt{t}+1) + \cdots + R_{\mathtt{L}} w(\mathtt{t}+\mathtt{L}) \in \mathbb{R}^{\mathtt{v}}$.

The combination of linearity, time-invariance, and completeness can be expressed in very many equivalent ways. In particular, the following are equivalent:

1. $\mathscr{B} \subseteq (\mathbb{R}^{\mathtt{w}})^{\mathbb{N}}$ is linear, time-invariant, and complete;
2. $\mathscr{B}$ is a linear, shift invariant ($:\Leftrightarrow \sigma\mathscr{B} \subseteq \mathscr{B}$), closed subset of $(\mathbb{R}^{\mathtt{w}})^{\mathbb{N}}$, with 'closed' understood in the topology of pointwise convergence;
3. $\exists\, R \in \mathbb{R}^{\bullet \times \mathtt{w}}[\xi]$ such that $\mathscr{B}$ consists of the solutions $w : \mathbb{N} \to \mathbb{R}^{\mathtt{w}}$ of

$$R(\sigma)w = 0. \tag{$\mathfrak{Ker}$}$$

The set of behaviors $\mathscr{B} \subseteq (\mathbb{R}^{\mathtt{w}})^{\mathbb{N}}$ that satisfy the equivalent conditions 1. to 3. is denoted by $\mathscr{L}^{\mathtt{w}}$, or, when the number of variables is unspecified, by $\mathscr{L}^{\bullet}$. Thus, in effect, $\mathscr{L}^{\bullet} = \cup_{\mathtt{w} \in \mathbb{N}} \mathscr{L}^{\mathtt{w}}$. Since $\mathscr{B} = \ker(R(\sigma))$ in ($\mathfrak{Ker}$), we call ($\mathfrak{Ker}$) a *kernel representation* of the behavior $\mathscr{B}$. We will meet other representations later. But first we introduce a characterization that is mathematically more abstract, but very pertinent in the context of SYSID. It identifies a behavior $\mathscr{B} \in \mathscr{L}^{\bullet}$ with an $\mathbb{R}[\xi]$-module.

Consider $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$. The polynomial vector $n \in \mathbb{R}^{1\times \mathtt{w}}[\xi]$ is called an *annihilator* (or a *consequence*) of $\mathscr{B}$ if $n(\sigma)\mathscr{B} = 0$, i.e. if $n(\sigma)w = 0$ for all $w \in \mathscr{B}$. Denote by $\mathscr{N}_{\mathscr{B}}$ the set of annihilators of $\mathscr{B}$. Observe that $\mathscr{N}_{\mathscr{B}}$ is an $\mathbb{R}[\xi]$-module. Indeed, $n \in \mathscr{N}_{\mathscr{B}}, n' \in \mathscr{N}_{\mathscr{B}}$, and $\alpha \in \mathbb{R}[\xi]$ imply $n + n' \in \mathscr{N}_{\mathscr{B}}$ and $\alpha n \in \mathscr{N}_{\mathscr{B}}$. Hence the map $\mathscr{B} \mapsto \mathscr{N}_{\mathscr{B}}$ associates with each $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$ a submodule of $\mathbb{R}^{1\times \mathtt{w}}[\xi]$. It turns out that this map is actually a bijection, i.e. to each submodule of $\mathbb{R}^{1\times \mathtt{w}}[\xi]$, there corresponds exactly one element of $\mathscr{L}^{\mathtt{w}}$. It is easy to see what the inverse map is. Let $\mathscr{K}$ be a submodule of $\mathbb{R}^{1\times \mathtt{w}}[\xi]$. Submodules of $\mathbb{R}^{1\times \mathtt{w}}[\xi]$ have nice properties. In particular, they are *finitely generated*, meaning that there exist elements (*'generators'*) $g_1, g_2, \ldots, g_{\mathtt{g}} \in \mathscr{K}$ such that $\mathscr{K}$ consists precisely of the linear combinations $\alpha_1 g_1 + \alpha_2 g_2 + \cdots + \alpha_{\mathtt{g}} g_{\mathtt{g}}$ where the $\alpha_{\mathtt{k}}$'s range over $\mathbb{R}[\xi]$. Now consider the system $(\mathfrak{Ker})$ with $R = \mathrm{col}(g_1, g_2, \ldots, g_{\mathtt{g}})$ and prove that

$$\mathscr{N}_{\ker\left(\mathrm{col}\left(g_1, g_2, \ldots, g_{\mathtt{g}}\right)(\sigma)\right)} = \mathscr{K}$$

($\supseteq$ is obvious, $\subseteq$ requires a bit of analysis). In terms of $(\mathfrak{Ker})$, we obtain the characterization

$$[\![\ \ker\left(R(\sigma)\right) = \mathscr{B}\ ]\!] \quad \Leftrightarrow \quad [\![\mathscr{N}_{\mathscr{B}} = \ \langle R \rangle\ ]\!]$$

where $\langle R \rangle$ denotes the $\mathbb{R}[\xi]$-module generated by the rows of R.

The observation that there is a bijective correspondence between $\mathscr{L}^{\mathtt{w}}$ and the $\mathbb{R}[\xi]$-submodules of $\mathbb{R}^{1\times \mathtt{w}}[\xi]$ is not altogether trivial. For instance, the surjectivity of the map

$$\mathscr{B} = \ker\left(R(\sigma)\right) \in \mathscr{L}^{\mathtt{w}} \ \ \mapsto \ \ \mathscr{N}_{\mathscr{B}} = \langle R \rangle$$

onto the $\mathbb{R}[\xi]$-submodules of $\mathbb{R}^{1\times \mathtt{w}}[\xi]$ depends on the solution concept used in $(\mathfrak{Ker})$. If we would have considered only solutions with compact support, or that are square integrable, this bijective correspondence is lost. Equations, in particular difference or differential equations, all by themselves, without a clear solution concept, i.e. without a definition of the corresponding behavior, are an inadequate specification of a mathematical model. Expressed otherwise: the theory of (linear) systems is not the domain of pure algebra. Analysis enters through the solution concept involved in the difference or differential equations, and guides the subsequent algebraic structure.

The characterization of $\mathscr{B}$ in terms of its module of annihilators is very useful in the context of deterministic SYSID. It shows precisely what we are looking for in order to identify a system in the model class $\mathscr{L}^{\bullet}$: (a set of generators of) the submodule $\mathscr{N}_{\mathscr{B}}$.

Behaviors in $\mathscr{L}^{\bullet}$ admit many other representations. The following two are exceedingly familiar to system theorists. In fact,

4. $[\![\mathscr{B} \in \mathscr{L}^{\mathtt{w}}]\!] \Leftrightarrow [\![\ \exists$ integers $\mathtt{m}, \mathtt{p} \in \mathbb{Z}_+$, with $\mathtt{m} + \mathtt{p} = \mathtt{w}$, polynomial matrices $P \in \mathbb{R}^{\mathtt{p}\times\mathtt{p}}[\xi], Q \in \mathbb{R}^{\mathtt{p}\times\mathtt{m}}[\xi]$, with $\det(P) \neq 0$, and a permutation matrix $\Pi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}$ such that $\mathscr{B}$ consists of all $w : \mathbb{N} \to \mathbb{R}^{\mathtt{w}}$ for which there exist $u : \mathbb{N} \to \mathbb{R}^{\mathtt{m}}$ and $y : \mathbb{N} \to \mathbb{R}^{\mathtt{p}}$ such that
$$P(\sigma)y = Q(\sigma)u \qquad \text{(I/O)}$$

and $w = \Pi \begin{bmatrix} u \\ y \end{bmatrix}$]]. The matrix of rational functions $G = P^{-1}Q \in (\mathbb{R}(\xi))^{\mathtt{p}\times\mathtt{m}}$ is called the *transfer function* of (I/O). Actually, for a given $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$, it is always possible to choose Π such that G is proper. If we would allow a basis change in $\mathbb{R}^{\mathtt{w}}$, i.e. allow any non-singular matrix for Π (instead of only a permutation matrix), then we could always take G to be strictly proper.

5. $[\![\mathscr{B} \in \mathscr{L}^{\mathtt{w}}]\!] \Leftrightarrow$ [[$\exists$ integers $\mathtt{m}, \mathtt{p}, \mathtt{n} \in \mathbb{Z}_+$ with $\mathtt{m} + \mathtt{p} = \mathtt{w}$, matrices $A \in \mathbb{R}^{\mathtt{n}\times\mathtt{n}}, B \in \mathbb{R}^{\mathtt{n}\times\mathtt{m}}, C \in \mathbb{R}^{\mathtt{p}\times\mathtt{n}}, D \in \mathbb{R}^{\mathtt{p}\times\mathtt{m}}$, and a permutation matrix $\Pi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}$ such that $\mathscr{B}$ consists of all $w : \mathbb{N} \to \mathbb{R}^{\mathtt{w}}$ for which there exist $u : \mathbb{N} \to \mathbb{R}^{\mathtt{m}}$, $x : \mathbb{N} \to \mathbb{R}^{\mathtt{n}}$, and $y : \mathbb{N} \to \mathbb{R}^{\mathtt{p}}$ such that

$$\sigma x = Ax + Bu, \quad y = Cx + Du \tag{$\mathfrak{S}$}$$

and $w = \Pi \begin{bmatrix} u \\ y \end{bmatrix}$]]. If we would allow a basis change in $\mathbb{R}^{\mathtt{w}}$, i.e. allow any non-singular matrix for Π, then we could always take $D = 0$.

(I/O) is called an *input/output* (i/o) and ($\mathfrak{S}$) an *input/state/output* (i/s/o) representation of the corresponding behavior $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$.

Why, if any element $\mathscr{B} \in \mathscr{L}^{\bullet}$ indeed admits a representation (I/O) or ($\mathfrak{S}$), should one not use one of these familiar representations ab initio? There are many good reasons for not doing so. To begin with, and most importantly, first principles models aim at describing a behavior, but are seldom in the form (I/O) or ($\mathfrak{S}$). Consequently, one must have a theory that supersedes (I/O) or ($\mathfrak{S}$) in order to have a clear idea what transformations are allowed in bringing a first principles model into the form (I/O) or ($\mathfrak{S}$). Secondly, as a rule, physical systems are simply not endowed with a signal flow direction. Adding a signal flow direction is often a figment of one's imagination, and when something is not real, it will turn out to be cumbersome sooner or later. A third reason, very much related to the second, is that the input/output framework is totally inappropriate for dealing with all but the most special system interconnections. We are surrounded by interconnected systems, but only very sparingly can these be viewed as input-to-output connections. Fourthly, the structure implied by (I/O) or ($\mathfrak{S}$) often needlessly complicates matters, mathematically and conceptually. A good theory of systems takes the behavior as the basic notion and the reference point for concepts and definitions, and switches back and forth between a wide variety of convenient representations. (I/O) or ($\mathfrak{S}$) have useful properties, but for many purposes other representations may be more convenient. For example, a kernel representation ($\mathfrak{Ker}$) is very relevant in SYSID. It suggests that we should look for (approximate) annihilators. On the other hand, when it comes to constructing trajectories, ($\mathfrak{S}$) is very convenient. It shows how trajectories are parameterized and generated : by the initial state $x(1) \in \mathbb{R}^{\mathtt{n}}$ and the input $u : \mathbb{N} \to \mathbb{R}^{\mathtt{m}}$.

Our next representation involves rational functions and is a bit more 'tricky'. Let $G \in (\mathbb{R}(\xi))^{\bullet\times\mathtt{w}}$ and consider the system of 'difference equations'

$$G(\sigma)w = 0. \tag{$\mathfrak{G}$}$$

What is meant by the behavior of ($\mathfrak{G}$)? Since G is a matrix of rational functions, it is not evident how to define solutions. This may be done in terms of co-prime

factorizations, as follows. G can be factored $G = P^{-1}Q$ with $P \in \mathbb{R}^{\bullet\times\bullet}[\xi]$ square, $\det(P) \neq 0, Q \in \mathbb{R}^{\bullet\times \mathtt{w}}[\xi]$ and (P,Q) left co-prime (meaning that $F = [P \;\; Q]$ is left prime, i.e.

$$[\![(U, F' \in \mathbb{R}^{\bullet\times\bullet}[\xi]) \wedge (F = UF')]\!] \Rightarrow [\![U \text{ is square and unimodular}]\!],$$

equivalently $\exists\, H \in \mathbb{R}^{\bullet\times\bullet}[\xi]$ such that $FH = I$). We *define* the behavior of $(\mathfrak{G})$ as that of

$$Q(\sigma)w = 0, \quad \text{i.e. as} \quad \ker(Q(\sigma))$$

Hence $(\mathfrak{G})$ defines a behavior $\in \mathscr{L}^{\mathtt{w}}$. It is easy to see that this definition is independent of which co-prime factorization is taken. There are other reasonable ways of approaching the problem of defining the behavior of $(\mathfrak{G})$, but they all turn out to be equivalent to the definition given. Note that, in a trivial way, since $(\mathfrak{Ker})$ is a special case of $(\mathfrak{G})$, every element of $\mathscr{L}^{\mathtt{w}}$ admits a representation $(\mathfrak{G})$.

Certain integer 'invariants' (meaning maps from $\mathscr{L}^{\bullet}$ to $\mathbb{Z}_+$) associated with systems in $\mathscr{L}^{\bullet}$ are important in SYSID. One is the *lag*, denoted by $\mathsf{L}(\mathscr{B})$, defined as the smallest $\mathtt{L} \in \mathbb{Z}_+$ such that $[\![w|_{[\mathtt{t},\mathtt{t}+\mathtt{L}]} \in \mathscr{B}|_{[\mathtt{1},\mathtt{t}+\mathtt{1}]} \text{ for all } \mathtt{t} \in \mathbb{N}]\!] \Rightarrow [\![w \in \mathscr{B}]\!]$. Equivalently, the smallest degree over the polynomial matrices R such that $\mathscr{B} = \ker(R(\sigma))$. A second integer invariant that is important is the *input cardinality*, denoted by $\mathsf{m}(\mathscr{B})$, defined as $\mathtt{m}$, the number of input variables in any (I/O) representation of $\mathscr{B}$. It turns out that $\mathtt{m}$ is an invariant (while the input/output partition, i.e. the permutation matrix Π in (I/O), is not). The number of output variables, $\mathtt{p}$, yields the *output cardinality* $\mathsf{p}(\mathscr{B})$. A third important integer invariant is the *state cardinality*, $\mathsf{n}(\mathscr{B})$, defined as the smallest number $\mathtt{n}$ of state variables over all i/s/o representations $(\mathfrak{S})$ of $\mathscr{B}$. The three integer invariants $\mathsf{m}(\mathscr{B})$, $\mathsf{n}(\mathscr{B})$, and $\mathsf{L}(\mathscr{B})$ can be nicely captured in one single formula, involving the growth as a function of $\mathtt{t}$ of the dimension of the subspace $\mathscr{B}|_{[\mathtt{1},\mathtt{t}]}$. Indeed, there holds

$$\dim(\mathscr{B}|_{[\mathtt{1},\mathtt{t}]}) \leq \mathsf{m}(\mathscr{B})\,\mathtt{t} + \mathsf{n}(\mathscr{B}) \quad \text{with equality iff} \quad \mathtt{t} \geq \mathsf{L}(\mathscr{B}).$$

State models $(\mathfrak{S})$ are an example of the more general, but very useful, class of latent variable models. Such models involve, in addition to the *manifest* variables (denoted by w in $(\mathfrak{L})$), the variables which the model aims at, also auxiliary, *latent* variables (denoted by ℓ in $(\mathfrak{L})$). For the case at hand this leads to behaviors $\mathscr{B}_{\text{full}} \in \mathscr{L}^{\mathtt{w}+\mathtt{l}}$ described by

$$R(\sigma)w = M(\sigma)\ell, \tag{$\mathfrak{L}$}$$

with $R \in \mathbb{R}^{\bullet\times\mathtt{w}}[\xi]$ and $M \in \mathbb{R}^{\bullet\times\mathtt{l}}[\xi]$.

Although the notion of observability applies more generally, we use it here for latent variable models only. We call $\mathscr{B}_{\text{full}} \in \mathscr{L}^{\mathtt{w}+\mathtt{l}}$ *observable* if

$$[\![(w,\ell_1) \in \mathscr{B}_{\text{full}} \text{ and } (w,\ell_2) \in \mathscr{B}_{\text{full}}]\!] \Rightarrow [\![\ell_1 = \ell_2]\!].$$

$(\mathfrak{L})$ defines an observable latent variable system iff $M(\lambda)$ has full row rank for all $\lambda \in \mathbb{C}$. For state systems (with x the latent variable), this corresponds to the usual observability of the pair (A,C).

An important result, the *elimination theorem*, states that $\mathscr{L}^{\bullet}$ is closed under projection. Hence $\mathscr{B}_{\text{full}} \in \mathscr{L}^{\mathtt{w}+\mathtt{l}}$ implies that the *manifest* behavior

$$\mathscr{B} = \{w : \mathbb{N} \to \mathbb{R}^{\mathtt{w}} \mid \exists\, \ell : \mathbb{N} \to \mathbb{R}^{\mathtt{l}} \text{ such that } (\mathfrak{L}) \text{ holds}\}$$

belongs to $\mathscr{L}^{\mathtt{w}}$, and therefore admits a kernel representation ($\mathfrak{Ker}$) of its own. So, in a trivial sense, ($\mathfrak{L}$) is yet another representation of $\mathscr{L}^{\mathtt{w}}$.

Latent variable representations (also unobservable ones) are very useful in all kinds of applications. This, notwithstanding the elimination theorem. They are the end result of modeling interconnected systems by *tearing and zooming*, with the interconnection variables viewed as latent variables. Many physical models (for example, in mechanics) express basic laws using latent variables. In the context of SYSID, the aim of most classical algorithms is in fact to arrive at a model ($\mathfrak{L}$), often unobservable, and with ℓ usually interpreted as an unobserved (stochastic) input – I will return to this later. But, of course, we are all most familiar with state models, states being the latent variables *par excellence*. In the next section we will see how latent variables can be used to express controllability.

4 Controllability

As in many areas of system theory, controllability often enters in SYSID as a regularizing assumption. In the behavioral theory, an appealing notion of controllability has been put forward. It expresses what is needed intuitively, it applies to any dynamical system, regardless of its representation, it has the classical state transfer definition as a special case, and it is readily generalized, for instance to distributed systems. It is somewhat strange that this definition has not been generally adopted. Adapted to the case at hand, it reads as follows. The time-invariant behavior $\mathscr{B} \subseteq (\mathbb{R}^{\bullet})^{\mathbb{N}}$ is said to be *controllable* if for any $w_1 \in \mathscr{B}$, $w_2 \in \mathscr{B}$, and $\mathtt{t}_1 \in \mathbb{N}$, there exists a $\mathtt{t}_2 \in \mathbb{N}$ and a $w \in \mathscr{B}$ such that $w(\mathtt{t}) = w_1(\mathtt{t})$ for $1 \leq \mathtt{t} \leq \mathtt{t}_1$, and $w(\mathtt{t}) = w_2(\mathtt{t} - \mathtt{t}_1 - \mathtt{t}_2)$ for $\mathtt{t} > \mathtt{t}_1 + \mathtt{t}_2$. For $\mathscr{B} \in \mathscr{L}^{\bullet}$, one can take without loss of generality $w_1 = 0$ in the above definition. The property just defined is hence akin to what is sometimes called *reachability*, but here I use the term controllability as synonymous. Denote the controllable elements of $\mathscr{L}^{\bullet}$ by $\mathscr{L}^{\bullet}_{\text{controllable}}$ and of $\mathscr{L}^{\mathtt{w}}$ by $\mathscr{L}^{\mathtt{w}}_{\text{controllable}}$.

($\mathfrak{Ker}$) defines a controllable system iff $R(\lambda)$ has the same rank for each $\lambda \in \mathbb{C}$. There is a very nice representation result that characterizes controllability: it is equivalent to the existence of an image representation. More precisely, $\mathscr{B} \in \mathscr{L}^{\bullet}_{\text{controllable}}$ iff there exists $M \in \mathbb{R}^{\bullet \times \bullet}[\xi]$ such that $\mathscr{B}$ equals the manifest behavior of the latent variable system

$$w = M(\sigma)\ell. \qquad (\mathfrak{Im})$$

In other words, iff $\mathscr{B} = \text{im}(M(\sigma))$. So, images, contrary to kernels, are always controllable. This image representation of a controllable system can always be taken to be observable.

For $\mathscr{B} \in \mathscr{L}^{\bullet}$, we define its *controllable part*, denoted by $\mathscr{B}_{\text{controllable}}$, as

$$\mathscr{B}_{\text{controllable}} := \{w \in \mathscr{B} \mid \ \forall \mathtt{t}' \in \mathbb{N}, \exists \mathtt{t}'' \in \mathbb{Z}_+, \text{ and } w' \in \mathscr{B} \text{ such that}$$
$$w'(t) = 0 \text{ for } 1 \leq \mathtt{t} \leq \mathtt{t}' \text{ and } w'(t) = w(t - t' - t'') \text{ for } \mathtt{t} > \mathtt{t}' + \mathtt{t}''\}.$$

Equivalently, $\mathscr{B}_{\text{controllable}}$ is the largest controllable subsystem contained in $\mathscr{B}$. It turns out that two systems of the form (I/O) (with the same input/output partition) have the same transfer function iff they have the same controllable part.

Consider $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$. The vector of rational functions $n \in \mathbb{R}^{1\times \mathtt{w}}(\xi)$ is called a *rational annihilator* of $\mathscr{B}$ if $n(\sigma)\mathscr{B} = 0$ (note that, since we gave a meaning to $(\mathfrak{G})$, this is well defined). Denote by $\mathscr{N}^{\text{rational}}_{\mathscr{B}}$ the set of rational annihilators of $\mathscr{B}$. Observe that $\mathscr{N}^{\text{rational}}_{\mathscr{B}}$ is a $\mathbb{R}(\xi)$-subspace of $\mathbb{R}^{1\times \mathtt{w}}(\xi)$. The map $\mathscr{B} \mapsto \mathscr{N}^{\text{rational}}_{\mathscr{B}}$ is not a bijection from $\mathscr{L}^{\mathtt{w}}$ to the $\mathbb{R}(\xi)$-subspaces of $\mathbb{R}^{1\times \mathtt{w}}(\xi)$. Indeed,

$$[\![\, \mathscr{N}^{\text{rational}}_{\mathscr{B}'} = \mathscr{N}^{\text{rational}}_{\mathscr{B}''} \,]\!] \Leftrightarrow [\![\, \mathscr{B}'_{\text{controllable}} = \mathscr{B}''_{\text{controllable}} \,]\!].$$

In fact, there exists a bijective correspondence between $\mathscr{L}^{\mathtt{w}}_{\text{controllable}}$ and the $\mathbb{R}(\xi)$-subspaces of $\mathbb{R}^{1\times \mathtt{w}}(\xi)$. Summarizing, $\mathbb{R}[\xi]$-submodules of $\mathbb{R}^{1\times \mathtt{w}}[\xi]$ stand in bijective correspondence with $\mathscr{L}^{\mathtt{w}}$, with each submodule corresponding to the set of polynomial annihilators, while $\mathbb{R}(\xi)$-subspaces of $\mathbb{R}^{1\times \mathtt{w}}(\xi)$ stand in bijective correspondence with $\mathscr{L}^{\mathtt{w}}_{\text{controllable}}$, with each subspace corresponding to the set of rational annihilators.

Controllability enters in a subtle way whenever a system is identified with its transfer function. Indeed, it is easy to prove that the system described by

$$w_2 = G(\sigma)w_1, \qquad w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \tag{$\mathfrak{T}$}$$

a special case of $(\mathfrak{G})$, is automatically controllable. This again shows the limitation of identifying a system with its transfer function. Two input/output systems (I/O) with the same transfer function are the same iff they are both controllable. In the end, transfer function thinking can deal with non-controllable systems only in contorted ways.

A property related to controllability is stabilizability. The behavior $\mathscr{B} \subseteq (\mathbb{R}^{\bullet})^{\mathbb{N}}$ is said to be *stabilizable* if for any $w \in \mathscr{B}$ and $\mathtt{t} \in \mathbb{N}$, there exists a $w' \in \mathscr{B}$ such that $w'(\mathtt{t}') = w(\mathtt{t}')$ for $1 \leq \mathtt{t}' \leq \mathtt{t}$, and $w'(\mathtt{t}) \to 0$ for $t \to \infty$. $(\mathfrak{Ker})$ defines a stabilizable system iff $R(\lambda)$ has the same rank for each $\lambda \in \mathbb{C}$ with $\text{Real}(\lambda) \geq 0$. An important system theoretic result (leading up to the Youla-Kučera parametrization of stabilizing controllers) states that $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$ is stabilizable iff it allows a representation $(\mathfrak{G})$ with $G \in (\mathbb{R}(\xi))^{\bullet\times \mathtt{w}}$ left prime over the ring $\mathbb{RH}_\infty$ $(:= \{f \in \mathbb{R}(\xi) \mid f$ is proper and has no poles in the closed right half of the complex plane $\})$. $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$ is controllable iff it allows a representation $w = G(\sigma)\ell$ with $G \in (\mathbb{R}(\xi))^{\mathtt{w}\times \bullet}$ right prime over the ring $\mathbb{RH}_\infty$.

Autonomous systems are on the other extreme of controllable ones. $\mathscr{B} \subseteq (\mathbb{R}^{\bullet})^{\mathbb{N}}$ is said to be *autonomous* if for every $w \in \mathscr{B}$, there exists a $\mathtt{t} \in \mathbb{N}$ such that $w|_{[1,\mathtt{t}]}$ uniquely specifies $w|_{[\mathtt{t}+1,\infty)}$, i.e. such that $w' \in \mathscr{B}$ and $w|_{[1,\mathtt{t}]} = w'|_{[1,\mathtt{t}]}$ imply $w' = w$. It can be shown that $\mathscr{B} \in \mathscr{L}^{\bullet}$ is autonomous iff it is finite dimensional. Autonomous systems and, more generally, uncontrollable systems are of utmost importance in

systems theory (including SYSID), in spite of much system theory folklore claiming the contrary. Controllability as a systems property is much more restrictive than is generally appreciated.

5 The MPUM

The SYSID algorithms which I will discuss associate with an observed time-series $\tilde{w}_{\mathtt{T}}$ an element of $\mathscr{L}^{\bullet}$, i.e. a complete linear time-invariant system. The first model class that comes to mind is $\mathscr{L}^{\mathtt{w}}$, i.e. models described by

$$R(\sigma)\, w = 0. \tag{$\mathfrak{Ker}$}$$

However, for reasons which I will go into more detail later, important flexibility is gained by aiming at a latent variable model

$$R(\sigma)w = M(\sigma)\ell, \tag{$\mathfrak{L}$}$$

with full behavior $\mathscr{B}_{\text{full}} \in \mathscr{L}^{\mathtt{w}+\mathtt{l}}$ and manifest behavior $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$. In this model we assume that the observed data is generated by a system that contains unobserved latent variables.

Assume that the model ($\mathfrak{L}$) is deduced from the observed data

$$\tilde{w}_{\mathtt{T}} \;=\; \big(\tilde{w}(\mathtt{1}), \tilde{w}(\mathtt{2}), \ldots, \tilde{w}(\mathtt{T})\big).$$

How should we assess this choice?

There are several conflicting quantitative measures which we wish to keep small, e.g. the complexity, the misfit, and the latency. The *complexity* roughly measures how many adjustable parameters the model has. The dimension of the subspaces $\dim\big(\mathscr{B}_{\text{full}}|_{[\mathtt{1},\mathtt{t}]}\big)$ for $\mathtt{t} \in \mathbb{N}$, or, more simply, the triple $\big(\mathsf{m}(\mathscr{B}_{\text{full}}), \mathsf{n}(\mathscr{B}_{\text{full}}), \mathsf{L}(\mathscr{B}_{\text{full}})\big)$ is a good measure of the complexity. The *misfit* measures the extent to which the model fails to explain the observations. The minimum of $||\tilde{w}_{\mathtt{T}} - w|_{[\mathtt{1},\mathtt{T}]}||_{\ell_2([\mathtt{1},\mathtt{T}],\mathbb{R}^{\mathtt{w}})}$ over $w \in \mathscr{B}$ is a good measure of the misfit. The *latency* measures the extent to which the model needs latent variables to explain the observations. The minimum of $||\ell|_{[\mathtt{1},\mathtt{T}]}||_{\ell_2(\mathbb{N},\mathbb{R}^{\mathtt{l}})}$ over ℓ such that $(w,\ell) \in \mathscr{B}_{\text{full}}$ and $w|_{[\mathtt{1},\mathtt{T}]} = \tilde{w}_{\mathtt{T}}$ is a good measure of the latency.

Another prevalent way for assessing an identification procedure is *consistency*. This means that if the data $\tilde{w}_{\mathtt{T}}$ is actually generated by an element of the model class, then the identification procedure should return the model that generated the data. Consistency is primarily used in a stochastic setting for infinite observed time series ($\mathtt{T} = \infty$), but one can apply it for deterministic SYSID and $\mathtt{T}$ finite as well. A related consideration for assessing an identification algorithm is the way the estimate based on $\tilde{w}_{\mathtt{T}}$ behaves as $T \to \infty$, if it is considered to be a truncation of an infinite time series.

In the remainder of this section, we discuss how we can associate a deterministic model ($\mathfrak{Ker}$) 'exactly' (with zero misfit) to an observed time-series. The idea is to

obtain the least complex model that explains the data exactly. The concept which implements this idea is the *Most Powerful Unfalsified Model* (MPUM). This notion was introduced in [35, Part II] for infinite observed time-series ($\mathtt{T} = \infty$ in $\tilde{w}_{\mathtt{T}}$). Here I adapt it to the case of finite time-series.

The behavior $\mathscr{B} \subseteq (\mathbb{R}^{\mathtt{w}})^{\mathbb{N}}$ is *unfalsified* by $\tilde{w}_{\mathtt{T}}$ if

$$\big(\tilde{w}(1), \tilde{w}(2), \ldots, \tilde{w}(\mathtt{T})\big) \in \mathscr{B}|_{[1,\mathtt{T}]}.$$

$\mathscr{B} \subseteq (\mathbb{R}^{\mathtt{w}})^{\mathbb{N}}$ is *more powerful* than $\mathscr{B}' \subseteq (\mathbb{R}^{\mathtt{w}})^{\mathbb{N}}$ if $\mathscr{B} \subset \mathscr{B}'$. Hence, following Karl Popper (1902–1994), *the more a model forbids, the better it is*. $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$ is the *MPUM in* $\mathscr{L}^{\mathtt{w}}$ *for* $\tilde{w}_{\mathtt{T}}$ if it is unfalsified by $\tilde{w}_{\mathtt{T}}$ and more powerful than any other behavior in $\mathscr{L}^{\mathtt{w}}$ that is also unfalsified by $\tilde{w}_{\mathtt{T}}$. It is easy to prove that for an *infinite* observed time-series ($\mathtt{T} = \infty$ in $\tilde{w}_{\mathtt{T}}$), this MPUM always exists. In fact, it is equal to the closure (in the topology of pointwise convergence) of $\mathrm{span}\{\sigma^{\mathtt{t}}\tilde{w}_{\mathtt{T}}, \mathtt{t} \in \mathbb{N}\}$. For $\mathtt{T}$ finite, the MPUM in $\mathscr{L}^{\mathtt{w}}$ may not exist, but it is not particularly useful anyway: when it exists, its behavior will be finite dimensional (corresponding to an autonomous behavior). However, it is desirable, also in the case $\mathtt{T} < \infty$, to obtain an MPUM that can also recover a behavior with some of the variables free inputs. This can be accomplished by looking for the MPUM in the class of systems with a restricted lag. This can be viewed as limiting the complexity of the model. Define

$$\mathscr{L}^{\mathtt{w}}_{\mathtt{L}} := \{\mathscr{B} \in \mathscr{L}^{\mathtt{w}} \mid \mathsf{L}(\mathscr{B}) \leq \mathtt{L}\}.$$

$\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$ is the *MPUM in* $\mathscr{L}^{\mathtt{w}}_{\mathtt{L}}$ *for* $\tilde{w}_{\mathtt{T}}$ if it is unfalsified by $\tilde{w}_{\mathtt{T}}$ and more powerful than any other behavior in $\mathscr{L}^{\mathtt{w}}_{\mathtt{L}}$ that is also unfalsified by $\tilde{w}_{\mathtt{T}}$. Denote the MPUM in $\mathscr{L}^{\mathtt{w}}_{\mathtt{L}}$ for $\tilde{w}_{\mathtt{T}}$ by $\mathscr{B}^{\star}_{\mathtt{L},\tilde{w}_{\mathtt{T}}}$. Questions that arise are: *When does* $\mathscr{B}^{\star}_{\mathtt{L},\tilde{w}_{\mathtt{T}}}$ *exist*? How can it be computed from the data $\tilde{w}_{\mathtt{T}}$? If $\tilde{w}_{\mathtt{T}}$ is generated by an element $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$, when will $\mathscr{B}^{\star}_{\mathtt{L},\tilde{w}_{\mathtt{T}}} = \mathscr{B}$?

Conditions for existence of this MPUM are readily deduced from the Hankel matrix formed by the data. We first introduce our notation for Hankel matrices (meaning block Hankel matrices). Exclusively Hankel matrices formed by vectors are needed. Let $f : [1, \mathtt{t}] \to \mathbb{R}^{\mathtt{n}}$. Define the *Hankel matrix formed by* f with *depth* Δ_1 and *width* Δ_2, $\Delta_1 + \Delta_2 - 1 \leq \mathtt{t}$, by

$$\mathscr{H}_{(\Delta_1,\Delta_2)}(f) := \begin{bmatrix} f(1) & f(2) & \cdots & f(\Delta_2) \\ f(2) & f(3) & \cdots & f(\Delta_2+1) \\ \vdots & \vdots & \ddots & \vdots \\ f(\Delta_1) & f(\Delta_1+1) & \cdots & f(\Delta_1+\Delta_2-1) \end{bmatrix}$$

The *left kernel* of a matrix $M \in \mathbb{R}^{\mathtt{n}_1 \times \mathtt{n}_2}$, denoted by $\mathsf{leftker}(M)$, consists of the row vectors $n \in \mathbb{R}^{1 \times \mathtt{n}_1}$ that are annihilated by postmultiplication by M: $nM = 0$. In the case of a Hankel matrix, it is often useful to view the elements in the left kernel as vector polynomials, as follows. Assume that $n = \begin{bmatrix} n_0 & n_1 \cdots n_{\Delta_1-2} & n_{\Delta_1-1} \end{bmatrix} \in \mathsf{leftker}\big(\mathscr{H}_{(\Delta_1,\Delta_2)}(f)\big)$ with the $n_{\mathtt{k}}$'s $\in \mathbb{R}^{1 \times \mathtt{n}}$. Associate with n the polynomial vector $n_0 + n_1\xi + \cdots + n_{\Delta_1-2}\xi^{\Delta_1-2} + n_{\Delta_1-1}\xi^{\Delta_1-1} \in \mathbb{R}^{1 \times \mathtt{n}}[\xi]$, and denote it, with slight

abuse of notation, also as $n(\xi)$. Note that it may *appear as if* $\text{degree}(n) = \Delta_1 - 1$, but since some of the coefficients of n may be zero, the actual degree may be lower. Call $\mathsf{leftker}\big(\mathscr{H}_{(\Delta_1,\Delta_2)}(f)\big)$ *module-like* if

$$[\![\ [\ n_0\ n_1\ \cdots\ n_{\Delta_1-3}\ n_{\Delta_1-2}\ 0\ ] \in \mathsf{leftker}\big(\mathscr{H}_{(\Delta_1,\Delta_2)}\big)\ ]\!] \Rightarrow [\![\ [\ 0\ n_0\ n_1\ \cdots\ n_{\Delta_1-3}\ n_{\Delta_1-2}\ ] \in \mathsf{leftker}\big(\mathscr{H}_{(\Delta_1,\Delta_2)}(f)(f)\big)\ ]\!].$$

Equivalently, if

$$[\![\ \big(n(\xi) \in \mathsf{leftker}\big(\mathscr{H}_{(\Delta_1,\Delta_2)}(f)\big)\big) \wedge \big(\text{degree}(n) < \Delta_1 - 1\big)\ ]\!] \Rightarrow [\![\ \xi n(\xi) \in \mathsf{leftker}\big(\mathscr{H}_{(\Delta_1,\Delta_2)}(f)\big)\ ]\!].$$

This last implication shows where the terminology 'module-like' refers to.

These notions are very useful for computing $\mathscr{B}^\star_{\mathtt{L},\tilde{w}_\mathtt{T}}$ (which may not exist – take $\tilde{w}_\mathtt{T} = (1,1,2,2,2)$ and $\mathtt{L} = 2$). Observe that if $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$ is unfalsified by $\tilde{w}_\mathtt{T}$, then

$$[\![\ (n \in \mathscr{N}_\mathscr{B}) \wedge (\text{degree}(n) \leq \mathtt{L})\ ]\!] \Rightarrow [\![\ n \in \mathsf{leftker}\big(\mathscr{H}_{(\mathtt{L}+1,\mathtt{T}-\mathtt{L})}(\tilde{w}_\mathtt{T})\big)\ ]\!].$$

It is logical to aim at the module generated by $\mathsf{leftker}\big(\mathscr{H}_{(\mathtt{L}+1,\mathtt{T}-\mathtt{L})}(\tilde{w}_\mathtt{T})\big)$ as the annihilators of the MPUM $\mathscr{B}^\star_{\mathtt{L},\tilde{w}_\mathtt{T}}$. However, for arbitrary $\tilde{w}_\mathtt{T}$, this module may contain falsified elements (consider, e.g. $\tilde{w}_\mathtt{T} = (0,0,\ldots,0,0,1)$). The module-like property ensures that all elements in the module generated by $\mathsf{leftker}\big(\mathscr{H}_{(\mathtt{L}+1,\mathtt{T}-\mathtt{L})}(\tilde{w}_\mathtt{T})\big)$ are unfalsified.

This leads to the following results.

(i) A sufficient condition for the existence of the MPUM $\mathscr{B}^\star_{\mathtt{L},\tilde{w}_\mathtt{T}}$ is that $\mathsf{leftker}\big(\mathscr{H}_{(\mathtt{L}+1,\mathtt{T}-\mathtt{L})}(\tilde{w}_\mathtt{T})\big)$ is module-like.

(ii) $\mathsf{leftker}\big(\mathscr{H}_{(\mathtt{L}+1,\mathtt{T}-\mathtt{L})}(\tilde{w}_\mathtt{T})\big)$ is module-like iff

$$\text{rank}\big(\mathscr{H}_{(\mathtt{L},\mathtt{T}-\mathtt{L})}(\tilde{w}_\mathtt{T})\big) = \text{rank}\big(\mathscr{H}_{(\mathtt{L},\mathtt{T}-\mathtt{L}+1)}(\tilde{w}_\mathtt{T})\big).$$

(iii) If $\mathsf{leftker}\big(\mathscr{H}_{(\mathtt{L}+1,\mathtt{T}-\mathtt{L})}(\tilde{w}_\mathtt{T})\big)$ is module-like, then $\mathscr{B}^\star_{\mathtt{L},\tilde{w}_\mathtt{T}} = \ker\big(N(\sigma)\big)$, with $N \in \mathbb{R}^{\bullet\times\mathtt{w}}[\xi]$ such that the $\mathbb{R}[\xi]$-module generated by its rows is equal to the $\mathbb{R}[\xi]$-module generated by $\mathsf{leftker}\big(\mathscr{H}_{(\mathtt{L}+1,\mathtt{T}-\mathtt{L})}(\tilde{w}_\mathtt{T})\big)$.

Under mild conditions, the MPUM in $\mathscr{L}^{\mathtt{w}}_{\mathtt{L}}$ for $\tilde{w}_\mathtt{T}$ recovers the system that generated the data. In other words, the procedure is consistent. The crucial condition here is that the input component in the observations must be persistently exciting. $f : [1,\mathtt{T}] \to \mathbb{R}^\bullet$ is said to be *persistently exciting of order* $\mathtt{P}$ if the rows of the Hankel matrix $\mathscr{H}_{(\mathtt{P},\mathtt{T}-\mathtt{P}+1)}(f)$ are linearly independent.

We have the following 'consistency' result. Consider $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$. Let $\begin{bmatrix} u \\ y \end{bmatrix} = \Pi^{-1} w$ be an input/output partition of $\mathscr{B}$ as in (I/O). Denote the corresponding observed input component of $\tilde{w}_\mathtt{T}$ by $\tilde{u}_\mathtt{T}$. It can be shown that $\mathscr{B}^\star_{\mathtt{L},\tilde{w}_\mathtt{T}} = \mathscr{B}$ if the following conditions are satisfied:

(i) $\tilde{w}_\mathtt{T} \in \mathscr{B}|_{[1,\mathtt{T}]}$,

(ii) $\mathsf{L}(\mathscr{B}) \leq \mathtt{L}$,
(iii) $\mathscr{B}$ is controllable,
(iv) $\tilde{u}_{\mathtt{T}}$ is persistently exciting of order $\mathtt{L} + \mathsf{n}(\mathscr{B}) + 1$.

Moreover, under these conditions, the left kernel of $\mathscr{H}_{(\mathtt{L}+1,\mathtt{T}-\mathtt{L})}(\tilde{w}_{\mathtt{T}})$ is module-like, and therefore $\mathscr{B} = \ker(N(\sigma))$, where $N \in \mathbb{R}^{\bullet\times\mathtt{w}}[\xi]$ is any polynomial matrix such that the $\mathbb{R}[\xi]$-module generated by its rows equals the $\mathbb{R}[\xi]$-module generated by $\mathsf{leftker}\left(\mathscr{H}_{(\mathtt{L}+1,\mathtt{T}-\mathtt{L})}(\tilde{w}_{\mathtt{T}})\right)$.

In a typical application of this result, assume that an infinite time-series $\tilde{w} : \mathbb{N} \to \mathbb{R}^{\mathtt{w}}$ is observed, generated by an unknown, but controllable, system $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$, for which upper bounds $\mathtt{L}$ for the lag $\mathsf{L}(\mathscr{B})$ and $\mathtt{n}$ for the state dimension $\mathsf{n}(\mathscr{B})$ are known (since $\mathsf{L}(\mathscr{B}) \leq \mathsf{n}(\mathscr{B})$, it suffices to have the upper bound $\mathtt{n}$). Then, if the observed input component $\tilde{u}$ of $\tilde{w}$ is persistently exciting of order $\mathtt{L} + \mathtt{n} + 1$, the MPUM $\mathscr{B}^{\star}_{\mathtt{L},\tilde{w}_{\mathtt{T}}}$ will be equal to $\mathscr{B}$ for all $\mathtt{T}$ such that $\tilde{u}_{\mathtt{T}}$ is persistently exciting of order $\mathtt{L} + \mathtt{n} + 1$. So, this exact deterministic SYSID algorithm yields consistency in finite time.

We now turn to the computation of $\mathscr{B}^{\star}_{\mathtt{L},\tilde{w}_{\mathtt{T}}}$. One possibility suggested by the above results is to compute a basis for the whole left kernel of $\mathscr{H}_{(\mathtt{L}+1,\mathtt{T}-\mathtt{L})}(\tilde{w}_{\mathtt{T}})$. This is feasible (also approximately, using SVD-like algorithms), but it is not necessary nor efficient to compute the whole left kernel, especially when the data is generated by a system in $\mathscr{L}^{\mathtt{w}}_{\mathtt{L}}$ for which the lags in the difference equations ($\mathfrak{Ker}$) vary widely, or when $\mathtt{L}$ is only a rough upper bound for $\mathsf{L}(\mathscr{B})$. It suffices to compute vector polynomials $n_1, n_2, \ldots, n_{\mathtt{g}} \in \mathbb{R}^{1\times\mathtt{w}}[\xi]$ such that the $\mathbb{R}[\xi]$-module spanned by $n_1, n_2, \ldots, n_{\mathtt{g}}$ equals the $\mathbb{R}[\xi]$-module spanned by the left kernel of $\mathscr{H}_{(\mathtt{L}+1,\mathtt{T}-\mathtt{L})}(\tilde{w}_{\mathtt{T}})$. The problem hence comes down to obtain a set of polynomial vectors in the left kernel of $\mathscr{H}_{(\mathtt{L}+1,\mathtt{T}-\mathtt{L})}(\tilde{w}_{\mathtt{T}})$ that generate this submodule. This is algorithmically much simpler that computing the whole left kernel, which may have a dimension that is much larger than the dimension (the number of generators) of this submodule.

This submodule may be computed with an algorithm that is recursive in the observation horizon $\mathtt{T}$ or in the lag $\mathtt{L}$. An algorithm that is recursive in $\mathtt{T}$ has been obtained in [36, 20]. This algorithm is based on a recursive computation of the MPUM for infinite time-series, and is a generalization of the Berlekamp-Massey decoding algorithm. We have recently also obtained an algorithm that is recursive in $\mathtt{L}$. The details will be reported elsewhere.

6 Subspace ID

The importance and usefulness of state models are beyond dispute. By explicitly displaying the memory of a system, state models bring the dynamical properties of a system to the surface, clearly into view. In fact, since the 1960's, input/state/output models are, *up to a fault*, used as the basic framework for studying open dynamical systems. Since the 1980's, it has become apparent that state models are also very well suited for model reduction. The highlight of this development is Keith Glover's classic paper [12] on balanced realizations and AAK type model reduction algorithms.

The one area in systems theory where state models play a somewhat secondary role is SYSID. In earlier sections, I have dealt extensively with the multitude of representations of elements of $\mathscr{L}^{\bullet}$. SYSID algorithms typically aim at one of these representations, usually a kernel representation ($\mathfrak{Ker}$), an input/output version (I/O), or a latent variable representation ($\mathfrak{L}$). One major exception to this are the *subspace identification* algorithms. These algorithms pass from the data $\tilde{w}_{\mathtt{T}}$ to an input/state/output representation ($\mathfrak{S}$). In this section, I deal with these algorithms. Contrary to many authors, I do not consider the classical realization algorithms as part of SYSID. In realization theory, one has a model at the very outset, a model that, for example, gives the output as a convolution of the input with the impulse response matrix, and the problem is to find an equivalent state representation. Finding the parameters of a state representation from the impulse response is a representation problem and does not have much to do with SYSID (although one may get many good ideas for SYSID from these algorithms). Much more SYSID oriented are algorithms that pass from $\tilde{w}_{\mathtt{T}}$ or $(\tilde{u}_{\mathtt{T}}, \tilde{y}_{\mathtt{T}})$ if an input/output partition is known, to a state representation of $\mathscr{B}^{\star}_{\mathtt{L},\tilde{w}_{\mathtt{T}}}$. This is the problem that I will discuss now.

In this section we consider SYSID algorithms that aim at i/s/o models ($\mathfrak{S}$). As always in state representations, we meet the issue of non-uniqueness, both of the state representation parameters, and of the state trajectory corresponding to a given w-trajectory. It is easy to show that every $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$ admits an observable representation ($\mathfrak{S}$) (but not necessarily a controllable one!). Moreover, for a fixed input/output partition Π, observable representations ($\mathfrak{S}$) of the same $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$ are unique up to a basis choice in the state space. If we allow also Π (i.e. the input/output partition) to vary, things become a bit more complicated, but we need not deal with this here. The discussion in this section is a bit informal: we assume, where needed, that $\tilde{w}_{\mathtt{T}}$ is such that the MPUM in $\mathscr{L}^{\mathtt{w}}_{\mathtt{L}}$ exists, that $\mathtt{L}$ and $\mathtt{T}$ are sufficiently large, controllability, observability, persistency of excitation, etc., and also that the input/output partition in $\mathscr{B}^{\star}_{\mathtt{L},\tilde{w}_{\mathtt{T}}}$ is known.

Assume that from the observations $\tilde{w}_{\mathtt{T}}$, we have somehow identified a model $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}$ such that $\tilde{w}_{\mathtt{T}} \in \mathscr{B}|_{[\mathtt{1},\mathtt{T}]}$ (for example, assume that we have identified the MPUM $\mathscr{B}^{\star}_{\mathtt{L},\tilde{w}_{\mathtt{T}}}$). We could then compute a state representation ($\mathfrak{S}$) for $\mathscr{B}$, and obtain a corresponding state trajectory

$$\tilde{x}_{\mathtt{T}} = \quad (\tilde{x}(\mathtt{1}), \tilde{x}(\mathtt{2}), \ldots, \tilde{x}(\mathtt{T})) \qquad \tilde{x}(\mathtt{t}) \in \mathbb{R}^{\mathtt{n}}.$$

This lays out the path to go from $\tilde{w}_{\mathtt{T}}$ to $\tilde{x}_{\mathtt{T}}$. $\tilde{x}_{\mathtt{T}}$ and $\tilde{w}_{\mathtt{T}} \cong (\tilde{u}_T, \tilde{y}_T)$ are related by the equations

$$\begin{bmatrix} \tilde{x}(\mathtt{2}) & \tilde{x}(\mathtt{3}) & \cdots & \tilde{x}(\mathtt{T}) \\ \tilde{y}(\mathtt{1}) & \tilde{y}(\mathtt{2}) & \cdots & \tilde{y}(\mathtt{T}-\mathtt{1}) \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \tilde{x}(\mathtt{1}) & \tilde{x}(\mathtt{2}) & \cdots & \tilde{x}(\mathtt{T}-\mathtt{1}) \\ \tilde{u}(\mathtt{1}) & \tilde{u}(\mathtt{2}) & \cdots & \tilde{u}(\mathtt{T}-\mathtt{1}) \end{bmatrix}. \qquad (\clubsuit)$$

However, if we could somehow obtain *first* the state trajectory $\tilde{x}_{\mathtt{T}}$, *directly from* $\tilde{w}_{\mathtt{T}}$, without using the underlying model, we could subsequently solve the linear system of equations ($\clubsuit$) for the unknown parameter matrices (A,B,C,D). This would yield a SYSID algorithm that identifies a state representation for $\mathscr{B}^{\star}_{\mathtt{L},\tilde{w}_{\mathtt{T}}}$. The advantage – and a major one indeed – of dealing with the equations ($\clubsuit$) is that, from a

numerical linear algebra point of view, they lend themselves very well to approximations. Indeed, one can first use SVD-like algorithms to reduce the state dimension by rank reduction of the matrix

$$\left[\, \tilde{x}(1)\ \tilde{x}(2)\ \cdots\ \tilde{x}(\mathtt{T}-1)\ \tilde{x}(\mathtt{T}) \,\right]$$

and subsequently solve the equations (♣) approximately for (A,B,C,D) in a least squares (LS) sense. These numerical features are very well explained and effectively exploited in [30].

The following question emerges:

How can we pass directly from $\tilde{w}_{\mathtt{T}}$ to the corresponding state trajectory $\tilde{x}_{\mathtt{T}}$ of a state representation of the MPUM $\mathscr{B}^{\star}_{\mathtt{L},\tilde{w}_{\mathtt{T}}}$?

One algorithm which achieves this is based on partitioning the Hankel matrix of the data into the 'past' and the 'future', as follows

$$\left[\begin{array}{cccc}
\tilde{w}(1) & \tilde{w}(2) & \cdots & \tilde{w}(\mathtt{T}-2\mathtt{L}-1) \\
\vdots & \vdots & \vdots & \vdots \\
\tilde{w}(\mathtt{L}+1) & \tilde{w}(\mathtt{L}+2) & \cdots & \tilde{w}(\mathtt{T}-\mathtt{L}-1) \\
\hline
\tilde{w}(\mathtt{L}+2) & \tilde{w}(\mathtt{L}+3) & \cdots & \tilde{w}(\mathtt{T}-\mathtt{L}) \\
\vdots & \vdots & \vdots & \vdots \\
\tilde{w}(2\mathtt{L}+2) & \tilde{w}(2\mathtt{L}+3) & \cdots & \tilde{w}(\mathtt{T})
\end{array}\right]
\begin{array}{l}
\\ \\ \leftarrow \text{past } \mathscr{P} \\ \hline \\ \leftarrow \text{future } \mathscr{F} \\ \\
\end{array}$$

i.e.

$$\mathscr{H}_{(2\mathtt{L}+2,\,\mathtt{T}-2\mathtt{L}-1)}(\tilde{w}_{\mathtt{T}}) = \left[\begin{array}{c}
\mathscr{H}_{(\mathtt{L}+1,\,\mathtt{T}-\mathtt{L}-1)}(\tilde{w}_{\mathtt{T}}) \\
\hline
\mathscr{H}_{(\mathtt{L}+1,\,\mathtt{T}-\mathtt{L}-1)}(\sigma^{\mathtt{L}+1}\tilde{w}_{\mathtt{T}})
\end{array}\right]
\begin{array}{l}
\leftarrow \text{past } \mathscr{P} \\ \hline \leftarrow \text{future } \mathscr{F}
\end{array}$$

Consider now the span of the rows of the 'past' matrix $\mathscr{P}$ and the span of the rows of the 'future' matrix $\mathscr{F}$. It turns out that the intersection of these spans is equal to the state space of $\mathscr{B}^{\star}_{\mathtt{L},\tilde{w}_{\mathtt{T}}}$, and that the linear combinations of the rows of $\mathscr{P}$ (or $\mathscr{F}$) contained in this intersection give the 'present' state trajectory! More precisely, assume that the span of the rows of the matrix $\mathscr{X}$ equals the intersection of the span of the rows of $\mathscr{P}$ with the span of the rows of $\mathscr{F}$. Then the columns of $\mathscr{X}$ are the 'present' state trajectory. Note that we may as well assume that the rows of $\mathscr{X}$ are linearly independent. The fact that $\mathscr{X}$ is unique up to pre-multiplication by a nonsingular matrix corresponds to the freedom of the choice of the basis of the state space of the underlying system. In other words,

$$\mathscr{X} = (\tilde{x}(\mathtt{L}+2), \tilde{x}(\mathtt{L}+3), \ldots, \tilde{x}(\mathtt{T}-\mathtt{L}-1), \tilde{x}(\mathtt{T}-\mathtt{L}))\,.$$

This fact, first noticed in [35, sections 15–17], and generalized in many directions by De Moor and co-workers (see e.g. [30]), allows to identify the system parameter matrices (A,B,C,D) of $\mathscr{B}^{\star}_{\mathtt{L},\tilde{w}_{\mathtt{T}}}$ in input/state/output form by solving equations as (♣).

This intersection result is reminiscent of the following observation, usually attributed to Akaike [1]. Consider a zero mean stationary ergodic gaussian vector process z, $z(\mathtt{t}) \in \mathbb{R}^{\mathtt{z}}, \mathtt{t} \in \mathbb{Z}$, with a rational spectral density.

Let $\tilde{z} : \mathbb{Z} \to \mathbb{R}^{\mathtt{z}}$ be a realization of it. Now form the Hankel matrix

$$\left[\begin{array}{ccccccc}
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\cdots & \tilde{z}(-2) & \tilde{z}(1) & \tilde{z}(0) & \cdots & \tilde{z}(\mathtt{t}-1) & \cdots \\
\cdots & \tilde{z}(-1) & \tilde{z}(0) & \tilde{z}(1) & \cdots & \tilde{z}(\mathtt{t}) & \cdots \\
\hline
\cdots & \tilde{z}(0) & \tilde{z}(1) & \tilde{z}(2) & \cdots & \tilde{z}(\mathtt{t}+1) & \cdots \\
\cdots & \tilde{z}(1) & \tilde{z}(2) & \tilde{z}(3) & \cdots & \tilde{z}(\mathtt{t}+2) & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{array}\right]
\begin{array}{l}
\\ \leftarrow \text{past} \\ \\ \\ \leftarrow \text{future} \\ \\
\end{array}$$

Then the orthogonal projection (in the inner product induced by the second order averages) of the span of the rows of the future onto the span of the rows of the past is finite dimensional, and the orthogonal projection of the rows of the future onto the span of the rows of the past yields a realization of the corresponding 'past-induced' Markovian trajectory $\tilde{x}$ associated with $\tilde{z}$.

But, how do we compute the common linear combinations spanned by the rows of M_1 and M_2 in the partitioned matrix $\left[\begin{smallmatrix} M_1 \\ M_2 \end{smallmatrix}\right]$? This can be done by computing a basis for the left kernel of $\left[\begin{smallmatrix} M_1 \\ M_2 \end{smallmatrix}\right]$. Indeed, if $\begin{bmatrix} n_1 & n_2 \end{bmatrix}$ is in the left kernel, then $n_1 M_1 = -n_2 M_2$ yields one of these linear combinations. By letting $\begin{bmatrix} n_1 & n_2 \end{bmatrix}$ range over a basis of the left kernel, all common linear combinations are obtained. This can be applied to $\left[\begin{smallmatrix} \mathscr{P} \\ \hline \mathscr{F} \end{smallmatrix}\right]$ and allows to compute the desired state trajectory by computing the left kernel of $\left[\begin{smallmatrix} \mathscr{P} \\ \hline \mathscr{F} \end{smallmatrix}\right]$. The disadvantage of this algorithm is that it does not make use of the Hankel structure of the matrices $\mathscr{P}$ and $\mathscr{F}$ (a common drawback of subspace algorithms). Equivalently, it makes no use of the module structure of the left kernel. For $\mathtt{L}$ large (say, much larger than $\mathsf{L}(\mathscr{B})$), the algorithms that compute the whole left kernel risk at being very inefficient. But, there is more. As we have shown in the previous section, the left kernel of $\left[\begin{smallmatrix} \mathscr{P} \\ \hline \mathscr{F} \end{smallmatrix}\right]$ can in fact be deduced from the left kernel of $\mathscr{P}$ alone by simple module operations, and therefore there is no need to consider the separation into 'past' and 'future'.

These considerations lead to the following algorithm for constructing the state trajectory from the observed data, assuming persistency of excitation, controllability, $\mathtt{L}$ sufficiently large, etc. Let $N \in \mathbb{R}^{\bullet \times \mathtt{w}}[\xi]$ be a polynomial matrix such that the $\mathbb{R}[\xi]$-module generated by its rows equals the $\mathbb{R}[\xi]$-module generated by $\mathsf{leftker}\big(\mathscr{H}_{(\mathtt{L}+1,\mathtt{T}-\mathtt{L})}(\tilde{w}_{\mathtt{T}})\big)$, and assume $\mathrm{degree}(N) \leq \mathtt{L}$. In the previous section, we have explained how to view elements in this left kernel as vector polynomials. The construction of the state trajectory involves the 'shift-and-cut' operator π on $\mathbb{R}[\xi]$, defined by

$$\pi : p_0 + p_1\xi + p_2\xi^2 + \cdots + p_{\mathtt{n}}\xi^{\mathtt{n}} \mapsto p_1 + p_2\xi + + \cdots + p_{\mathtt{n}}\xi^{\mathtt{n}-1}.$$

π can be extended in an obvious way to act on polynomial vectors and matrices. Now define X_N as

$$X_N := \mathrm{col}(\pi N, \pi^2 N, \ldots, \pi^{\mathrm{degree}(N)-1} N),$$

and view X_N as a matrix by reversing the process by which we viewed elements in the left kernel of $\mathscr{H}_{(\mathtt{L}+1,\mathtt{T}-\mathtt{L})}(\tilde{w}_\mathtt{T})$ as polynomial vectors. Then

$$X_N \mathscr{H}_{(\mathtt{L}+1,\mathtt{T}-\mathtt{L})}(\tilde{w}_\mathtt{T}) = \left[\tilde{x}(1)\ \tilde{x}(2)\ \cdots\ \tilde{x}(\mathtt{T}-\mathtt{L}-1)\ \tilde{x}(\mathtt{T}-\mathtt{L})\right].$$

This yields a very effective construction of a state trajectory. This algorithm is actually an implementation of the 'cut-and-shift' algorithm for constructing a state representation starting from a kernel representation developed in [27].

There exist many related subspace identification type algorithms, for example based on oblique projections and various LS methods [30], [34], [22].

We have already mentioned that subspace ID is very well suited, pragmatically, through rank reduction of the state trajectory $\mathscr{X}$, followed by LS methods for solving (♣), for identifying a model that fits the observed data $\tilde{w}_\mathtt{T}$ approximately. Of course, the most satisfying way theoretically of approaching SYSID through approximate modeling is by defining a *misfit* criterion, say

$$\mathrm{misfit}(\tilde{w}_\mathtt{T}, \mathscr{B}) = \underset{w \in \mathscr{B}}{\mathrm{minimum}} \, ||\tilde{w}_\mathtt{T} - w|_{[1,\mathtt{T}]}||_{\ell_2([1,\mathtt{T}],\mathbb{R}^\mathtt{w})}.$$

and then minimizing the misfit $\mathrm{misfit}(\tilde{w}_\mathtt{T}, \mathscr{B})$ over a model class, for example the elements of $\mathscr{B} \in \mathscr{L}^\mathtt{w}$ with a limited complexity, with complexity defined as something like $(\mathtt{m}(\mathscr{B}), \mathtt{n}(\mathscr{B}), \mathtt{L}(\mathscr{B}))$. This is the approach taken in [16], [28], and [22]. This last reference shows how to approach this problem using STLS (structured total least squares) algorithms.

7 Latency minimization

The usual model class considered in SYSID lead to equations of the form

$$R(\sigma)\, w = M(\sigma)\, \varepsilon \tag{$\mathfrak{E}$}$$

with R and M real polynomial matrices of suitable sizes. These equations involve, in addition to the time-series $w : \mathbb{N} \to \mathbb{R}^\mathtt{w}$, also the time-series $\varepsilon : \mathbb{N} \to \mathbb{R}^\mathtt{e}$. In ($\mathfrak{E}$), w is the time-series which is observed on the interval $[1, \mathtt{T}]$ through $\tilde{w}_\mathtt{T}$, and ε consists of unobserved latent variables, which serve to help to explain the data. In the behavioral language, this means that we start with a latent variable representation and aim at explaining the observed manifest variables w with the help of the unobserved latent variables ε. It is often useful to consider a more structured version of ($\mathfrak{E}$), for example,

$$P(\sigma)\, y = Q(\sigma)\, u + M(\sigma)\, \varepsilon, \quad w = \begin{bmatrix} u \\ y \end{bmatrix}, \tag{$\mathfrak{E}'$}$$

with $\det(P) \neq 0$, ensuring that both u and ε are inputs, and y is the output. The question then basically is to identify the behavior $\mathscr{B}_{\mathrm{full}}$, or the polynomial matrices

(R,M) or (P,Q,M), from observations (of a finite sample) of w, or u and y, entangled by an unobserved latent variable trajectory ε.

The performance, from a SYSID point of view, of the model $(\mathfrak{E})$ with behavior $\mathscr{B}_{\text{full}} \in \mathscr{L}^{\mathtt{w}+\mathtt{e}}$ can be assessed by the *latency*, say

$$\text{latency}(\tilde{w}_{\mathtt{T}}, \mathscr{B}_{\text{full}}) := \underset{[\![(w,\varepsilon)\in\mathscr{B}_{\text{full}}]\!] \wedge [\![w|_{[1,\mathtt{T}]}=\tilde{w}_{\mathtt{T}}]\!]}{\text{minimum}} ||\varepsilon||_{\ell_2(\mathbb{N},\mathbb{R}^{\mathtt{e}})}.$$

Subsequently minimizing the latency over the model class (a family of $\mathscr{B}_{\text{full}}$'s) leads to a SYSID algorithm. Note that because the models $(\mathfrak{E})$ are usually unobservable (in the sense that ε cannot be deduced from $w,R,$ and M), it is in general not possible to reformulate the latency criterion in terms of the w variables alone.

The following questions arise. Given the behavior $\mathscr{B}_{\text{full}} \in \mathscr{L}^{\mathtt{w}+\mathtt{e}}$ how do we compute the latency in terms of the polynomial matrices (R,M)? Given a model class $\mathscr{M} \subseteq \mathscr{L}^{\mathtt{w}+\mathtt{e}}$, how do we minimize the latency over $\mathscr{M}$? The computation of the latency is a deterministic Kalman filtering problem. The minimization over the model class usually leads to a non-convex minimization problem.

It turns out, in fact, that from a numerical (but not from a conceptual) point of view, latency minimization is *precisely* what is done in prediction error minimization methods (PEM) [21] (strongly related to maximum likelihood (ML)). The difference is 'merely' the interpretation of the latent variables ε. In latency minimization, ε is introduced in order to explain the observed data $\tilde{w}_{\mathtt{T}}$ (but no physical reality is attached to it), while in PEM and ML, ε is interpreted as a stochastic 'disturbance' input which, together with the input component in the w variable in $(\mathfrak{E})$ (see $(\mathfrak{E}')$) and the initial conditions, produces the data. Under suitable stochastic assumptions, as ε independent of the input component u of $(\mathfrak{E}')$, one obtains precisely the same estimate for (R,M) by using latency minimization as by using (PEM) or (ML). In latency minimization, one wants to keep ε small, while explaining the observations, in PEM one minimizes the a posteriori prediction error in a stochastic sense, and in ML one maximizes the likelihood of the observations, given the model.

Assume now that (R,M) in $(\mathfrak{E})$, equivalently (P,Q,M) in $(\mathfrak{E}')$, have been identified from the observations $\tilde{w}_{\mathtt{T}}$, how should one proceed? Which model does this identification procedure entail? If these estimates have been obtained by latency minimization, then it is natural to take $R(\sigma)w = 0$ i.e. $P(\sigma)y = Q(\sigma)u$ as the model to be employed for prediction, control, or what have you. This is different, of course, from taking the manifest behavior of $(\mathfrak{E})$. Setting $\varepsilon = 0$ in $(\mathfrak{E})$ yields a much smaller behavior than eliminating ε from $(\mathfrak{E})$. However, if the estimates have been obtained by (PEM) or (ML), then it is natural to stick to $(\mathfrak{E})$, and take into consideration the identified stochastic characteristics of the variables ε in $(\mathfrak{E})$ or $(\mathfrak{E}')$. This stochastic model can then be used for predicting future outputs from future inputs, for control, etc.

Does $(\mathfrak{E})$ provide additional flexibility compared to $(\mathfrak{Ker})$? Is the set of w trajectories that can be modelled this way larger than what can be achieved without the ε's? This may appear so at first sight, but it is not. Indeed, by the elimination theorem the 'manifest' behavior of $(\mathfrak{E})$, i.e.

$$\mathscr{B} = \{w :\to \mathbb{R}^{\mathtt{w}} \mid \exists\, \varepsilon \to \mathbb{R}^{\mathtt{e}} \text{ such that } (\mathfrak{E}) \text{ holds}\}$$

in fact also belongs to $\mathscr{L}^{\mathtt{w}}$. In other words, there exists $R' \in \mathbb{R}^{\bullet \times \mathtt{w}}[\xi]$ such that $\mathscr{B} = \ker(R'(\sigma))$, yielding the kernel representation $R'(\sigma)w = 0$. *What is then the rationale for introducing ε and using the model class ($\mathfrak{E}$)?* We can intuitively interpret the use of latency minimization, followed by setting the latent variables $\varepsilon = 0$ and using $R(\sigma)w = 0$ as the identified model, as a way of minimizing some sort of misfit. But it is a misfit that involves *unobservable* latent variables, with 'unobservable' interpreted in the technical sense for latent variable systems explained earlier. The advantage of introducing latent variables above straight misfit minimization can be seen by considering the case $\mathtt{w} = 1$. $\mathscr{B} \in \mathscr{L}^1$ implies that either $\mathscr{B} = \mathbb{R}^{\mathbb{N}}$, not useful as a model, or that $\mathscr{B}$ is an autonomous system, and therefore finite dimensional, consisting of trajectories that are a finite sum of polynomial exponentials. If we assume in addition that $\mathscr{B}$ is stable, i.e. $[\![w \in \mathscr{B}]\!] \Rightarrow [\![w(t) \to 0 \text{ for } t \to \infty]\!]$, then we see that an autonomous behavior $\mathscr{B}$ cannot adequately capture the dynamic features of a persistent trajectory $\tilde{w}_{\mathtt{T}}$. The assumption of the presence of unobservable latent inputs ε in ($\mathfrak{E}$) indeed offers better modeling possibilities for data fitting. Of course, one can also use a combination of the straight misfit minimization and latency minimization.

It is an open question if a model obtained by latency minimization and setting $\varepsilon = 0$, i.e. $R(\sigma)w = 0$, or $P(\sigma)y = Q(\sigma)u$, equivalently, obtained by (PEM) and keeping the deterministic part, or a model obtained by misfit minimization, will do significantly better than a model (in a corresponding model class) obtained from, say, a heuristic subspace ID type algorithm obtained by reduction of the state trajectory $\mathscr{X}$ followed by a LS solution (♣), as explained earlier. Extensive testing on industrial data [23] suggests that models obtained by latency minimization or (PEM) give a marginally better fit in predicting future outputs from future inputs as compared to subspace ID, but at the expense of a *much* larger computational effort. Systematic methods to fit a linear dynamical system to the data will lead to a reasonable model. It makes sense to expect that a linear approximation found during the learning stage will prevail during the validation stage. But there is no reason to expect that statistical features which happen to be present in the data during the learning stage to prevail during the validation stage.

Which interpretation, latency minimization, or the stochastic interpretation of (PEM) or (ML) should one prefer? My own preference lies squarely with the latency minimization. The main argument is that, in my opinion, *deterministic* approximation articulates in a much more satisfying way the essence of the SYSID problem. It seems to me that the lack of fit between the observed data and the chosen model will in most practical situations (both in engineering and in economics) be due to the fact that the data has been produced by a nonlinear, high order, time varying system, influenced by spurious inputs and other unknown interactions with the environment. Minimization of the latency or the misfit pointedly articulates the fitting problem. Stochastic methods assume that the lack of fit is due to some unobserved input *with stochastic regularity*. It is unclear how to justify such an assumption, let alone the statistical independence of ε and the driving input component of u, etc. Furthermore, it is awkward how to deal with the additive noise. Obviously, in many applications this noise term cannot be justified when the system is at rest, it is incompatible with

physical properties as dissipativeness, it often assumes an infinite supply of energy, etc. In typical applications it is simply artificial to insist on a stochastic interpretation, even as an approximation. It is even hard to imagine how such stochastic regularity could come about in reality. The whole issue is obviously a very complex one, that can not be discussed without examining the interpretation of probability.

8 Interpretations of probability

There is scarcely any field of science for which the foundations and the interpretation have stirred as much controversy and debate as probability. It shares this doubtful honor with the rules of inference (*tertium non datur?*) and with quantum mechanics, but even there the interpretation question is to a large extent tied up with the meaning of probability. It is ironic that both quantum mechanics and probability, perhaps the most successful scientific developments of the twentieth century, appear to have such shaky foundations.

The foundations and interpretation of probability have been discussed by some of the finest scientific minds in history. Jakob Bernoulli, Borel, Carnap, de Finetti, Jeffreys, Fréchet, Keynes, Kolmogorov, Laplace, von Mises, Poincaré, Popper, Ramsey, Venn, are only a few of the many scientists who have written books on this topic (see [8, 14, 32] for the history and the status of this subject). In this section, I describe my own impressions about this, briefly, and perhaps too superficially. The subject area is quite subtle, unsettled, and very much *la mer à boire*.

Four main views have emerged, in combination with a seemingly uncountable number of intermediate nuances. The first two interpretations are considered objective, empirical, physical. The first one is the *relative frequency* point of view, formalized by von Mises [31]. Popper [25, 26] devised a second view, in which probability is interpreted as *'propensity'*. The third and fourth main interpretations are epistemological and lead to probability as *degree of belief*. They often go under the name 'Bayesian', but this certainly does not nail down the viewpoint uniquely [13]. The third interpretation, called the 'logical' theory, deduces probability from relations among basic propositions. Keynes [19] is usually considered the person who introduced this approach, later put in a more definitive form by Carnap [3]. The fourth interpretation, championed by De Finetti [5], is radically subjective: probability articulates one's own personal degree of belief. Nothing more, nothing less. Between these interpretations, we find many subtle nuances leading, in the words of Galavotti [8], to *"a whole array of seemingly irreconcilable perspectives"*.

The relative frequency interpretation assumes that we have an ensemble of objects, and that each individual object has certain characteristics, which differ from object to object. This characteristic then becomes a random variable. For example, if a characteristic x is real-valued, then its distribution function $F(x)$ is equal to the relative frequency of objects for which the characteristic is less than or equal to x. If we apply this to a well-defined existing finite ensemble and a well-defined set of characteristics, there seems no apparent interpretational difficulty. However, for

applications in which the ensemble involves events that are still to come, as is unavoidable in time-series analysis (and hence in stochastic SYSID), or that are only potentially realized (flips of a coin, throws of a die), or are too numerous to be realizable (bridge hands), or when the ensemble is infinite, the frequency interpretation poses severe interpretation problems.

Popper formulated the propensity interpretation in order to accommodate for the fact that quantum mechanics asks for a physical interpretation of a single event probability. In this interpretation, certain physical objects, even if prepared in *identical* ways and operating in *identical* environments, may react differently, with outcomes that have fixed probabilities. This probability will then be the relative frequency obtained from repeated identical trials. Since, in Feynman's words, *nobody* understands quantum mechanics, I must (reluctantly) accept that this view may be appropriate, perhaps unavoidable, for accommodating the present day orthodox view of quantum mechanics. But I find it hard to accept propensity as a description of our day to day (seemingly) random experiences. However, some physicists appear to suggest that the propensity interpretation may even apply to mundane things like coin tossing, and there are many publications concerning the transition from 'determinism' to 'randomness' of a wobbling coin. In fact, recently experiments have been conducted which show [6] that a coin, repeatedly flipped by means of a carefully constructed mechanical device, lands practically each time on the same side. These experimenters summarize their results by stating – ironically, I hope – *"We conclude that coin tossing is 'physics', not 'random'"*, as if there was a reason to expect otherwise. These experiments have been avidly reported in the popular press, whose main concern seems to be that the toss at the beginning of a football match may not be fair.

The logical theory is based on rules of inductive logic, leading to relations between propositions, for example that certain hypotheses make other hypotheses more probable. Or that certain basic 'symmetric' events (as any pair of cards being on top of a well-shuffled deck) demand equal probabilities. This is useable for games of chance, with discrete outcomes, where probability quickly reduces to combinatorics. However, it is hard to apply this principle to continuous random variables, as the many discussions surrounding the principles of indifference or insufficient reason demonstrate.

The personal belief interpretation is in a sense the clearest of them all. By stating a probability, an expert quantifies his or her degree of belief in the occurrence of an event. The difficulty is to describe how these degrees of belief are arrived at, certainly when it concerns complex phenomena, in turn influenced by many uncertainties, leading to a seemingly infinite regress. Think, for example, of the problem of trying to arrive, in a systematic way, at the probability as a degree of belief of the Dutch soccer team winning the coming World Cup. It is clear what probability as degree of belief means, but it is difficult to obtain it in a scientific manner, let alone to falsify. And, of course, it leads to the question why *I* should care about *your* degree of belief.

There are frequentists who maintain that the frequency interpretation is the only viable one. Propensity states a law about a limiting frequency which is yet to be realized, belief expresses the relative frequency of what would happen if a situation presented itself over and over again. For example, when we say that the probability

of the Dutch team winning the World Cup is $\frac{1}{2}$, a frequentist would say that we mean that if this tournament were repeated indefinitely, under identical circumstances, the team would be victorious in 50% of the repeats. However, it is hard to take the latter earnestly, in part because, in my opinion, identical circumstances lead to identical outcomes. I am more sympathetic with Carnap's admission of two concepts of probability [2], relative frequency on the one hand, and belief on the other. And I am also sympathetic with complaints that it is simply confusing and bad mathematical etiquette to use the same term 'probability' for distinct ideas as 'frequency' and 'belief'.

An important point that is unfortunately all too often absent or insufficiently emphasized in discussions about mathematical models is the issue of *open* versus *closed* systems. In open systems, we aim for a law that describes the relation between system variables, but does not go to the point of stating what will actually happen. The interaction with the (unmodeled) environment is explicitly part of the model, but what the environment will be, is not specified. The paradigmatic example of an open system is an input/output system, which explains the output in terms of the input, but leaves the input as imposed by the environment, unexplained and unspecified. The paradigmatic example of a closed system is an autonomous system. Systems and control theory, signal processing, and computing science seem to be the only areas which clearly address open systems. Mathematicians and physicists, for example, unfortunately usually end up viewing dynamical systems as closed systems.

The very definition of a random variable or a random process suggests a closed system, and the influence of the environment is not explicitly taken into account in the mathematical set-up. Compare this, for example, with our notion of a (deterministic) behavior, which clearly accounts for the interaction with the environment.

The literature on coin tossing or dice throwing, often viewed as the paradigmatic examples of randomness, suffers from a lack of understanding of the fundamental difference between open and closed systems. All too often, it is suggested that the randomness in a coin toss or a dice throw comes from the coin or the die itself. For example, in the textbook by Cramér [4, page 154], it is stated *"The numbers* $p_{\mathtt{r}}$ [the probability that the die shows outcome `r`] *should in fact be regarded as physical constants of the particular die that we are using [⋯] experience shows that for a well-made die the frequency of any event* `r` *in a long series of throws usually approaches 1/6 [⋯]"*. Obviously, in dice throwing, the source of randomness, if present, comes not from the die, but from the actions of the dice thrower (including the way the die is prepared for a throw). In other words, we are dealing with an open, not a closed system. So, it is absurd to insist that the probability of the outcomes is a property of the die, well-made or not.

When adherents of the propensity interpretation state (see Hacking [14, page 14]) *"The propensity to give heads is as much a property of the coin as its mass, and the stable long run frequency found on repeated trials is an objective fact of nature independent of anyone's knowledge of it, or evidence for it"*, they are stating something about the *environment* of the coin. And this something is either *untrue* or *trivially true*. If the coin is experimented with in a fixed environment, i.e. if it is tossed repeatedly in exactly the same way, the coin will each time fall on the

same side, in accordance with the laws of mechanics and as the experimenters in [6] have demonstrated. Whether it shows always heads or always tails, depends on the experimental set-up – the environment – and on the physical properties of the coin. Some fixed environments will give a run with all heads, some will give a run with all tails. For some experimental set-ups, it is trivial to predict which run will happen, for others, it may be difficult – perhaps impossible, given the present state of knowledge, sensor technology, and computer power – to predict what will happen *before* the toss, but that we will see a fixed run is certain. However, if the coin is experimented with in an environment that changes from trial to trial, then the coin may sometimes fall on the one side and sometimes on the other, each time in accordance with the laws of mechanics. The relative frequency (including the question of whether the limit exists) depends of how the experimenter changes the environment from trial to trial. There need not be a stable long run frequency found on repeated trials, by properly selecting the trials, any relative frequency can be found, and so it is not an objective fact of nature. If we see a stable long run frequency for the coin, it is only because we choose elements in the different trials according to some stable long run frequency, or better, according to some kind of randomized law that generates a stable long run frequency. The fact that needs to be explained, thou, is why and when a little bit of randomness introduced in the trials generates a robust stable long run frequency in the outcomes. It looks reasonable to assume that untrained, uncoordinated individual who enthusiastically throws a well-made die will produce a stable long frequency, but this is more a matter of belief than an objective fact of nature. Or, worse yet, a truism in the sense it is by definition what is meant by untrained, uncoordinated, enthusiastic throws.

The many experiments that have been carried out to prove that coin tossing is random at best serve to prove that the coin transmits, and perhaps amplifies or regularizes, randomness that is external to the coin. So, one may as well use a random number generator to drive a mechanism that tosses the coin, avoid the conclusion that *coin tossing is 'random', not 'physics'*, but instead opt for the punchline that *a random number generator generates random numbers*. It is also instructive to deconstruct *Buffon's needle* or *Bertrand's 'paradox'* as originating by trying to explain the outcomes of variables of an open system using the limited language of closed systems. Of course, we get different probabilities if we throw a needle using different mechanisms. There is nothing paradoxical in this whatsoever.

This also applies to stochastic processes. As such, I find ARMAX systems (with the exogenous variables left unexplained) a much more convincing model class (e.g. for econometric modeling) than ARMA systems.

9 Stochastics in SYSID

In engineering education in general, and in systems and control in particular, the issue of the interpretation of probability is invariably glossed over completely. At best, there is some passing reference to the relative frequency interpretation. The mathematical structure and algorithms invariably take the central place in textbooks. The

saving grace for this nonchalance is the fact that the mathematical measure theoretic framework appears to be useable by and large regardless of the interpretation.

The extent to which the interpretation of probability is relevant in the use of probabilistic models and methods depends, in my opinion, to a large measure on the field of application. In descriptive sciences as physics and economics, it is indispensable to articulate the interpretation used. In mathematics, the interpretation usually does not really matter at all. Deep and useful theorems have been derived about the mathematical theory of probability without worrying about the interpretation.

In engineering (and prescriptive aspects of economics) one can, it seems to me, take the following intermediate position. An algorithm-based engineering device, say in signal processing, communication, or control, comes with a set of 'certificates', i.e. statements that guarantee that the device or the algorithm will work well under certain specified circumstances. These circumstances need not be the ones under which the device will operate in actual practice. They may not even be circumstances which can happen in the real world. These certificates are merely quality specifications. Examples of such performance guarantees may be that an error correcting code corrects an encoded message that is received with on the average not more than a certain percentage of errors, or that a filter generates the conditional expectation of an unobserved signal from an observed one under certain prescribed stochastic assumptions, or that a controller ensures robust stability if the plant is in a certain neighborhood of a nominal one, etc.

Which interpretation of probability prevails in SYSID? Of course, in this case the aim is to model reality, and one cannot avoid the interpretation issue all together. But one can view SYSID as a design of algorithms, and as such, it is reasonable to test a proposed algorithm under certain 'idealized' assumptions, i.e. for a 'test case' in order to obtain a certificate, in the spirit of what has been explained in the previous paragraph. For example, one could assume that (u, y, ε) in $(\mathfrak{E}')$ is a stationary ergodic gaussian process with u and ε independent and satisfying $(\mathfrak{E}')$ for some (P, Q, M) from a model class, and to examine how the algorithm that chooses an (P, Q, M) on the basis of $\tilde{w}_{\mathtt{T}}$ performs for $\mathtt{T} \to \infty$. This, of course, points to a relative frequency interpretation. As a 'testbed' this has appealing features: it is well-founded mathematically and algorithmically, and the probabilistic structure will have an 'averaging effect' on the data, avoiding excessive influence of unusual features. As a testbed, there is nothing wrong with this, to the contrary. As long as one avoids the claim that it yields an (approximate) description of the reality that has produced the data. However, it is just as reasonable to test the algorithms for signals (w, ε) for which the correlations like

$$\lim_{\mathtt{T}\to\infty} \frac{1}{\mathtt{T}} \sum_{\mathtt{t}'=1}^{\mathtt{T}} \begin{bmatrix} \sigma^{\mathtt{t}} w(\mathtt{t}'+\mathtt{t}) \\ \sigma^{\mathtt{t}} \varepsilon(\mathtt{t}'+\mathtt{t}) \end{bmatrix} \begin{bmatrix} w(\mathtt{t}') \\ \varepsilon(\mathtt{t}') \end{bmatrix}^{\top}$$

exist for certain $\mathtt{t} \in \mathbb{N}$. Or demand good SYSID features for certain classes of signals $(w, \varepsilon) \in \ell_2(\mathbb{N}, \mathbb{R}^{\mathtt{w}+\mathtt{e}})$, thereby avoiding stochastic complexities.

The relation of probability with the real world is a serious problem that should be addressed in SYSID, and it is a pity that attempts to put this on the systems theory agenda [17] have fallen on deaf ears.

The development of probability as limiting frequency owes much to John Venn, best known to us through the 'Venn diagrams' for testing syllogisms. John Venn (1834–1923) was president of Gonville and Caius college, fortuitously the same college which will host our symposium dinner. Venn wrote the first book [33] devoted to the frequency interpretation of probability. He complains that *"probability has been very much abandoned to mathematicians, who as mathematicians have generally been unwilling to treat it thoroughly. In the work of mathematicians we would search in vain for anything like a critical discussion of the fundamental principles upon which its rules rest, the class of enquiries to which it is most properly applicable, or the relation it bears to logic and the general rules of inductive evidence"*. Little seems to have changed in this respect since these words were written, except that the unwillingness to which John Venn refers has now been extended beyond mathematics. It certainly includes SYSID. None of my favorite books on the subject [15, 18, 21, 29, 30] discusses this issue at all, except [21, page 22], where it is merely stated that *"information about past disturbances could, however, be important for making qualified guesses about future values. It is therefore natural to employ a probabilistic framework to describe future disturbances"*. Surely, an important area of system theory with an ambition to be widely applicable in engineering practice deserves a more balanced justification for the mathematical modeling framework in which it chooses to operate.

Acknowledgments

This research is supported by the Belgian Federal Government under the DWTC program Interuniversity Attraction Poles, Phase V, 2002–2006, Dynamical Systems and Control: Computation, Identification and Modelling, by the KUL Concerted Research Action GOA AM-BioRICS, and by several grants the Flemish Fund for Scientific Research: ICCoS, ANMMM, MLDM.

I would also like to thank Ivan Markovsky (K.U. Leuven), Paolo Rapisarda (Un. of Southampton), and Bart De Moor (K.U. Leuven) for their collaboration on many aspects of the problems discussed in this paper, and Gert de Cooman (Ghent Un.) and Manfred Deistler (T.U. Vienna) for remarks concerning the latter sections of this article.

References

1. H. Akaike, Markovian representation of stochastic processes by canonical variables, *SIAM Journal on Control*, volume 13, pages 162–173, 1975.
2. R. Carnap, The two concepts of probability, *Philosophy and Phenomenological Research*, volume V, pages 513–532, 1945.
3. R. Carnap, *Logical Foundations of Probability* , Chicago University Press, 1950.
4. H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946.
5. B. De Finetti, *Theory of Probability, Volume 1 and 2*, Wiley, 1974.
6. P. Diaconis, S. Holmes and R. Montgomery, Dynamical bias in the coin toss, manuscript, 2004.

7. J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, State-space solutions to standard $\mathscr{H}_2$ and $\mathscr{H}_\infty$ control problems, *IEEE Transcations on Automatic Control*, volume 43, pages 831–847, 1989.
8. M. C. Galavotti, *Philosophical Introduction to Probability*, CSLI Publications, 2005.
9. K. Glover and J. C. Willems, On the stability of numerical integration routines for ordinary differential equations, *Journal of the IMA*, volume 11, pages 171–180, 1973.
10. K. Glover, *Structural Aspects of System Identification*, Ph.D. dissertation, M.I.T., 1973.
11. K. Glover and J. C. Willems, Parametrizations of linear dynamical systems: Canonical forms and identifiability, *IEEE Transcations on Automatic Control*, volume 19, pages 640–646, 1974.
12. K. Glover, All optimal Hankel-norm approximations of linear multivariable systems and their $\mathscr{L}_\infty$ error bounds, *International Journal of Control*,volume 39, pages 1115–1193, 1984.
13. I. J. Good, 46656 varieties of Bayesians, *American Statistician*, volume 25, pages 62–63, 1971. Reprinted in *Good Thinking. The Foundations of Probability and its Applications*, University of Minnesota Press, pages 20–21, 1983.
14. I. Hacking, *The Emergence of Probability*, Cambridge University Press, 1975.
15. E. J. Hannan and M. Deistler, *The Statistical Theory of Linear Systems*, Academic Press, 1979.
16. C. Heij, *Deterministic Identification of Dynamical Systems*, Lecture Notes in Control and Information Sciences, volume 127, Springer Verlag, 1989.
17. R. E. Kalman, Randomness reexamined, *Modeling, Identification and Control*, volume 15, pages 141–151, 1994.
18. T. Katayama, *Subspace Methods for System Identification*, Springer Verlag, 2005.
19. J. M. Keynes, *A Treatise on Probability*, MacMillan, 1963 (originally published in 1921).
20. M. Kuijper, An algorithm for constructing a minimal partial realization in the multivariable case, *Systems & Control Letters*, volume 31, pages 225–233, 1997.
21. L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, 1999.
22. I. Markovsky, J. C. Willems, S. Van Huffel, and B. L. M. De Moor, *Exact and Approximate Modeling of Linear Systems: A Behavioral Approach*, SIAM Publications, to appear.
23. I. Markovsky, J. C. Willems, and B. L. M. De Moor, Comparison of identification algorithms on the database for system identification DAISY, Report 05–227, Department of Electrical Engineering, K.U. Leuven, 2005.
24. J. W. Polderman and J. C. Willems, *Introduction to Mathematical Systems Theory: A Behavioral Approach*, Springer Verlag, 1998.
25. K. R. Popper, The propensity interpretation of probability, *British Journal for the Philosophy of Science*, volume X, pages 25–42, 1959.
26. K. J. Popper, *A World of Propensities*, Thoemmes, 1990.
27. P. Rapisarda and J. C. Willems, State maps for linear systems, *SIAM Journal on Control and Optimization*, 35, pages 1053–1091, 1997.
28. B. Roorda, Algorithms for global least squares modelling of finite multivariable time series, *Automatica*, volume 31, pages 391–404, 1995.
29. T. Söderström and P. Stoica, *System Identification*, Prentice Hall, 1989.
30. P. Van Overschee and B. L. M. De Moor, *Subspace Identification for Linear Systems: Theory, Implementation, Applications*, Kluwer Academic Press, 1996.
31. R. von Mises, *Warscheinlichkeit, Statistik und Warheit*, Springer, 1928.
32. J. von Plato, *Creating Modern Probability*, Cambridge University Press, 1994.
33. J. Venn, *The Logic of Chance*, MacMillan, 1866.

34. M. Verhaegen, Identification of the deterministic part of MIMO state space models given in innovations form from input-output data, *Automatica*, volume 30, pages 61–74, 1994.
35. J. C. Willems, From time series to linear system — Part I. Finite dimensional linear time invariant systems, Part II. Exact modelling, Part III. Approximate modelling, *Automatica*, volume 22, pages 561–580 and 675–694, 1986, volume 23, pages 87–115, 1987.
36. J. C. Willems, Fitting data sequences to linear systems, *Systems and Control in the Twenty-First Century* (Edited by C. I. Byrnes, B. N. Datta, D. S. Gilliam, and C. F. Martin), Birkhäuser, pages 405–416, 1996.