WORKING PAPER

A DETERMINISTIC APPROACH TO APPROXIMATION MODELLING

C. Heij
J.C. Willems *

October 1989
WP-89-076

IIASA
International Institute
for Applied Systems Analysis

A DETERMINISTIC APPROACH TO APPROXIMATION MODELLING

C. Heij
J.C. Willems *

October 1989
WP-89-076

Econometrics Institute, Erasmus University Rotterdam, The Netherlands
* Department of Mathematics, Groningen University, The Netherlands

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

FOREWORD

This is a contribution to the activity on the topic *From Data to Model* initiated at the Systems and Decision Sciences Program of IIASA by Professor J. C. Willems.

A. Kurzhanski
Program Leader
System and Decision Sciences Program.

A DETERMINISTIC APPROACH TO APPROXIMATE MODELLING

C. HEIJ AND J.C. WILLEMS

Abstract

In this paper we will describe a deterministic approach to time series analysis. The central problem consists of approximate modelling of an observed time series by means of a deterministic dynamical system. The quality of a model with respect to data will depend on the purpose of modelling. We will consider the purpose of description and that of prediction. We define the quality by means of complexity and misfit measures, expressed in terms of canonical parametrizations of dynamical systems. We give algorithms to determine optimal models for a given time series and investigate some consistency properties. Finally we present some simulations of these modelling procedures.

Keywords

Approximate modelling, time series analysis, dynamical systems, canonical forms, complexity, misfit, consistency.

1. INTRODUCTION

1.1. Modelling: specification and identification

The purpose of this paper is to describe a deterministic approach to time series analysis. This means that within the realm "from data to model", we will pay special attention to the case where the data consist of a sequence of observations over time and where the models consist of deterministic dynamical systems. Our approach to this particular modelling problem forms part of a more general modelling philosophy, which we will now describe.

Some of the essential factors which play a role in the problem of modelling data are depicted in figure 1. Two of the main aspects in approaching this problem are *specification* of the problem and, subsequently, *identification* of the model.

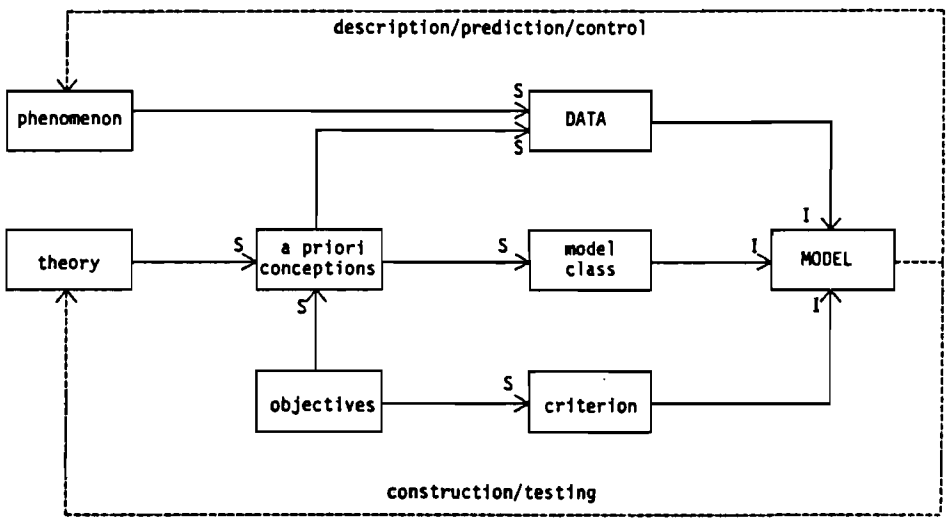

figure 1: modelling (S: specification; I: identification)

In general terms, the problem of modelling *data* consists of constructing a good *model* on the basis of these data. So the class of candidate models, i.e., the *model class*, has to be specified. Moreover, the quality of candidate models for modelling the data has to be assessed. This assessment, by means of a *criterion*, depends on the *objectives* underlying the modelling problem. An *identification procedure* describes the way a model is chosen (identified) from the model class, given the data. The aim is to construct the procedure in such a way that the identified models are of good quality with respect to the data, as measured by the criterion.

So in order to investigate the identification aspect of the data modelling problem it is necessary to specify the model class and the objectives. In modelling problems in general it is not known a priori which data will be included for identification of a model. This leads us to the specification aspect.

Often the primary objective of constructing a model is not only to model the data, but also to model a *phenomenon*. It then is supposed that the data somehow reflect the phenomenon. The phenomenon is then considered as a system which produces the data.

In the specification of the modelling problem one can incorporate prior knowledge concerning the phenomenon. This prior knowledge partly can be given by a *theory* concerning the phenomenon. Apart from this, one will impose restrictions partly based upon the objectives of modelling and partly for convenience. This leads to a collection of *a priori conceptions*, on the basis of which one decides which variables will be included in the model and what models will be considered. The identification problem is then specified.

Some of the main objectives of modelling are given in figure 1. On the one hand, an objective could be to model the phenomenon. One can think of description, prediction or control of the phenomenon. On the other hand, another objective could be to construct or test theories concerning the phenomenon.

It is beyond the scope of this paper to discuss fundamental problems of data, like the relationship between the phenomenon and the data and problems of data collection.

In the practice of modelling one often considers the specification aspect as part of the relevant scientific discipline and the identification aspect as a problem of constructing mathematical procedures. However, especially the choice of the model class also implies prior conceptions of a mathematical nature. The choice between deterministic and stochastic models forms a particular example.

We will illustrate the foregoing general description of the data modelling problem by means of five simple examples.

1.2. Example 1: a resistor

Suppose one wants to describe a resistor. On the basis of physical theory ("Ohm's law") one postulates a linear relationship between the voltage

(V) across and the electrical current (I) through the resistor, i.e., $V = I.R$ with $R \geq 0$ the resistance. A resistor is then described by a model R. So the model class is $\mathbb{R}_+$. To identify R, suppose one performs a number (n) of experiments with resulting voltage and current measurements $(\tilde{V}_i, \tilde{I}_i)$, $i = 1, \ldots, n$. See figure 2.

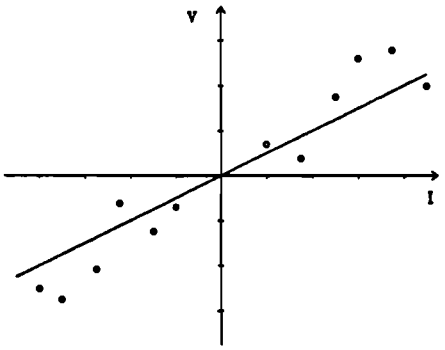

figure 2

The identification problem consists of choosing R on the basis of these data. In general there will exist no R such that $\tilde{V}_i = \tilde{I}_i.R$ for all $i = 1, \ldots, n$. This can be due to inaccurate measurements and to the fact that the linear relationship is an idealization – though it may be an accurate one. A reasonable criterion could be total least squares.

So in this case, in order to describe the resistor, one uses physical theory to specify the model class and the data to be collected.

1.3. Example 2: eye colour

Suppose one wants to predict the colour of the eyes of a person. On the basis of biological theory (genetics) one postulates a specific probabilistic relationship between this colour and the colour of the eyes of the ancestors. Assume that the colour is either brown (1) or blue (0). As model class one could take [0,1], where a particular model $p \in [0,1]$ means that p is the probability that the person has brown eyes. To identify p one collects data on the colour of the eyes of the parents, grandparents and so on. One then identifies p by means of elementary probabilistic calculations. See figure 3.

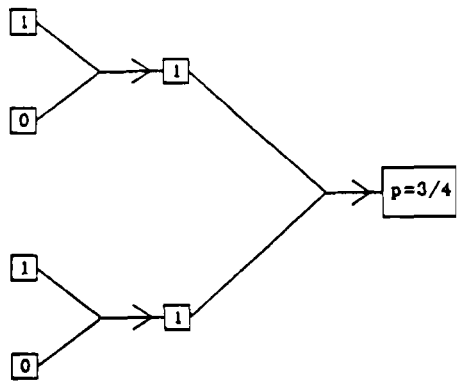

figure 3

One could now make a prediction for example by maximum likelihood, i.e., predict the colour to be brown if and only if $p > \frac{1}{2}$.

So in this case, in order to predict the eye colour, one uses biological theory to specify the identification and prediction problem.

1.4. Example 3: consumption

Suppose one wants to predict the consumption C_{t_0+1} for the coming year. On the basis of an economic theory one postulates that the dominant factor determining C_{t_0+1} is the income Y_{t_0} in the current year. Suppose data for consumption and income, $(\tilde{C}_t, \tilde{Y}_t)$, $t = s, s+1, \ldots, t_0$, are available. For convenience one could postulate an affine relationship between consumption in a year and income in the preceding year. The model class for example

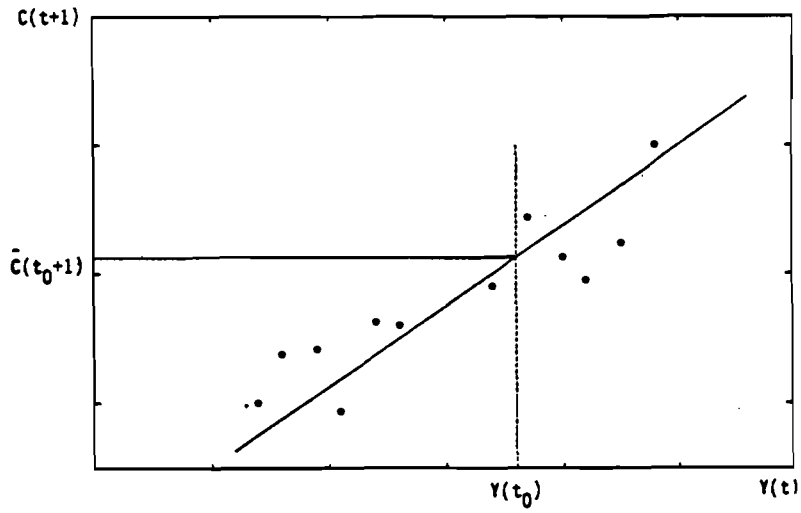

figure 4

could be $\mathbb{R}_+^2$, where the model (a,b) with $a,b \geq 0$ describes the postulated relationship $C_{t+1} = a + b.Y_t$. In order to identify a model one could use the data to estimate a and b for example by means of ordinary least squares. If the resulting estimates $\hat{a},\hat{b}$ indeed are nonnegative, one could predict C_{t_0+1} by means of $\hat{a} + \hat{b}.\tilde{Y}_{t_0}$. See figure 4.

So in this case, in order to predict consumption, one uses economic theory to specify the data. The choice of the model class is entirely a matter of convenience. If the estimated values $\hat{a},\hat{b}$ are not accepted as a reasonable description of consumptive behaviour one is ready to specify a different class of models, e.g., $C_{t+1} = \alpha + \beta.\log Y_t$.

1.5. Example 4: rainfall

Suppose one wants to control the water supply from a reservoir. The water of the reservoir is supplied to customers and replenished by rain. Suppose that one can construct a reasonable control strategy, once the rainfall is modelled.

If the climatological conditions are rather stable the rainfall could be viewed as a stationary stochastic process. As model class one could consider the class of Gaussian ARMA processes. Suppose that rainfall data $\{\tilde{r}(t);\ t_1 \leq t \leq t_2\}$ are available. To identify a model on the basis of these data one could consider the objective of simultaneous prediction of the rainfall for a number of periods in the future.

So in this case, in order to formulate the water supply problem in terms of only the rainfall, one has used prior knowledge of e.g. the demand pattern for water and of (stochastic) control theory. It is assumed that the rainfall can be modelled as a stationary stochastic process. This assumption is of a mathematical nature. It can be supported by arguing that the mechanism producing the rainfall is rather stable. This for example means that, although the rainfall is uncertain, some time averages of the rainfall are less uncertain.

1.6. Example 5: realization

Suppose one wants to interpolate n points $(x_i,y_i) \in \mathbb{R}^2$, $i = 1,...,n$, by means

of a polynomial p of lowest possible degree. So the data consists of n points in $\mathbb{R}^2$ and the model class consists of polynomials. As a criterion to choose p one requires $y_i = p(x_i)$, $i = 1,\ldots,n$, and the degree of p has to be minimal.

So in this case the objective is to give an exact description of the data in a most simple way. This is an example of exact modelling or realization. The concepts of phenomenon or theory do not play a role in the specification of the modelling problem. The criterion is inspired by aesthetics or the desire to give a compact representation of the data.

1.7. Choice of model class

The foregoing examples especially are intended to illustrate the various considerations which can play a role in specifying the model class. In examples 1 and 2 well-established theories are used to choose the model class, one deterministic and the other probabilistic. In example 5 the choice is inspired by aesthetics. In examples 3 and 4 the choice of the model class reflects an aim of simplicity.

One of the crucial elements of the specification of modelling problems is the choice whether the model class should consist of *stochastic* or of *deterministic* models. In examples 1 and 2 the choice is based on a relevant scientific theory. In examples 3 and 4, like in the majority of modelling problems outside of the natural sciences, the choice is inspired by convenience. Moreover, the current practice seems to be to take the model to be stochastic. This implies that one introduces disturbances (noise) to explain the fact that in general the data do not satisfy simple, exact relationships. Moreover, it is nearly invariably assumed that the noise has a stable distribution over time, i.e., the disturbances form a stationary process.

This explanation of the discrepancy between the data and simple (deterministic) relations has two important implications. First, the model error is caused by disturbances of a stable nature, i.e., the relative frequency of the disturbance terms is assumed to be rather constant over time. Second, and based on this, the quality of proposed identification procedures is assessed on the basis of statistical criteria like unbiasedness, consistency and efficiency.

Clearly, this paradigm of stochastics often is a reasonable and convenient one. However, especially for complex phenomena, the fact that

the data do not exactly satisfy simple deterministic relationships is often not due to disturbances or observation noise. Often the phenomenon simply is too complex to be modelled exactly within the model class. The models even deliberately are chosen to be simple. Both for human understanding and for practical implementation a simple, slightly inaccurate model of the phenomenon often is preferred above a complex, more accurate one. The central issue then is not noise or stochastics, but approximation.

1.8 Overview of the paper

To conclude the introduction we give an overview of the contents of the paper.

In section 2 we give a formal framework for approximate modelling, using the concepts of complexity and misfit. We illustrate this framework by some examples which play an important role in the sequel. In section 3 we describe the model class which we will consider in this paper, i.e., the class of deterministic dynamical systems. We will consider the objectives of description and prediction. Corresponding identification procedures are presented in section 6. These procedures solve an optimal approximate modelling problem, defined in terms of a utility of models. This utility depends on complexity and misfit measures, which are described in section 5. The complexity and misfit measures are expressed in terms of canonical representations of dynamical systems. These canonical forms reflect the objectives of description or prediction and are defined in section 4.

Section 7 describes the numerical algorithms corresponding to the modelling procedures of section 6. In section 8 we investigate some of the consistency properties of the procedures. The procedures have a clear optimality property as data modelling procedures. However, consistency analysis deals with the question whether the models identified by a procedure also are good models of the phenomenon. It is assumed that the phenomenon belongs to a certain class of systems, which does not need to coincide with the model class.

Section 9 contains some numerical simulations illustrating the deterministic approximate modelling procedures of section 6. Section 10 concludes the paper by summarizing the main results and indicating some topics of current research.

The main reference for the deterministic approach to approximate modelling as presented in this paper is Willems [15].

2. APPROXIMATE MODELLING

2.1 Complexity, misfit, utility

In the sequel of this paper we restrict attention to the identification aspect of the modelling problem. So we assume that one has specified the objectives of modelling, denoted by π, the model class, denoted by $\mathbf{M}$, and a set of conceivable data, denoted by D.

Definition 2-1 A data modelling *procedure* is a map $P: D \to 2^{\mathbf{M}}$.

In other words, a procedure associates with any data a set of models. Usually $P(d)$ will be a singleton, but it need not be.

The aim now is to construct procedures which are optimal in view of the objectives π. This means that for $d \in D$ the identified model(s) $P(d)$ should, within $\mathbf{M}$, reflect the data in a way which is optimal with respect to π.

A general objective is to construct models which are both simple and accurate. We will assume that the objectives π can be specified by a *complexity map* $c: \mathbf{M} \to C$ and a *misfit map* $\varepsilon: D \times \mathbf{M} \to E$. We assume the spaces C and E to be partially ordered. It is desirable to have models for which both the complexity and the misfit are small. However, these desires in general are competitive. We will therefore assume that π can be expressed by means of a *utility map* $u: C \times E \to U$, with U a partially ordered set. The aim then is to choose a model for which the complexity and misfit are such that the corresponding utility is maximal. For a partial ordering $\leq$ on U, $m \in U' \subset U$ is said to be a maximal element of U' if $\{u' \in U',\ m \leq u'\} \Rightarrow \{u' = m\}$.

Definition 2-2 The procedure $P_u: D \to 2^{\mathbf{M}}$ *corresponding* to the utility $u: C \times E \to U$ is defined by $P_u(d) := \operatorname{argmax}\{u(c(M), \varepsilon(d,M));\ M \in \mathbf{M}\}$ for $d \in D$.

So P_u assigns to data the set of models for which the utility is maximal. This clearly raises questions of existence and unicity of maximal elements.

In the remainder of this section we illustrate this approach by means of several examples. It will turn out that many classical identification procedures can be formalized in this context.

2.2. Exact modelling

In *exact modelling* one does not allow any misfit and wants to minimize the complexity. We consider three examples.

2.2.1. Synthesis problem

As a first example, consider a synthesis problem of electrical circuit theory. Suppose one wants to construct an electrical circuit with one external port with a prescribed current/voltage behaviour B. Here $B \subset (\mathbb{R}^2)^{\mathbb{R}}$ describes which current/voltage trajectories over time at the external port are compatible with the circuit. Moreover, suppose one wants to realize B by means of an RLC-network, i.e., only using resistors, inductors and capacitors. For an RLC-network with one external port, let $B(RLC)$ denote the current/voltage behaviour at the port and let $n(RLC)$ denote the total number of resistors, inductors and capacitors of the network.

The synthesis problem consists of finding an RLC-network with external behaviour B and such that $n(RLC)$ is as small as possible. So one allows no misfit and wants to minimize the complexity, measured by the number of constituent elements. This can be formulated in terms of a utility. Let $D = \mathbb{M}$ consist of the external current/voltage behaviours of RLC-networks with one external port. Define the complexity by $c(B(RLC)) := n(RLC)$ and the misfit by $\varepsilon(B,B') := +\infty$ if $B \neq B'$, $\varepsilon(B,B') := 0$ if $B = B'$. The synthesis problem then corresponds to the utility $u(n,\varepsilon) := -n-\varepsilon$.

2.2.2. Undominated unfalsified modelling

Let S be a set and let the set of conceivable data consist of finite tuples of observations in S, i.e., $D := \cup\{S^n; n \geq 1\}$. Let a model M consist of a subset $M \subset S$ and let $\mathbb{M} \subset 2^S$ denote a class of models.

A model M is called ***unfalsified*** by a measurement $d \in D$ if $d \subset M$. A model M is called ***undominated*** unfalsified in $\mathbb{M}$ for d if $d \subset M \in \mathbb{M}$ and $\{d \subset M' \in \mathbb{M}, M' \subset M\} \Rightarrow \{M' = M\}$. Define $P(d)$ as the collection of undominated unfalsified models in $\mathbb{M}$ for d. So P models d by models which are as small as possible in the sense of set inclusion. This could be expressed by means of the following utility. Let $\varepsilon(d,M) := 1$ if $d \not\subset M$, $\varepsilon(d,M) := 0$ if $d \subset M$ and define $c(M) := M$. Let $\underline{u} \notin \mathbb{M}$, $U := \mathbb{M} \cup \{\underline{u}\}$ and define the utility by $uu(M,1) := \underline{u}$ and $uu(M,0) := M$. Define a partial ordering $\leq$ on U as follows: $\underline{u} \leq M$ for all $M \in \mathbb{M}$ and for $M_1, M_2 \in \mathbb{M}$, $M_1 \leq M_2$ if and only if $M_1 \supset M_2$. Then P coincides with the

procedure P_{uu} corresponding to the utility uu.

A special case of this arises if $S=(\mathbb{R}^q)^{\mathbb{Z}}$, so the data consists of a finite number of infinite time series in q real-valued variables. We will briefly return to this case in section 3.2. For a more thorough discussion we refer to Willems [16]. Here we only discuss a particular instance, known as the *minimal realization* problem.

In the minimal realization problem of linear systems theory the data set is $D=(\mathbb{R}^{p\times m})^{\mathbb{N}}$, where $\mathbb{N}:=\{1,2,3,\dots\}$. In this case the data $d\in D$ consists of an (impulse response) sequence $(G_k;\ k\in\mathbb{N})$ with $G_k\in\mathbb{R}^{p\times m}$, $k\in\mathbb{N}$. The model set consists of triples (A,B,C) with $A\in\mathbb{R}^{n\times n}$, $B\in\mathbb{R}^{n\times m}$, $C\in\mathbb{R}^{p\times n}$ for some $n\in\mathbb{N}$. The triple (A,B,C) is called a realization of $(G_k;\ k\in\mathbb{N})$ if $CA^{k-1}B=G_k$ for all $k\in\mathbb{N}$. It is called a minimal realization if n is as small as possible. For $d=(G_k;\ k\in\mathbb{N})$ and $M=(A,B,C)\in\mathbb{R}^{n\times n}\times\mathbb{R}^{n\times m}\times\mathbb{R}^{p\times n}$ define the misfit by $\varepsilon(d,M):=0$ if M is a realization of d and $\varepsilon(d,M):=1$ otherwise. Moreover define the complexity of M by $c(M):=n$. Let $U:=\{-1,-2,-3,\dots\}\cup\{-\infty\}$. Define a utility by $u(n,1):=-\infty$ and $u(n,0):=-n$ for $n\in\mathbb{N}$. The procedure corresponding to this utility solves the minimal realization problem. The number n has the interpretation of the dimension of the state space. In case a solution exists, it is unique up to a choice of a basis in the state space. See e.g. Kalman, Falb and Arbib [7].

2.2.3. Minimum description length principle

As a final example of exact modelling we mention the minimum description length principle of Rissanen, see e.g. Rissanen [14]. In this case the data set D consists of finite sequences of (finite precision) real numbers. The model class $\mathbf{M}$ consists of finite sequences of binary digits. A model represents data exactly by means of an injective code $C:D\to\mathbf{M}$. It is assumed that C codes the data d by means of an auxiliary (countable) class $\mathbf{P}=\{P_\theta;\ \theta\in\Theta\}$ of probability distributions on D, in the following way. The binary sequence $C(d)$ consists of an initial part describing the parameter θ and a remaining part describing the data in a way which is optimal in P_θ (minimum mean description length code for P_θ).

The complexity of a model is defined as the length of the binary sequence. Given the class $\mathbf{P}$, the minimum description length principle corresponds to the procedure which consists of coding the data by means of the shortest possible binary string, i.e., by the model of least complexity. This minimum description length principle balances the desire

for a small number of parameters (in θ) and a simple description of the data by means of P_θ (maximal likelihood). It is interesting to note that this approach gives a deterministic interpretation, in terms of exact modelling, of e.g. maximum likelihood estimation and modelling by means of minimizing prediction errors.

2.3. Minimal complexity, given tolerated misfit

Suppose that the complexity space C and the misfit space E both are totally ordered. We denote the orderings by $\leq$. A possible reconciliation between the objectives of low complexity and of low misfit is to specify a *maximal tolerated misfit* and to minimize the complexity under this constraint. Given $\varepsilon_{tol} \in E$, we define the utility $u_{\varepsilon_{tol}}$ as follows. Let $\underline{u} \notin C \times E$ and $U := (C \times E) \cup \{\underline{u}\}$. For $\varepsilon \geq \varepsilon_{tol}$ let $u_{\varepsilon_{tol}}(c,\varepsilon) := \underline{u}$, and for $\varepsilon < \varepsilon_{tol}$ $u_{\varepsilon_{tol}}(c,\varepsilon) := (c,\varepsilon)$. On U we impose the following total ordering: $\underline{u} < (c,\varepsilon)$ for all $(c,\varepsilon) \in C \times E$, and $(c_1,\varepsilon_1) < (c_2,\varepsilon_2)$ if $c_1 > c_2$ or if $c_1 = c_2$ and $\varepsilon_1 > \varepsilon_2$. So misfits of ε_{tol} or higher are not allowed. Further, models of low complexity are preferred, and for models of equal complexity low misfit is preferred. The procedure $P_{\varepsilon_{tol}}$ now is defined as the procedure corresponding to $u_{\varepsilon_{tol}}$.

Definition 2-3 $P_{\varepsilon_{tol}}(d) := \text{argmax}\{u(c(M),\ \varepsilon(d,M));\ M \in \mathbf{M}\}$, where $\{u(c_1,\varepsilon_1) = u(c_2,\varepsilon_2)\}$: $\leftrightarrow$ $\{\varepsilon_1,\varepsilon_2 \geq \varepsilon_{tol}$ or $(c_1,\varepsilon_1) = (c_2,\varepsilon_2)\}$ and $\{u(c_1,\varepsilon_1) < u(c_2,\varepsilon_2)\}$: $\leftrightarrow$ $\{\varepsilon_1 \geq \varepsilon_{tol} > \varepsilon_2$, or $\varepsilon_1,\varepsilon_2 < \varepsilon_{tol}$, $c_1 > c_2$, or $\varepsilon_1,\varepsilon_2 < \varepsilon_{tol}$, $c_1 = c_2$, $\varepsilon_1 > \varepsilon_2\}$.

Two of the procedures described in section 6 are of this type. These procedures are based upon the ones which will be presented in sections 2.6 and 2.7.

The procedure corresponding to the requirement $\varepsilon \leq \varepsilon_{tol}$ (instead of $\varepsilon < \varepsilon_{tol}$) will be denoted by $\bar{P}_{\varepsilon_{tol}}$.

Here we illustrate the approach by a simple geometric example.

Let D consist of the bounded convex subsets of $\mathbf{R}^2$ and $\mathbf{M}$ of the convex polyhedral subsets of $\mathbf{R}^2$. For $M \in \mathbf{M}$ define the complexity $c(M)$ as the number of extremal points of M. For $C \in D$ and $M \in \mathbf{M}$ define the misfit $\varepsilon(C,M)$ as the Lebesgue measure of the symmetric difference $(C \setminus M) \cup (M \setminus C)$. Let ε_{tol} be given. Then $P_{\varepsilon_{tol}}$ models C by means of the convex hull of a minimal number of points under the misfit restriction, and chooses among solutions those with minimal misfit. See figure 5 for an illustration.

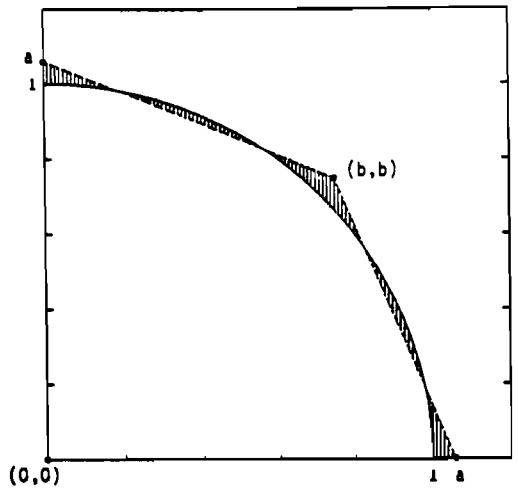

figure 5: $C=\{(x,y)\in\mathbb{R}^2; x^2+y^2\le 1,\ x\ge 0,\ y\ge 0\}$, $\varepsilon_{tol}=0.05$; $P_{\varepsilon_{tol}}(C)$ is convex hull of $(0,0)$, $(0,a)$, $(a,0)$ and (b,b), with $a:=2(\alpha^2+1)^{1/2}/(4\alpha^2+1)^{1/2}$ and $b:=\alpha a/(1+\alpha)$, where $\alpha:=\tan(\frac{3}{8}\pi)$

Another example is speech processing. Let S denote the set of binary strings of finite length. The problem is to code, transmit and decode a signal $s\in S$ in the simplest way possible, given a tolerated misfit and an auxiliary class of models $\mathsf{M}_{aux}\subset S$. A coder is a map $f:S\to\mathsf{M}_{aux}\times S$ transforming a signal s into a transmitted signal $t\in S$. The signal t consists of an initial part describing the auxiliary model and a remaining part describing the signal s in an approximate way by means of the auxiliary model. A decoder is a map g: $\mathsf{M}_{aux}\times S\to S$ transforming a signal t into a decoded signal $\hat{s}$. See figure 6.

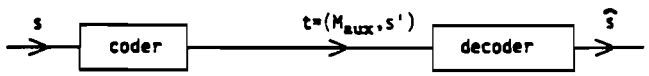

figure 6

For example, M_{aux} could be chosen to be the (set of parameters of the) class of autoregressive systems. The initial part of t then describes the order and the numerical values of the parameters of the auxiliary system. The remaining part of t could be used to describe the prediction errors of the estimates generated by the auxiliary system with respect to the signal

s. The decoder could construct a signal $\hat{s}$ based upon the estimates generated by the auxiliary system and the transmitted prediction errors. See e.g. Jayant and Noll [6].

Here the set of conceivable data is $D=S$ and the model class is $\mathbb{M}=\mathbb{M}_{aux}\times S$. Define the complexity of a model $t\in\mathbb{M}_{aux}\times S$ as the length of the string t. Let $\delta(s,\hat{s})$ denote a measure of the error of $\hat{s}$ with respect to s. Define the misfit of a model $t=(M_{aux},s')$ with respect to data s by $\varepsilon(s,(M_{aux},s')):=\delta(s,\hat{s})$ where $\hat{s}:=g(M_{aux},s')$. Given a tolerated misfit, one wants to minimize the complexity of the transmitted signal, i.e., of the model.

This approach resembles the minimum description length principle, though in speech processing it is not required that the data can be reconstructed exactly from the transmitted signal.

2.4. Minimal misfit, given tolerated complexity

Again suppose that C and E are totally ordered. Another possible reconciliation between the objectives of low complexity and of low misfit is to specify a *maximal tolerated complexity* and to minimize the misfit under this constraint. Given $c_{tol}\in C$, we define the utility $u_{c_{tol}}$ as follows. Let $\underline{u}\notin C\times E$ and $U:=(C\times E)\cup\{\underline{u}\}$. For $c>c_{tol}$ let $u_{c_{tol}}(c,\varepsilon):=\underline{u}$, and for $c\le c_{tol}$ define $u_{c_{tol}}(c,\varepsilon):=(c,\varepsilon)$. On U we impose the following total ordering: $\underline{u}<(c,\varepsilon)$ for all $(c,\varepsilon)\in C\times E$, and $(c_1,\varepsilon_1)<(c_2,\varepsilon_2)$ if $\varepsilon_1>\varepsilon_2$ or if $\varepsilon_1=\varepsilon_2$ and $c_1>c_2$. So a complexity above c_{tol} is not allowed. Further, models of low misfit are preferred, and for models of equal misfit low complexity is preferred. The procedure $P_{c_{tol}}$ now is defined as the procedure corresponding to $u_{c_{tol}}$.

Definition 2-4 $P_{c_{tol}}(d):=\operatorname{argmax}\{u(c(M),\varepsilon(d,M));\ M\in\mathbb{M}\}$, where $\{u(c_1,\varepsilon_1)=u(c_2,\varepsilon_2)\}:\leftrightarrow\{c_1,c_2>c_{tol}$ or $(c_1,\varepsilon_1)=(c_2,\varepsilon_2)\}$ and $\{u(c_1,\varepsilon_1)<u(c_2,\varepsilon_2)\}:\leftrightarrow\{c_1>c_{tol}\ge c_2,$ or $c_1,c_2\le c_{tol},\ \varepsilon_1>\varepsilon_2,$ or $c_1,c_2\le c_{tol},\ \varepsilon_1=\varepsilon_2,\ c_1>c_2\}$.

Again two of the procedures described in section 6 are of this type, along with procedures presented in sections 2.6 and 2.7.

Returning to the geometrical example of section 2.3, suppose c_{tol} is given. Then $P_{c_{tol}}$ models C by means of the convex hull of at most c_{tol} points in such a way that the resulting measure of the symmetric difference

is minimal. Among solutions it chooses those with minimal number of extremal points. It can be shown that the last step in fact never will be invoked.

In the next section we give another example of modelling with given tolerated complexity.

2.5. Simultaneous equation models

We consider a modelling procedure which is sometimes followed in macro-econometrics and other disciplines dealing with complex dynamical phenomena. See e.g. Maddala [12].

Suppose one wants to describe the relationship between two groups of variables, one consisting of n_1 variables collected in $x \in \mathbb{R}^{n_1}$ and the other consisting of n_2 variables collected in $y \in \mathbb{R}^{n_2}$. For example, x could consist of the values of n_1 variables of interest at time t and y of values of these and possibly some other, auxiliary variables at times $s < t$.

Suppose one wants to use linear models. In general, no simple linear relationship will be exactly satisfied by the data. It is assumed that this misfit can be adequately modelled by means of a (Gaussian) disturbance term.

The model class of simultaneous equation models in this case can be parametrized by $\{(A,B,\Sigma);\ A \in \mathbb{R}^{n_1 \times n_1}$ nonsingular, $B \in \mathbb{R}^{n_1 \times n_2}, \Sigma \in \mathbb{R}^{n_1 \times n_1}, \Sigma = \Sigma^T \geq 0\}$. The parameter (A,B,Σ) corresponds to the model $Ax + By = \varepsilon$, where ε is a Gaussian random variable with mean zero and covariance matrix Σ.

Let data $\{(\tilde{x}_i, \tilde{y}_i);\ i = 1,\ldots,n\}$ be available. One possible approach to identify a model on the basis of these data, i.e., to estimate (A,B,Σ), is the following. Suppose the data are generated by a stochastic system $A_o x_i + B_o y_i = \varepsilon_i$, $i = 1,\ldots n$, where the ε_i are independent identically distributed zero mean Gaussian random variables with covariance matrix Σ_o. First estimate $(-A_o^{-1}B_o\ ,\ A_o^{-1}\Sigma_o(A_o^{-1})^T)$, e.g. by least squares (maximum likelihood). Denote the resulting estimates by $(\hat{\Pi}, \hat{S})$. Impose restrictions on the parameter (A,B) in order to make the map $f:(A,B) \to -A^{-1}B$ injective. The injectivity of f is called identifiability in the literature. In this case the model could be estimated as $(\hat{A},\hat{B}) := f^{-1}(\hat{\Pi})$ and $\hat{\Sigma} := \hat{A}\hat{S}\hat{A}^T$.

We want to state some of the essential elements in this approach.

First, identifiability often is obtained by imposing prior restrictions on A and B, declaring certain elements of these matrices to be zero. The interpretation is that every equation corresponds to a part of

the phenomenon which only incorporates certain variables. These zero restrictions are often inspired by theory. Imposing the restrictions resembles fixing the tolerated complexity, interpreted as the number of non-zero coefficients.

Second, it is not so much the least squares misfit as the variance of the estimated parameters which determines the confidence in the model. In a strict sense, every observation fits any model for which $\Sigma > 0$. However, inspection of the variability of the parameter estimates corresponds to some intuitive concept of misfit.

Finally, both the complexity and the "confidence" are defined in terms of parametrizations of models. In particular, every equation is investigated independent of the other ones. For example, declaring a parameter in a particular equation to be zero does not imply the absence of a direct relationship between the corresponding variables, as such a relationship can be due to the other equations.

In section 6 we decribe two modelling procedures for modelling dynamical phenomena which do not make use of stochastic assumptions. This in particular avoids the assumption of a stable distribution generating disturbances. Moreover, complexity and misfit measures are explicitly defined in terms of canonical parametrizations of dynamical models. These canonical forms are directly inspired by the objectives of modelling and do not depend on a theory concerning the phenomenon. The resulting measures have a clear interpretation in terms of model quality, as opposed to parameter quality. Moreover, the measures take the simultaneous nature of the model equations explicitly into account.

The procedures of section 6 for modelling dynamical phenomena make use of static modelling procedures. We will now describe these static procedures in sections 2.6 and 2.7.

2.6. Static descriptive modelling

Suppose we want to describe a finite number of points in $\mathbf{R}^n$ by means of a linear subspace. So D consists of the finite subsets of $\mathbf{R}^n$ and $\mathbf{M}$ consists of the linear subspaces of $\mathbf{R}^n$. A model M declares $x \in \mathbf{R}^n$ to be compatible with the phenomenon if and only if $x \in M$. As complexity we take $c^D: \mathbf{M} \to \{0,1,\ldots,n\}$ defined as follows.

Definition 2-5 The *descriptive complexity* of a model $M \in \mathbf{M}$ is defined as

its dimension, i.e., $c^D(M) := \dim(M)$.

So a simple model is one which excludes much.

Let $\mathbb{R}^n$ be equipped with e.g. the Euclidean inner product, denoted by $<\cdot,\cdot>$. To define a descriptive misfit, first consider models of codimension 1, i.e., there is $0 \neq a \in \mathbb{R}^n$ with $M = (\text{span}\{a\})^\perp$. Such a model claims the law $<x,a> = 0$ to hold true for the phenomenon. A measure of the quality of this law with respect to data $d = (\tilde{x}_1,\ldots,\tilde{x}_N) \in (\mathbb{R}^n)^N$ is $\epsilon_1^D(d,M) := e^D(d,a)$, which is defined as follows.

Definition 2-6 For data $d = (\tilde{x}_1,\ldots,\tilde{x}_N) \in (\mathbb{R}^n)^N$ and $a \in \mathbb{R}^n$, the *descriptive misfit* of the law $<x,a> = 0$ with respect to d is defined as $e^D(d,a) := \{\frac{1}{N} \sum_{i=1}^{N} <\tilde{x}_i,a>^2 / \|a\|^2\}^{1/2}$.

If $\text{codim}(M) > 1$, then $\epsilon_1^D(d,M)$ is defined as the descriptive misfit of the worst law claimed by M, i.e., $\epsilon_1^D(d,M) := \max\{\epsilon_1^D(d,M');\ M \subset M',\ \text{codim}(M') = 1\}$. Note that the model M claims that $\tilde{x}_i \in M$, so in particular $\tilde{x}_i \in M'$ for $M' \supset M$, $i = 1,\ldots,n$.

Definition 2-7 For $d \in (\mathbb{R}^n)^N$, $M \in \mathbb{M}$, the *first descriptive misfit* is $\epsilon_1^D(d,M) := \max\{e^D(d,a);\ 0 \neq a \in M^\perp\}$.

Note that M claims that $<\tilde{x}_i,a> = 0$ for all $i = 1,\ldots,n$, $a \in M^\perp$. The *second* descriptive misfit is defined as the worst-but-one claimed law, i.e., if $\epsilon_1^D(d,M) = e^D(d,a_1)$, $a_1 \in M^\perp$, then $\epsilon_2^D(d,M) := \max\{e^D(d,a);\ 0 \neq a \in M^\perp \cap (\text{span}\{a_1\})^\perp\}$. So $\epsilon_2^D(d,M)$ measures the quality of the laws claimed by M and orthogonal to the worst law a_1. For $k = 3,\ldots,n-c(M)$ the k-th descriptive misfit is inductively defined as follows: if for $j<k$ $\epsilon_j^D(d,M) = e^D(d,a_j)$, $a_j \in M^\perp \cap (\text{span}\{a_1,\ldots,a_{j-1}\})^\perp$, then $\epsilon_k^D(d,M) := \max\{e^D(d,a);\ 0 \neq a \in M^\perp \cap (\text{span}\{a_1,\ldots,a_{k-1}\})^\perp\}$. It can be shown that $\epsilon_k^D(d,M)$ is well-defined this way, even if the a_j are not unique. For $k = n-c(M)+1,\ldots,n$ we define $\epsilon_k^D(d,M) := 0$. In this way the misfit is a map $\epsilon^D: D \times \mathbb{M} \to \mathbb{R}_+^n$.

On the complexity space $\{0,1,\ldots,n\}$ we take the natural ordering, as well as on $\mathbb{R}_+$. The misfit space $\mathbb{R}_+^n$ we order lexicographically, i.e., $(\epsilon_1,\ldots,\epsilon_n) \geq (\tilde{\epsilon}_1,\ldots,\tilde{\epsilon}_n)$ if and only if $\epsilon_k = \tilde{\epsilon}_k$ for all $k = 1,\ldots,n$ or if there is a k such that $\epsilon_i = \tilde{\epsilon}_i$ for $i<k$ and $\epsilon_k > \tilde{\epsilon}_k$.

We remark that complexity and misfit are defined on the level of

models, not on the parameter level.

In the next propositions we give explicit algorithms for the procedures $P^D_{\epsilon_{tol}}$ corresponding to minimizing complexity, given a tolerated misfit, and $P^D_{c_{tol}}$ corresponding to minimizing misfit, given a tolerated complexity, as described in sections 2.3 and 2.4 respectively.

For data $d=(\tilde{x}_1,\ldots,\tilde{x}_N)$ let $\frac{1}{N}\sum_{i=1}^{N}\tilde{x}_i\tilde{x}_i^T$ have singular value decomposition (S.V.D.) $\frac{1}{N}\sum_{i=1}^{N}\tilde{x}_i\tilde{x}_i^T=U\Sigma U^T$. Here U is orthogonal, i.e., $UU^T=U^TU=I_n$, the identity matrix in $\mathbb{R}^{n\times n}$. Σ is diagonal, $\Sigma=\mathrm{diag}(\sigma_1,\ldots,\sigma_n)$ with $\sigma_1\geq\ldots\geq\sigma_n\geq 0$. Let r:=rank$(\frac{1}{N}\sum_{i=1}^{N}\tilde{x}_i\tilde{x}_i^T)$, then $\sigma_{r+1}=\ldots=\sigma_n=0$. Let u_j denote the j-th column of U. Define M^*_k:= span$\{u_1,\ldots,u_k\}$ and $M(\sigma)$:= span$\{u_j;\ \sigma_j=\sigma\}$.

Proposition 2-8 For given data $d=(\tilde{x}_1,\ldots\tilde{x}_N)\in(\mathbb{R}^n)^N$ and tolerated complexity c_{tol}, $P^D_{c_{tol}}(d)$ is given by

(*i*) $P^D_{c_{tol}}(d)=\{0\}$ if $c_{tol}=0$;

(*ii*) $P^D_{c_{tol}}(d)=\mathrm{span}\{\tilde{x}_1,\ldots,\tilde{x}_N\}$ if $c_{tol}\geq r$;

(*iii*) $P^D_{c_{tol}}(d)=M^*_{c_{tol}}$ if $0<c_{tol}<r$ and $\sigma_{c_{tol}}>\sigma_{c_{tol}+1}$;

(*iv*) if $\sigma_1\geq\ldots\geq\sigma_{c_1}>\sigma_{c_1+1}=\ldots=\sigma_{c_{tol}}=\sigma_{c_{tol}+1}\geq\sigma_{c_{tol}+2}\geq\ldots\geq\sigma_n$ then
$P^D_{c_{tol}}(d)=\{M^*_{c_1}+L;\ L\subset M(\sigma_{c_{tol}}),\ \dim(L)=c_{tol}-c_1\}$.

Proposition 2-9 Let data $d=(\tilde{x}_1,\ldots,\tilde{x}_N)\in(\mathbb{R}^n)^N$ be given. Assume moreover that a maximal misfit level is given with $\epsilon_{tol}=\epsilon_1^{tol}.(1,\ldots,1)$, so the misfit restriction concerns only the worst law claimed by a model. Then

(*i*) $P^D_{\epsilon_{tol}}(d)=\{0\}$ if $\epsilon_1^{tol}>\sigma_1$;

(*ii*) $P^D_{\epsilon_{tol}}(d)=\mathrm{span}\{\tilde{x}_1,\ldots,\tilde{x}_N\}$ if $\epsilon_1^{tol}\leq\sigma_r$;

(*iii*) if $\sigma_r<\epsilon_1^{tol}\leq\sigma_1$, then $P^D_{\epsilon_{tol}}(d)=M^*_k$ with k such that $\sigma_k\geq\epsilon_1^{tol}>\sigma_{k+1}$.

We also refer to Willems [15].

We finally remark that there is a close relationship between these procedures and total least squares. See e.g. Golub and Van Loan [1].Consider as a simple example the case $c_{tol}=n-1$. For $0\neq a\in\mathbb{R}^n$ let $M(a)$:= $(\mathrm{span}\{a\})^{\perp}$:= $\{x\in\mathbb{R}^n;\ <x,a>=0\}$ and let π_a denote the orthogonal projection operator onto $M(a)$. For given data $d=(\tilde{x}_1,\ldots,\tilde{x}_N)\in(\mathbb{R}^n)^N$, in total

least squares one determines a such that $\delta(d,a) := \frac{1}{N} \sum_{i=1}^{N} \|\tilde{x}_i - \pi_a \tilde{x}_i\|^2$ is minimal. See figure 7 for the case $n=2$.

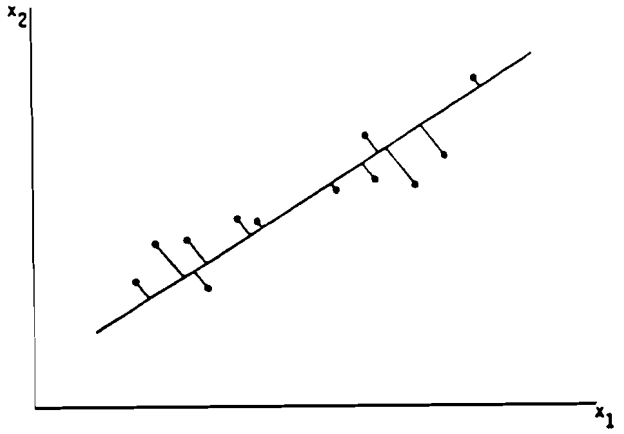

figure 7

It is easily shown that $\delta(d,a) = \{a^T(\frac{1}{N}\sum_{i=1}^{N} \tilde{x}_i \tilde{x}_i^T)a\}/\|a\|^2 = \{\varepsilon_1^D(d, M(a))\}^2$. So in this case of $c_{tol} = n-1$ the procedure $P^D_{c_{tol}}$ corresponds exactly to total least squares. Analogous results can be obtained for $c_{tol} < n-1$ and for $P^D_{\epsilon_{tol}}$.

2.7. Static predictive modelling

Suppose we want to predict (or estimate) n_2 variables $y \in \mathbb{R}^{n_2}$ on the basis of n_1 other variables $x \in \mathbb{R}^{n_1}$ by means of a linear subspace of $\mathbb{R}^{n_1+n_2}$.

Let N observations $(\tilde{x}_i, \tilde{y}_i)$, $\tilde{x}_i \in \mathbb{R}^{n_1}$, $\tilde{y}_i \in \mathbb{R}^{n_2}$, $i = 1,\ldots,N$ be available, so the data set is $D = (\mathbb{R}^{n_1+n_2})^N$.

Let M be a linear subspace of $\mathbb{R}^{n_1+n_2}$. The model M has the interpretation that, given x, it is predicted that y will belong to the set $M(x) := \{y \in \mathbb{R}^{n_2};\ (x,y) \in M\}$. Stated otherwise, let $x \in \mathbb{R}^{n_1}$ be observed. The model M amounts to predicting that the with x associated, but unobserved, y will be such that $<a_1, x> + <a_2, y> = 0$ for all $(a_1, a_2) \in M^{\perp}$, $a_1 \in \mathbb{R}^{n_1}$, $a_2 \in \mathbb{R}^{n_2}$. As model class $\mathbb{M}$ we will take the class of those linear subspaces M of $\mathbb{R}^{n_1+n_2}$ for which the projection on the x coordinate is surjective, i.e., $\{x;\ \exists y$ such that $(x,y) \in M\} = \mathbb{R}^{n_1}$. This means that prediction is possible for every $x \in \mathbb{R}^{n_1}$.

It is easily seen that $M(x) = y + M(0)$ for any $x \in \mathbb{R}^{n_1}$, $y \in M(x)$. So for given model $M \in \mathbb{M}$, the dimension of the (affine) predicted set is independent of the observation x. We define the predictive complexity $c^P: \mathbb{M} \to \{0,1,\ldots,n_2\}$ as follows.

Definition 2-10 The *predictive complexity* of a model $M \in \mathbb{M}$ is defined as the dimension of the affine predicted set, i.e., $c^P(M) := \dim(M(0))$.

So a simple model corresponds to predictions with few degrees of freedom.

To define a predictive misfit we again consider first models of codimension 1. Let $0 \neq a = (a_1, a_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and $M = (\text{span}\{a\})^\perp$. Note that $M \in \mathbb{M}$ implies $a_2 \neq 0$. The model M predicts that, given x, y will satisfy $< a_2, y > = - < a_1, x >$. For data $d = \{(\tilde{x}_i, \tilde{y}_i);\ i = 1, \ldots, N\}$ the relative mean prediction error of this model is $\epsilon_1^P(d, M) := e^P(d, a)$, which is defined as follows.

Definition 2-11 For data $d = \{(\tilde{x}_i, \tilde{y}_i);\ i = 1, \ldots, N\} \in (\mathbb{R}^{n_1} \times \mathbb{R}^{n_2})^N$ and $a = (a_1, a_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ with $a_2 \neq 0$, the *relative mean prediction error* is defined by $e^P(d, a) := [\ \{\frac{1}{N} \sum_{i=1}^{N} (< a_1, \tilde{x}_i > + < a_2, \tilde{y}_i >)^2\} / \{\frac{1}{N} \sum_{i=1}^{N} < a_2, \tilde{y}_i >^2\}\ ]^{1/2}$.

If $\text{codim}(M) > 1$, then $\epsilon^P(d, M)$ is defined in analogy with the misfit in section 2.6, i.e., $\epsilon_1^P(d, M)$ measures the predictive misfit of the worst prediction made by M, $\epsilon_2^P(d, M)$ the misfit of the prediction worst-but-one, and so on.

Formally, let $M_2^\perp := \{a_2; \exists a_1 \text{ such that } (a_1, a_2) \in M^\perp\}$, so $M_2^\perp$ consists of the space of predicted functionals on y. There holds $\dim(M_2^\perp) = n_2 - c(M)$. For $k = 1, \ldots, \dim(M_2^\perp)$ we define $\epsilon_k^P(d, M)$ inductively as follows.

Definition 2-12 For $d \in (\mathbb{R}^{n_1} \times \mathbb{R}^{n_2})^N$, $M \in \mathbb{M}$, the *first predictive misfit* is $\epsilon_1^P(d, M) := \max\{e^P(d, a);\ a \in M^\perp\}$.

Further, if for $j = 1, \ldots, k-1$ $\epsilon_j^P(d, M) = e^P(d, a^{(j)})$, $a_2^{(j)} \in M_2^\perp \cap (\text{span}\{a_2^{(1)}, \ldots, a_2^{(j-1)}\})^\perp$, then $\epsilon_k^P(d, M) := \max\{e^P(d, a);\ a_2 \in M_2^\perp \cap (\text{span}\{a_2^{(1)}, \ldots, a_2^{(k-1)}\})^\perp\}$. For $k = \dim(M_2^\perp) + 1, \ldots, n_2$ we define $\epsilon_k^P(d, M) := 0$. In this way the misfit $\epsilon^P : D \times \mathbb{M} \to \mathbb{R}_+^{n_2}$ is well-defined, provided $N \geq n_2$ and provided that the data are generic in the sense that $\text{span}\{\tilde{y}_1, \ldots, \tilde{y}_N\} = \mathbb{R}^{n_2}$.

We order the complexity and misfit spaces as in section 2.6, i.e., naturally and lexicographically respectively.

Note that again complexity and misfit are defined on the level of models, not on the parameter level.

Next we will give explicit algorithms for the procedures $P^P_{\varepsilon_{tol}}$ corresponding to minimizing complexity, given a tolerated misfit, and $P^P_{c_{tol}}$ corresponding to minimizing predictive misfit, given a tolerated complexity.

Let the data be $d=\{(\tilde{x}_i,\tilde{y}_i);\ i=1,\dots,N\}$. Suppose that $N\geq\max\{n_1,n_2\}$ and that the data are generic in the sense that $\operatorname{span}\{\tilde{x}_1,\dots,\tilde{x}_N\}=\mathbb{R}^{n_1}$ and $\operatorname{span}\{\tilde{y}_1,\dots,\tilde{y}_N\}=\mathbb{R}^{n_2}$. Let $\begin{bmatrix}S_{xx} & S_{xy}\\ S_{yx} & S_{yy}\end{bmatrix}:=\frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix}\tilde{x}_i\\ \tilde{y}_i\end{bmatrix}\begin{bmatrix}\tilde{x}_i\\ \tilde{y}_i\end{bmatrix}^T\in\mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$ and let $S_{xx}^{-1/2}S_{xy}S_{yy}^{-1/2}$ have S.V.D. $U\Lambda V^T$, with $U\in\mathbb{R}^{n_1\times n_1}$ and $V\in\mathbb{R}^{n_2\times n_2}$ both orthogonal matrices and $\Lambda=\begin{bmatrix}\Sigma & 0\\ 0 & 0\end{bmatrix}\in\mathbb{R}^{n_1\times n_2}$, $\Sigma=\operatorname{diag}(\sigma_1,\dots,\sigma_r)$, $\sigma_1\geq\dots\geq\sigma_r>0$. There holds $\sigma_1\leq 1$ and $r=\operatorname{rank}(S_{xy})$. Let r^* denote the number of singular values equal to 1. Denote the columns of $S_{xx}^{-1/2}U$ by $a_1^{(i)}$, $i=1,\dots,n_1$, and those of $S_{yy}^{-1/2}V$ by $a_2^{(i)}$, $i=1,\dots,n_2$. For $k=1,\dots,r$ define $M_k^*:=\{(x,y);\ a_2^{(i)}y=\sigma_i\cdot a_1^{(i)}x,\ i=1,\dots,k\}$. Then $c(M_k^*)=n_2-k$ and $\varepsilon^P(d,M_k^*)=((1-\sigma_k^2)^{1/2},\dots,(1-\sigma_1^2)^{1/2},0,\dots,0)$. Finally, let $M(\sigma):=\{(x,y);\ a_2^{(i)}y=\sigma a_1^{(i)}x$ for all i with $\sigma_i=\sigma\}$.

Proposition 2-13 For generic data $d=\{(\tilde{x}_i,\tilde{y}_i);\ i=1,\dots,N\}$ and tolerated complexity c_{tol}, $P^P_{c_{tol}}$ is given by

(*i*) $P^P_{c_{tol}}(d)=\{M\in\mathbb{M};\ M\subset M_r^*,\ \dim(M_2^\perp)=n_2-c_{tol}\}$ if $c_{tol}<n_2-r$;

(*ii*) $P^P_{c_{tol}}(d)=M_{r^*}^*$ if $c_{tol}\geq n_2-r^*$;

(*iii*) $P^P_{c_{tol}}(d)=M^*_{n_2-c_{tol}}$ if $r^*<n_2-c_{tol}\leq r$ and $\sigma_{n_2-c_{tol}}>\sigma_{n_2-c_{tol}+1}$;

(*iv*) if $\sigma_1\geq\dots\geq\sigma_{c_1}>\sigma_{c_1+1}=\dots=\sigma_{n_2-c_{tol}}=\sigma_{n_2-c_{tol}+1}=\dots=\sigma_{c_2}>\sigma_{c_2+1}\geq\dots\geq\sigma_r>0$, then $P^P_{c_{tol}}(d)=\{M^*_{c_1}\cap L;\ L\supset M(\sigma_{n_2-c_{tol}+1}),\ c(L)=c_{tol}+c_1\}$.

Proposition 2-14 Let data $d=\{(\tilde{x}_i,\tilde{y}_i);\ i=1,\dots,N\}$ be generic. Assume moreover that a maximal misfit level is given with $\varepsilon_{tol}=\varepsilon_1^{tol}\cdot(1,\dots,1)$, so the misfit restriction concerns only the worst prediction made by a model. Then

(*i*) $P^P_{\varepsilon_{tol}}(d)=M^*_{n_2}$ if $\varepsilon_1^{tol}>(1-\sigma^2{}_{n_2})^{1/2}$;

(*ii*) $P^P_{\varepsilon_{tol}}(d)=\mathbb{R}^{n_1+n_2}$ if $\varepsilon_1^{tol}\leq(1-\sigma_1^2)^{1/2}$;

(*iii*) $P^P_{\varepsilon_{tol}}(d)=M_r^*$ if $r<n_2$ and $(1-\sigma_r^2)^{1/2}<\varepsilon_1^{tol}\leq 1$;

(*iv*) if $(1-\sigma_1^2)^{1/2}<\varepsilon_1^{tol}\leq(1-\sigma_r^2)^{1/2}$, then $P^P_{\varepsilon_{tol}}(d)=M_k^*$ where k is such that $(1-\sigma_k^2)^{1/2}<\varepsilon_1^{tol}\leq(1-\sigma_{k+1}^2)^{1/2}$.

We also refer to Heij [4].

We remark that for $n_2 = 1$ and $c_{tol} = 0$ the procedure $P^P_{c_{tol}}$ reduces to ordinary least squares fitting. See figure 8.

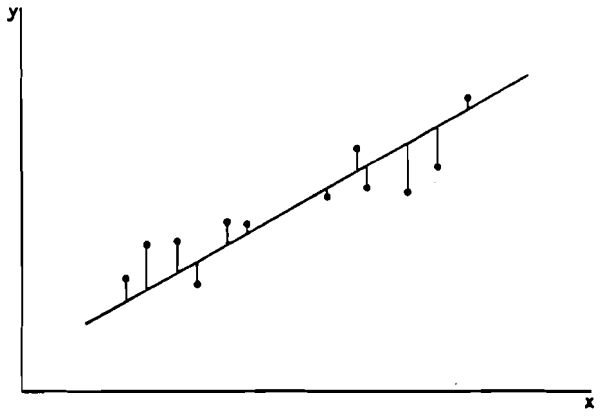

figure 8

The special (vertical) way of measuring the error in this case reflects the purpose of predicting y on the basis of x.

This concludes our section on approximate modelling. The procedures for static modelling in sections 2.6 and 2.7 are used for approximate modelling of time series by means of dynamical models in section 6. In order to do this, we introduce the concept of a dynamical system and a class of dynamical models in section 3. We define complexity and misfit in section 5 in terms of canonical parametrizations of these models. These canonical forms are described in section 4.

3. DYNAMICAL SYSTEMS

3.1. Definition of a dynamical system

Definition 3-1 A *dynamical system* is a triple $(T, \mathbb{W}, B)$ with $T \subset \mathbb{R}$ the time set, $\mathbb{W}$ the signal set and $B \subset \mathbb{W}^T$ the behaviour of the system.

The behaviour B we will sometimes call a system or a model.

A dynamical system describes the relationships between variables of interest in the following way. Let $\mathbb{W}$ be the set in which the variables on every time instant take their values, and let T denote the time set under

consideration. The behaviour B then consists of a set of time series $w:T\to\mathbb{W}$ with the interpretation that time series $w\in B$ are compatible with the laws of the system, while time series $w\notin B$ are not compatible with these laws. This gives a deterministic description of the system.

For some illustrative examples we refer to Willems [15], [16].

3.2. AR-systems

In the sequel we will restrict attention to a special class of dynamical systems, namely those describable by a finite number of autoregressive equations. We will invariably consider discrete time systems with $T=\mathbb{Z}$ and with signal set $\mathbb{W}=\mathbb{R}^q$. So there are q variables of interest which take on real values.

We will use the following notation. Let $R_k\in\mathbb{R}^{g\times q}$ for $k=d_1,d_1+1,\ldots,d_2$, where $d_1,d_2\in\mathbb{Z}$, $d_1\le d_2$. Define $R\in\mathbb{R}^{g\times q}[s,s^{-1}]$ by $R(s,s^{-1}):=\sum_{k=d_1}^{d_2}R_k s^k$, so R is a finite Laurent series in s with coefficients in $\mathbb{R}^{g\times q}$. By a slight abuse of language we will call R a polynomial matrix in s and s^{-1}. By σ we denote left shift, i.e., if $w:\mathbb{Z}\to\mathbb{R}^q$ then $\sigma w:\mathbb{Z}\to\mathbb{R}^q$ is defined by $(\sigma w)(t):=w(t+1)$, $t\in\mathbb{Z}$. By σ^{-1} we denote the inverse of σ. The autoregressive system $B(R)$ then is defined as $\ker(R(\sigma,\sigma^{-1}))$, i.e., $B(R)$ is the set of those time series $w:\mathbb{Z}\to\mathbb{R}^q$ for which $R(\sigma,\sigma^{-1})w=0$, i.e., $\sum_{k=d_1}^{d_2}R_k w(t+k)=0$ for all $t\in\mathbb{Z}$.

Definition 3-2 Let $R\in\mathbb{R}^{g\times q}[s,s^{-1}]$. Then the *autoregressive system (AR-system)* $B(R)$ is defined by $B(R):=\{w\in(\mathbb{R}^q)^{\mathbb{Z}};\ R(\sigma,\sigma^{-1})w=0\}$.

We will denote the class of all AR-systems by $\mathfrak{B}$, i.e., $\mathfrak{B}:=\{B\subset(\mathbb{R}^q)^{\mathbb{Z}};\ \exists g\ \exists R\in\mathbb{R}^{g\times q}[s,s^{-1}]$ such that $B=B(R)\}$.

This class of systems is interesting for a number of reasons. First, it forms a class of models often used in practical modelling situations where one wants to describe linear relationships between the variables and their lagged values, as e.g. in econometrics, signal processing and linear control. Second, this class of systems includes some widely used systems as, for example, linear input/output systems with finite dimensional state space. Third, there exists a nice interpretation of AR-systems on the behavioural level of sets of time series, which we will now describe.

It can be shown that a system $B\subset(\mathbb{R}^q)^{\mathbb{Z}}$ is an AR-system, i.e., there is

a polynomial matrix R such that $B=B(R)$, if and only is B is a *linear*, *time invariant*, *complete* system. B is called *linear* if it is a linear subspace of $(\mathbb{R}^q)^{\mathbb{Z}}$. It is called *time invariant* if $\sigma B=B$, i.e., shifted time series of the system also satisfy the laws of the system. This means that the laws of the system are time invariant. B is called *complete* if $\{w\in B\} \leftrightarrow \{w|_{[t_0,t_1]} \in B|_{[t_0,t_1]}$ for all $-\infty<t_0\leq t_1<+\infty\}$. This means that in order to check whether a time series $w\in(\mathbb{R}^q)^{\mathbb{Z}}$ belongs to B or not it suffices to consider only windows $[t_0,t_1]$ of arbitrary finite length. Moreover it can be shown that if B is linear and time invariant, then B is complete if and only if there exists a $\Delta\geq 0$ such that $\{w\in B\} \leftrightarrow \{w|_{[t,t+\Delta]} \in B|_{[0,\Delta]}$ for all $t\in\mathbb{Z}\}$. So in this case the laws which are imposed by B are local in time.

We finally mention that the class of AR-systems exactly consists of those subsets $B\subset(\mathbb{R}^q)^{\mathbb{Z}}$ which are linear, shift invariant and closed in the topology of pointwise convergence in $(\mathbb{R}^q)^{\mathbb{Z}}$. We will illustrate the use of this characterization by briefly returning to section 2.2.2 on undominated unfalsified modelling. Let $D=(\mathbb{R}^q)^{\mathbb{Z}}$, so the data consists of an infinite time series, and let $\mathbb{M}=\mathbb{B}$, so the model class consists of the AR-systems. The property of closedness of AR-systems implies that for every $\tilde{w}\in D$ there exists a unique $B^*(\tilde{w})\in\mathbb{B}$ such that $\tilde{w}\in B^*(\tilde{w})$ and $\{\tilde{w}\in B\in\mathbb{B}\} \Rightarrow \{B^*(\tilde{w})\subset B\}$. The procedure P_{uu} corresponding to undominated unfalsified modelling hence models $\tilde{w}$ by means of $B^*(\tilde{w})$. It is called the most powerful unfalsified model. In the sequel we will not consider exact modelling of an infinite time series, but approximate modelling of a finite time series.

3.3. Modelling a time series

Suppose we want to model a dynamical phenomenon. In terms of figure 1 in section 1.1, we assume that the objective is either description or prediction of the phenomenon. So we do not discuss control problems or objectives corresponding to theories concerning the phenomenon. Moreover, it is supposed that it is reasonable to model the phenomenon by means of a system which is linear, time invariant and complete. The interpretation is that the model gives a description of the phenomenon which is local, both in space (linearity) and in time (time invariance and completeness). The model class hence is $\mathbb{B}$. It is assumed that q real-valued variables have been specified which have to be included in the model and that data on these variables is available in the form of a finite time series. We denote the variables by $w:=(w_1,\ldots,w_q)^T$, the time interval of observation by

$\mathcal{T} := [t_0, t_1]$ for some $-\infty < t_0 \leq t_1 < +\infty$, and the data by $\tilde{w} := (\tilde{w}(t);\ t \in \mathcal{T})$, an ordered sequence of observations. It is assumed that the data are directly related to the variables of interest and that there are no "missing observations".

In this case the data set is $D = \cup\{(\mathbb{R}^q)^n; n \in \mathbb{N}\}$, so the data consists of a time series of length n in $\mathbb{R}^q$. The model class is $\mathbf{M} = \mathbf{B} := \{B \subset (\mathbb{R}^q)^{\mathbb{Z}};\ B$ linear, time invariant, complete$\}$. The objective π is description or prediction. The modelling problem consists of choosing a procedure $P_\pi : D \to 2^{\mathbf{B}}$, corresponding to a utility u_π reflecting the purpose π of modelling. We will follow the approximate modelling approach described in section 2.1. Therefore we will define complexity maps $c_\pi : \mathbf{B} \to C_\pi$ and misfit maps $\epsilon_\pi : D \times \mathbf{B} \to E_\pi$ and impose orderings on C_π and E_π. The resulting identification problem is depicted in figure 9.

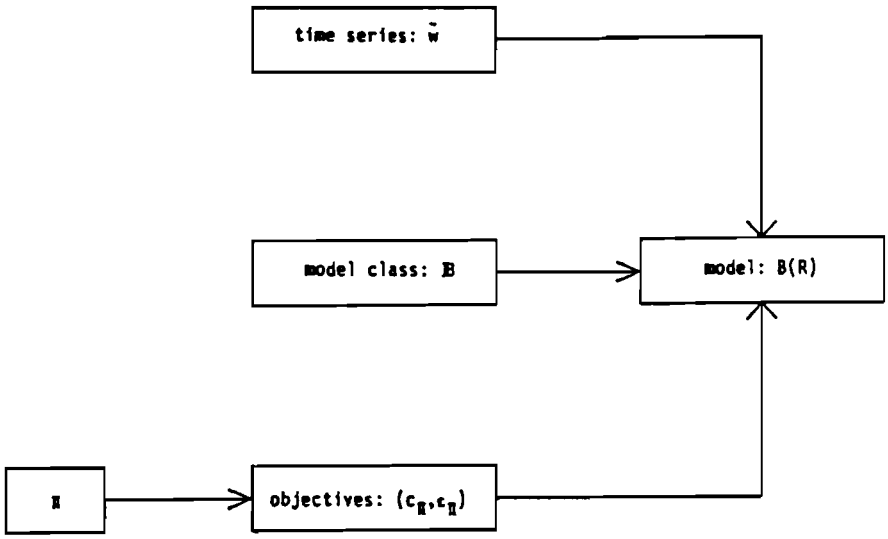

figure 9: modelling a time series

In order to implement procedures algorithmically it is desirable to express the utility not only in terms of the sets $B \subset (\mathbb{R}^q)^{\mathbb{Z}}$ but also in terms of a finite number of parameters parametrizing B, i.e., in terms of an AR-representation R such that $B = B(R)$. However, defining a utility in terms of R need not automatically be compatible with a utility in terms of B, as the map $f : \cup\{\mathbb{R}^{g \times q}[s, s^{-1}];\ g \in \mathbb{N}\} \to \mathbf{B}$ with $f(R) := B(R)$ is not injective. The representation of B by means of R such that $B = B(R)$ is highly non-unique.

In section 4 we will describe the nature of the equivalence relation $\sim$ defined on $\cup\{ \mathbb{R}^{g\times q}[s,s^{-1}];\ g\in\mathbb{N}\}$ by $\{R_1\sim R_2\}:\ \leftrightarrow\ \{B(R_1)=B(R_2)\}$. Moreover we will define two canonical forms under this relation $\sim$, which are inspired by the objectives of modelling. In section 5 we will define complexity and misfit maps for the problem of modelling time series by means of AR-systems. These maps are defined in terms of the canonical forms, i.e., in terms of special AR-representations, and induce well-defined complexity and misfit measures for systems in $\mathfrak{B}$. The corresponding modelling procedures defined in section 2.3 and 2.4 are described in section 6. In section 7 we give the resulting algorithms.

4. CANONICAL FORMS

4.1. Equivalent parametrizations

Let $\mathfrak{B}$ denote the class of models $B\subset(\mathbb{R}^q)^{\mathbb{Z}}$ which are linear, time invariant and complete. As stated before, $B\in\mathfrak{B}$ if and only if there exist $g\in\mathbb{N}$, $d_1,d_2\in\mathbb{Z}$, $d_1\le d_2$, and a polynomial matrix $R=\sum_{k=d_1}^{d_2} R_k s^k\in\mathbb{R}^{g\times q}[s,s^{-1}]$ such that $B=B(R):=\{ w\in(\mathbb{R}^q)^{\mathbb{Z}};\ R(\sigma,\sigma^{-1})w=0\}$.

We will use the following notation. R_1 is called equivalent to R_2, notation $R_1\sim R_2$, if $B(R_1)=B(R_2)$. For $B\in\mathfrak{B}$ let $B^\perp$ denote the family of laws which are satisfied by the behaviour B, i.e., $B^\perp:=\{r\in\mathbb{R}^{1\times q}[s,s^{-1}];\ r(\sigma,\sigma^{-1})w=0$ for all $w\in B\}$. Let $R\in\mathbb{R}^{g\times q}[s,s^{-1}]$ have rows $r_i\in\mathbb{R}^{1\times q}[s,s^{-1}]$, $i=1,\dots,g$, then the polynomial module generated by $r_1,\dots,r_g$ is denoted by $M(R):=\{r\in\mathbb{R}^{1\times q}[s,s^{-1}];\ \exists p_i\in\mathbb{R}[s,s^{-1}],\ i=1,\dots,g$, such that $r=\sum_{i=1}^{g} p_i r_i\}$. Let $\mathfrak{B}^\perp$ denote the class of these (finitely generated) submodules of $\mathbb{R}^{1\times q}[s,s^{-1}]$. By $\dim(M^\perp)$ we denote the dimension of $M^\perp\in\mathfrak{B}^\perp$ as a module, i.e., $\dim(M^\perp)$ is the minimal number of elements of $M^\perp$ which generate $M^\perp$. Finally, $U\in\mathbb{R}^{g\times g}[s,s^{-1}]$ is called unimodular if it is invertible in $\mathbb{R}^{g\times g}[s,s^{-1}]$.

The next proposition summarizes some results on AR-representations of models in $\mathfrak{B}$.

Proposition 4-1 (*i*) For every $B\in\mathfrak{B}$, $B^\perp\in\mathfrak{B}^\perp$; the map $f:\mathfrak{B}\to\mathfrak{B}^\perp:B\to B^\perp$ is a *bijection* of $\mathfrak{B}$ onto $\mathfrak{B}^\perp$; (*ii*) $\{B=B(R)\}\ \leftrightarrow\ \{B^\perp=M(R)\}$; (*iii*) if $\dim(B^\perp)=p$, then there exists $R\in\mathbb{R}^{p\times q}[s,s^{-1}]$ with $B=B(R)$; moreover, this R is unique up to left multiplication by a unimodular matrix.

This implies that the equivalence class of AR-parametrizations of a given model $B\in\mathbf{B}$ consists of those polynomials $R\in\mathbb{R}^{g\times q}[s,s^{-1}]$, for some $g\in\mathbb{N}$, for which the rows generate $B^{\perp}$. So the (autoregressive) laws which are satisfied for any time series in B consist of the rows of R and (polynomial) combinations of them.

We will use these results on equivalent parametrizations to define two canonical forms. A *canonical form* is defined as any subset $C\subset \cup\{\mathbb{R}^{g\times q}[s,s^{-1}];\ g\in\mathbb{N}\}$ which contains at least one element of every equivalence class, i.e., for any $g\in\mathbb{N}$ and $R\in\mathbb{R}^{g\times q}[s,s^{-1}]$ there exists an $R_c\in C$ such that $R\sim R_c$. C is called minimal if it contains exactly one element of every equivalence class, i.e., $R_1,R_2\in C$ with $R_1\sim R_2$ implies that $R_1=R_2$. The two canonical forms defined in sections 4.3 and 4.4 are not minimal. This non-minimality is rather intrinsic, i.e., forcing a reduction of the canonical form so that it would become minimal would require arguments which are not related to the objectives of modelling.

4.2. Preliminaries

In order to describe the canonical forms it is useful to introduce some vocabulary and notation.

For $r\in\mathbb{R}^{1\times q}[s,s^{-1}]$, $r=\sum_{k=-\infty}^{\infty} r_k s^k$, $r_k\in\mathbb{R}^{1\times q}$, define the order of r by $d(r):=\max\{k;\ r_k\neq 0\}-\min\{k;\ r_k\neq 0\}$. Let $R=\mathrm{col}(r_1,\ldots,r_g)\in\mathbb{R}^{g\times q}[s,s^{-1}]$ denote the polynomial matrix with rows $r_1,\ldots,r_g$, then the order of R is defined as $d(R):=\max\{d(r_i);\ i=1,\ldots,g\}$. Suppose $r_i=\sum_{k=d_i'}^{d_i''} r_k^{(i)}s_k$ with $d_i''\geq d_i'$, $r_{d_i'}^{(i)}\neq 0\neq r_{d_i''}^{(i)}$, so $d(r_i)=d_i''-d_i'$. Let $L_+:=\mathrm{col}(r_{d_i''}^{(i)};\ i=1,\ldots,g)$ and $L_-:=\mathrm{col}(r_{d_i'}^{(i)};\ i=1,\ldots,g)$ be the leading and trailing coefficient matrices of R. Then R is called bilaterally row proper if L_+ and L_- both have full row rank g.

Let $R=\mathrm{col}(r_1,\ldots,r_g)\in\mathbb{R}^{g\times q}[s,s^{-1}]$, then $(d(r_1),\ldots,d(r_g))$ is called the lag structure of R. In the sequel we will make use of the equation structure of R, which is defined in terms of the lag structure, as follows.

Definition 4-2 If $R\in\mathbb{R}^{g\times q}[s,s^{-1}]$ has lag structure $(d_1,\ldots,d_g)$, then the *equation structure* of R is defined as $e(R):=(e_t;\ t\geq 0)$, where $e_t:=\#\{i;d_i=t\}$ is the number of rows in R of order t.

For lag structures we define a total ordering by $\{(d'_1,\dots,d'_{g'}) \le (d''_1,\dots,d''_{g''})\}$: $\leftrightarrow$ $\{(d'_1,\dots,d'_{g'}) = (d''_1,\dots,d''_{g''})$ or $g' < g''$ or there is a $g \le g' = g''$ such that $d'_g < d''_g$ and $d'_i = d''_i$ for all $i < g\}$. So few equations and short lags are preferred. We order equation structures by $\{e' \le e''\} :\leftrightarrow \{e' = e''$ or $\sum_{t=0}^{\infty} e'_t < \sum_{t=0}^{\infty} e''_t$ or $\sum_{t=0}^{\infty} e'_t = \sum_{t=0}^{\infty} e''_t$ and there is a t_0 such that $e'_{t_0} > e''_{t_0}$ and $e'_t = e''_t$ for all $t < t_0\}$. For $B \in \mathbf{B}$ we call R a *shortest lag* or *tightest equation* representation of B if $B = B(R)$ and the lag or equation structure respectively is minimal in the class of AR-representations of B. Clearly, every $B \in \mathbf{B}$ has shortest lag and tightest equation representations. The following proposition characterizes these minimal descriptions.

Proposition 4-3 Let $B = B(R)$. Then the following statements are equivalent:

(*i*) R is bilaterally row proper;

(*ii*) R is a tightest equation representation of B;

(*iii*) there exists a permutation matrix Π such that ΠR is a shortest lag representation of B.

We will finally characterize shortest lag representations in terms of matrices. Let $B \in \mathbf{B}$ and $B^{\perp} := \{r \in \mathbb{R}^{1\times q}[s,s^{-1}];\ r(\sigma,\sigma^{-1})w = 0$ for all $w \in B\}$. Let $\mathbb{R}^{1\times q}_t[s]$ denote the class of polynomials in s of power at most t, i.e., $\mathbb{R}^{1\times q}_t[s] := \{r \in \mathbb{R}^{1\times q}[s];\ r = \sum_{k=-\infty}^{\infty} r_k s^k,\ r_k = 0$ for $k < 0$ and $k > t\}$. Let $B^{\perp}_t := B^{\perp} \cap \mathbb{R}^{1\times q}_t[s]$, then $B^{\perp}_t$ describes the family of laws of order at most t which are satisfied by the behaviour B. We will identify $B^{\perp}_t$ with a subspace of $(\mathbb{R}^{1\times q})^{t+1}$ as follows.

Definition 4.4 The bijection $v_t : \mathbb{R}^{1\times q}_t[s] \to (\mathbb{R}^{1\times q})^{t+1}$ is defined as follows. Let $r = \sum_{k=0}^{t} r_k s^k \in \mathbb{R}^{1\times q}_t[s]$, then $v_t(r) \in (\mathbb{R}^{1\times q})^{t+1}$ is defined by $v_t(r) := (r_0, r_1, \dots, r_t)$.

It can be shown that $v_t(B^{\perp}_t)$ is the (Euclidean) orthogonal complement in $(\mathbb{R}^q)^{t+1}$ of $B_t := B|_{[-t,0]} = B|_{[s,s+t]}$ for any $s \in \mathbf{Z}$, i.e., the behaviour on an interval of length $t+1$.

Next we define spaces $L_t \subset B^{\perp}$ as follows. Let $L_0 := B^{\perp}_0$ consist of the

zero order laws for B. Define $V_0 := v_0(L_0)$. Observe that $B_0^\perp + sB_0^\perp \subset B_1^\perp$. We will say that the first order laws in $B_0^\perp + sB_0^\perp$ are *implied* by zero order laws. *Truly* first order laws for B, collected in $L_1 \subset B_1^\perp$, are required to be independent of those implied laws. Formally, let V_1 be a *complementary space* of $v_1(B_0^\perp + sB_0^\perp)$ in $v_1(B_1^\perp)$, i.e., $V_1 \cap v_1(B_0^\perp + sB_0^\perp) = \{0\}$ and $V_1 + v_1(B_0^\perp + sB_0^\perp) = v_1(B_1^\perp)$. Then $L_1 := v_1^{-1}(V_1)$. Analogously, the t-th order laws in $B_{t-1}^\perp + sB_{t-1}^\perp \subset B_t^\perp$ are implied by lower order laws. Truly t-th order laws are collected in $L_t \subset B_t^\perp$, defined as $L_t := v_t^{-1}(V_t)$ for a complementary space V_t of $v_t(B_{t-1}^\perp + sB_{t-1}^\perp)$ in $v_t(B_t^\perp)$, i.e., $V_t \cap v_t(B_{t-1}^\perp + sB_{t-1}^\perp) = \{0\}$ and $V_t + v_t(B_{t-1}^\perp + sB_{t-1}^\perp) = v_t(B_t^\perp)$.

Clearly, the spaces V_t and L_t in general are not uniquely defined. Let $n_t := \dim(V_t)$ and let $\{v_1^{(t)}, \ldots, v_{n_t}^{(t)}\}$ be an arbitrary basis of V_t. Moreover define $r_i^{(t)} := v_t^{-1}(v_i^{(t)})$, $i = 1, \ldots, n_t$. The following proposition establishes the relationship between the sets L_t and shortest lag representations of a model $B \in \mathbb{B}$.

Proposition 4-5 Let $B \in \mathbb{B}$. Then there exists a d such that $n_d \neq 0$ and $n_t = 0$ for all $t > d$. Any tightest equation representation R of B has equation structure $e(R) = (n_0, \ldots, n_d, 0, 0, \ldots)$. Finally, R is a tightest equation representation of B if and only if there exists a choice of the complementary spaces V_t, of bases $\{v_i^{(t)};\ i = 1, \ldots, n_t\}$ of V_t, and of numbers $k_i(t) \in \mathbb{Z}$ for $i = 1, \ldots, n_t$, $t = 0, \ldots, d$, such that the rows of R consist of $\{\sigma^{k_i(t)} \cdot r_i^{(t)};\ i = 1, \ldots, n_t,\ t = 0, \ldots, d\}$.

The canonical forms will correspond to a special choice of the complementary spaces V_t, which we will describe in the next two sections.

4.3. Canonical descriptive form

In section 5 we will define the descriptive complexity and misfit of models in terms of tightest equation representations of a special type. Note that proposition 4-5 characterizes the non-unicity of tightest equation representations in terms of the choice of the complementary spaces V_t and bases of these spaces. The canonical descriptive form selects particular complementary spaces, but the choice of bases is left arbitrary. The complexity and misfit in section 5 will be defined independent of this

choice of bases.

We choose truly t-th order laws of B such that they are (Euclidean) orthogonal to the t-th order laws which are implied by lower order ones. Formally, we define $L_t^D \subset B_t^\perp$ as follows. $L_0^D := B_0^\perp$, and $L_t^D := v_t^{-1}\{ [v_t(B_{t-1}^\perp + sB_{t-1}^\perp)]^\perp \cap [v_t(B_t^\perp)] \}$. So, intuitively, the laws $r \in L_t^D$ are orthogonal to those in $B_{t-1}^\perp + sB_{t-1}^\perp$. The orthogonality is imposed to ensure that laws in L_t^D are "far" from being implied by laws of lower order. Of course, in some cases it could be sensible to choose other inner products than the Euclidean one.

Now R is defined to be in canonical descriptive form if it is itself a tightest equation description of the corresponding behaviour $B(R)$ and if the laws of truly order t are contained in L_t^D. We will then say that laws of different order are orthogonal.

Definition 4-6 R is in *canonical descriptive form* (CDF) if

(*i*) R is a tightest equation representation of $B(R)$;

(*ii*) laws of different order are orthogonal.

Proposition 4-7 (CDF) is a canonical form.

Note that for R in (CDF) $R \in \mathbb{R}^{g \times q}[s]$, i.e., R is a polynomial matrix in s.

We will describe (CDF) in terms of matrices as follows. Let $R \in \mathbb{R}^{g \times q}[s]$ and let $R^{(t)} := \mathrm{col}(r_i^{(t)};\ i = 1, \ldots, n_t)$ consist of the rows of R of order t, $t \geq 0$, $n_t \geq 0$, $\sum_{t=0}^{\infty} n_t = g$. Let d be the highest power of s in R and for $t \geq 0$ let $N_t := \mathrm{col}(v_d(r_i^{(t)});\ i = 1, \ldots, n_t) \in \mathbb{R}^{n_t \times (d+1)q}$ correspond to the t-th order laws in R. Let $N_t = [R_0^{(t)} \ldots R_d^{(t)}]$ with $R_i^{(t)} \in \mathbb{R}^{n_t \times q}$, $i = 0, \ldots, d$. Let $k_t := \max\{i;\ R_i^{(t)} \neq 0\}$. Let $L_- := \mathrm{col}(R_0^{(0)}, \ldots, R_0^{(d)}) \in \mathbb{R}^{g \times q}$ and $L_+ := \mathrm{col}(R_{k_0}^{(0)}, \ldots, R_{k_d}^{(d)}) \in \mathbb{R}^{g \times q}$. Define s: $\mathbb{R}^{1 \times (d+1)q} \to \mathbb{R}^{1 \times (d+1)q}$ as follows. If $v = (v_0, \ldots, v_{d-1}, v_d)$ with $v_i \in \mathbb{R}^{1 \times q}$, $i = 0, \ldots, d$, then $s(v) := (0, v_0, \ldots, v_{d-1})$. Let $V_0 := N_0$ and define $\bar{V}_t$ for $t = 1, \ldots, d$ inductively by $\bar{V}_t := \mathrm{col}(\bar{V}_{t-1}, s\bar{V}_{t-1}, N_t)$. Finally, for matrices A_1 and A_2 let $A_1 \perp A_2$ denote that every row of A_1 is orthogonal to any row of A_2.

Proposition 4-8 R is in canonical descriptive form if and only if

(*i*) L_+ and L_- have full row rank; (this implies $k_t = t$)

(*ii*) $N_t \perp \mathrm{col}(\bar{V}_{t-1}, s\bar{V}_{t-1})$ for all $t = 1, \ldots, d$.

So, whether R is in (CDF) or not can be checked by means of proposition 4-8 in terms of matrices which can be easily calculated from R. These algebraic conditions will play a role in the algorithms of section 7.

The next proposition describes the non-unicity of (CDF) representations of systems $B \in \mathfrak{B}$.

Proposition 4-9 Let $B \in \mathfrak{B}$, $B = B(R)$ with $d(R) = d$ and R in (CDF). Let the rows of R be ordered with increasing degree. Then $B = B(R')$ with R' in (CDF) if and only if there exists a permutation matrix Π and a blockdiagonal matrix $A = \mathrm{diag}(A_{00}, \ldots, A_{dd})$ with $A_{tt} \in \mathbb{R}^{n_t \times n_t}$ nonsingular such that $R' = \Pi A R$.

4.4. Canonical predictive form

The canonical predictive form also corresponds to a particular tightest equation representation of the AR-equations describing a behaviour. Again, the complementary spaces V_t of section 4.2 are chosen in a particular way and the choice of bases is left arbitrary. The spaces are intimately connected with the purpose of prediction and corresponding complexity and misfit maps, which will be defined in section 5.

To define the canonical predictive form, we consider the (forward) predictive interpretation of a law $r \in \mathbb{R}^{1 \times q}[s]$. Let $d(r) = d$, $r = \sum_{k=-\infty}^{\infty} r_k s^k$ with $r_k = 0$ for $k < 0$ and $k > d$. The law r corresponding to $r(\sigma)w = 0$ predicts that, given $w(s)$ for $s = t-d, \ldots, t-1$, $w(t)$ will be such that $r_d w(t) = -\sum_{k=0}^{d-1} r_k w(t-d+k)$, $t \in \mathbb{Z}$. We call r a predictive law of order d, r_d a predicted functional of order d, and $-\sum_{k=0}^{d-1} r_k s^k$ a prediction polynomial of order d. Intuitively speaking, we will choose the complementary spaces V_t such that the predicted functionals of different order are orthogonal and such that prediction polynomials of a certain order are orthogonal to predictive laws of lower order. This ensures that predictive laws of different order are "far" from each other.

Formally, for $B \in \mathfrak{B}$ define $L_t^P \subset B_t^\perp$ as follows. Let $F_t := \{\tilde{r} \in \mathbb{R}^{1 \times q};\ \exists r \in B_t^\perp,\ r = \sum_{k=0}^{t} r_k s^k$, such that $r_t = \tilde{r}\}$ denote the set of predicted functionals of order at most t. Then $L_0^P := B_0^\perp$ and $L_t^P := v_t^{-1}\{ [v_t(F_{t-1} \cdot s^t) + v_t(B_{t-1}^\perp)]^\perp \cap [v_t(B_t^\perp)] \}$. R is said to be in canonical predictive form if it is itself a tightest equation representation of the corresponding behaviour $B(R)$ and if the

predictive laws of order t are contained in L_t^P. We will then say that predicted functionals of different order are orthogonal, corresponding to $v_t(L_t^P) \perp v_t(F_{t-1}.s^t)$, and that the prediction polynomials are orthogonal to predictive laws of lower order, corresponding to $v_t(L_t^P) \perp v_t(B_{t-1}^{\perp})$.

Definition 4-10 R is in *canonical predictive form* (CPF) if

(*i*) R is a tightest equation representation of $B(R)$;

(*ii*) predicted functionals of different orders are orthogonal;

(*iii*) prediction polynomials are orthogonal to predictive laws of lower order.

Proposition 4-11 (CPF) is a canonical form.

Using the notation of section 4.3, proposition 4-12 gives simple algebraic conditions for R to be in (CPF). These conditions will be used in the algorithms of section 7.

Proposition 4-12 R is in canonical predictive form if and only if

(*i*) L_+ and L_- have full row rank; (this implies $k_t = t$)

(*ii*) $R_t^{(t)} \perp R_s^{(s)}$ for all $t \neq s$, $t, s = 0, \ldots, d$;

(*iii*) $N_t \perp \bar{V}_{t-1}$ for all $t = 1, \ldots, d$.

The non-unicity of (CPF) representations is exactly of the same kind as described for (CDF) in proposition 4-9, i.e., the representation is unique up to a permutation of the rows and a choice of bases in the spaces L_t^P.

We conclude this section by giving a simple example illustrating the canonical forms (CDF) and (CPF). Consider $B \in \mathfrak{B}$ defined by $B := \{w \in (\mathbb{R}^3)^{\mathbb{Z}}; w_1(t) + w_2(t-1) = 0,\ w_1(t) + w_3(t) + w_2(t-2) = 0,\ t \in \mathbb{Z}\}$. Then $B = B(R)$ with $R := \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.s + \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix}.s^2$. R is neither in (CDF) nor in (CPF). Let $U_1 := \begin{bmatrix} 1 & 0 \\ -\frac{1}{2} & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ -\frac{1}{2} & 0 \end{bmatrix}.s$, $U_2 := \begin{bmatrix} 1 & 0 \\ -\frac{1}{2} & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix}.s$, $R_1 := U_1 \cdot R$ and $R_2 := U_2 \cdot R$. Then $B = B(R_1) = B(R_2)$, $R_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & -\frac{1}{2} & 0 \end{bmatrix}.s + \begin{bmatrix} 0 & 0 & 0 \\ \frac{1}{2} & 0 & 1 \end{bmatrix}.s^2$ is in (CDF) and $R_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & -1 & 0 \end{bmatrix}.s + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.s^2$ is in (CPF).

5. COMPLEXITY AND MISFIT

5.1. Complexity

As before, let $\mathfrak{B}$ denote the class of linear, time invariant, complete systems in $(\mathbb{R}^q)^{\mathbb{Z}}$. Intuitively, a system is more complex if more time series are compatible with the system, i.e., if the system imposes less restrictions on the behaviour. A simple system is one with a few degree of freedom. In particular, if $B_1, B_2 \in \mathfrak{B}$ and $B_1 \subset B_2$, $B_1 \neq B_2$, then we call B_1 less complex than B_2. More general, we will call B_1 less complex than B_2 if it allows less time series. The complexity of a system will express the magnitude of the set of time series compatible with the system. For $B \in \mathfrak{B}$, let $B_t := B|_{[0,t]}$ denote the space of time series of length $t+1$ which are compatible with the system. By $\mathbb{Z}_+$ we denote the set $\mathbb{Z}_+ := \{0,1,2,3,\ldots\}$. We now define the complexity as a sequence of numbers $c_t(B)$, $t \in \mathbb{Z}_+$, where $c_t(B)$ measures the magnitude of B_t.

Definition 5-1 The *complexity* of dynamical systems is defined by $c: \mathfrak{B} \to (\mathbb{R}_+)^{\mathbb{Z}_+}$, $c(B) := (c_t(B);\ t \in \mathbb{Z}_+)$, where $c_t(B) := \frac{1}{t+1} \cdot \dim(B_t)$.

It can be shown that the limits $\lim_{t\to\infty} c_t(B) =: m$ and $\lim_{t\to\infty} t.\{c_t(B) - m\} =: n$ exist and that m is the number of inputs in B and n the (minimal) number of state variables.

A natural ordering of complexities is the partial ordering defined by $\{c^{(1)} \geq c^{(2)}\} :\leftrightarrow \{c_t^{(1)} \geq c_t^{(2)}$ for all $t \in \mathbb{Z}_+\}$. This ordering is related to tightest equation representations. For $B \in \mathfrak{B}$ let $e^* = (e_t^*;\ t \geq 0)$ denote the equation structure of a tightest equation representation of B. If $B_1, B_2 \in \mathfrak{B}$ with equation structures $e^{*(1)}$ and $e^{*(2)}$ respectively, then $\dim(B_i|_{[0,t]}) = (t+1)q - \sum_{k=0}^{t}(t+1-k)e_k^{*(i)}$, so $c(B_1) \geq c(B_2)$ if and only if for all $t \in \mathbb{Z}_+$ $\sum_{k=0}^{t}(t+1-k)e_k^{*(1)} \leq \sum_{k=0}^{t}(t+1-k)e_k^{*(2)}$. So systems are complex if their behaviour is restricted by few laws which are of high order.

In the approximate modelling procedures of section 6 we will use utility functions involving the complexity. These utility functions will be based on a total (lexicographic) ordering of complexities which is a refinement of the natural ordering, and which is defined by $\{c^{(1)} \geq c^{(2)}\} :\leftrightarrow \{c^{(1)} = c^{(2)}$ or there is a $t_0 \in \mathbb{Z}_+$ such that $c_{t_0}^{(1)} > c_{t_0}^{(2)}$ and $c_t^{(1)} = c_t^{(2)}$ for all $t < t_0\}$.

We want to make some remarks on this ordering.

First, in assessing the complexity of a system the number of short lag equations is decisive. Indeed, as $c_t = q - \frac{1}{t+1} \cdot \sum_{k=0}^{t} (t+1-k)e_k^*$, it follows that $\{c^{(1)} \geq c^{(2)}\} \leftrightarrow \{e^{*(1)} = e^{*(2)}$ or there is a $t_0 \in \mathbf{Z}_+$ such that $e_{t_0}^{*(1)} < e_{t_0}^{*(2)}$ and $e_t^{*(1)} = e_t^{*(2)}$ for all $t < t_0\}$. Note that this ordering of equation structures differs from the one described in section 4.2.

Second, it can be shown that for a system $B \in \mathbf{B}$ there holds $m = q - \sum_{t=0}^{\infty} e_t^*$ and $n = \sum_{t=0}^{\infty} t.e_t^*$, where m denotes the number of inputs or unrestricted variables, n the number of states and $(e_t^*;\ t \in \mathbf{Z}_+)$ the tightest equation structure of B. A simple model is one which leaves little unrestricted, i.e., for which the total number of laws $\sum_{t=0}^{\infty} e_t^*$ is large, and which has small memory, i.e., for which $\sum_{t=0}^{\infty} t.e_t^*$ is small. This amounts to preference of many equations and of short lag, i.e., of small values of $c_t(B)$ for t small. This is reflected by the lexicographic ordering of complexities. Note that the complexity is related to the system considered as a set of trajectories and not to the number of parameters needed to represent the system.

Third, this lexicographic ordering allows for simple recursive algorithms, as will be seen in section 7.

Finally, the reverse lexicographic ordering defined by $\{c^{(1)} \geq c^{(2)}\}$: $\leftrightarrow \{c^{(1)} = c^{(2)}$ or there is a $t_0 \in \mathbf{Z}_+$ such that $c_{t_0}^{(1)} > c_{t_0}^{(2)}$ and $c_t^{(1)} \geq c_t^{(2)}$ for all $t > t_0\}$ seems more appealing. It is directly connected with m and n, as for this ordering $\{m_1 > m_2\} \Rightarrow \{c^{(1)} > c^{(2)}\}$ and $\{m_1 = m_2,\ n_1 > n_2\} \Rightarrow \{c^{(1)} > c^{(2)}\}$. This does not hold true for the lexicographic ordering. However, the construction of algorithms for modelling procedures based on the reverse lexicographic ordering seems to be difficult.

We conclude this section by defining the (total) complexity ordering which we will use in the sequel and by expressing this ordering in terms of equation structures.

Definition 5-2 The *ordering* of complexities of systems in $\mathbf{B}$ is defined by $\{c(B_1) \geq c(B_2)\}$: $\leftrightarrow \{c(B_1) = c(B_2)$ or there is a $t_0 \in \mathbf{Z}_+$ such that $c_{t_0}(B_1) > c_{t_0}(B_2)$ and $c_t(B_1) = c_t(B_2)$ for all $t < t_0\}$.

Proposition 5-3 Let $B_i \in \mathbf{B}$ have tightest equation structure $e^*(B_i) := (e_t^*(B_i);\ t \in \mathbf{Z}_+)$, $i = 1,2$. Then $c(B_1) \geq c(B_2)$ if and only if $e^*(B_1) \leq e^*(B_2)$ in the lexicographic ordering, i.e., $e^*(B_1) = e^*(B_2)$ or

there is a $t_0 \in \mathbb{Z}_+$ such that $e^*_{t_0}(B_1) < e^*_{t_0}(B_2)$ and $e^*_t(B_1) = e^*_t(B_2)$ for all $t < t_0$.

The complexity ordering can easily be characterized in terms of the canonical forms of sections 4.3 and 4.4 by using proposition 4-3.

Corollary 5-4 Let $B_i \in \mathbb{B}$, $B_i = B(R_d^{(i)}) = B(R_p^{(i)})$ with $R_d^{(i)}$ in (CDF) and $R_p^{(i)}$ in (CPF), $i = 1,2$. Let $e_d^{(i)}$ and $e_p^{(i)}$ denote the equation structure of $R_d^{(i)}$ and $R_p^{(i)}$ respectively, $i = 1,2$. Then $\{c(B_1) \geq c(B_2)\} \Leftrightarrow \{e_p^{(1)} = e_d^{(1)} \leq e_d^{(2)} = e_p^{(2)}$ in lexicographic ordering$\}$.

5.2. Descriptive misfit

In this section we define the misfit of a model $B \in \mathbb{B}$ in describing data consisting of a finite time series $\tilde{w} := (\tilde{w}(t);\ t \in \mathcal{T})$ on an interval $\mathcal{T} = [t_0, t_1]$. As in section 2.6 we first consider the case where B imposes one restriction, in the sense that $B = B(r)$ for some $r \in \mathbb{R}^{1\times q}[s, s^{-1}]$.

As descriptive misfit we consider the average equation error. Let $n \in \mathbb{Z}$, $d \in \mathbb{Z}_+$, $r = \sum_{k=n}^{n+d} r_k s^k$ with $r_k \in \mathbb{R}^{1\times q}$, $r_n \neq 0 \neq r_{n+d}$. We define $\|r\|^2 := \sum_{k=n}^{n+d} \|r_k\|^2$ and $\|r\tilde{w}\|^2 := \frac{1}{t_1 - t_0 - d + 1} \cdot \sum_{t=t_0-n}^{t_1-n-d} \{ \sum_{k=n}^{n+d} r_k \tilde{w}(t+k) \}^2$. So $\|r\tilde{w}\|$ measures in how far $\tilde{w}$ satisfies the restriction imposed by $B(r)$ that $(r\tilde{w})(t) = 0$ for $t = t_0 - n, \ldots, t_1 - n - d$. It is assumed that $d(r) = d \leq t_1 - t_0$.

Definition 5-5 The *descriptive misfit* of $r \in \mathbb{R}^{1\times q}[s, s^{-1}]$ with respect to data $\tilde{w} \in (\mathbb{R}^q)^{\mathcal{T}}$ is defined as the mean *equation error*, i.e., $e^D(\tilde{w}, r) := \|r\tilde{w}\| / \|r\|$.

We define the misfit of $B(r)$ by $\epsilon^D_{d,1}(\tilde{w}, B(r)) := e^D(\tilde{w}, r)$.

Next let $\dim(B^\perp) \geq 2$. For $r \in B^\perp$ we measure the descriptive misfit by $e^D(\tilde{w}, r)$. The problem is to define the misfit of B, which imposes an infinite number of laws on the phenomenon. We will define the misfit of B by choosing a canonical basis in $B^\perp$, using the canonical descriptive form (CDF). The idea is to define a sequence of misfits, measuring the quality of laws of different order claimed by B. Note that using (CDF) guarantees that laws of different order are orthogonal, so loosely speaking these quality measures become more or less *independent*. By this we mean that e.g.

a first order law should not be judged as being of small misfit if this is due to the fact that this first order law is ("near" to being) implied by good zero order laws. This is made explicit by the orthogonality conditions in (CDF) as stated in section 4.3 and will be illustrated by means of examples in section 9.

To define $\varepsilon^D(\tilde{w},B)$, consider the spaces L_t^D of truly t–th order decriptive laws as defined in section 4.3. Let $n_t := \dim(v_t(L_t^D))$, then $n_t = e_t$ where $(e_t;\ t\in\mathbf{Z}_+)$ is the tightest equation structure of AR–representation of B. For $n_t>0$ define $\varepsilon_{t,1}^D(\tilde{w},B)$ as the worst fit of the truly t–th order laws claimed by B, i.e. $\varepsilon_{t,1}^D(\tilde{w},B) := \max\{e^D(\tilde{w},r);\ r\in L_t^D\}$.

Definition 5-6 For $B\in\mathbf{B}$, let L_t^D denote the space of truly t–th order descriptive laws of B. For data $\tilde{w}\in(\mathbf{R}^q)^{\mathcal{T}}$, the *main* t–th *descriptive misfit* is defined by $\varepsilon_{t,1}^D(\tilde{w},B) := \max\{e^D(\tilde{w},r);\ r\in L_t^D\}$ if $\dim(v_t(L_t^D))>0$, else $\varepsilon_{t,1}^D(\tilde{w},B) := 0$.

If $n_t>1$, then we define $\varepsilon_{t,2}^D(\tilde{w},B)$ as the misfit of the worst–but–one t–th order law, i.e., if $\varepsilon_{t,1}(\tilde{w},B) = e^D(\tilde{w},r_1)$, $r_1\in L_t^D$, then $\varepsilon_{t,2}^D(\tilde{w},B) := \max\{e^D(\tilde{w},r);\ r\in v_t^{-1}\{\ v_t(L_t^D)\cap[v_t(r_1)]^{\perp}\}\}$. For $k=2,\ldots,n_t$, $\varepsilon_{t,k}^D(\tilde{w},B)$ is inductively defined as the worst–but–$(k-1)$ t–th order misfit, as follows. If $\varepsilon_{t,j}^D(\tilde{w},B) = e^D(\tilde{w},r_j)$, $r_j\in v_t^{-1}\{\ v_t(L_t^D)\cap[\mathrm{span}(v_t(r_1),\ldots,v_t(r_{j-1}))]^{\perp}\}$ for $j=1,2,\ldots,k-1$, then $\varepsilon_{t,k}^D(\tilde{w},B) := \max\{e^D(\tilde{w},r);\ r\in v_t^{-1}\{\ v_t(L_t^D)\cap[\mathrm{span}(v_t(r_1),\ldots,v_t(r_{k-1}))]^{\perp}\}\}$. For $k=n_t+1,\ldots,q$, $\varepsilon_{t,k}^D(\tilde{w},B) := 0$. It can be shown that $\varepsilon_{t,k}^D$ is well–defined in this way, i.e., independent of the maximizing arguments r_j.

Definition 5-7 The *descriptive misfit* is a map $\varepsilon^D:(\mathbf{R}^q)^{\mathcal{T}}\times\mathbf{B}\to(\mathbf{R}_+^{1\times q})^{\mathbf{Z}_+}$, where $\varepsilon_{t,k}^D(\tilde{w},B)$ is the descriptive misfit of the worst–but–$(k-1)$ law of the truly t–th order descriptive laws in L_t^D claimed by B, $t\in\mathbf{Z}_+$, $k=1,\ldots,q$.

We remark that both the complexity and the descriptive misfit are defined in terms of the spaces L_t^D, hence in terms of (CDF), but independent of a choice of basis in L_t^D. A convenient basis for L_t^D could be $\{r_1,\ldots,r_{n_t}\}$ as defined above.

Note that there are at most $\sum_{t=0}^{\infty} e_t = q-m \le q$ misfit numbers unequal to zero. These numbers give the equation error of a suitably chosen basis of

all the equations which are claimed by the model. The numbers $\{\varepsilon^D_{t,k}\,;\ k=1,\ldots,q\}$ measure the quality of the t-th order equations, which are orthogonal to the lower order ones.

We will impose the following lexicographic ordering on misfits.

Definition 5-8 $\{\varepsilon'=(\varepsilon'_{t,k}) \geq \varepsilon''=(\varepsilon''_{t,k})\}$: $\Leftrightarrow$ $\{\varepsilon'=\varepsilon''$; or there exists $t_0\in\mathbf{Z}_+$, $k_0\leq q$ such that $\varepsilon'_{t_0,k_0}>\varepsilon''_{t_0,k_0}$ and $\varepsilon'_{t,k}=\varepsilon''_{t,k}$ for all $t<t_0$, $k=1,\ldots,q$ and for $t=t_0$, $k=1,\ldots,k_0-1$; or there exists $t_0\in\mathbf{Z}_+$ such that $\varepsilon'_{t_0,1}>\varepsilon''_{t_0,1}$ and $\varepsilon'_{t,k}=\varepsilon''_{t,k}$ for all $t<t_0$, $k=1,\ldots,q\}$.

Note that if B_1 has lower order laws than B_2, then the misfit of B_1 in general will be larger than that of B_2. On the other hand the complexity of B_1 is smaller than that of B_2. In section 6 we will describe two procedures to balance the desires for low misfit and low complexity by fixing a maximal tolerated level for one of the objectives and optimizing with respect to the other one. These procedures correspond to the utilities defined in sections 2.3 and 2.4. We will do the same for predictive misfit, defined in the next section.

5.3. Predictive misfit

The one-step-ahead predictive misfit of a dynamical system in predicting a time series is based on the prediction error defined in section 2.7 for static prediction. Now the data consists of a finite time series $\tilde{w}=(\tilde{w}(t);\ t\in\mathcal{T}=[t_0,t_1])$ and the model class consists of the class of linear, time invariant, complete systems $\mathfrak{B}$.

Again we first consider the case where $B=B(r)$ with $r\in\mathbf{R}^{1\times q}[s,s^{-1}]$. Let $n\in\mathbf{Z}$, $d\in\mathbf{Z}_+$, $r=\sum_{k=n}^{n+d} r_k s^k$ with $r_k\in\mathbf{R}^{1\times q}$, $r_n\neq 0\neq r_{n+d}$. Then $B(r)$ predicts that $r_{n+d}w(t+n+d) = -\sum_{k=n}^{n+d-1} r_k w(t+k)$. Let $r_{n+d}\tilde{w}(t+n+d) = -\sum_{k=n}^{n+d-1} r_k\tilde{w}(t+k)+e(t+n+d)$ for $t=t_0-n,\ldots,t_1-n-d$. So $e(t)$ is the error made at time t in the prediction of $r_{n+d}w(t)$. Let $\|e\|^2:=\frac{1}{t_1-t_0-d+1}\sum_{t=t_0+d}^{t_1} e^2(t)$ denote the average prediction error and let $\|r_{n+d}\tilde{w}\|_d^2:=\frac{1}{t_1-t_0-d+1}\sum_{t_1=t_0+d}^{t_1}\{r_{n+d}\tilde{w}(t)\}^2$ denote the average magnitude of the predicted functional. It is assumed that $d\leq t_1-t_0$.

Definition 5-9 The *predictive misfit* of $\tau \in \mathbb{R}^{1\times q}[s,s^{-1}]$, with $1 \le d(\tau) \le t_1 - t_0$ and with leading coefficient vector $\tau^* \in \mathbb{R}^{1\times q}$, with respect to data $\tilde{w} \in (\mathbb{R}^q)^{\mathcal{T}}$ is defined as the *relative mean prediction error*, i.e., $e^P(\tilde{w},\tau) := \|\tau\tilde{w}\| \,/\, \|\tau^*\tilde{w}\|_d = \|e\| \,/\, \|\tau^*\tilde{w}\|_d$.

We define the predictive misfit of $B(\tau)$ by $\varepsilon^P_{d,1}(\tilde{w},B(\tau)) := e^P(\tilde{w},\tau)$.

Next we define the misfit for models with $\dim(B^\perp) \ge 2$. Again we will measure the predictive quality of a model by means of a sequence of numbers which measure the quality of predictive laws of different order. The quality assessment for laws of different orders is made independently by using the canonical predictive form (CPF). First of all we require the t–th order laws to be truly t–th order, i.e., the t–th order laws should not be implied by lower order ones. Second, we require predicted functionals of different order to be orthogonal. This is essential to guarantee that good quality of one predictive law is not due to good quality of another predictive law. This is made explicit by the orthogonality conditions of (CPF) in section 4.4 and will be illustrated by means of examples in section 9.

To define $\varepsilon^P(\tilde{w},B)$, consider the spaces L^P_t defined in section 4.4 and let $n_t := \dim(v_t(L^P_t)) = e_t$. We give the definition of predictive misfit in analogy with the definition of descriptive misfit in section 5.2 and with the same motivation. For $t=0$ we define $\varepsilon^P_{t,k}(\tilde{w},B) := \varepsilon^D_{t,k}(\tilde{w},B)$, as for $d(\tau)=0$ $e^P(\tilde{w},\tau)=1$ for any $\tilde{w}$, so the predictive misfit makes no sense for these static laws. In this case we measure the misfit simply by $\|e\|/\|\tau\|$.

Definition 5-10 For $B \in \mathfrak{B}$, let L^P_t denote the space of truly t–th order predictive laws of B. For data $\tilde{w} \in (\mathbb{R}^q)^{\mathcal{T}}$, the *main t–th order predictive misfit* for $t \ge 1$ is defined by $\varepsilon^P_{t,1}(\tilde{w},B) := \max\{e^P(\tilde{w},\tau);\ \tau \in L^P_t\}$ if $\dim(v_t(L^P_t)) > 0$, else $\varepsilon^P_{t,1}(\tilde{w},B) := 0$.

Moreover, $\varepsilon^P_{t,k}$ measures the predictive misfit of the worst–but–$(k-1)$ law of the truly t–th order predictive laws in L^P_t claimed by B. If $t \ge 1$ and $n_t > 1$, then $\varepsilon^P_{t,k}(\tilde{w},B)$ for $k = 2,\dots,n_t$ is inductively defined as follows. If $\varepsilon^P_{t,j}(\tilde{w},B) = e^P(\tilde{w},\tau_j)$ with $\tau_j \in v_t^{-1}\{ v_t(L^P_t) \cap [\mathrm{span}(v_t(\tau_1),\dots,v_t(\tau_{j-1}))]^\perp\}$ for $j = 1,\dots,k-1$, then $\varepsilon^P_{t,k}(\tilde{w},B) := \max\{e^P(\tilde{w},\tau);\ \tau \in v_t^{-1}\{ v_t(L^P_t) \cap [\mathrm{span}(v_t(\tau_1),\dots, v_t(\tau_{k-1}))]^\perp\}\}$. For $k = n_t+1,\dots,q$ we define $\varepsilon^P_{t,k}(\tilde{w},B) := 0$. It can be shown that $\varepsilon^P_{t,k}(\tilde{w},B)$ is well–defined.

Definition 5-11 The *predictive misfit* is a map $\varepsilon^P:(\mathbb{R}^q)^{\mathcal{T}}\times\mathfrak{B}\to(\mathbb{R}_+^{1\times q})^{\mathbb{Z}_+}$ where $\varepsilon^P_{0,k}(\tilde{w},B):=\varepsilon^D_{0,k}(\tilde{w},B)$ and for $t\geq 1$ $\varepsilon^P_{t,k}(\tilde{w},B)$ is the predictive misfit of the worst-but-$(k-1)$ law of the truly t-th order predictive laws in L^P_t claimed by B, $k=1,\ldots,q$.

We order the predictive misfit sequences in the same way as the descriptive misfit sequences, i.e., lexicographically. Corresponding modelling procedures are described in the next section.

6. MODELLING PROCEDURES

6.1. Introduction

In this section we describe four modelling procedures. Both for the purpose of description and for that of prediction we define two utility functions, corresponding to fixing the tolerated misfit or the tolerated complexity and optimizing complexity and misfit respectively. The corresponding procedures lead to relatively simple algorithms, the details of which are given in section 7.

6.2. Deterministic descriptive modelling procedures

Let $\mathfrak{B}$ consist of the class of AR-systems $B\subset(\mathbb{R}^q)^{\mathbb{Z}}$ and let the set of conceivable data be $D:=\cup\{(\mathbb{R}^q)^n;\ n\in\mathbb{N}\}$, so the data consists of a finite time series $\tilde{w}\in(\mathbb{R}^q)^{\mathcal{T}}$ for some $\mathcal{T}=[t_0,t_1]$.

First consider the case that a maximal tolerated complexity $c_{tol}:=(c_t^{tol};\ t\in\mathbb{Z}_+)$ is given. Fixing c_{tol} is interpreted as requiring that allowable models should satisfy $c_t(B)\leq c_t^{tol}$ for all $t\in\mathbb{Z}_+$. As $c_t=q-\frac{1}{t+1}\sum_{k=0}^{t}(t+1-k)e_k^*$ this amounts to requiring $\sum_{k=0}^{t}(t+1-k)e_k^*(B)\geq(q-c_t^{tol}).(t+1)$ for all $t\in\mathbb{Z}_+$, where $(e_t^*(B);\ t\in\mathbb{Z}_+)$ is the equation structure of a tighest equation representation of B. So a maximal tolerated complexity amounts to requiring that B imposes a minimal tolerated number of (truly) t-th order restrictions. Under this requirement the descriptive misfit will be minimized. The misfit of B is the sequence $\varepsilon^D(\tilde{w},B)\in(\mathbb{R}_+^{1\times q})^{\mathbb{Z}_+}$ with lexicographic ordering as defined in section 5.2. The procedure $P^D_{c_{tol}}:D\to 2^{\mathfrak{B}}$ then is defined as in section 2.4, i.e., as follows.

Definition 6-1 For $\tilde{w}\in D$, $P^D_{c_{tol}}(\tilde{w}) := \operatorname{argmax}\{ u_{c_{tol}}(c(B),\varepsilon^D(\tilde{w},B));\ B\in\mathbb{B}\}$, where the ordering for $u := u_{c_{tol}}$ is defined by

(i) $\{u(c^{(1)},\varepsilon^{(1)}) = u(c^{(2)},\varepsilon^{(2)})\}$: $\leftrightarrow$ $\{\exists t_i\in\mathbb{Z}_+\ c^{(i)}_{t_i} > c^{tol}_{t_i},\ i=1,2;$ or $(c^{(1)},\varepsilon^{(1)}) = (c^{(2)},\varepsilon^{(2)})\}$;

(ii) $\{u(c^{(1)},\varepsilon^{(1)}) < u(c^{(2)},\varepsilon^{(2)})\}$: $\leftrightarrow$ $\{\exists t_0\in\mathbb{Z}_+\ c^{(1)}_{t_0} > c^{tol}_{t_0}$ and $\forall t\in\mathbb{Z}_+\ c^{(2)}_t \le c^{tol}_t$; or $\forall t\in\mathbb{Z}_+\ c^{(1)}_t, c^{(2)}_t \le c^{tol}_t$ and $\exists t_0\in\mathbb{Z}_+$ such that $\varepsilon^{(1)}_{t_0} > \varepsilon^{(2)}_{t_0}$ and $\varepsilon^{(1)}_t = \varepsilon^{(2)}_t$ for all $t<t_0$; or $\forall t\in\mathbb{Z}_+\ c^{(1)}_t, c^{(2)}_t \le c^{tol}_t, \varepsilon^{(1)} = \varepsilon^{(2)}$ and $\exists t_0\in\mathbb{Z}_+$ such that $c^{(1)}_{t_0} > c^{(2)}_{t_0}$ and $c^{(1)}_t = c^{(2)}_t$ for all $t<t_0\}$. Here the vectors $\varepsilon_t\in\mathbb{R}^{1\times q}$ are ordered lexicographically.

Note that the requirement $c(B)\le c_{tol}$ is not interpreted in the lexicographic ordering, but in the pointwise ordering, i.e., $c(B)\le c_{tol}$ if and only if $c_t(B)\le c^{tol}_t$ for all $t\in\mathbb{Z}_+$.

Next suppose that a maximal tolerated misfit $\varepsilon_{tol} := (\varepsilon^{tol}_t;\ t\in\mathbb{Z}_+)\in(\mathbb{R}^{1\times q})^{\mathbb{Z}_+}$ is given. We will invariably assume that $\varepsilon^{tol}_t = \bar{\varepsilon}^{tol}_t.(1,\ldots,1)$ with $\bar{\varepsilon}^{tol}_t\in\mathbb{R}$. The requirement $\varepsilon^D(\tilde{w},B)<\varepsilon_{tol}$ also is not interpreted in the lexicographical sense, but pointwise. As $\varepsilon^D_{t,k}(\tilde{w},B)\le\varepsilon^D_{t,l}(\tilde{w},B)$ for $k\ge l$, this means that a model $B\in\mathbb{B}$ is tolerated if and only if $\varepsilon^D_{t,1}(\tilde{w},B)<\bar{\varepsilon}^{tol}_t$ for all $t\in\mathbb{Z}_+$. So fixing ε_{tol} amounts to requiring that the misfit of (truly) t-th order laws should be smaller than $\bar{\varepsilon}^{tol}_t$. One can impose an upper bound L on the order of equations by taking $\bar{\varepsilon}^{tol}_t<0$ for $t>L$.

Under the requirement $\varepsilon^D_{t,1}(\tilde{w},B)<\bar{\varepsilon}^{tol}_t$ the complexity has to be minimized. The complexity of a system is $c(B)\in(\mathbb{R}_+)^{\mathbb{Z}_+}$ with lexicographic ordering, as defined in section 5.1. Equivalently, under the misfit restriction the equation structure $(e^*_t(B);\ t\ge 0)$ has to be maximized lexicographically. So the purpose is to find as many relationships of small order as possible.

The procedure $P^{*D}_{\varepsilon_{tol}}: D\to 2^{\mathbb{B}}$ corresponding to the one described in section 2.3 for minimizing complexity given a misfit restriction is defined as follows. For $\tilde{w}\in D$, $P^{*D}_{\varepsilon_{tol}}(\tilde{w}) := \operatorname{argmax}\{u(c(B),\varepsilon^D(\tilde{w},B));\ B\in\mathbb{B}\}$, with the ordering $\{u(c^{(1)},\varepsilon^{(1)}) = u(c^{(2)},\varepsilon^{(2)})\}$: $\leftrightarrow$ $\{\exists t_i\in\mathbb{Z}_+\ \varepsilon^{(i)}_{t_i,1}\ge\bar{\varepsilon}^{tol}_{t_i},\ i=1,2;$ or $(c^{(1)},\varepsilon^{(1)}) = (c^{(2)},\varepsilon^{(2)})\}$, and $\{u(c^{(1)},\varepsilon^{(1)})<u(c^{(2)},\varepsilon^{(2)})\}$: $\leftrightarrow$ $\{\exists t_0\in\mathbb{Z}_+$ $\varepsilon^{(1)}_{t_0,1}\ge\bar{\varepsilon}^{tol}_{t_0}$ and $\forall t\in\mathbb{Z}_+\ \varepsilon^{(2)}_{t,1}<\bar{\varepsilon}^{tol}_t$; or $\forall t\in\mathbb{Z}_+\ \varepsilon^{(1)}_{t,1}, \varepsilon^{(2)}_{t,1}<\varepsilon^{tol}_t$ and $\exists t_0\in\mathbb{Z}_+$ such that $c^{(1)}_{t_0} > c^{(2)}_{t_0}$ and $c^{(1)}_t = c^{(2)}_t$ for all $t<t_0$; or $\forall t\in\mathbb{Z}_+\ \varepsilon^{(1)}_{t,1}, \varepsilon^{(2)}_{t,1}<\bar{\varepsilon}^{tol}_t$, $c^{(1)} = c^{(2)}$ and $\varepsilon^{(1)} > \varepsilon^{(2)}$ in lexicographic ordering}.

However, $P^{*D}_{\varepsilon_{tol}}$ is difficult to implement algorithmically. We will

consider a slight variation $P^D_{\epsilon_{tol}}$ of $P^{*D}_{\epsilon_{tol}}$. We will illustrate the difference between these two procedures by means of a simple example in section 9. The procedure $P^D_{\epsilon_{tol}}$ allows for a relatively simple algorithm, described in section 7.

We now first define $P^D_{\epsilon_{tol}}$ and subsequently give an interpretation.

Definition 6-2 For $\tilde{w}\in D$, $P^D_{\epsilon_{tol}}(\tilde{w}) := \text{argmax}\{ u_{\epsilon_{tol}}(c(B),\varepsilon^D(\tilde{w},B));\ B\in\mathbb{B}\}$ where the ordering for $u := u_{\epsilon_{tol}}$ is defined by

(i) $\{u(c^{(1)},\varepsilon^{(1)}) = u(c^{(2)},\varepsilon^{(2)})\}$: $\leftrightarrow$ $\{\exists t_i\in\mathbb{Z}_+\ \varepsilon^{(i)}_{t_i,1} \geq \bar{\varepsilon}^{tol}_{t_i},\ i=1,2;$ or $(c^{(1)},\varepsilon^{(1)}) = (c^{(2)},\varepsilon^{(2)})\}$;

(ii) $\{u(c^{(1)},\varepsilon^{(1)}) < u(c^{(2)},\varepsilon^{(2)})\}$: $\leftrightarrow$ $\{\exists\, t_0\in\mathbb{Z}_+\ \varepsilon^{(1)}_{t_0,1}\geq\bar{\varepsilon}^{tol}_{t_0}$ and $\forall t\in\mathbb{Z}_+$ $\varepsilon^{(2)}_{t,1}<\bar{\varepsilon}^{tol}_t$; or $\forall t\in\mathbb{Z}_+$ $\varepsilon^{(1)}_{t,1}, \varepsilon^{(2)}_{t,1} < \bar{\varepsilon}^{tol}_t$ and $(c^{(1)}_0,\varepsilon^{(1)}_{0,1},\dots,\varepsilon^{(1)}_{0,e_0(1)}, c^{(1)}_1, \varepsilon^{(1)}_{1,1},\dots,\varepsilon^{(1)}_{1,e_1(1)}, c^{(1)}_2, \varepsilon^{(1)}_{2,1},\dots,\varepsilon^{(1)}_{2,e_2(1)}, c^{(1)}_3,\dots) > (c^{(2)}_0, \varepsilon^{(2)}_{0,1},\dots,\varepsilon^{(2)}_{0,e_0(2)}, c^{(2)}_1, \varepsilon^{(2)}_{1,1},\dots,\varepsilon^{(2)}_{1,e_1(2)}, c^{(2)}_2, \varepsilon^{(2)}_{2,1},\dots,\varepsilon^{(2)}_{2,e_2(2)}, c^{(2)}_3,\dots)$ in the lexicographic ordering, where $e^{(i)}$ is the tightest equation structure corresponding to $c^{(i)}$, $i=1,2\}$.

This means that $P^D_{\epsilon_{tol}}$ maximizes the number of zero order relations under the misfit constraint. Among solutions, which in general are highly non-unique, it chooses the one with minimal misfit. Subsequently the number of first order relations is maximized, and then the first order misfit is minimized, and so on. Note that these first order relations should be orthogonal to the zero order ones, as the utility is defined in terms of $\varepsilon^D(\tilde{w},B)$ which involves (CDF). The resulting model is optimal with respect to the utility $u_{\epsilon_{tol}}$. Proposition 5-3 indicates a close relationship between $P^D_{\epsilon_{tol}}$ and $P^{*D}_{\epsilon_{tol}}$. However, $P^D_{\epsilon_{tol}}$ need not always minimize the complexity with respect to the lexicographic ordering on $(c_t(B);\ t\in\mathbb{Z}_+)$, as will be illustrated by means of an example in section 9. This is due to the auxiliary minimization of misfits, which is essential for obtaining simple (recursive) algorithms.

Proposition 6-3 The procedures $P^D_{c_{tol}}$ and $P^D_{\epsilon_{tol}}$ are well-defined maps from D into $2^{\mathbb{B}}$.

Finally, by $\bar{P}^D_{\epsilon_{tol}}(\tilde{w})$ we denote the procedure which is defined in analogy with $P^D_{\epsilon_{tol}}$, but requiring $\varepsilon^D_{t,1}(\tilde{w},B)\leq\bar{\varepsilon}^{tol}_t$ in contrast with $P^D_{\epsilon_{tol}}$ which

requires $\varepsilon^D_{t,1}(\tilde{w},B) < \bar{\varepsilon}^{tol}_t$.

6.3. Two deterministic predictive modelling procedures

In this section we briefly describe two predictive procedures, corresponding to fixing a maximal tolerated complexity or misfit and minimizing misfit and complexity respectively. These procedures are analogues of the descriptive procedures defined in section 6.2 and are obtained by replacing the descriptive misfit ε^D by the predictive misfit ε^P.

Again, fixing a maximal tolerated complexity amounts to requiring of an allowable model B that it imposes a minimal tolerated number of (truly) t–th order restrictions on the phenomenon, $t \in \mathbf{Z}_+$. Under this requirement the relative mean prediction error ε^P is minimized lexicographically. So first the misfit of the zero order laws (in L^P_0) is minimized, then the misfit of the truly first order laws (in L^P_1, hence orthogonal to the zero order laws), and so on.

On the other hand, one can fix a maximal tolerated relative mean prediction error $\bar{\varepsilon}^{tol}_t \in \mathbf{R}$ for predictive laws of (truly) order t. The procedure $P^{*P}_{\varepsilon_{tol}}(\tilde{w})$ corresponding to minimizing the complexity lexicographically under the constraint $\varepsilon^P_{t,1}(\tilde{w},B) < \bar{\varepsilon}^{tol}_t$, $t \in \mathbf{Z}_+$, again is difficult to implement algorithmically. Therefore we will consider a slightly different procedure $P^P_{\varepsilon_{tol}}$, in analogy with $P^D_{\varepsilon_{tol}}$. This procedure corresponds to first finding a maximal number of zero order relations, then minimizing the misfit of these, subsequently maximizing the number of first order relations and minimizing their predictive misfit, and so on. Due to proposition 5–3 there is a close relationship between $P^P_{\varepsilon_{tol}}$ and $P^{*P}_{\varepsilon_{tol}}$. However, they are not equivalent, due to the auxiliary minimizatiòn of the misfit.

We define $\bar{P}^P_{\varepsilon_{tol}}$ in analogy with $P^P_{\varepsilon_{tol}}$, replacing the constraints $\varepsilon^P_{t,1}(\tilde{w},B) < \bar{\varepsilon}^{tol}_t$ by $\varepsilon^P_{t,1}(\tilde{w},B) \leq \bar{\varepsilon}^{tol}_t$.

For completeness we define $P^P_{c_{tol}}$ and $P^P_{\varepsilon_{tol}}$ explicitly.

Definition 6–4 For given $c_{tol} \in (\mathbf{R}_+)^{\mathbf{Z}_+}, \varepsilon_{tol} \in (\mathbf{R}^{1xq})^{\mathbf{Z}_+}$ with $\varepsilon^{tol}_t = \bar{\varepsilon}^{tol}_t.(1,\ldots,1)$, $\bar{\varepsilon}^{tol}_t \in \mathbf{R}$, the procedures $P^P_{c_{tol}}: D \to 2^{\mathbf{B}}$ and $P^P_{\varepsilon_{tol}}: D \to 2^{\mathbf{B}}$ are defined as follows. For $\tilde{w} \in D$, $P^P_{c_{tol}}(\tilde{w}) := \text{argmax}\{ u_{c_{tol}}(c(B), \varepsilon^P(\tilde{w},B)); B \in \mathbf{B} \}$ and $P^P_{\varepsilon_{tol}}(\tilde{w}) := \text{argmax}\{ u_{\varepsilon_{tol}}(c(B), \varepsilon^P(\tilde{w},B)); B \in \mathbf{B} \}$, with

the orderings for $u_{c_{tol}}$ and $u_{\epsilon_{tol}}$ defined as in the definition of $P^D_{c_{tol}}$ and $P^D_{\epsilon_{tol}}$ respectively.

We finally remark that for univariate time series, i.e., $q=1$, the descriptive and predictive procedures are equivalent. That is, for $\tilde{w}\in\mathbb{R}^{\mathcal{T}}$ $P^D_{c_{tol}}(\tilde{w})=P^P_{c_{tol}}(\tilde{w})$ for all c_{tol}, and $P^D_{\epsilon_{tol}}(\tilde{w})=P^P_{\epsilon_{tol}}(\tilde{w})$, $\bar{P}^D_{\epsilon_{tol}}(\tilde{w})=\bar{P}^P_{\epsilon_{tol}}(\tilde{w})$ for all ϵ_{tol}.

7. ALGORITHMS

7.1. Introduction

In this section we describe algorithms for the four deterministic approximate modelling procedures of section 6. These algorithms basically consist of sequential application of the results stated in propositions 2–8 and 2–9 in section 2.6 and propositions 2–13 and 2–14 in section 2.7. Before giving a detailed description of the algorithms we first introduce some concepts and notation and illustrate the approach by describing $P^D_{c_{tol}}$ in general terms.

Let the data consist of a finite time series $\tilde{w}\in(\mathbb{R}^q)^{\mathcal{T}}$ with $\mathcal{T}=[t_0,t_1]$. Let $0\le d\le t_1-t_0$ and $\tau(\mathcal{T},d):=t_1-t_0-d+1$, then for $r\in\mathbb{R}^{1\times q}[s,s^{-1}]$, $r=\sum_{k=n}^{n+d} r_k s^k$, $r_k\in\mathbb{R}^{1\times q}$, $r_n\neq 0\neq r_{n+d}$, there holds $\|r\tilde{w}\|^2:=\frac{1}{\tau(\mathcal{T},d)}\cdot\sum_{t=t_0-n}^{t_1-n-d}\{\sum_{k=n}^{n+d} r_k\tilde{w}(t+k)\}^2=$ $v_d(r).S(w,d).v_d(r)^T$ where $S(\tilde{w},d):=\frac{1}{\tau(\mathcal{T},d)}\cdot\sum_{t=t_0}^{t_1-d}(\tilde{w}(t)^T,\ldots,\tilde{w}(t+d)^T)^T.(\tilde{w}(t)^T,\ldots,\tilde{w}(t+d)^T)$ is the empirical covariance matrix of order d.

The algorithms consist of constructing complementary spaces $\{V_t;\ t\in\mathbb{Z}_+\}$. The corresponding models $B\in\mathfrak{B}$ are then defined in terms of $L_t:=v_t^{-1}(V_t)$ by $B:=\{w\in(\mathbb{R}^q)^{\mathbb{Z}};\ r(\sigma)w=0$ for all $r\in L_t,\ t\in\mathbb{Z}_+\}$. Here $L_t=\{0\}$ for t sufficiently large.

The models identified by the algorithms coincide with the models corresponding to the procedures of section 6 for specifications of c_{tol} and ϵ_{tol} which are in accordance with the number of observations and for generic data. In general terms, one should not allow laws for which the order is too large in comparison with the number of data. Moreover, the algorithms generate optimal models for λ–generic data, i.e., non–optimality

only can arise in a subset N of $(\mathbb{R}^q)^{\mathcal{T}}$ for which $(\mathbb{R}^q)^{\mathcal{T}} \backslash N$ contains an open set of full Lebesgue measure in $(\mathbb{R}^q)^{\mathcal{T}}$.

We will illustrate the foregoing by considering $P^D_{c_{tol}}$. We will make a sensibility assumption on c_{tol} which is related to the number of observations. Moreover we will make some generic assumptions on the data.

First, in order that the descriptive misfit $e^D(\tilde{w},r) := \|r\tilde{w}\|/\|r\|$ is well-defined, it is required that $d := d(r) \le t_1 - t_0$. Moreover, $\{\varepsilon^D(\tilde{w},r)\}^2 = \|r\|^{-2}.v_d(r).S(\tilde{w},d).v_d(r)^T$, with $\operatorname{rank}(S(\tilde{w},d)) \le \min\{t_1-t_0-d+1,\ q(d+1)\}$. If $t_1-t_0-d+1 < q(d+1)$, then for any $\tilde{w} \in (\mathbb{R}^q)^{\mathcal{T}}$ there exists an r with $d(r) \le d$ and $e^D(\tilde{w},r) = 0$. To prevent overparametrization it is reasonable at least to require $t_1-t_0-d+1 \ge q(d+1)$,i.e., $d \le \bar{d}(\mathcal{T}) := (t_1-t_0+1-q)/(q+1)$. This restricts the set of laws for which the quality can be reasonably assessed, and implies restrictions on the requirements in c_{tol} to be sensible. In order to state this exactly as well as some generic assumptions on the data, we consider for given $c_{tol} \in (\mathbb{R}_+)^{\mathbb{Z}_+}$ the class of allowable models $B \in \mathbb{B}$ for which $c_t(B) \le c_t^{tol}$ for all $t \in \mathbb{Z}_+$ and the corresponding class of tightest equation structures $E(c_{tol}) := \{(e_t^*;\ t \in \mathbb{Z}_+);\ \exists B \in \mathbb{B},\ c_t(B) \le c_t^{tol}$ for all $t \in \mathbb{Z}_+$, such that $(e_t^*;\ t \in \mathbb{Z}_+)$ is the tightest equation structure of $B\}$. Equip $E(c_{tol})$ with the lexicographic ordering, and let $e(c_{tol})$ be the corresponding minimal element of $E(c_{tol})$.

Definition 7-1 For given tolerated complexity c_{tol}, the *equation structure corresponding to* c_{tol} is defined as the minimal achievable tightest equation structure of tolerated models in $\mathbb{B}$ with respect to the lexicographic ordering.

We will now first state the assumptions and then comment on them.

Assumption 7-2 Let $c_{tol} \in (\mathbb{R}_+)^{\mathbb{Z}_+}$ and $\tilde{w} \in (\mathbb{R}^q)^{\mathcal{T}}$ be given.

(*i*) $\max\{t;\ e_t(c_{tol}) \ne 0\} \le \bar{d}(\mathcal{T}) := (t_1-t_0+1-q)/(q+1)$;

(*ii*) $P^D_{c_{tol}}(\tilde{w}) = \{B\}$, i.e., a singleton;

(*iii*) B has tightest equation structure $e(c_{tol})$.

Proposition 7-3 Given (i), then (ii) and (iii) hold true for generic data $\tilde{w}$.

Assumption 7-2(i) expresses a sensibility requirement for c_{tol}, as

equations of order more than $\bar{d}(\mathcal{T})$ are not sensible. Assumption 7-2(iii) also expresses a sensibility requirement which we only illustrate in detail for $e_0^*(B) = e_0(c_{tot})$, as the other requirements have a similar interpretation. The condition $c_0(B) \leq c_0^{tot}$ implies that at least $q - c_0^{tot}$ zero order laws need to be accepted. Let n_0 denote the number of independent equations of order zero which are exactly satisfied by the data $\tilde{w}$. It is reasonable to suppose that $q - c_0^{tot} \geq n_0$. In this case any optimal model B has a tightest equation structure $(e_t^*(B); t \in \mathbb{Z}_+)$ with $e_0^*(B) = q - c_0^{tot}$, which is minimal in view of the requirement $c_0(B) \leq c_0^{tot}$. That $e_0^*(B) = q - c_0^{tot}$ for optimal models B is seen as follows. Let $e_0^*(B) > q - c_0^{tot} \geq n_0$. It follows from the definition of ε^D in section 5.2 that $\varepsilon^D_{0, q-c_0^{tot}-n_0+1}(\tilde{w}, B) > 0$. As the ordering on ε^D is lexicographic, an optimal model should satisfy $e_0^*(B) = q - c_0^{tot}$, because models with $e_0^*(B) < q - c_0^{tot}$ are not allowed and models with $e_0^*(B) > q - c_0^{tot}$ can be improved by deleting an equation. Similarly, once $B_{t-1}^{\perp}$ has been identified, the requirements in c_{tot} imply a minimal required number e_t of truly t-th order laws in the space $v_t^{-1}\{ [v_t(B_{t-1}^{\perp} + sB_{t-1}^{\perp})]^{\perp} \}$. Let n_t denote the number of independent t-th order equations in this space which are exactly satisfied by the data. Under the reasonable assumption that $e_t \geq n_t$ it follows that for optimal models $e_t^*(B) = e_t$. Roughly stated, due to the lexicographic ordering it is preferable to accept as few low order equations as possible, given the complexity constraint.

It can be shown that for generic data $\tilde{w}$ there holds $n_t = 0$ for all $t \leq \bar{d}(\mathcal{T})$. So in this case assumption (iii) is satisfied

Under assumption 7-2, due to the lexicographic ordering on ε^D we first have to identify $e_0(c_{tot})$ zero order equations of minimal misfit. In the following section it will be assumed that this problem has a unique solution. This holds true for generic data. Let the solution be L_0 and define $B_0^{\perp} := L_0$, $V_0 := v_0(L_0)$. Next we have to identify $e_1(c_{tot})$ equations of first order and minimal misfit, under the restriction that the equations are truly first order, i.e., orthogonal to $B_0^{\perp} + sB_0^{\perp}$. A second (generically satisfied) assumption is that this problem also has a unique solution, say L_1. Let $V_1 := v_1(L_1) \perp v_1(B_0^{\perp} + sB_0^{\perp})$ and $B_1^{\perp} := B_0^{\perp} + sB_0^{\perp} + L_1$. In the same way we identify $e_t(c_{tot})$ equations of truly t-th order of minimal misfit. It is assumed that this problem has a unique solution L_t. Let $V_t := v_t(L_t)$ and $B_t^{\perp} := B_{t-1}^{\perp} + sB_{t-1}^{\perp} + L_t$. The resulting model is then defined by $B := \{ w \in (\mathbb{R}^q)^{\mathbb{Z}} ; r(\sigma)w = 0 \text{ for all } r \in \bigcup_{t \geq 0} B_t^{\perp} \}$. For this B there holds $L_t = L_t^D$ of (CDF). Moreover, for generic data $\tilde{w}$ the model B is uniquely defined by $\tilde{w}$

and gives the optimal model $P^D_{c_{tol}}(\tilde{w})$.

Note that the foregoing consists of sequential optimal choice of $e_t(c_{tol})$ descriptive equations of minimal misfit. Every step of this sequential optimization will be solved by means of an algorithm corresponding to proposition 2–8.

In the next sections we describe computational details of this algorithm and the other ones. We specify input, initialization, recursive part, termination and output of the algorithms. Moreover, we state the optimality properties of the resulting models in terms of assumptions on the data which are generically satisfied. We refer also to Willems [15] and Heij [4].

In the algorithms we will use the notation $A = \mathrm{col}(A_1, \ldots, A_n)$ to indicate the matrix $A \in \mathbb{R}^{l \times m}$ with blockrows $A_i \in \mathbb{R}^{l_i \times m}$, $i = 1, \ldots, n$, where $l := \sum_{i=1}^{n} l_i$.

7.2. Descriptive modelling, given tolerated complexity

In this section we describe an algorithm which for generic data $\tilde{w} \in (\mathbb{R}^q)^{\mathcal{T}}$ and sensible tolerated complexity c_{tol} generates the model $\{B\} = P^D_{c_{tol}}(\tilde{w})$ as defined in section 6.2. We first give the algorithm and subsequently state the generic conditions on the data.

Algorithm for $P^D_{c_{tol}}$.

1. *Input.*

1.1. Data $\tilde{w} = (\tilde{w}(t);\ t \in \mathcal{T} = [t_0, t_1]) \in (\mathbb{R}^q)^{\mathcal{T}}$.

1.2. Tolerated complexity $c_{tol} = (c_t^{tol};\ t \in \mathbb{Z}_+) \in (\mathbb{R}_+)^{\mathbb{Z}_+}$.

Let $e_{tol} := e(c_{tol})$ denote the equation structure corresponding to c_{tol}.

2. *Initialization (step 0).*

2.1. Let $S(\tilde{w}, 0) := \frac{1}{t_1 - t_0 + 1} \cdot \sum_{t=t_0}^{t_1} \tilde{w}(t)\tilde{w}(t)^T$, the empirical covariance matrix of order 0, have singular value decomposition (SVD) $S(\tilde{w}, 0) = U_0 \Sigma_0 U_0^T$, $\Sigma_0 = \mathrm{diag}(\sigma_1^{(0)}, \ldots, \sigma_q^{(0)})$, $\sigma_1^{(0)} \geq \ldots \geq \sigma^{(0)}_{q - e_0^{tol}} \geq \sigma^{(0)}_{q - e_0^{tol}+1} \geq \ldots \geq \sigma_q^{(0)} \geq 0$.

2.2. If $U_0 = (u_1^{(0)}, \ldots, u_q^{(0)})$, $u_k^{(0)} \in \mathbb{R}^q$, $k = 1, \ldots, q$, then define $V_0 := \mathrm{span}\{u_k^{(0)T};\ k \geq q - e_0^{tol} + 1\}$ and $B_0^{\perp} := v_0^{-1}(V_0)$.

2.3. Define $p_1 := 2e_0^{tol}$ and let $\{v_k^{(1)T};\ k = 1, \ldots, p_1\}$ be an orthonormal basis

of $v_1(B_0^{\perp}+sB_0^{\perp})\subset \mathbb{R}^{1\times 2q}$, e.g., $v_k^{(1)T}$ is the $k-$th row of $\begin{bmatrix} \bar{U}_0 & 0 \\ 0 & \bar{U}_0 \end{bmatrix}$ where $\bar{U}_0 := \mathrm{col}(u_k^{(0)T};\ k=q-e_0^{t\alpha l}+1,\ldots,q)$.

3. *Recursion (step t).*

3.0. Input from step $t-1$: an orthonormal basis $\{v_k^{(t)T};\ k=1,\ldots,p_t\}$ of $v_t(B_{t-1}^{\perp}+sB_{t-1}^{\perp}) \subset \mathbb{R}^{1\times q(t+1)}$, where $p_t = \dim(v_t(B_{t-1}^{\perp}+sB_{t-1}^{\perp})) = \sum_{k=0}^{t-1}(t+1-k).\ e_k^{t\alpha l}$.

SVD: $\sum_{k=1}^{p_t} v_k^{(t)} v_k^{(t)T} = \bar{V}_t \bar{\Sigma}_t \bar{V}_t^T$, $\bar{\Sigma}_t = \mathrm{diag}(\bar{\sigma}_1^{(t)},\ldots,\bar{\sigma}_{q(t+1)}^{(t)})$, $1=\bar{\sigma}_1^{(t)}=\ldots=\bar{\sigma}_{p_t}^{(t)} > \bar{\sigma}_{p_t+1}^{(t)}=\ldots=\bar{\sigma}_{q(t+1)}^{(t)}=0$, $\bar{V}_t = (v_1^{(t)},\ldots,v_{p_t}^{(t)},v_{p_t+1}^{(t)},\ldots,v_{q(t+1)}^{(t)})$. Let $q_t := q(t+1)-p_t$ and define $P_t := \mathrm{col}(v_k^{(t)T}; k=p_t+1,\ldots,q(t+1)) \in \mathbb{R}^{q_t\times q(t+1)}$. So the rows of P_t form an orthonormal basis for $[v_t(B_{t-1}^{\perp}+sB_{t-1}^{\perp})]^{\perp} \subset \mathbb{R}^{1\times q(t+1)}$.

3.1. Let $S(\tilde{w},t) := \frac{1}{t_1-t_0-t+1}\cdot\sum_{k=t_0}^{t_1-t}(\tilde{w}(k)^T,\ldots,\tilde{w}(k+t)^T)^T.(\tilde{w}(k)^T,\ldots,\tilde{w}(k+t)^T)$, the empirical covariance matrix of order t, and let $P_tS(\tilde{w},t)P_t^T$ have SVD $P_tS(\tilde{w},t)P_t^T = U_t\Sigma_tU_t^T$, $\Sigma_t = \mathrm{diag}(\sigma_1^{(t)},\ldots,\sigma_{q_t}^{(t)})$, $\sigma_1^{(t)} \geq \ldots \geq \sigma_{q_t-e_t^{t\alpha l}}^{(t)} \geq \sigma_{q_t-e_t^{t\alpha l}+1}^{(t)} \geq \ldots \geq \sigma_{q_t}^{(t)} \geq 0$.

3.2. If $U_t = (u_1^{(t)},\ldots,u_{q_t}^{(t)})$, $u_k^{(t)} \in \mathbb{R}^{q_t}$, $k=1,\ldots,q_t$, then define $V_t := \mathrm{span}\{u_k^{(t)T}.P_t;\ k \geq q_t - e_t^{t\alpha l}+1\}$, $L_t := v_t^{-1}(V_t) \subset \{r\in\mathbb{R}^{1\times q}[s];\ r=\sum_{k=0}^{t} r_k s^k,\ r_k\in\mathbb{R}^{1\times q},\ k=0,\ldots,t\}$ and $B_t^{\perp} := B_{t-1}^{\perp}+sB_{t-1}^{\perp}+L_t$.

3.3. Output to step $t+1$: an orthonormal basis $\{v_k^{(t+1)T};\ k=1,\ldots,p_{t+1}\}$ of $v_{t+1}(B_t^{\perp}+sB_t^{\perp})$, $p_{t+1} := \sum_{k=0}^{t}(t+2-k).e_k^{t\alpha l}$.

Note that $O_t := \{v_k^{(t)T};\ k=1,\ldots,p_t\} \cup \{u_k^{(t)T}.P_t;\ k=q_t-e_t^{t\alpha l}+1,\ldots,q_t\}$ forms an orthonormal basis of $v_t(B_t^{\perp})$, with $\dim(O_t) = \sum_{k=0}^{t}(t+1-k)e_k^{t\alpha l}$. Let $O_t^0 := \{(v,0);\ v\in O_t,\ 0\in\mathbb{R}^{1\times q}\}$ and ${}^0O_t := \{(0,v);\ 0\in\mathbb{R}^{1\times q},\ v\in O_t\}$, then it suffices to choose $\sum_{k=0}^{t} e_k^{t\alpha l}$ orthonormal vectors in span ${}^0O_t$, orthogonal to O_t^0.

4. *Termination (at step t^*).*

Either at $t^* = \bar{d}(\mathcal{T}) := (t_1-t_0+1-q)/(q+1)$, or at $t^* < \bar{d}(\mathcal{T})$ when $\sum_{t=0}^{t^*} e_t^{t\alpha l} = q$.

5. *Output.*

Bases for V_t, $t \le t^*$, and $B^{\perp}_{t^*}$. Define $B := \{w \in (\mathbb{R}^q)^{\mathbb{Z}};\ r(\sigma)w = 0,\ r \in B^{\perp}_{t^*}\}$.

We remark that the algorithm basically consists of *sequential* application of proposition 2–8 in section 2.6. In the initialization the data is $x_i := \tilde{w}(t_0+i)$, $i = 0,\dots,t_1-t_0$. In step t of the recursion the data consists of $x_i := P_t.\mathrm{col}(\tilde{w}(t_0+i),\dots,\tilde{w}(t_0+i+t))$, $i = 0,\dots,t_1-t_0-t$. The operators P_t take care of the requirement that the new laws should be orthogonal to the old ones. Concerning step 3.1 note that for laws r with $d(r) = t$ and $v_t(r) \in [v_t(B^{\perp}_{t-1} + sB^{\perp}_{t-1})]^{\perp}$ there holds $\|r\tilde{w}\|^2 = v_t(r).P_t.S(\tilde{w},t).P_t^T.v_t(r)^T$.

Next we state the assumptions on $\tilde{w}$ and c_{tol}.

Assumption 7–4 ($P^D_{c_{tol}}$). Let $c_{tol} \in (\mathbb{R}_+)^{\mathbb{Z}_+}$ and $\tilde{w} \in (\mathbb{R}^q)^{\mathcal{T}}$ be given.

(*i*) assumption 7–2(i);

(*ii*) $\sigma^{(0)}_{q-e_0^{tol}} > \sigma^{(0)}_{q-e_0^{tol}+1}$; in step t $\sigma^{(t)}_{q_t-e_t^{tol}} > \sigma^{(t)}_{q_t-e_t^{tol}+1}$;

(*iii*) for step t, let $u_k^{(t)T}.P_t = (u_{k,0},\dots,u_{k,t})$, $u_{k,j} \in \mathbb{R}^{1\times q}$, and $U_0 := \mathrm{col}\{u_{k,0};\ k \ge q_t - e_t^{tol}+1\}$, $U_t := \mathrm{col}\{u_{k,t};\ k \ge q_t - e_t^{tol}+1\}$; assume $\mathrm{rank}(U_0) = \mathrm{rank}(U_t) = e_t^{tol}$.

Assumption (i) expresses a sensibility requirement for c_{tol}. Assumption (ii) is satisfied for generic data and guarantees the existence of a unique solution for the problem of optimal choice of e_t^{tol} equations of order t, orthogonal to $B^{\perp}_{t-1} + sB^{\perp}_{t-1}$. Assumption 7–4(ii) implies assumption 7–2(ii) and (iii). Assumption 7–4(iii) is satisfied for generic data and corresponds to requiring that the laws, identified in step t, really have order t, i.e., $\{0 \ne r \in L_t\} \Rightarrow \{d(r) = t\}$.

Theorem 7–5 Suppose assumption 7–4 is satisfied, then

(*i*) $P^D_{c_{tol}}(\tilde{w}) = \{B\}$, the model generated by the algorithm;

(*ii*) $e^*(B) = e_{tol}$;

(*iii*) $\varepsilon^D_{t,k}(\tilde{w},B) = \{\sigma^{(t)}_{q_t-e_t^{tol}+k}\}^{1/2}$, $k = 1,\dots,e_t^{tol}$;

(*iv*) $L_t = L_t^D$ for B, so the algorithm gives a CDF representation of B.

Optimality of the model generated by the algorithm follows from proposition 2–8, due to the lexicographic ordering on ε^D and assumption 7–4(ii).

It can be shown that the algorithm always generates an allowable

model, i.e., $c_t(B) \le c_t^{tol}$ for all $t \in \mathbb{Z}$. However, the generated model may be suboptimal in case assumption 7–4 is not satisfied, i.e., for non–generic data.

7.3. Descriptive modelling, given tolerated misfit

Next we describe an algorithm which for generic data $\tilde{w} \in (\mathbb{R}^q)^{\mathcal{T}}$ and sensible tolerated misfit generates the model $P^D_{\varepsilon_{tol}}(\tilde{w})$ as defined in section 6.2. The algorithm basically consists of sequential application of proposition 2–9. The (generic) optimality of the model generated by the algorithm is a consequence of proposition 2–9 and the special utility $u_{\varepsilon_{tol}}$ as defined in definition 6–2.

Algorithm for $P^D_{\varepsilon_{tol}}$.

1. *Input.*

1.1. Data $\tilde{w} = (\tilde{w}(t);\ t \in \mathcal{T} = [t_0, t_1]) \in (\mathbb{R}^q)^{\mathcal{T}}$.

1.2. Tolerated misfit $\varepsilon_{tol} = (\varepsilon_t^{tol};\ t \in \mathbb{Z}_+)$, $\varepsilon_t^{tol} = \bar{\varepsilon}_t^{tol}.(1,\ldots,1) \in \mathbb{R}^{1\times q}$, $\bar{\varepsilon}_t^{tol} \in \mathbb{R}$.

2. *Initialization (step 0).*

2.1. SVD: $S(\tilde{w},0) = U_0 \Sigma_0 U_0^T$, $\Sigma_0 = \mathrm{diag}(\sigma_1^{(0)},\ldots,\sigma_q^{(0)})$, $\sigma_1^{(0)} \ge \ldots \ge \sigma_{q-e_0}^{(0)} \ge (\bar{\varepsilon}_0^{tol})^2 > \sigma_{q-e_0+1}^{(0)} \ge \ldots \ge \sigma_q^{(0)} \ge 0$.

2.2. If $U_0 = (u_1^{(0)},\ldots,u_q^{(0)})$, $u_k^{(0)} \in \mathbb{R}^q$, $k = 1,\ldots,q$, then define $V_0 := \mathrm{span}\{u_k^{(0)T}; k \ge q - e_0 + 1\}$ and $B_0^{\perp} := v_0^{-1}(V_0)$.

2.3. Define $p_1 := 2e_0$ and let $\{v_k^{(1)T};\ k = 1,\ldots,p_1\}$ be an orthonormal basis of $v_1(B_0^{\perp} + sB_0^{\perp}) \subset \mathbb{R}^{1\times 2q}$, e.g., $v_k^{(1)T}$ is the k–th row of $\begin{bmatrix} \bar{U}_0 & 0 \\ 0 & \bar{U}_0 \end{bmatrix}$ where $\bar{U}_0 := \mathrm{col}(u_k^{(0)T};\ k = q - e_0 + 1,\ldots,q)$.

3. *Recursion (step t).*

3.0. Input from step $t-1$: an orthonormal basis $\{v_k^{(t)T};\ k = 1,\ldots,p_t\}$ of $v_t(B_{t-1}^{\perp} + sB_{t-1}^{\perp}) \subset \mathbb{R}^{1\times q(t+1)}$, where $p_t = \dim(v_t(B_{t-1}^{\perp} + sB_{t-1}^{\perp})) = \sum_{k=0}^{t-1}(t+1-k).e_k$, where e_k is the number of accepted k–th order laws. Let $q_t := q(t+1) - p_t$, $e_t' := q - \sum_{k=0}^{t-1} e_k$ and define P_t as in step 3.0 of the algorithm for $P^D_{\varepsilon_{tol}}$.

3.1. SVD: $P_t S(\tilde{w},t) P_t^T = U_t \Sigma_t U_t^T$, $\Sigma_t = \mathrm{diag}(\sigma_1^{(t)},\ldots,\sigma_{q_t}^{(t)})$, $\sigma_1^{(t)} \ge \ldots \ge \sigma_{q_t - e_t''}^{(t)} \ge (\bar{\varepsilon}_t^{tol})^2 > \sigma_{q_t - e_t'' + 1}^{(t)} \ge \ldots \ge \sigma_{q_t}^{(t)} \ge 0$.

3.2. If $U_t = (u_1^{(t)},\ldots,u_{q_t}^{(t)})$, $u_k^{(t)} \in \mathbb{R}^{q_t}$, $k = 1,\ldots,q_t$, then with $e_t := \min\{e_t', e_t''\}$ define $V_t := \mathrm{span}\{u_k^{(t)T}.P_t;\ k \ge q_t - e_t + 1\}$, $L_t := v_t^{-1}(V_t)$ and $B_t^{\perp} :=$

$B^{\perp}_{t-1} + sB^{\perp}_{t-1} + L_t$.

3.3. Output to step $t+1$: an orthonormal basis $\{v_k^{(t+1)T};\ k=1,\ldots,p_{t+1}\}$ of $v_{t+1}(B^{\perp}_t + sB^{\perp}_t)$, $p_{t+1} := \sum_{k=0}^{t}(t+2-k).e_k$. See also step 3.3 of the algorithm for $P^D_{c_{tol}}$.

4. *Termination (at step t^*).*

Either at $t^* = \bar{d}(\mathcal{T})$, or at $t^* < \bar{d}(\mathcal{T})$ when $\sum_{t=0}^{t^*} e_t = q$ or $\bar{\varepsilon}_t^{tol} \leq 0$ for $t > t^*$.

5. *Output.*

Bases for V_t, $t \leq t^*$, and $B^{\perp}_{t^*}$. Define $B := \{w \in (\mathbb{R}^q)^{\mathbb{Z}};\ r(\sigma)w = 0,\ r \in B^{\perp}_{t^*}\}$.

We will make the following assumptions on $\tilde{w}$ and ε_{tol}.

Assumption 7-6 ($P^D_{\varepsilon_{tol}}$). Let $(\bar{\varepsilon}_t^{tol};\ t \in \mathbb{Z}_+) \in \mathbb{R}^{\mathbb{Z}_+}$ and $\tilde{w} \in (\mathbb{R}^q)^{\mathcal{T}}$ be given.

(*i*) $\bar{\varepsilon}_t^{tol} \leq 0$ for all $t > \bar{d}(\mathcal{T})$;

(*ii*) if at t^* $e''_{t^*} > e'_{t^*} (> 0)$, then assume $\sigma^{(t^*)}_{q_t - e'_{t^*}} > \sigma^{(t^*)}_{q_t - e'_{t^*}+1}$;

(*iii*) assumption 7-4(iii), with e_t^{tol} replaced by e_t.

Here (i) expresses a sensibility requirement for ε_{tol}, (ii) is satisfied for generic data and guarantees the uniqueness of $P^D_{\varepsilon_{tol}}(\tilde{w})$, and (iii) is satisfied for generic data and amounts to requiring that the laws, identified in step t, really have order t.

Theorem 7-7 Suppose assumption 7-6 is satisfied, then

(*i*) $P^D_{\varepsilon_{tol}}(\tilde{w}) = \{B\}$, the model generated by the algorithm;

(*ii*) $e^*(B) = (e_t;\ t \in \mathbb{Z}_+)$;

(*iii*) $\varepsilon^D_{t,k}(\tilde{w}, B) = \{\sigma^{(t)}_{q_t - e_t + k}\}^{1/2}$, $k = 1,\ldots,e_t$;

(*iv*) $L_t = L^D_t$ for B, so the algorithm gives a CDF representation of B.

7.4. Predictive modelling, given tolerated complexity

In this section we give an algorithm which for generic data $\tilde{w} \in (\mathbb{R}^q)^{\mathcal{T}}$ and sensible tolerated complexity c_{tol} generates the model $\{B\} = P^P_{c_{tol}}(\tilde{w})$ as defined in section 6.3. We first give the algorithm and subsequently state the generic conditions on the data.

Algorithm for $P^P_{c_{tol}}$.

1. *Input.*

As for $P^D_{c_{tol}}$.

2. *Initialization (step 0).*

2.1. As for $P^D_{c_{tol}}$.

2.2. As for $P^D_{c_{tol}}$.

2.3. Define $p_0 := e_0^{tol}$, $n_0 := e_0^{tol}$ and let $\{v_k^{(0)T};\ k \geq q - e_0^{tol} + 1\}$, $v_k^{(0)} := u_k^{(0)}$, $k \geq q - e_0^{tol} + 1$, be an orthonormal basis of $v_0(B_0^{\perp})$ and $F_0 = v_0(B_0^{\perp})$, where F_0 is as defined in section 4.4.

3. *Recursion (step t).*

3.0. Input from step $t-1$: an orthonormal basis $\{v_k^{(t-1)T};\ k = 1, \ldots, p_{t-1}\}$, $p_{t-1} := \sum_{k=0}^{t-1}(t-k)e_k^{tol}$, of $v_{t-1}(B_{t-1}^{\perp}) \subset \mathbb{R}^{1 \times qt}$, and an orthonormal basis $\{f_k^{(t-1)T};\ k = 1, \ldots, n_{t-1}\}$, $n_{t-1} := \sum_{k=0}^{t-1} e_k^{tol}$, of $F_{t-1} := \{\tilde{r} \in \mathbb{R}^{1 \times q};\ \exists r \in B_{t-1}^{\perp},\ r = \sum_{k=0}^{t-1} r_k s^k$, such that $r_{t-1} = \tilde{r}\}$.

SVD: $\sum_{k=1}^{p_{t-1}} v_k^{(t-1)} v_k^{(t-1)T} = \bar{V}_{t-1} \bar{\Sigma}_{t-1} \bar{V}_{t-1}^T$, $\bar{\Sigma}_{t-1} = \mathrm{diag}(\bar{\sigma}_1^{(t-1)}, \ldots, \bar{\sigma}_{q.t}^{(t-1)})$, $1 = \bar{\sigma}_1^{(t-1)} = \ldots = \bar{\sigma}_{p_{t-1}}^{(t-1)} > \bar{\sigma}_{p_{t-1}+1}^{(t-1)} = \ldots = \bar{\sigma}_{q.t}^{(t-1)} = 0$, $\bar{V}_{t-1} = (v_1^{(t-1)}, \ldots, v_{p_{t-1}}^{(t-1)}, v_{p_{t-1}+1}^{(t-1)}, \ldots, v_{q.t}^{(t-1)})$. Let $q_t := q.t - p_{t-1}$ and define $P_{1t} := \mathrm{col}(v_k^{(t-1)T};\ k = p_{t-1}+1, \ldots, q.t) \in \mathbb{R}^{q_t \times q.t}$.

Similarly, SVD: $\sum_{k=1}^{t-1} f_k^{(t-1)} f_k^{(t-1)T} = \bar{\bar{V}}_{t-1} \bar{\bar{\Sigma}}_{t-1} \bar{\bar{V}}_{t-1}^T$, $\bar{\bar{\Sigma}}_{t-1} = \mathrm{diag}(\bar{\bar{\sigma}}_1^{(t-1)}, \ldots, \bar{\bar{\sigma}}_q^{(t-1)})$, $1 = \bar{\bar{\sigma}}_1^{(t-1)} = \ldots = \bar{\bar{\sigma}}_{n_{t-1}}^{(t-1)} > \bar{\bar{\sigma}}_{n_{t-1}+1}^{(t-1)} = \ldots = \bar{\bar{\sigma}}_q^{(t-1)} = 0$, $\bar{\bar{V}}_{t-1} = (f_1^{(t-1)}, \ldots, f_q^{(t-1)})$. Define $P_{2t} := \mathrm{col}(f_k^{(t-1)T};\ k = n_{t-1}+1, \ldots, q) \in \mathbb{R}^{(q-n_{t-1}) \times q}$.

Finally let $P_t := \begin{bmatrix} P_{1t} & 0 \\ 0 & P_{2t} \end{bmatrix}$. Then the rows of P_t form an orthonormal basis for $[v_t(F_{t-1} \cdot s^t) + v_t(B_{t-1}^{\perp})]^{\perp} \subset \mathbb{R}^{1 \times q(t+1)}$.

3.1 Let $P_t S(\tilde{w}, t) P_t^T = \begin{bmatrix} S_-^{(t)} & S_{-+}^{(t)} \\ S_{+-}^{(t)} & S_+^{(t)} \end{bmatrix}$ with $S_-^{(t)} \in \mathbb{R}^{q_t \times q_t}$,

$S_+^{(t)} \in \mathbb{R}^{(q-n_{t-1})\times(q-n_{t-1})}$, $S_{-+}^{(t)} = S_{+-}^{(t)T} \in \mathbb{R}^{q_t\times(q-n_{t-1})}$.

SVD: $(S_-^{(t)})^{-1/2}.S_{-+}^{(t)}.(S_+^{(t)})^{-1/2} = U_t^- \Lambda_t U_t^{+T}$, $\Lambda_t = \begin{bmatrix} \Sigma_t \\ 0 \end{bmatrix} \in \mathbb{R}^{q_t\times(q-n_{t-1})}$,

$\Sigma_t = \operatorname{diag}(\sigma_1^{(t)},\dots,\sigma_{q-n_{t-1}}^{(t)})$, $\sigma_1^{(t)} \geq \dots \geq \sigma_{e_t^{tol}}^{(t)} \geq \sigma_{e_t^{tol}+1}^{(t)} \geq \dots \geq \sigma_{q-n_{t-1}}^{(t)} \geq 0$.

3.2. If $(S_-^{(t)})^{-1/2}.U_t^- = (\bar{u}_1^{(t)},\dots,\bar{u}_{q_t}^{(t)})$ and $(S_+^{(t)})^{-1/2}.U_t^+ = (\bar{\bar{u}}_1^{(t)},\dots,\bar{\bar{u}}_{q-n_t}^{(t)})$, then for $k \leq e_t^{tol}$ let $u_k^{(t)T} := (-\sigma_k^{(t)}.\bar{u}_k^{(t)T}, \bar{\bar{u}}_k^{(t)T}).P_t \in \mathbb{R}^{1\times q(t+1)}$.

Define $V_t := \operatorname{span}\{u_k^{(t)T};\ k \leq e_t^{tol}\}$, $L_t := v_t^{-1}(V_t)$ and $B_t^\perp := B_{t-1}^\perp + sB_{t-1}^\perp + L_t$.

3.3. Output to step $t+1$: orthonormal bases $\{v_k^{(t)};\ k=1,\dots,p_t\}$ of $v_t(B_t^\perp)$ and $\{f_k^{(t)T};\ k=1,\dots,n_t\}$ of F_t. Here $p_t := p_{t-1} + \sum_{k=0}^{t} e_k^{tol}$ and $n_t := n_{t-1} + e_t^{tol}$.

Note that a basis for F_t is $\{f_k^{(t-1)T};\ k=1,\dots,n_{t-1}\} \cup \{\bar{\bar{u}}_k^{(t)T}.P_{2t};\ k \leq e_t^{tol}\}$. Further, let $O_{t-1} := \{v_k^{(k-1)T};\ k=1,\dots,p_{t-1}\}$, $O_{t-1}^0 := \{(v,0);\ v\in O_{t-1},\ 0\in\mathbb{R}^{1\times q}\}$ and ${}^0O_{t-1} := \{(0,v);\ 0\in\mathbb{R}^{1\times q},\ v\in O_{t-1}\}$. For $v_t(B_t^\perp)$ it then suffices to take O_{t-1}^0, V_t, and n_{t-1} orthonormal vectors in span ${}^0O_{t-1}$, orthogonal to $O_{t-1}^0 + V_t$.

4. *Termination (at step t^*).*

As for $P_{c_{tol}}^D$.

5. *Output.*

Bases for V_t, $t \leq t^*$, and $B_{t^*}^\perp$. Define $B := \{w\in(\mathbb{R}^q)^{\mathbb{Z}};\ r(\sigma)w = 0,\ r\in B_{t^*}^\perp\}$.

We remark that the algorithm basically consists of *sequential* application of proposition 2–13 of section 2.7. As a rough outline, $P_{c_{tol}}^P$ models data by successively minimizing the misfit of a required number e_0^{tol} of zero order laws, then minimizing the predictive misfit of a required number e_1^{tol} of first order laws, and so on. In order to measure the misfit more or less independently, as made precise in section 5.3, the newly identified laws r of order t have to be elements of the space $[v_t(F_{t-1}.s^t) + v_t(B_{t-1}^\perp)]^\perp$, see section 4.4. The operator P_t takes care of this requirement. The resulting optimization problem of step t of the recursion is of a static nature as described in section 2.7. The data consists of (x_i, y_i), $i = 0,\dots,t_1 - t_0 - t$, with $y_i := P_{2t}\tilde{w}(t_0+t+i)$ and $x_i := P_{1t}.\operatorname{col}(\tilde{w}(t_0+i),\dots,\tilde{w}(t_0+t-1+i))$.

Next we state the assumption on $\tilde{w}$ and c_{tol}.

Assumption 7–8 ($P_{c_{tol}}^P$). Let $c_{tol}\in(\mathbb{R}_+)^{\mathbb{Z}_+}$ and $\tilde{w}\in(\mathbb{R}^q)^{\mathcal{T}}$ be given.

(*i*) assumption 7–2(i);

(*ii*) $\sigma^{(0)}_{q-e_0^{tol}} > \sigma^{(0)}_{q-e_0^{tol}+1}$; in step t $\sigma^{(t)}_{e_t^{tol}} > \sigma^{(t)}_{e_t^{tol}+1}$;

(*iii*) for step t, let $u_k^{(t)T} = (u_{k,0}, \ldots, u_{k,t})$, $u_{k,j} \in \mathbb{R}^{1\times q}$, and $U_0 := \mathrm{col}\{u_{k,0};\ k \le e_t^{tol}\}$, $U_t := \mathrm{col}\{u_{k,t};\ k \le e_t^{tol}\}$; assume $\mathrm{rank}(U_0) = \mathrm{rank}(U_t) = e_t^{tol}$;

(*iv*) for step t, $S_-^{(t)}$ and $S_+^{(t)}$ have full rank.

Here (i) is a sensibility requirement for c_{tol}. Assumption (ii) is satisfied for generic data and implies assumption 7–2(ii) and (iii). Assumption (iii) also is satisfied for generic data and corresponds to requiring that the laws, identified in step t, really have order t, i.e., $\{0 \neq r \in L_t\} \Rightarrow \{d(r) = t\}$. Also, given assumption (i), assumption (iv) is satisfied for generic data, which is seen as follows. For step t, the number of data is $t_1 - t_0 - t + 1$ and $S_-^{(t)} \in \mathbb{R}^{q_t \times q_t}$, $S_+^{(t)} \in \mathbb{R}^{(q-n_{t-1})\times(q-n_{t-1})}$. As $q_t \le q.t$, $q - n_{t-1} \le q.t$, $S_-^{(t)}$ and $S_+^{(t)}$ generically have full rank if $t_1 - t_0 - t + 1 \ge q.t$, i.e., $t \le (t_1 - t_0 + 1)/(q+1)$, which is implied by assumption (i).

The following theorem is a consequence of proposition 2–13 and the lexicographic ordering of ε^P.

Theorem 7–9 Suppose assumption 7–8 is satisfied, then

(*i*) $P^P_{c_{tol}}(\tilde{w}) = \{B\}$, the model generated by the algorithm;

(*ii*) $e^*(B) = e_{tol}$;

(*iii*) $\varepsilon^P_{t,k}(\tilde{w}, B) = \{1 - (\sigma^{(t)}_{e_t^{tol}-k+1})^2\}^{1/2}$, $k = 1, \ldots, e_t^{tol}$;

(*iv*) $L_t = L_t^P$ for B, so the algorithm gives a CPF representation of B.

7.5. Predictive modelling, given tolerated misfit

Finally we give an algorithm which for generic data $\tilde{w} \in (\mathbb{R}^q)^{\mathcal{T}}$ and sensible ε_{tol} generates the model $P^P_{\varepsilon_{tol}}(\tilde{w})$ as defined in section 6.3. The algorithm basically consists of sequential application of proposition 2–14 of section 2.7. The (generic) optimality of the model generated by the algorithm is a consequence of proposition 2–14 and the special utility $u_{\varepsilon_{tol}}$ as defined in definition 6–2.

Algorithm for $P^P_{\varepsilon_{tol}}$.

1. *Input.*

As for $P^D_{\varepsilon_{tol}}$

2. *Initialization* (*step* 0).

2.1. As for $P^D_{\varepsilon_{tol}}$.

2.2. As for $P^D_{\varepsilon_{tol}}$.

2.3. As for $P^P_{c_{tol}}$, with e_0^{tol} replaced by e_0.

3. *Recursion* (*step* t).

3.0. As for $P^P_{c_{tol}}$, with e_k^{tol} replaced e_k, $k \le t-1$; let $e_t' := q - \sum_{k=0}^{t-1} e_k$.

3.1. As for $P^P_{c_{tol}}$. Let $0 \le 1-(\sigma_1^{(t)})^2 \le \ldots \le 1-(\sigma_{e_t''}^{(t)})^2 < (\bar{\varepsilon}_t^{tol})^2 \le 1-(\sigma_{e_t''+1}^{(t)})^2 \le \ldots \le 1-(\sigma_{q-n_{t-1}}^{(t)})^2 \le 1$.

3.2. As for $P^P_{c_{tol}}$, with e_t^{tol} replaced by $e_t := \min\{e_t', e_t''\}$.

3.3. As for $P^P_{c_{tol}}$, with e_t^{tol} replaced by e_t.

4. *Termination* (*at step* t^*).

As for $P^D_{\varepsilon_{tol}}$.

5. *Output*.

Bases for V_t, $t \le t^*$, and $B^{\perp}_{t^*}$. Define $B := \{w \in (\mathbb{R}^q)^{\mathbb{Z}};\ \tau(\sigma)w = 0,\ \tau \in B^{\perp}_{t^*}\}$.

Assumption 7–10 ($P^P_{\varepsilon_{tol}}$).

(*i*) assumption 7–6(i);

(*ii*) assumption 7–6(ii);

(*iii*) assumption 7–8(iii) with e_t^{tol} replaced by e_t;

(*iv*) assumption 7–8(iv).

Again (i) is a sensibility requirement for ε_{tol}. Given (i), the assumptions (ii), (iii) and (iv) are satisfied for generic data.

Theorem 7–11 Suppose assumption 7–10 is satisfied, then

(*i*) $P^P_{\varepsilon_{tol}}(\tilde{w}) = \{B\}$, the model generated by the algorithm;

(*ii*) $e^*(B) = (e_t;\ t \in \mathbb{Z}_+)$;

(*iii*) $\varepsilon^P_{t,k}(\tilde{w}, B) = \{1-(\sigma^{(t)}_{e_t-k+1})^2\}^{1/2}$, $k = 1, \ldots, e_t$;

(*iv*) $L_t = L_t^P$ for B, so the algorithm gives a CPF representation of B.

7.6. Comments

The algorithms described in the foregoing sections allow for a simple numerical implementation of the procedures of section 6. The computational complexity is mainly determined by singular value analysis of empirical covariance matrices and, in the case of predictive modelling, determination of the square root of positive definite matrices. The algorithms have been numerically implemented and employed, e.g., for the simulations described in section 9.

The essential part of the algorithms is the construction of the complementary spaces V_t, either generating a canonical descriptive form or a canonical predictive form. The operators P_t guarantee that newly identified laws are "far" from being implied by the already identified laws. In this way the misfit is measured according to the principles of section 5. This perhaps is one of the main contributions of the paper. In assessing the quality of a model, the simultaneous nature of AR-equations representing a system is fully taken into account. The quality is measured by means of canonical parametrizations, which are not determined by (scientific) theory, but which are based upon the purpose of modelling, i.e. here, description or prediction.

The identified models may be rather sensitive for changes in c_{tol}. For changes in ε_{tol} the identified models only change at discrete critical values. This indicates that fixing the complexity (the structural form) leads to non-robust identified models. Minimizing misfit of a given parametrized model hence often leads to models which are less robust than models obtained by minimizing complexity under the constraint of a maximal tolerated misfit. So in cases where one has no strong reasons to postulate the structure of a phenomenon, it seems preferable to infer approximate structure from the data by imposing a pragmatic requirement of fit.

8. CONSISTENCY

8.1. Definition of consistency

The procedures of section 6 have a clear optimality property as *data modelling* procedures. The identified models are optimal with respect to the utility $u_{c_{tol}}$ or $u_{\varepsilon_{tol}}$. The procedures give a solution for the

identification problem, i.e., given data and the model class $\mathfrak{B}$, a model is chosen from the model class which is optimal in view of a criterion, based on the objective of modelling. It need not be assumed that the data are generated by a phenomenon of a certain structure. This pure data modelling is of interest e.g. in data compression, speech processing, econometrics, and so on.

However, in other cases one wants to construct a good model of the *phenomenon* which generates the data. The identified model then should not only be good with respect to the particular data, but it should be good with respect to the generating system.

In this section we will define a general concept of consistency, reflecting the purpose of constructing models which approximate the generating system in an optimal way. The approach is inspired by Ljung [9], [10]. We also refer to Heij and Willems [5].

Intuitively, a procedure is called *consistent* if the model, identified by the procedure, converges to an *optimal approximation of the generating system* when the number of observations tends to infinity. So in the limit a consistent procedure identifies a model which, within the given model class, is as close as possible to the phenomenon. In this sense a consistent procedure gives a good model of the phenomenon, provided the number of observations is large enough.

To define consistency we introduce some additional concepts. Let the set of conceivable data be $D := \cup\{(\mathbb{R}^q)^n;\ n \in \mathbb{N}\}$, so data $\tilde{w} \in D$ consists of a finite time series $\tilde{w} = (\tilde{w}(t);\ t \in \mathcal{T} = [t_0, t_1])$ in q variables. Let $\#(\mathcal{T}) := t_1 - t_0 + 1$ denote the number of observations. Let $\mathbf{M}$ be a class of models and $\mathbf{G}$ a class of generating systems. It is assumed that the phenomenon generating the data corresponds to a system $G \in \mathbf{G}$. This means that there is a time series $w \in (\mathbb{R}^q)^{\mathbb{Z}}$ compatible with G from which we observe $\tilde{w} = w|_{\mathcal{T}}$.

Suppose that the objectives π have been used to construct a procedure $P: D \to 2^{\mathbf{M}}$. Moreover, assume that π induces an optimal approximation map $A: \mathbf{G} \to 2^{\mathbf{M}}$. This means that, with respect to π, $A(G)$ is the set of optimal approximations within the class $\mathbf{M}$ of the system $G \in \mathbf{G}$. Often $A(G)$ will consist of a singleton. Further, let $\to$ be a concept of convergence in $2^{\mathbf{M}}$, possibly also related to π. Finally, let n.a. denote a concept of "nearly always" for systems $G \in \mathbf{G}$. Such a concept is crucial, as optimal properties of procedures can fail to hold true for nasty data which nearly never occur.

Consistency now is defined as follows.

Definition 8-1 P is called *consistent* if for all $G \in \mathbb{G}$, n.a. in $w \in G$, $P(w|_{\mathcal{T}}) \to A(G)$ if $\#(\mathcal{T}) \to \infty$.

This means that, if the length of the observed time series tends to infinity, the set of models identified by a consistent procedure converges "nearly always" to the set of optimal approximations within $\mathbb{M}$ of the generating system G.

In this paper, $A(\mathbb{G})$ will consist of singleton , i.e., for $G \in \mathbb{G}$ there exists a unique approximation $a(G) \in \mathbb{M}$, so $A(G) = \{a(G)\}$. In this case, let $\to$ be a concept of convergence in $\mathbb{M}$. Then $P: D \to 2^{\mathbb{M}}$ is called *consistent* if for all $G \in \mathbb{G}$, n.a. in $w \in G$, $P(w|_{\mathcal{T}}) = \{M(w|_{\mathcal{T}})\}$, i.e., a singleton, for $\#(\mathcal{T})$ sufficiently large, and $M(w|_{\mathcal{T}}) \to a(G)$ for $\#(\mathcal{T}) \to \infty$. By slight abuse of notation we will indicatate this by $P(w|_{\mathcal{T}}) \to A(G)$.

The consistency problem is depicted in figure 10.

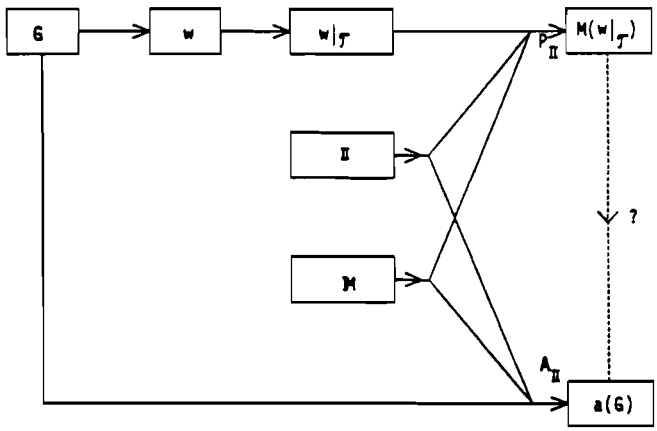

figure 10: consistency

This concept of model consistency differs in some important aspects from the concept of parameter consistency in statistics, see e.g. Kendall and Stuart [8]. In the latter case $\mathbb{M} = \mathbb{G} = \{M(\theta);\ \theta \in \Theta\}$ for some parametrized class of models (probability distributions). The data modelling problem is formulated as an estimation problem, and a modelling procedure is a map $E: D \to \Theta$. The procedure is called consistent if (n.a.) $E(w|_{\mathcal{T}}) \to \theta$ when $\#(\mathcal{T}) \to \infty$, where θ parametrizes the generating system. Model consistency differs in four main respects from this parameter consistency. First, it need not be

assumed that $\mathbf{M}=\mathbf{G}$, i.e., that the generating system belongs to the model class. Second, convergence is defined in terms of models, not in terms of parametrizations. Third, parameter consistency raises problems in case of non-unique parametrizations, model consistency avoids these problems. Fourth, the models need not be stochastic.

For the case of time series analysis, see e.g. Hannan, Dunsmuir and Deistler [3] for parameter consistency and e.g. Ljung and Caines [11] for model consistency.

In the next two sections we investigate consistency of some of the procedures of section 6 for certain classes of generating systems $\mathbf{G}$. In section 8.2. we suppose $\mathbf{G}=\mathbf{B}$, i.e. the phenomenon itself is a linear, time invariant, complete (deterministic) dynamical system. In section 8.3 we consider the case where $\mathbf{G}$ consists of stochastic ARMA models and the purpose π is prediction. For this case we define optimal deterministic approximations of stochastic systems.

8.2. Deterministic generating AR-systems

Let the model class $\mathbf{M}$ again consist of the AR-models, i.e., $\mathbf{M}:=\mathbf{B}$. Suppose that the data are generated by a system $G\in\mathbf{G}=\mathbf{B}$, i.e., the generating system itself is an AR-system, so there exists an exact model of the phenomenon in the model class. In this case it is assumed that there is a system $B\in\mathbf{B}$ such that the data $\tilde{w}\in(\mathbb{R}^q)^{\mathcal{T}}$ is a finite observation of a time series $w\in(\mathbb{R}^q)^{\mathbb{Z}}$ generated by B, i.e., there is $w\in B$ with $\tilde{w}=w|_{\mathcal{T}}$. We restrict attention to so-called controllable systems B, cf. Willems [15].

Let $D:=\cup\{(\mathbb{R}^q)^n;\ n\in\mathbb{N}\}$ and $P:D\to 2^{\mathbf{M}}$ a procedure. To define consistency we specify an optimal approximation map $A:\mathbf{G}\to\mathbf{B}$ and a concept of convergence on $\mathbf{B}$. As $\mathbf{G}=\mathbf{B}$, an obvious choice for A is the identity map. Moreover, we take the discrete topology on $\mathbf{B}$. A procedure P then is consistent if for all $B\in\mathbf{B}$, n.a. in $w\in B$, there holds $P(w|_{\mathcal{T}})=\{B\}$ for $\#(\mathcal{T})$ sufficiently large. In this case, nearly always after observing a sufficiently large finite part of the time series the procedure identifies the generating system exactly.

To define n.a., we use the concept of genericity. Let $V\subset(\mathbb{R}^q)^{\mathcal{T}}$ be a linear subspace. A subset $V'\subset V$ is called generic in V if there is a polynomial $p:V\to\mathbb{R}$, $p\neq 0$ such that the complement of V' in V is contained in $p^{-1}(0)$. For $B\in\mathbf{B}$ we call $B'\subset B$ generic in B if $B'|_{\mathcal{T}}\subset B|_{\mathcal{T}}$ is generic in $B|_{\mathcal{T}}$ for $\#(\mathcal{T})$ sufficiently large. A property now is said to hold true n.a. for B if

the set of points $w \in B$ where the property holds true is generic in B.

In this setting of consistency we first consider the exact modelling procedure P_{uu} as described in section 2.2.2, i.e., the procedure corresponding to undominated unfalsified modelling. So $P_{uu}: D \to 2^{\mathbb{B}}$, where for $\tilde{w} \in (\mathbb{R}^q)^{\mathcal{T}}$ $B \in P_{uu}(\tilde{w})$ if and only if $B \in \mathbb{B}$, B is unfalsified, i.e., $\tilde{w} \in B|_{\mathcal{T}}$, and B is undominated, i.e., $\{\tilde{w} \in B'|_{\mathcal{T}},\ B' \in \mathbb{B},\ B' \subset B\} \Rightarrow \{B' = B\}$.

Proposition 8-2 P_{uu} is not consistent.

As a simple example, take $B = (\mathbb{R}^q)^{\mathbb{Z}}$. For any $w \in B$ and any $\mathcal{T}$ of finite length there exist $B' \in \mathbb{B}$ such that $w|_{\mathcal{T}} \in B'|_{\mathcal{T}}$ and $\dim(B') \leq q.\#(\mathcal{T})$, hence $B \notin P_{uu}(w|_{\mathcal{T}})$.

Next we consider the procedures described in section 6. We define two exact and sensible modelling procedures as follows. For $k \in \mathbb{Z}_+$ let $\bar{\varepsilon}_{tol}(k) = (\bar{\varepsilon}_t^{tol}(k);\ t \in \mathbb{Z}_+) \in \mathbb{R}^{\mathbb{Z}_+}$ be defined by $\bar{\varepsilon}_t^{tol}(k) := 0$ for $0 \leq t \leq \bar{d}(k) := (k-q)/(q+1)$ and $\bar{\varepsilon}_t^{tol}(k) := -1$ for $t > \bar{d}(k)$. Let $\varepsilon_{tol}(k) := (\varepsilon_t^{tol}(k);\ t \in \mathbb{Z}_+)$ with $\varepsilon_t^{tol}(k) := \bar{\varepsilon}_t^{tol}(k).(1,\ldots,1)$. The procedures $\bar{P}^D_{\varepsilon_{tol}(k)}$ and $\bar{P}^P_{\varepsilon_{tol}(k)}$ as defined in sections 6.2 and 6.3 correspond to accepting only exact laws of order at most $\bar{d}(k)$. Now define $P^D(w|_{\mathcal{T}}) := \bar{P}^D_{\varepsilon_{tol}(\#(\mathcal{T}))}(w|_{\mathcal{T}})$ and $P^P(w|_{\mathcal{T}}) := \bar{P}^P_{\varepsilon_{tol}(\#(\mathcal{T}))}(w|_{\mathcal{T}})$. So P^D and P^P accept the exact laws which are significant, given the number of data.

Proposition 8-3 P^D and P^P are consistent on controllable systems.

For fixed c_{tol} or ε_{tol}, i.e., independent of the number of data, the procedures $P^D_{c_{tol}}$, $P^D_{\varepsilon_{tol}}$, $P^P_{c_{tol}}$, $P^P_{\varepsilon_{tol}}$, $\bar{P}^D_{\varepsilon_{tol}}$ and $\bar{P}^P_{\varepsilon_{tol}}$ are not consistent, in the strict sense of exact identification for generic finite time series. We illustrate this for $P^D_{c_{tol}}$ and $P^D_{\varepsilon_{tol}}$. Similar arguments hold true for the other procedures. First suppose c_{tol} is given. Let $e_{tol} := e(c_{tol})$, be the equation structure corresponding to c_{tol}. If $e_{tol} = 0$, then $P^D_{c_{tol}}$ is not consistent for the same reasons as given for P_{uu}. If there is $t \in \mathbb{Z}_+$ with $e_t^{tol} \geq 1$, then $B \in \mathbb{B}$ with $e_t^*(B) = 0$ cannot be exactly identified, hence $P^D_{c_{tol}}$ is not consistent. Next suppose ε_{tol} is given. If $\varepsilon_{t,1}^{tol} \leq 0$ for some $t \in \mathbb{Z}_+$, then exact identification of $B \in \mathbb{B}$ with $e_t^*(B) \geq 1$ is impossible. If $\varepsilon_{t,1}^{tol} > 0$ for all $t \in \mathbb{Z}_+$, then ε_{tol} does not satisfy the sensibility assumption 7-6(i) for any $\mathcal{T}$. Moreover, as $\varepsilon_{0,1}^{tol} > 0$ $P^D_{\varepsilon_{tol}}(w|_{\mathcal{T}})$ will accept laws of order 0 for $w|_{\mathcal{T}} \in (\mathbb{R}^q)^{\mathcal{T}}$ of sufficiently small norm. Not having this sufficiently small norm is not a generic property for any $B \in \mathbb{B}$ with $B \neq \{0\}$. If $B \in \mathbb{B}$ with $e_0^*(B) = 0$, then $P^D_{\varepsilon_{tol}}$

in this case cannot exactly identify B generically, hence $P^D_{\epsilon_{tol}}$ is not consistent.

An interesting question is the relationship between consistency of $P^D_{\epsilon_{tol}}$ and $P^P_{\epsilon_{tol}}$ and a definition of n.a. in terms of *"sufficient excitation"*. Without going into details, the procedures are consistent for the class of controllable systems if n.a. is defined in terms of sufficient excitation of the inputs with respect to ϵ_{tol}. Exact identification then is guaranteed provided the inputs are sufficiently rich with respect to ϵ_{tol}.

8.3. Stochastic generating ARMA-systems

8.3.1. Introduction

In this section we will consider the predictive procedures $P^P_{c_{tol}}$ and $P^P_{\epsilon_{tol}}$ in case the data consist of a finite part of a realization of a stochastic process. In section 8.3.4 we will define the optimal approximation of a stochastic process by a deterministic system, given c_{tol} or ϵ_{tol}. Roughly speaking, the optimal deterministic approximation is described by the predictive relationships corresponding to c_{tol} or ϵ_{tol} in case the stochastic process were known. Note that both deterministic and stochastic systems generally can be given an interpretation in terms of (optimal) one–step–ahead prediction by means of deterministic equations.

A similar exposition could be given for the descriptive procedures $P^D_{c_{tol}}$ and $P^D_{\epsilon_{tol}}$. However, in general it seems difficult to give an interpretation of stochastic systems in terms of deterministic descriptive relationships. Therefore we restrict attention to $P^P_{c_{tol}}$ and $P^P_{\epsilon_{tol}}$.

In the following we introduce a concept of convergence on $\mathbb{B}$, describe a class of generating ARMA–systems, define optimal approximation maps $A^P_{c_{tol}}$and $A^P_{\epsilon_{tol}}$and state consistency results.

8.3.2. Convergence

Let $B_k \in \mathbb{B}$, $k \in \mathbb{N}$, and $B_\infty \in \mathbb{B}$. Then B_k is defined to converge to B_∞ for $k \to \infty$ if there exist parametrizations $B_k = B(R_k)$, $k \in \mathbb{N}$, and $B_\infty = B(R_\infty)$ with the following properties. R_∞ has full row rank over the polynomials, $\{d(R_k);\ k \in \mathbb{N}\}$ is bounded, and $R_k \to R_\infty$ for $k \to \infty$ in Euclidean sense. By this we mean that for k sufficiently large R_k has as many rows as R_∞, and if $R_k =$

$\sum_{j=-\infty}^{\infty} R_j^{(k)} s^j$, $R_j^{(k)} = (r_{lm}^{jk}) \in \mathbb{R}^{p \times q}$, $k \in \mathbb{N} \cup \{\infty\}$, then $\sum_{j=-\infty}^{\infty} \sum_{l=1}^{p} \sum_{m=1}^{q} (r_{lm}^{jk} - r_{lm}^{j\infty})^2 \to 0$ if $k \to \infty$.

This concept of convergence is analysed by Nieuwenhuis and Willems [13]. There it is shown that this convergence in terms of parametrizations is equivalent to a natural concept of convergence of systems, considered as subsets of $(\mathbb{R}^q)^{\mathbb{Z}}$.

8.3.3. Generating stochastic systems

We assume that the generating system belongs to the class $\mathbf{G}$ of stochastic processes $w = \{w(t);\ t \in \mathbb{Z}\}$ which satisfy the following assumption.

Assumption 8-4 (*i*) w is second order stationary with for all $t \in \mathbb{Z}$ $Ew(t) = 0$, $C_k := Ew(t)w(t+k)^T$; (*ii*) almost surely for realizations w_r of w there holds for all $k \in \mathbb{Z}_+$ $\frac{1}{t_1 - t_0 + 1} \cdot \sum_{t=t_0}^{t_1 - k} w_r(t) w_r(t+k)^T \to C_k$ if $|t_1 - t_0| \to \infty$.

A sufficient condition for the assumption to be satisfied is that w is strictly stationary and ergodic, e.g., that w is Gaussian with a spectral distribution Φ which is continuous on the unit circle. We refer to Hannan [2]. This especially holds true for Gaussian ARMA-processes, in which case $\Phi(z) := \sum_{k=-\infty}^{\infty} C_k z^{-k}$ is a rational function with no poles on the unit circle. The process w then has a representation of the following form. There exist $m \in \mathbb{N}$, polynomial matrices $N \in \mathbb{R}^{q \times m}[s]$ and $M \in \mathbb{R}^{q \times q}[s]$ with $\det(M(s)) \neq 0$ on $|s| \leq 1$, and an m-dimensional Gaussian white noise process n, i.e., $En(t) = 0$ and $En(t)n(s)^T = 0$ for $t \neq s$, such that $M(\sigma^{-1})w = N(\sigma^{-1})n$.

The consistency result stated in section 8.3.5 is in terms of generic subclasses of $\mathbf{G}$ which we will define in section 8.3.4. Here genericity is defined as follows. Define $\mathbf{C} \subset (\mathbb{R}^{q \times q})^{\mathbb{Z}}$ as the collection of $(C_k;\ k \in \mathbb{Z})$ for which there exist $w \in \mathbf{G}$ with $C_k = Ew(t)w(t+k)^T$, $k \in \mathbb{Z}$. A subset $\mathbf{C}' \subset \mathbf{C}$ is called generic if for all $-\infty < t_0 \leq t_1 < +\infty$ $\mathbf{C}'|_{[t_0, t_1]}$ is a λ-generic set in $\mathbf{C}|_{[t_0, t_1]}$, i.e., it contains an open subset of full Lebesgue measure in $\mathbf{C}|_{[t_0, t_1]}$. A class of stochastic systems $\mathbf{G}' \subset \mathbf{G}$ is called generic if $\mathbf{C}' := \{(C_k;\ k \in \mathbb{Z});\ \exists w \in \mathbf{G}'$ with $C_k = Ew(t)w(t+k)^T$ for all $k \in \mathbb{Z}\}$ is generic, i.e., if the set of covariance sequences in $\mathbf{G}'$ is λ-generic.

The classes $\mathbf{G}_{c_{tol}}$ and $\mathbf{G}_{\varepsilon_{tol}}$ of section 8.3.4 are generic. Moreover, the Gaussian ARMA-processes in $\mathbf{G}_{c_{tol}}$ and $\mathbf{G}_{\varepsilon_{tol}}$ are generic in the class of all Gaussian ARMA-processes in $\mathbf{G}$. So the consistency results of section 8.3.5 in particular hold true for generic ARMA-processes.

8.3.4. Approximation maps and the classes $\mathbf{G}_{c_{tol}}$, $\mathbf{G}_{\varepsilon_{tol}}$

In this section we construct for a given stochastic process w optimal approximations in $\mathbf{B}$. The optimality has to be understood in the sense of a utility corresponding to the purpose of modelling. For w we define the optimal approximations $A^P_{c_{tol}}(w)$ and $A^P_{\varepsilon_{tol}}(w)$ as the models of optimal prediction of w for c_{tol} and ε_{tol} respectively in case the generating system w were known.

The foregoing is made precise as follows. For $r \in \mathbb{R}^{1\times q}[s,s^{-1}]$ with $d(r)>0$ define the relative expected prediction error in analogy with section 5.3 as $e^P(w,r) := \{ (E\|rw\|^2) / (E\|r^*w\|^2) \}^{1/2}$, where r^* is the leading coefficient vector of r and $E\|rw\|^2 := E\{(r(\sigma,\sigma^{-1})w)(t)\}^2$ which does not depend on t due to stationarity. If $d(r)=0$ then define $e^P(w,r) := \{ E\|rw\|^2 / \|r\|^2 \}^{1/2}$. For $B \in \mathbf{B}$ we define $\varepsilon^P(w,B) \in (\mathbb{R}_+^{1\times q})^{\mathbb{Z}_+}$ exactly analogous to $\varepsilon^P(\tilde{w},B)$ in section 5.3. Hence $\varepsilon^P_{t,1}(w,B)$ measures the largest relative expected prediction error of the truly t-th order predictive laws claimed by B, $t \in \mathbb{Z}_+$, and so on. We now define $A^P_{c_{tol}}(w)$ and $A^P_{\varepsilon_{tol}}(w)$ as the predictive models which are optimal for c_{tol} and ε_{tol} respectively, in case w were known.

Definition 8-5 For $w \in \mathbf{G}$, $A^P_{c_{tol}}(w) := \operatorname{argmax}\{ u_{c_{tol}}(c(B),\ \varepsilon^P(w,B));\ B \in \mathbf{B} \}$ and $A^P_{\varepsilon_{tol}}(w) := \operatorname{argmax}\{ u_{\varepsilon_{tol}}(c(B),\ \varepsilon^P(w,B));\ B \in \mathbf{B} \}$.

So $A^P_{c_{tol}}$ and $A^P_{\varepsilon_{tol}}$ give deterministic approximations of stochastic processes which are optimal in terms of a utility on complexity and predictive quality of models described by (deterministic) autoregressive equations.

In the sequel we will restrict attention to subclasses of $\mathbf{G}$ for which $A^P_{c_{tol}}$ and $A^P_{\varepsilon_{tol}}$ consist of singletons. For $w \in \mathbf{G}$ define $S(w,t) := E[\mathrm{col}(w(t),\ldots,w(t+k)).\mathrm{col}(w(t),\ldots,w(t+k))^T]$, $t \in \mathbb{Z}_+$. Now consider the algorithms of sections 7.4 and 7.5 with $S(\tilde{w},t)$ replaced by $S(w,t)$. Note that any c_{tol} satisfies assumption 7-2(i) for $\#(\mathcal{T})$ sufficiently large. Suppose that ε_{tol} is such that there is a t such that $\varepsilon^{tol}_{s,1} \le 0$ for $s > t$.

Definition 8-6 $\mathbf{G}_{c_{tol}} := \{w \in \mathbf{G};$ assumption 7-8(ii), (iii), (iv) is satisfied$\}$; $\mathbf{G}_{\varepsilon_{tol}} := \{w \in \mathbf{G};$ assumption 7-10(ii), (iii), (iv) is satisfied

and $\sigma^{(0)}_{q-e_0+1} < (\bar{\varepsilon}^{tol}_0)^2 < \sigma^{(0)}_{q-e_0}$, $1-(\sigma^{(t)}_{e'_t})^2 < (\bar{\varepsilon}^{tol}_t)^2 < 1-(\sigma^{(t)}_{e'_t+1})^2$ }.

Proposition 8-7 (*i*) $\mathbf{G}_{c_{tol}}$ and $\mathbf{G}_{\epsilon_{tol}}$ are generic in $\mathbf{G}$;

(*ii*) for $w \in \mathbf{G}_{c_{tol}}$ $A^P_{c_{tol}}(w)$ is a singleton, generated by the algorithm of section 7.4 with $S(\tilde{w},t)$ replaced by $S(w,t)$;

(*iii*) for $w \in \mathbf{G}_{\epsilon_{tol}}$ $A^P_{\epsilon_{tol}}(w)$ is a singleton, generated by the algorithm of section 7.5 with $S(\tilde{w},t)$ replaced by $S(w,t)$.

Moreover, the Gaussian ARMA–processes in $\mathbf{G}_{c_{tol}}$ and $\mathbf{G}_{\epsilon_{tol}}$ are generic in the class of all Gaussian ARMA–processes in $\mathbf{G}$.

8.3.5. Consistency results

Assume that the data $\tilde{w}$ consist of a (finite) observation on $\mathcal{T}$ of a realization $w_\tau \in (\mathbb{R}^q)^{\mathbb{Z}}$ of a stochastic process w. As definition of n.a. in w we take a.s., i.e., "almost sure" with respect to the process . The next theorem states consistency results for $P^P_{c_{tol}}$ and $P^P_{\epsilon_{tol}}$, with the approximation maps as in section 8.3.4 and the concept of convergence as defined in section 8.3.2. It is assumed that for ϵ_{tol} there is a t such that $\epsilon^{tol}_{s,1} \leq 0$ for $s > t$, in which case we call ϵ_{tol} finite.

Theorem 8–8 For every c_{tol}, $P^P_{c_{tol}}$ is consistent on $\mathbf{G}_{c_{tol}}$. For every finite ϵ_{tol}, $P^P_{\epsilon_{tol}}$ is consistent on $\mathbf{G}_{\epsilon_{tol}}$.

This means the following. Let w_τ be a realization of a stochastic process $w \in \mathbf{G}_{c_{tol}}$ and let $\tilde{w} = w_\tau|_{\mathcal{T}}$. Let $A^P_{c_{tol}}(w) = B \in \mathbf{B}$ with corresponding predictive spaces $V^P_t := v_t(L^P_t)$, where L^P_t is as defined in section 4.4. Then almost sure $P^P_{c_{tol}}(\tilde{w})$ is a singleton for $\#(\mathcal{T})$ sufficiently large. Denote the corresponding predictive spaces by $V^P_t(\mathcal{T})$, the complexity by $c(\mathcal{T})$ and the predictive misfit by $\varepsilon(\mathcal{T})$. Then for $\#(\mathcal{T}) \to \infty$ there holds a.s. that $c_t(\mathcal{T}) \to c_t(B)$, $V^P_t(\mathcal{T}) \to V^P_t$ in the Grassmannian topology (i.e., there exist choices of bases of $V^P_t(\mathcal{T})$ which converge to a basis of V^P_t), and $\varepsilon_{t,k}(\mathcal{T}) \to \varepsilon^P_{t,k}(w,B)$, $k = 1,\ldots,q$, $t \in \mathbb{Z}_+$. A similar result holds true for $P^P_{\epsilon_{tol}}$. The convergence $V^P_t(\mathcal{T}) \to V^P_t$ implies convergence of AR–relations and of the corresponding models. So if the number of observations tends to infinity, the identified model a.s. converges to the optimal (prediction) model B which would be identified in case w were known.

Proof of the theorem consists of using the ergodic properties of w and

establishing continuity properties of the steps of the algorithms in sections 7.4 and 7.5 with respect to changes in $S(\tilde{w},t)$, $t\in\mathbf{Z}_+$.

We remark that also the procedure $\bar{P}^P_{\epsilon_{tol}}$ is consistent on $\mathbf{G}_{\epsilon_{tol}}$. Moreover, $P^P_{\epsilon_{tol}}$ is not consistent if ϵ_{tol} is not finite. Note that such ϵ_{tol} is not sensible.

We conclude this section by commenting on the optimality. Consider e.g. $P^P_{\epsilon_{tol}}$ and suppose that $w\in\mathbf{G}_{\epsilon_{tol}}$ is such that $B:=A^P_{\epsilon_{tol}}(w)$ satisfies $\sum_{t=0}^{\infty} e_t^*(B)=q$. Then use of B leads to one-step-ahead pointpredictions, which we indicate by $\hat{w}^*$. In this case a.s. and for $\#(\mathcal{T})$ sufficiently large $P^P_{\epsilon_{tol}}(\tilde{w})$ also leads to pointpredictions, indicated by $\hat{w}(\mathcal{T})$. There holds $E\|\hat{w}^*-\hat{w}(\mathcal{T})\|\to 0$ if $\#(\mathcal{T})\to\infty$. In this sense the one-step-ahead predictions converge to the optimal ones. However, if $q>1$ in general there does not exist a choice of ϵ_{tol} such that $\hat{w}^*$ (and hence $\hat{w}(\mathcal{T})$) is close to the least squares (causal) predictor for w. So the optimality has to be interpreted in terms of $u_{\epsilon_{tol}}$, not in terms of minimal mean square prediction error. It is not unreasonable to be slightly non-optimal in accuracy if the predictions can be made by much simpler models.

9. SIMULATIONS

9.1. Introduction

In this section we will illustrate the modelling procedures of section 6 by means of four simple numerical examples.

In section 9.2 we consider exact modelling. In this case only exactly satisfied laws are accepted. This corresponds to applying the procedures $\bar{P}^D_{\epsilon_{tol}}$ and $\bar{P}^P_{\epsilon_{tol}}$ with $\epsilon_{tol}=0$. The data consists of an exact observation of a time series generated by an AR-system.

Section 9.3 gives an example of descriptive modelling of a time series, given a maximal tolerated complexity, i.e., of the procedure $P^D_{c_{tol}}$. The data consists of a noisy observation of a signal generated by an AR-system. We will compare the (non-causal) impulse response of the generating system with that of the identified model.

In section 9.4 we illustrate the difference between descriptive and predictive modelling. For a given time series we compare the models identified by the procedures $P^D_{\epsilon_{tol}}$ and $P^P_{\epsilon_{tol}}$.

Finally section 9.5 contains a simulation illustrating the fact that the procedures for modelling, given a maximal tolerated misfit, need not generate models of minimal complexity. This indicates the difference between the procedures $P^D_{\epsilon_{tol}}$ ($P^P_{\epsilon_{tol}}$) and $P^{*D}_{\epsilon_{tol}}$ ($P^{*P}_{\epsilon_{tol}}$) as defined in sections 6.2 and 6.3 respectively. We also illustrate consistency of $P^P_{\epsilon_{tol}}$.

9.2. Exact modelling

9.2.1. Data

In the first simulation we consider exact modelling of a signal generated by an AR–system. The signal consists of two components, each being a sum of two sinusoids. To be specific, let $f_1 := 2\pi/100$, $f_2 := 2\pi/120$ and $f_3 := 2\pi/150$. Define $s_k(t) := \sin(f_k \cdot t)$, $k = 1,2,3$, $t \in \mathbb{R}$, and $w_1(t) := s_1(t) + s_2(t)$, $w_2(t) := s_1(t) + s_3(t)$. The data consists of observations of the signals w_1 and w_2 on times $t = 1,\ldots,300$, i.e., $\tilde{w} = (\begin{bmatrix} w_1(t) \\ w_2(t) \end{bmatrix};\ t = 1,\ldots,300) \in (\mathbb{R}^2)^{300}$. The signals are given in figure 11.

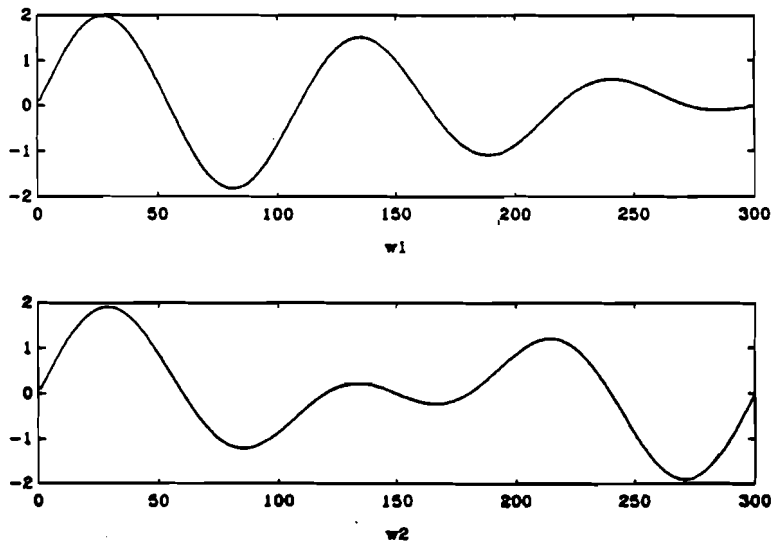

figure 11: data for simulation 9.2.

9.2.2. System

Both w_1 and w_2 are periodic, with period 600 and 300 respectively. Hence $w \in B(R)$ with $R := \begin{bmatrix} \sigma^{600} - 1 & 0 \\ 0 & \sigma^{300} - 1 \end{bmatrix}$. However, there are more powerful models for

w. Observe that for $s(t)=\sin(f.t)$ there holds $s(t+2)+s(t)=2\cos(f).s(t+1)$, hence $s\in B(r)$ with $r(s):=s^2-2\cos(f).s+1=(s-e^{if})(s-e^{-if})$. Defining $p_k(s):=(s-e^{if_k})(s-e^{-if_k})$, $k=1,2,3$, we conclude that $\tilde{w}\in B(R_0)$ with $R_0:=$

$$\begin{bmatrix} p_1\cdot p_2 & 0 \\ 0 & p_1\cdot p_3 \end{bmatrix}.$$

9.2.3. Model identification

Exact models for the data $\tilde{w}$ are obtained by applying the procedures $\bar{P}^D_{\varepsilon_{tol}}$ and $\bar{P}^P_{\varepsilon_{tol}}$ with $\varepsilon_{tol}=0$. We denote the resulting models by $B(R^D):=\bar{P}^D_0(\tilde{w})$ and $B(R^P):=\bar{P}^P_0(\tilde{w})$. These models are identified by using the algorithms of section 7 with $\varepsilon_{tol}=0$. Both models consist of one second order laws and one fourth order law. Let R^D and R^P have elements r^D_{lm} and r^P_{lm} respectively, $l,m=1,2$. The identified laws are given in table 1.

	coefficients of: σ^0	σ^1	σ^2	σ^3	σ^4
laws:					
r^D_{11}	0.5007	-1.0000	0.5007	0	0
r^D_{12}	-0.2754	0.5502	-0.2754	0	0
r^D_{21}	0.4637	-0.9568	0.5746	-0.1319	0.0507
r^D_{22}	-0.0352	-0.3517	1.0000	-0.8055	0.1920
r^P_{11}	1.2392	-2.4750	1.2392	0	0
r^P_{12}	-0.6815	1.3618	-0.6815	0	0
r^P_{21}	0.6815	-2.7224	4.0818	-2.7223	0.6815
r^P_{22}	1.2392	-4.9490	7.4196	-4.9489	1.2391

table 1: identified AR-laws for simulation 9.2.

9.2.4. Model validation

Two questions arise, namely, whether these AR-laws are equivalent and whether they are equivalent to R_0, i.e., if $B(R^D)=B(R^P)=B(R_0)$.

Direct calculation shows that there exist a constant $\alpha\neq 0$ and unimodular matrices U^D and U^P such that $U^DR^D=U^PR^P=R_I:=\begin{bmatrix} p_2 & \alpha p_3 \\ p_1p_2 & 0 \end{bmatrix}$. So

indeed $B(R^D)=B(R^P)$. As $\begin{bmatrix}0 & 1\\ p_1 & -1\end{bmatrix}R_I=\begin{bmatrix}1 & 0\\ 0 & \alpha\end{bmatrix}R_0$ it follows that $B(R_I)\subset B(R_0)$, but $B(R_I)\neq B(R_0)$. So the identified laws R^D and R^P are equivalent, but not equivalent to R_0. This is due to the fact that $B(R_0)$ is not the most powerful unfalsified model for $\tilde{w}$. Indeed, a short calculation gives that $p_2+\alpha p_3=\alpha' p_1$, where $\alpha:=\{\cos(f_1)-\cos(f_2)\}/\{\cos(f_3)-\cos(f_1)\}$ and $\alpha':=\{\cos(f_3)-\cos(f_2)\}/\{\cos(f_3)-\cos(f_1)\}$. Stated otherwise, the space of polynomials $\{s^2+c.s+1;\ c\in\mathbb{R}\}$ has dimension two. The most powerful unfalsified model for the generating system is $B(R_0^*)$ with $R_0^*:=\begin{bmatrix}p_1p_2 & 0\\ 0 & p_1p_3\\ p_2 & \alpha p_3\end{bmatrix}$. It easily follows that $B(R^D)=B(R^P)=B(R_I)=B(R_0^*)$.

The foregoing shows that the identified models correspond to the (most powerful unfalsified) model for the generating system. Hence the generating system is exactly identified. This illustrates the consistency result stated in proposition 8–3.

9.3. Descriptive modelling

9.3.1. Introduction

In the second simulation we model a time series by minimizing the descriptive misfit, given a maximal tolerated complexity, i.e., we use the procedure $P^D_{c_{tol}}$. We will first describe the data and the system generating it, then present the identified model and finally compare this model with the generating system.

9.3.2. Data

The data consists of a two–dimensional time series $\tilde{w}=\begin{pmatrix}w_1\\ w_2\end{pmatrix}\in(\mathbb{R}^2)^{1000}$ and is depicted in figure 12.

9.3.3. System

The data $\tilde{w}$ is generated by the system shown in figure 13. Here s_1 is the noise–free input, n_1 the noise on the input, and $w_1:=s_1+n_1$ the exactly observed input. The signal s_2 is the output generated by the input w_1. The observed output is $w_2:=s_2+n_2$.

The signals s_1,s_2 and the noise n_1,n_2 are given in figure 14. For a

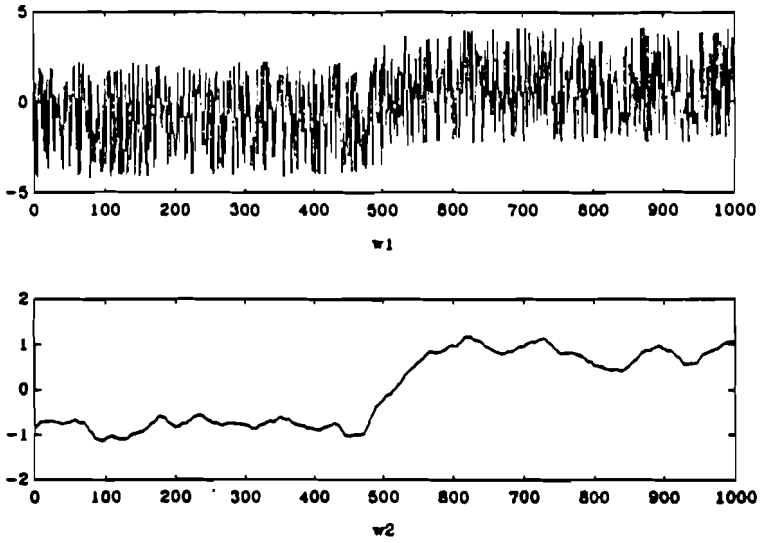

figure 12: data for simulation 9.3.

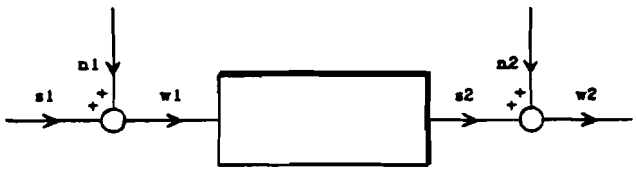

figure 13: generating system for simulation 9.3.

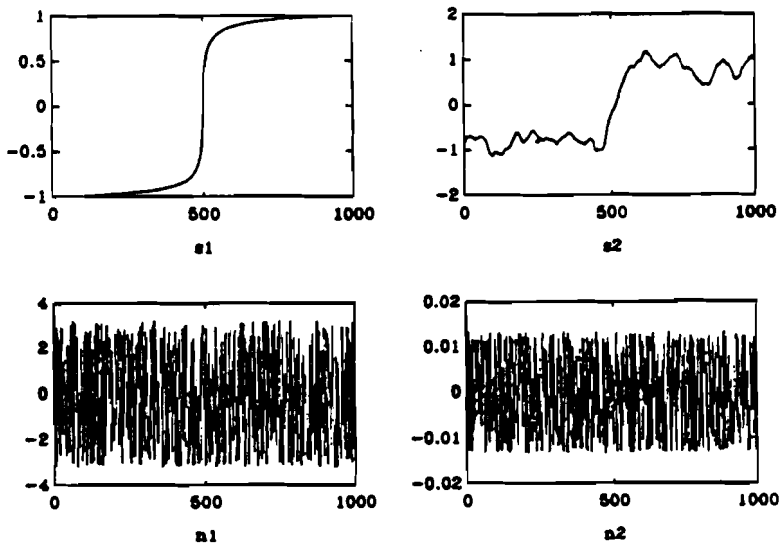

figure 14: signals and noise for simulation 9.3.

signal $s \in \mathbb{R}^T$ and noise $n \in \mathbb{R}^T$ we define the signal to noise ratio in $s+n$ as $\|s\|/\|n\| := \{ \sum_{t=1}^{T} s(t)^2 / \sum_{t=1}^{T} n(t)^2 \}^{1/2}$. In this simulation the signal to noise ratio for w_1 is ½, for w_2 100.

The system generating s_2 from w_1 is a (symmetric) exponential smoother. For $0<\alpha<1$ we define the exponential smoother e_α as follows. Let l_∞ denote the set of bounded sequences, i.e., $l_\infty := \{w \in \mathbb{R}^{\mathbb{Z}}; \sup(|w(t)|; t \in \mathbb{Z}) < \infty\}$. Then $e_\alpha: l_\infty \to l_\infty$ is defined by $e_\alpha(u) := y$, where $y(t) := \frac{1-\alpha}{1+\alpha} \cdot \sum_{\tau=-\infty}^{\infty} \alpha^{|\tau|} u(t+\tau)$. Note that for u a constant signal, $u(t) = c$ for all $t \in \mathbb{Z}$, the output is $y = u$.

We will embed the graph of e_α $\mathrm{gr}(e_\alpha) := \{(u,y) \in l_\infty^2; y = e_\alpha(u)\}$ in an AR-system $B_\alpha \subset (\mathbb{R}^2)^{\mathbb{Z}}$. In order to describe B_α, let $y = e_\alpha(u) = \frac{1-\alpha}{1+\alpha} \cdot (y_- + u + y_+)$, where $y_-(t) := \sum_{\tau=1}^{\infty} \alpha^\tau u(t-\tau)$ and $y_+(\tau) := \sum_{\tau=1}^{\infty} \alpha^\tau u(t+\tau)$. Then $(\sigma-\alpha)y_- = \alpha u$ and $(1-\alpha\sigma)y_+ = \alpha\sigma u$, hence $(\sigma-\alpha)(1-\alpha\sigma)(y_- + u + y_+) = [(1-\alpha\sigma)\alpha + (\sigma-\alpha)(1-\alpha\sigma) + (\sigma-\alpha)\alpha\sigma]u = (1-\alpha^2)\sigma u$. Define $p_\alpha := (s-\alpha)(1-\alpha s)$ and $q_\alpha := \frac{1-\alpha}{1+\alpha} \cdot (1-\alpha^2)s = (1-\alpha)^2 s$, then $\mathrm{gr}(e_\alpha) \subset B_\alpha := B(R_\alpha)$ where $R_\alpha := (-q_\alpha, p_\alpha)$.

In the simulation the signal s_2 is the exponential smoothing of w_1 with $\alpha = 0.95$. Hence the (most powerful unfalsified model of the) generating system is $B(R_g)$ with $R_g = (-q_g, p_g) := (-q_{0.95}, p_{0.95})$. We remark that in identifying the model there is no prior knowledge that w_1 is the input and w_2 the output.

9.3.4. Model identification

Next we analyse the data $\tilde{w}$ by means of $P^D_{c_{tot}}$. We consider models of decreasing complexity, corresponding to requiring one AR-relation of order 5,4,3,2,1 and 0 respectively. For order k the resulting model is indicated by $B_k := B((-q^{(k)}, p^{(k)})) := \{(u,y) \in (\mathbb{R}^2)^{\mathbb{Z}}; p^{(k)}(\sigma)y = q^{(k)}(\sigma)u\}$, $k = 5,4,3,2,1,0$. See table 2. This table also contains the roots of the polynomials $p^{(k)}$, $q^{(k)}$, and the descriptive error $\varepsilon^D_{k,1}(\tilde{w}, B_k)$.

The results in table 2 indicate that little descriptive power is lost by reducing the order from 5 to 2. Moreover, two of the roots of the identified polynomial p turn out to be rather invariant under different orders, while the roots of the identified polynomial q seem to be quite random, although generally one of them is close to 0. It seems reasonable to take c_{tot} such that the corresponding equation structure is $e(c_{tot}) = (0,0,1,0,0,0,\dots)$, i.e., to require one second order relation.

	coefficients of: σ^0	σ^1	σ^2	σ^3	σ^4	σ^5	roots p	q	error
order 5: $p^{(5)}$	0.4475	0.0893	-0.5333	-0.5563	0.1161	0.4295	0.9536	0.21	0.0154
$q^{(5)}$	0.0003	-0.0010	-0.0023	-0.0025	-0.0014	-0.0003	1.0548	-0.64±1.071	
							-1.05	-1.53±0.831	
							-0.61±0.781	-18	0.0155
order 4: $p^{(4)}$	0.5482	-0.3488	-0.4063	-0.3417	0.5440		1.0514	0.15	
$q^{(4)}$	0.0003	-0.0014	-0.0016	-0.0017	-0.0001		-6.69±0.731	-0.56±0.881	
order 3: $p^{(3)}$	0.5427	-0.6713	-0.2884	0.4144			0.9501	0.037	0.0159
$q^{(3)}$	0.0001	-0.0014	-0.0009	-0.0003			1.0537	-1.31±1.551	
order 2: $p^{(2)}$	0.4061	-0.8168	0.4099				-1.31 0.9529	5.24	0.0159
$q^{(2)}$	0.0002	-0.0011	0.0002				1.0396	0.15	
order 1: $p^{(1)}$	0.7073	-0.7069					1.0006	1.20	0.0176
$q^{(1)}$	0.0011	-0.0009							
order 0: $p^{(0)}$	0.9806								0.7190
$q^{(0)}$	0.1962								

table 2: identified AR-laws for simulation 9.3.

9.3.5. Model validation

The identified model $B((-q_I, p_I)) := B_2$ is compared with the generating system $B((-q_g, p_g))$ in table 3. This indicates that the AR-law of the identified system is close to the law of the generating system.

	coefficients of: σ^0	σ^1	σ^2	roots	
system: p_g	1	-2.0026	1	0.95	1.0526
q_g	0	-0.0026	0	0	
model: p_I	0.9906	-1.9925	1	0.9529	1.0396
q_I	0.0004	-0.0028	0.0005	0.1537	5.2435

table 3: system and identified model.

We next want to compare the model and the system with respect to their input–output behaviour. So we now will use the prior knowledge that w_1 is

an input and w_2 an output. We will compare the impulse responses of the model and the system.

For $B=\{(u,y)\in(\mathbb{R}^2)^{\mathbb{Z}};\ p(\sigma)y=q(\sigma)u\}$ we define the impulse response of B with respect to u as $B^{\delta}:=\{(u,y)\in B;\ u=\delta\}$, where $\delta(0):=1$ and $\delta(t):=0$ for all $t\neq 0$. It can be shown that B^{δ} contains exactly one bounded element if $q\neq 0$, $p\neq 0$ and p has no roots on the unit circle. In this case we call the time series $i\in\mathbb{R}^{\mathbb{Z}}$ such that $(\delta,i)\in B^{\delta}\cap l_{\infty}$ the stable impulse response. The models $B((-q_g,p_g))$ and $B((-q_I,p_I))$ satisfy these conditions. We denote their stable impulse responses by i_g and i_I respectively.Here $i_g(t)=\frac{1-\alpha}{1+\alpha}.\alpha^{|t|}$ and i_I is determined as follows. There exist unique real numbers a_1, a_2, b_1, b_2, d with $|a_1|<1$, $|a_2|>1$ such that $\frac{q_I}{p_I}=\frac{b_1}{s-a_1}+\frac{b_2}{s-a_2}+d$. Define $i_I(0):=d-\frac{b_2}{a_2}$, $i_I(t):=b_1.a_1^{t-1}$ for $t>0$ and $i_I(t):=-b_2a_2^{t-1}$ for $t<0$. It then is a matter of simple calculation to verify that $p_I(\sigma)i_I=q_I(\sigma)\delta$. This corresponds to a causal interpretation of the transferfunction $\frac{b_1}{s-a_1}$ and an anticausal one for $\frac{b_2}{s-a_2}$.

The stable impulse responses i_g of the system and i_I of the identified system are given in figure 15.

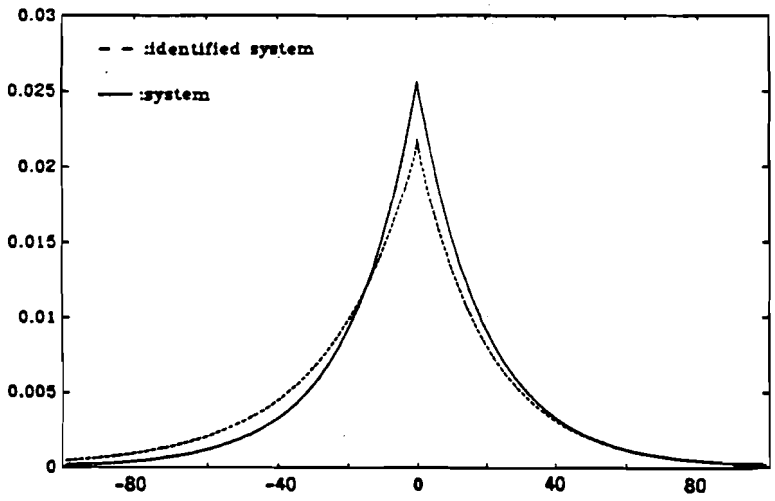

figure 15: impulse responses for simulation 9.3.

9.3.6. Scaling and sampling

We conclude this section with some remarks.

First, the stable impulse response of a system is a highly sensitive function of the AR–coefficients describing the system. For example, in the system $(\sigma-1-\varepsilon)y=u$ with $|\varepsilon|<1$ the stable impulse response is causal if $\varepsilon<0$,

anticausal if $\varepsilon > 0$.

Second, the result of the identification algorithm depends on scaling of the variables. In order to illustrate this, consider scaling of the output in the system $B(R_g)$ by a factor $c \neq 0$. Let $B_c := \{(u,y) \in (\mathbb{R}^2)^{\mathbb{Z}};$ $p_g(\sigma)y = c.q_g(\sigma)u\}$. Let $\varepsilon := e^D(\tilde{w}, (-q_I, p_I))$ denote the descriptive misfit of the identified law $(-q_I, p_I)$ with respect to the data $\tilde{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$. Denote the transformed data by $\tilde{w}_c := \begin{bmatrix} w_1 \\ c w_2 \end{bmatrix}$. From definition 5-4 it follows that $e^D(\tilde{w}_c, (-cq_I, p_I)) = \varepsilon.(\|q_I\|^2 + \|p_I\|^2)^{1/2} / (\|q_I\|^2 + c^{-2}.\|p_I\|^2)^{1/2}$. Using the results in table 3, it follows that the descriptive misfit of $(-cq_I, p_I)$ with respect to the scaled data $\tilde{w}_c$ is approximately $c.\varepsilon$. So, e.g., if c is very large then the law $u = 0$ has smaller error. In the next section we will illustrate that the predictive procedures prevent these problems of scaling.

Finally, autoregressive modelling is subject to problems of fast sampling. Consider the case that a continuous time system is sampled at a certain sample rate Δ^{-1}. The magnitudes of the AR-coefficients of the sampled system depend on this sample rate. This affects the descriptive quality of the AR-laws, as indicated above. The constant c is related to Δ as $c = \Delta$. It especially seems difficult to identify good approximations of infinite dimensional systems by means of autoregressive modelling in case of high sample rate and small noise. This is only partly due to the smoothness of the resulting signals. It seems contradictory that having a large amount of data, i.e., fast sampling, and good data, i.e., small noise, would be undesirable in identification.

To illustrate this we refer to table 2, where the best AR-law of order 1 is close to $(\sigma - 1)w_2 = 0$ with a small descriptive misfit of 0.0176. If we scale the output appropriately this effect is reduced. For example, $e^D(\tilde{w}_c, (0, \sigma - 1)) = c.00176$, while $e^D(\tilde{w}_c, (-c.q_I, p_I)) = 0.0159.(\|q_I\|^2 + \|p_I\|^2)^{1/2} / (\|q_I\|^2 + c^{-2}\|p_I\|^2)^{1/2}$. So for c sufficiently large the law $(-cq_I, p_I)$ has much better descriptive fit than the law corresponding to smoothness. We remark that decrease in the signal to noise ratio of the output hardly helps in discriminating $(-q_I, p_I)$ from $(0, \sigma - 1)$. This is due to the fact that $\|p_I\| / (\|q_I\|^2 + \|p_I\|^2)^{1/2} \approx 1$. If $\tilde{w}' = \begin{bmatrix} w_1 \\ w_2' \end{bmatrix}$ with $w_2' = s_2 + c.n_2$, $c > 1$, then $e^D(\tilde{w}', (-q_I, p_I)) \approx 0.0159 + (c-1).\|n_2\|.\|p_I\| / (\|q_I\|^2 + \|p_I\|^2)^{1/2}$ and $e^D(\tilde{w}', (0, \sigma - 1)) \approx 0.0176 + (c-1).\|n_2\|$, so for c large the errors are nearly the same.

9.4. Predictive modelling

9.4.1. Introduction

In the third simulation we illustrate the difference between descriptive and predictive modelling. We will see that the predictive procedures suffer less from scaling problems. On the other hand, the imposed asymmetry in time, due to the one-step-ahead prediction criterion, sometimes is artificial, in which case the descriptive procedures seem preferable.

We will now first describe the data and the generating system and subsequently analyse the data by means of descriptive and predictive procedures.

9.4.2. Data

The data consists of a three-dimensional time series $\tilde{w} = \mathrm{col}(w_1, w_{21}, w_{22}) \in (\mathbb{R}^3)^{200}$. We will investigate the effect of scaling. In order to illustrate this we will scale w_{22} and identify models for the scaled data $\tilde{w}^{(k)} := \mathrm{col}(\tilde{w}_1^{(k)}, \tilde{w}_2^{(k)}, \tilde{w}_3^{(k)}) := \mathrm{col}(w_1, w_{21}, k.w_{22})$, $k \in \mathbb{R}_+$.

9.4.3. System

The data is generated by the system shown in figure 16.

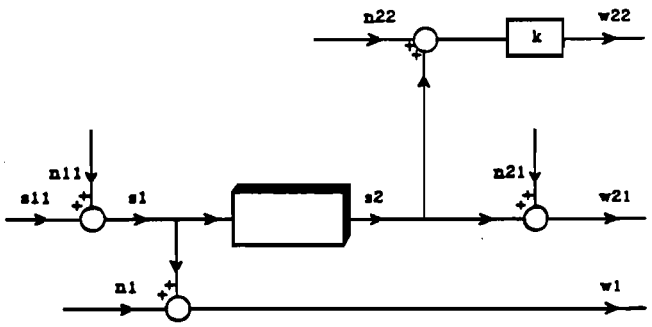

figure 16: generating system for simulation 9.4.

Here s_{11} is the noise-free input, n_{11} noise on the system input, $s_1 := s_{11} + n_{11}$ the input for the system, n_1 noise on the observed input, $w_1 := s_1 + n_1$ the observed input, s_2 the output of the system, n_{21} and n_{22} noise on observed outputs, $w_{21} := s_2 + n_{21}$ and $w_{22} := s_2 + n_{22}$ the observed outputs. The signal to noise ratios are $\|s_{11}\|/\|n_{11}\| = 10$, $\|s_1\|/\|n_1\| = 20$, $\|s_2\|/\|n_{21}\| = 10$ and

$\|s_2\|/\|n_{22}\| = 2$.

The signals, observed data and noise are given in figure 17 for the case $k = 1$ (no scaling on w_{22}).

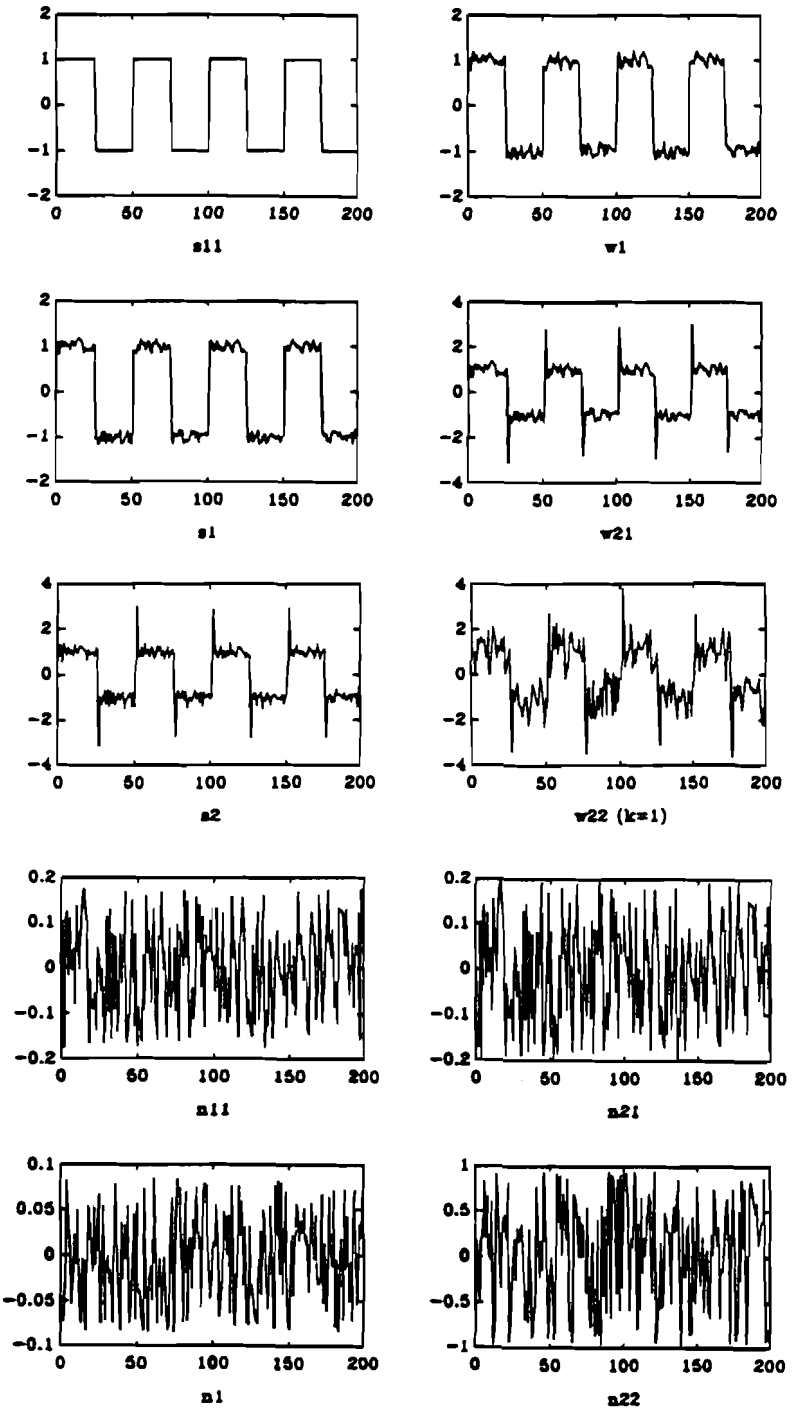

figure 17: data, signals and noise for simulation 9.4.

The system relating s_2 to s_1 is described by $\sigma^2 s_2 = (2\sigma - 1)s_1$. This corresponds to a simple linear extrapolator $s_2(t) := s_1(t-1) + \{s_1(t-1) - s_1(t-2)\}$.

9.4.4. Model identification and validation

In order to identify a model, we have to reconcile the desires for low complexity and for low misfit. In the simulation we identified the AR-models with best descriptive and predictive fit for orders from 0 up to 4 and for data $\tilde{w}^{(k)}$ corresponding to various scaling constants k. In order to choose a model we compared the increase in fit due to increase in complexity. It turns out that the descriptive misfit decreases only slightly for orders larger than two. Moreover, the results for $k>1$ nearly coincide with those for $k=1$.

The main results of the simulation are summarized in tables 4 and 5. Table 4 contains the best predictive models of orders from 0 up to 4 and for various values of k. Table 5 contains the best descriptive models of orders 0 and 2 and for various k. Specified are the AR-coefficients in $r_1(\sigma)\tilde{w}_1^{(k)} + r_{21}(\sigma)\tilde{w}_2^{(k)} + r_{22}(\sigma)\tilde{w}_3^{(k)} = 0$, some of the roots of r_1, r_{21}, r_{22}, and the misfits.

From table 4 it is clear that the model identified by the predictive procedure does not depend on scaling of w_{22}. Moreover, considering the predictive misfits it seems very reasonable to choose a second order model, with predictive misfit 0.12. The model for data $\tilde{w}^{(k)}$ then becomes $r_1^{(k)}(\sigma)w_1^{(k)} + r_{21}^{(k)}(\sigma)w_2^{(k)} + r_{22}^{(k)}(\sigma)w_3^{(k)} = 0$, where $r_1^{(k)}(s) = 0.08s^2 - 1.99s + 0.96$, $r_{21}^{(k)}(s) = s^2 - 0.05s + 0.01$, $r_{22}^{(k)} \approx k^{-1}(0.01s - 0.03)$. So this law is close to the generating system $(-2\sigma + 1)s_1 + \sigma^2 s_2 = 0$. The procedure identifies the relation between w_1 and w_{21} as its misfit is due to the noise on w_1 and w_{21}, which is much smaller than the noise on w_{22}. Note finally that, even if $\tilde{w}^{(k)}$ is observed instead of $\tilde{w} = \tilde{w}^{(1)}$, the predictive procedure for all k identifies the same AR-relation for the unscaled variables (w_1, w_{21}, w_{22}).

On the other hand, as shown in table 5, the model identified by the descriptive procedures depends strongly on scaling of w_{22}. Roughly speaking, for values of k larger than 0.1 it seems reasonable to choose a model of order 2, which model turns out to be relatively close to the generating system. For values of k smaller than 0.1 it seems reasonable to choose a model of order 0, approximately corresponding to $w_3^{(k)} = k.w_2^{(k)}$.

	order 0			order 1			order 2			order 3			order 4		
	r_1	r_{21}	r_{22}	r_1	r_{21}	r_{22}	r_1	r_{21}	r_{22}	r_1	r_{21}	r_{22}	r_1	r_{21}	r_{22}
k=1															
coeff. σ^0	-0.60	1	-0.44	-1.82	0.48	-0.05	0.96	0.01	-0.03	0.18	0.01	-0.02	-0.18	0.05	-0.00
σ^1				0.40	1	-0.04	-1.99	-0.05	0.01	0.69	-0.02	-0.02	0:30	0.07	-0.02
σ^2							0.08	1	0.00	-1.99	0.09	0.01	0.73	-0.09	-0.02
σ^3										0.08	1	-0.00	-1.99	0.07	0.01
σ^4													0.07	1	-0.00
roots		-		4.62	-0.48	-1.25	0.49	0.02± 0.111	1.65	0.53			0.41±0.211		
							25.2		-10.3	-0.17			-0.43		
										24.4			26.6		
misfit	0.3250			0.2153			0.1168			0.1149			0.1134		
k=0.1															
coeff. σ^0	-0.60	1	-0.44	-1.82	0.48	-0.46	0.96	0.01	-0.28	0.18	0.01	-0.19	-0.18	0.05	-0.00
σ^1				0.40	1	-0.37	-1.99	-0.05	0.14	0.69	-0.02	-0.21	0.30	0.07	-0.18
σ^2							0.08	1	0.02	-1.99	0.09	0.14	0.73	-0.09	-0.19
σ^3										0.08	1	-0.02	-1.99	0.07	0.13
σ^4													0.07	1	-0.04
roots		-		4.62	-0.48	-1.25	0.49	0.02± 0.111	1.65	0.53			0.41±0.211		
							25.2		-10.3	-0.17			-0.43		
										24.4			26.6		
misfit	0.3250			0.2153			0.1168			0.1149			0.1134		
k=0.01															
coeff. σ^0	-0.60	1	-0.44	-1.82	0.48	-4.57	0.96	0.01	-2.75	0.18	0.01	-1.89	-0.18	0.05	-0.08
σ^1				0.40	1	-3.65	-1.99	-0.05	1.40	0.69	-0.02	-2.14	0.30	0.07	-1.81
σ^2							0.08	1	0.16	-1.99	0.09	1.40	0.73	-0.09	-1.93
σ^3										0.08	1	-0.17	-1.99	0.07	1.29
σ^4													0.07	1	-0.42
roots		-		4.62	-0.48	-1.25	0.49	0.02± 0.111	1.65	0.53			0.41±0.211		
							25.2		-10.3	-0.17			-0.43		
										24.4			26.6		
misfit	0.3250			0.2153			0.1168			0.1149			0.1134		

table 4: predictive AR-laws for simulation 9.4.

	order 0	misfit	coeff. order 2: σ^0	σ^1	σ^2	roots	misfit
k=1:							
r_1	1.36	0.3250	1.13	-1.99	0.02	0.57 ; 87.7	0.0561
r_{21}	-2.28		-0.03	-0.12	1	0.24 ; -0.12	
r_{22}	1		-0.03	0.02	-0.00	4.92 ; 1.99	
k=0.2:							
r_1	-0.00	0.1137	1.13	-1.99	0.02	0.57 ; 89.5	0.0559
r_{21}	-0.21		-0.02	-0.13	1	0.20 ; -0.08	
r_{22}	1		-0.19	0.14	-0.02	3.06 ; 2.49	
k=0.14:							
r_1	-0.01	0.0804	1.11	-1.98	0.02	0.57 ; 91.8	0.0555
r_{21}	-0.14		0.01	-0.14	1	0.07±0.091	
r_{22}	1		-0.43	0.33	-0.08	1.98±1.091	
k=0.12:							
r_1	-0.01	0.0691	1.08	-1.95	0.02	056; 89.6	0.0547
r_{21}	-0.12		0.06	-0.17	1	0.08±0.231	
r_{22}	1		-0.80	0.68	-0.24	1.43±1.151	
k=0.11:							
r_1	-0.01	0.0634	1.02	-1.88	0.02	0.55 ; 76.9	0.0535
r_{21}	-0.11		0.13	-0.22	1	0.11±0.341	
r_{22}	1		-1.37	1.29	-0.59	1.10±1.061	
k=0.1:							
r_1	-0.01	0.0577	0.90	-1.72	0.03	0.53 ; 49.4	0.0505
r_{21}	-0.10		0.26	-0.33	1	0.17±0.481	
r_{22}	1		-2.54	2.71	-1.54	0.88±0.941	
k=0.09:							
r_1	-0.01	0.0520	0.76	-1.52	0.05	051; 30.3	0.0461
r_{21}	-0.09		0.40	-0.47	1	0.24±0591	
r_{22}	1		-4.06	4.66	-2.96	0.79±0.871	
k=0.01:							
r_1	-0.00	0.0058	-0.01	0.01	-0.00	0.40 ; 8.64	0.0052
r_{21}	-0.01		-0.01	0.01	-0.02	0.44±0.771	
r_{22}	1		1.10	-1.39	1	0.70±0.781	

table 5: descriptive AR-laws for simulation 9.4.

In this way the simulation clearly indicates the effect of scaling of data on the resulting model identified by the descriptive procedures. The model identified by the predictive procedures is invariant under scaling.

9.4.5. Effects of scaling for SISO systems

We conclude this section with a few remarks on the effect of scaling on the identification of single-input single-output (SISO) systems.

In table 6 we give the main results of the simulation experiment consisting of modelling the data $\tilde{\tilde{w}}^{(k)} := \mathrm{col}(w_1, k.w_{21})$ for various k by means of the descriptive procedures. From the table of misfits it seems reasonable to accept a second order law , as the second order laws have considerably better fit than lower order laws and nearly as good fit as higher order laws. The table indicates that scaling has little influence on the model for (w_1, w_{21}), as for scaling constant k the identified AR-law $(r_1^{(k)}, r_{21}^{(k)})$ is approximately equal to $(kr_1^{(1)}, r_{21}^{(1)})$.

On the other hand, it turns out that by decreasing the signal to noise ratio for w_{21}, the identified model becomes more sensitive to scaling. Moreover, in section 9.3 we concluded that for the exponential weighting system the identified model is sensitive to scaling. It hence appears that scaling sometimes has influence on the identified model, but that the effect need not always be large. Here we only will give a sketch of an explanation.

For simplicity, consider a second order system $B = \{(w_1, w_2);\ p(\sigma)w_2 = q(\sigma)w_1\}$ with degrees $d((p,q)) = d(p) = 2$. Assume that w_2 is scaled in such a way that $\|p\|^2 = \|q\|^2 = \frac{1}{2}$. Let the data consist of $\tilde{w} = (\tilde{w}_1, \tilde{w}_2)$, $\tilde{w}_1 = w_1 + \varepsilon_1$, $\tilde{w}_2 = w_2 + \varepsilon_2$, where ε_1 and ε_2 are uncorrelated white noise with $\sigma_1 := \|\varepsilon_1\|$ and $\sigma_2 := \|\varepsilon_2\|$. To investigate the effect of scaling, suppose we observe $(c_1\tilde{w}_1, c_2\tilde{w}_2)$, $c_1.c_2 \neq 0$. As the identified models are invariant under a data transformation $(\pm c\tilde{w}_1, \pm c\tilde{w}_2)$, $c \neq 0$, we may consider $\tilde{w}^{(k)} := (\tilde{w}_1, k.\tilde{w}_2)$, with $k := |c_2/c_1|$.

First let $k = 1$ and let α denote the descriptive misfit of $(-q, p)$, i.e., $\alpha := \|p\tilde{w}_2 - q\tilde{w}_1\| \approx \frac{1}{2}\sqrt{2}.(\sigma_1^2 + \sigma_2^2)^{\frac{1}{2}}$. Moreover, let β and γ denote the descriptive misfit of the best first order law for $\tilde{w}_1$ and $\tilde{w}_2$ respectively. For k let e_k^1 denote the descriptive misfit of the best first order law for $\tilde{w}^{(k)}$, and α_k the misfit of $(-kq, p)$, i.e., $\alpha_k := e^D(\tilde{w}^{(k)}, (-kq, p)) = \alpha.k\sqrt{2}/(1+k^2)^{\frac{1}{2}}$. A relevant indication for the sensitivity to scaling is the influence of k on α_k and

misfit	order 0	1	2	3	4
k=100	0.4812	0.1587	0.0616	0.0564	0.0554
k=10	0.4798	0.1585	0.0616	0.0564	0.0554
k=1	0.3726	0.1370	0.0565	0.0528	0.0520
k=0.1	0.0544	0.0245	0.0134	0.0127	0.0125
k=0.01	0.0055	0.0025	0.0014	0.0013	0.0013

AR-law		coeff. of: σ^0	σ^1	σ^2	roots	
k=100:	r_1	118	-202	3.37	0.59	59.1
	r_{21}	-0.07	-0.12	1	0.33	-0.21
k=10:	r_1	11.8	-20.2	0.34	0.59	59.3
	r_{21}	-0.07	-0.12	1	0.33	-0.21
k=1:	r_1	1.15	-2.00	0.02	0.58	80.0
	r_{21}	-0.06	-0.11	1	0.31	-0.20
k=0.1:	r_1	0.10	-0.19	-0.00	0.52	-111
	r_{21}	-0.03	-0.05	1	0.19	-0.14
k=0.01:	r_1	0.01	-0.02	-0.00	0.51	-98.0
	r_{21}	-0.02	-0.05	1	0.18	-0.13
k=1: predictive:	r_1	0.97	-1.99	0.08	0.50	23.8
	r_{21}	-0.02	-0.04	1	0.17	-0.13

table 6: descriptive misfit and AR-laws for $\tilde{\tilde{w}}^{(k)}$.

e_k^1. We assume that for small k $e_k^1 \approx k.\gamma$ and for large k $e_k^1 \approx \beta$. This seems often to be the case. Now if $\alpha\sqrt{2} < \min\{\beta,\gamma\}$ we may expect little sensitivity to scaling, as it seems probable that in this case $e_k^1 > \alpha_k$ for all $k \in \mathbb{R}_+$.

In the case of data $\tilde{\tilde{w}}^{(k)} := \mathrm{col}(w_1, kw_{21})$ in this section the underlying system is described by $p(s) = s^2$ and $q(s) = 2s - 1$. So for $k = 1/\sqrt{5}$ we have

$\|kq\| = \|p\|$. Form this we get $\alpha \approx 0.04$, $\beta \approx 0.28$, $\gamma \approx 0.27$. So indeed $\alpha\sqrt{2} < \min\{\beta,\gamma\}$.

On the other hand, for the exponential weighting system of section 9.3 we have $\|p_g\| >> \|q_g\|$. It can be calculated that for $c = 850$ we have $\|cq_g\| \approx \|p_g\|$ and $\alpha \approx 9.5$, $\beta \approx 1.82$, $\gamma \approx 15.3$. So in this case $\beta < \alpha\sqrt{2} < \gamma$. For large values of k we will be unable to identify the generating system. The simulation of section 9.3 corresponds to small k ($k \approx 1/850$).

Finally, if w_1 and w_2 are very smooth we will always have problems in identifying the relationship between w_1 and w_2. In this case $\beta \approx e^D(\tilde{w}_1, \sigma - 1) \approx \sigma_1$ and $\gamma \approx e^D(\tilde{w}_2, \sigma - 1) \approx \sigma_2$, while $\alpha_k \approx (\sigma_1^2 + \sigma_2^2)^{1/2} \cdot k/(1+k^2)^{1/2}$. In this case we may expect $e_k^1 < \alpha_k$ for all k.

9.5. An example illustrating non-optimality

9.5.1. Introduction

In the fourth and final simulation we illustrate the fact that the procedures for modelling, given a maximal tolerated misfit, need not generate models of minimal complexity. This then shows that the procedures $P^D_{\epsilon_{tol}}$ and $P^P_{\epsilon_{tol}}$ differ from the (optimal) procedures $P^{*D}_{\epsilon_{tol}}$ and $P^{*P}_{\epsilon_{tol}}$ respectively, as indicated in sections 6.2 and 6.3.

We first describe the data and the generating system, then analyse the data by means of the procedures $P^D_{\epsilon_{tol}}$ and $P^P_{\epsilon_{tol}}$, and comment on the identified models. We finally illustrate the consistency of $P^P_{\epsilon_{tol}}$.

9.5.2. Data and system

The data $\tilde{w} = \mathrm{col}(\tilde{w}_1, \tilde{w}_2, \tilde{w}_3) \in (\mathbb{R}^3)^{400}$ is generated by an ARMA-system $M(\sigma^{-1})w = N(\sigma^{-1})n$, where $n = \mathrm{col}(n_1, n_2, n_3)$ consists of three uncorrelated white noise processes with $En_k = 0$, $En_k^2 = 1$, $k = 1,2,3$. The matrices M and N are given by $M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1-\alpha\sigma^{-1} \end{bmatrix}$ and $N = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & 1 \end{bmatrix}$ with $\alpha := 1/\sqrt{11}$ and $\beta := \sqrt{1.1}$.

This corresponds to $w_1 = 1/2 \cdot n_1$, $\sigma w_3 = \alpha w_3 + \sigma n_3$, $w_2 = w_3 + \beta n_2$. Figure 18 shows the data $\tilde{w}$, generated by a realization of n.

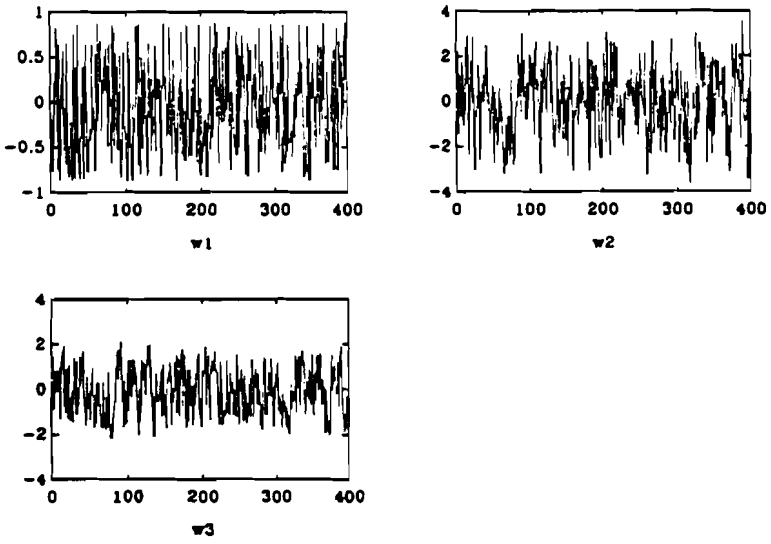

figure 18: data for simulation 9.5.

9.5.3. Model identification

We will identify a model for $\tilde{w}$ by means of descriptive and predictive procedures with (unfavourable) given tolerated misfits.

First we consider $P^D_{\epsilon_{tol}}$ with $\epsilon_{tol}=(\ \bar{\epsilon}^{tol}_t.(1,1,1)\ ;\ t\in\mathbf{Z}_+)$, $\bar{\epsilon}^{tol}_0 := e^D_0 := 1.6$, $\bar{\epsilon}^{tol}_1 := e^D_1 := 1.2$, and $\bar{\epsilon}^{tol}_t := -1$ for $t>1$. This means that only zero order and first order laws may be used in the identification of a model. The identified model is given in table 7, along with the best (not–allowable) first order law.

Next we consider $P^P_{\epsilon_{tol}}$ with $\epsilon_{tol}=(\ \bar{\epsilon}^{tol}_t.(1,1,1)\ ;\ t\in\mathbf{Z}_+)$, $\bar{\epsilon}^{tol}_0 := e^P_0 := 1.6$, $\bar{\epsilon}^{tol}_1 := e^P_1 := 0.95$, and $\bar{\epsilon}^{tol}_t := -1$ for $t>1$. The identified model is given in table 7, along with the best (not–allowable) first order law.

9.5.4. Model validation

The identified models are not of minimal complexity, given the maximal tolerated misfit. This is also indicated in table 7. It turns out that both for descriptive and predictive tolerated misfit as given before the model $B^* := \{w\in(\mathbf{R}^3)^{\mathbf{Z}};\ w_1=0, w_2=0,\ (\sigma-\alpha)w_3=0\}$ satisfies the misfit constraint. This model has complexity $c(B^*)=(1,\frac{1}{2},\frac{1}{3},\frac{1}{4},\dots)$, which is smaller than the complexity of the identified models, which is $(1,1,1,1,\dots)$. It easily follows that $c(B^*)$ is the lowest achievable complexity, given the misfit constraints. However, among these allowable models of lowest complexity there exists none of minimal misfit. For the procedures $\bar{P}^D_{\epsilon_{tol}}$ and $\bar{P}^P_{\epsilon_{tol}}$

	identified model				model B^*			
	w_1	w_2	w_3	misfit	w_1	w_2	w_3	misfit
descr. AR								
order 0	0.9978	-0.0364	0.0552	0.4992	1	0	0	0.5000
	-0.0661	-0.5347	0.8425	0.6562	0	1	0	1.4938
				1.7197				
order 1 : σ^0	-0.0012	-0.8443	-0.5359	1.4470	0	0	$-\alpha$	0.9574
σ^1	0.0012	0.8439	0.5356		0	0	1	
pred. AR								
order 0	0.9978	-0.0364	0.0552	0.4992	1	0	0	0.5000
	-0.0661	-0.5347	0.8425	0.6562	0	1	0	1.4938
				1.7197				
order 1 : σ^0	-0.0004	-0.2937	-0.1865	0.9559	0	0	$-\alpha$	0.9301
σ^1	0.0014	1	0.6348		0	0	1	

table 7: descriptive and predictive AR-laws for simulation 9.5.

there exist models of lowest complexity and minimal misfit, but they seem difficult to compute. Their identification involves the question what is the lowest possible zero order misfit such that there exist first order relations, satisfying the misfit constraint and the orthogonality conditions of the (descriptive or predictive) canonical form.

The procedures $P^D_{\epsilon_{tol}}$ and $P^P_{\epsilon_{tol}}$ first determine as many zero order laws as possible. Requiring three of those laws results in a zero order misfit (1.7197, 0.6562, 0.4992), which is more than tolerated. Hence two zero order laws are accepted. Moreover, the best two laws are chosen. This implies conditions, due to the canonical form, on first order laws. In this simulation there is no allowable first order law satisfying these conditions. The model B^* shows that it is profitable not to take the best two zero order laws in order to get allowable first order laws, i.e., with misfit less than ϵ^D_1 or ϵ^P_1.

9.5.5. Consistency

We finally consider increase of the number of data generated by the ARMA-system. In table 8 we summarize results for the procedure $P^P_{\epsilon_{tol}}$ in

	identified models T=50	T=100	T=400	T=800	$A^P_{\epsilon_{tol}}$
order 0: AR-coeff.					
w_1	0.9999	0.9824	0.9978	0.9961	1
w_2	0.0019	0.1422	-0.0364	-0.0234	0
w_3	0.0161	-0.1210	0.0552	-0.0346	0
misfit	0.5620	0.5161	0.4992	0.4994	0.5000
AR-coeff.					
w_1	-0.0127	0.1797	-0.0661	-0.0547	0
w_2	-0.5286	-0.5440	-0.5347	-0.5246	-0.5257
w_3	0.8488	0.8196	0.8425	0.8471	0.8507
misfit	0.6593	0.6621	0.6562	0.6429	0.6482
AR-coeff.					
w_1	-0.0102				
w_2	0.8489				
w_3	0.5285				
misfit	1.5920	>1,6	>1.6	>1.6	1.6970
order 1: AR-coeff.	-				
σ^0: w_1		0.0228	-0.0004	-0.0004	0
w_2		-0.3708	-0.2937	-0.2874	-0.2182
w_3		-0.2511	-0.1865	-0.1772	-0.1348
σ^1: w_1		-0.0614	0.0014	0.0014	0
w_2		1	1	1	1
w_3		0.6771	0.6348	0.6164	0.6180
misfit		0.9296	0.9559	0.9578	0.9759

table 8: consistency of $P^P_{\epsilon_{tol}}$.

case of $T = 50, 100, 400$ and 800 observations. We also calculated the best first order laws. Observe that for $T = 50$ the procedure for this simulation would accept three zero order laws, while for $T = 100$ it would accept a first order law. We also give the optimal approximation $A^P_{\varepsilon_{tot}}$, corresponding to the optimal predictive model for ε_{tot} in case the generating system were known. This model can be calculated from covariance matrices, derived from M and N.

The results in table 8 illustrate consistency, as defined in section 8. Note especially that in the limit the best first order law which satisfies the orthogonality conditions of the canonical predictive form has predictive misfit $0.9759 > \varepsilon^P_1 = 0.95$. Hence, almost sure, for a sufficiently large number of observations the procedure $P^P_{\varepsilon_{tot}}$ will only accept two zero order laws.

10. CONCLUSION

In this paper we have described some procedures for approximate modelling of a time series, along with corresponding algorithms. The procedures have been illustrated by means of some numerical simulations.

The procedures determine a deterministic dynamical system which for given data is optimal with respect to a utility of models, depending on the objective of modelling. This utility is expressed in terms of a complexity of models and a measure of fit between data and models. The utility reflects a compromise between the generally conflicting objectives of identifying a simple model and a model which fits the data well. The utility is numerically expressed in terms of canonical parametrizations of dynamical systems. These canonical forms are determined in accordance with the objective of modelling.

The procedures form part of a more general deterministic approach to approximate modelling, as extensively discussed and illustrated in the paper.

The procedures have a clear optimality property as data modelling procedures, in terms of the corresponding utility. A procedure also has an optimal performance as a method of modelling phenomena if it is consistent. This means that nearly optimal models of the phenomenon are identified if the number of observations generated by the phenomenon is sufficiently large. This has been investigated for certain classes of data generating

systems and some of the procedures.

We finally mention some topics for future research.

(i) The construction of algorithms for utilities other then $u_{c_{tol}}$ and $u_{\varepsilon_{tol}}$, especially for minimizing the number of unexplained variables (inputs) under a misfit constraint.

(ii) Utilities and algorithms when the purpose of modelling is control.

(iii) Consistency analysis for generating systems of ARMAX type, i.e., with inputs, and the related issue of sufficient excitation.

(iv) Definition of approximate structure of a phenomenon, and corresponding interpretation of stochastic systems, especially of ARMAX type.

(v) Definition of the amount of confidence in identified models, sensitivity with respect to changes in data and tolerated levels of complexity or misfit, and robustness.

REFERENCES

[1] Golub, G.H., and C.F. Van Loan, An analysis of the total least squares problem, *SIAM Journal on Numerical Analysis* 17(6), pp. 883–893, 1980.

[2] Hannan, E.J., *Multiple Time Series*, John Wiley, New York, 1970.

[3] Hannan, E.J., W.T.M. Dunsmuir and M. Deistler, Estimation of vector ARMAX models, *Journal of Multivariate Analysis* 10, pp. 275–295, 1980.

[4] Heij, C., Approximate modelling of deterministic systems, in Curtain, R.F. (ed.), *Modelling, Robustness and Sensitivity Reduction in Control Systems*, pp. 271–283, NATO ASI Series, Springer, Berlin, 1987.

[5] Heij, C., and J.C. Willems, Consistency analysis of approximate modelling procedures, in Byrnes, C.I., C.F. Martin and R.E. Saeks (eds.), *Linear Circuits, Systems and Signal Processing: Theory and Application*, pp. 445–456, North Holland, Amsterdam, 1988.

[6] Jayant, N.S., and P. Noll, *Digital Coding of Waveforms*, Prentice Hall, Englewood Cliffs, New Jersey, 1984.

[7] Kalman, R.E., P.L. Falb and M.A. Arbib, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.

[8] Kendall, M.G., and A. Stuart, *The Advanced Theory of Statistics*, Griffin, London, 1964.

[9] Ljung, L., Convergence analysis of parametric identification methods, *IEEE AC*-23(5), pp. 770–783, 1978.

[10] Ljung, L., *System Identification – Theory for the User*, Prentice Hall, Englewood Cliffs, New Jersey, 1987.

[11] Ljung, L., and P.E. Caines, Asymptotic normality of prediction error estimators for approximate system models, *Stochastics* 3, pp. 29–46, 1979.

[12] Maddala, G.S., *Econometrics*, McGraw-Hill, New York, 1976.

[13] Nieuwenhuis, J.W., and J.C. Willems, Continuity of dynamical systems: a system theoretic approach, *Mathematics of Control, Signals, and Systems* 1(2), pp. 147–165, 1988.

[14] Rissanen, J., Stochastic complexity and modelling, *The Annals of Statistics* 14 (3), pp. 1080–1100, 1986.

[15] Willems, J.C., From time series to linear system. Part I: Finite dimensional linear time invariant systems. Part II: Exact modelling. Part III: Approximate modelling. *Automatica* 22, pp. 561–580, 1986; 22, pp. 675–694, 1986; 23, pp. 87–115, 1987.

[16] Willems, J.C., Models for dynamics, *Dynamics Reported* 2, pp. 171–269, 1989.