Lecture Notes in
Control and
Information Sciences

Edited by A.V. Balakrishnan and M. Thoma

10

Jan M. Maciejowski

The Modelling of Systems with Small Observation Sets

Springer-Verlag
Berlin · Heidelberg · New York

Lecture Notes in Control and Information Sciences

Edited by A.V. Balakrishnan and M. Thoma

10

Jan M. Maciejowski

The Modelling of Systems with Small Observation Sets

Springer-Verlag
Berlin Heidelberg New York 1978

Series Editors
A. V. Balakrishnan · M. Thoma

Advisory Board
A. G. J. MacFarlane · H. Kwakernaak · Ya. Z. Tsypkin

Author
Dr. Jan Marian Maciejowski
Maudsley Research Fellow, Pembroke College, Cambridge
also with the
Control and Management Systems Group,
Cambridge University Engineering Department
Mill Lane, Cambridge CB2 1RX, England

ISBN 3-540-09004-5 Springer-Verlag Berlin Heidelberg NewYork
ISBN 0-387-09004-5 Springer-Verlag NewYork Heidelberg Berlin

Printing and binding: Beltz Offsetdruck, Hemsbach/Bergstr.
2061/3020-543210

SUMMARY

The problem of assessing and interpreting models of systems, when only small sets of observations are available, is introduced and discussed. System identification is defined as the progression from a set of observations of the behaviour of a system, to a theory which accounts for that behaviour. The concepts of algorithmic information theory are drawn on to develop a characterisation of modelling, which constitutes a partial solution to the problem of system identification, while taking account of the size of the set of available observations. A model is defined to be an algorithm for computing the output observation set of a system under specified restrictions.

A general criterion of the quality of a model, its "information gain", is proposed, and its consistency with more conventional criteria is discussed. It is proved that no "universal modelling algorithm" can exist, in the sense that it is not possible, in general, to find the model with the highest information gain.

Information gain is a suitable criterion for a wide class of models, including nonlinear dynamical stochastic models, and its computation is straightforward. The use of information gain for the assessment of rival models is demonstrated.

The calculation of information gain requires that the model be expressed as a computer program. The choice of programming language is associated with the modeller's *a priori* beliefs about the system. It is shown that this choice becomes insignificant as the observation sets become large. A detailed

investigation shows that it is possible to speak precisely of "the smallest language" required to run a particular program. *A priori* knowledge assumed about a system can therefore be considered to be defined by the smallest language required to run the model.

Finally, the effect on model assessment of the manner in which system observations are coded is examined. It is found that a "safe" coding exists, which often leads to the same assessment as would the use of most other codings.

ACKNOWLEDGEMENTS

The idea of examining modelling in the light of algorithmic information theory is due to Professor A.G.J. MacFarlane. His constant encouragement and enthusiasm, as well as detailed criticism, has been an essential ingredient of this work.

I have also benefited from discussions with many members of the Control and Management Systems Group, of whom Dr.F.P. Kelly, Dr. S.R. Watson and Dr. M.B. Beck deserve special mention. The quotation from Newton in the last chapter was pointed out to me by Dr.A.T.Fuller.

Financial support for this research came from the Science Research Council, and in the final stages from Pembroke College.

Roberta Hill has produced her usual excellent standard of typing, but special thanks are due to her for struggling so successfully through chapter 5.

My wife has asked me not to write one of those embarassing acknowledgements, saying how impossible this research would have been without her constant encouragement and support; consequently I shall leave this to the reader's imagination.

CONTENTS

1. INTRODUCTION

1.1 Motivation

The areas in which the scientific method has been demonstrably and spectacularly successful are characterised by the possibility of performing experiments, or making observations, more or less freely whenever these are deemed desirable. The result of this has been that explicit consideration of the size of the set of observations from which a model is hypothesised, and to which a model is fitted, has been neglected. Any doubts which arise about the model can be resolved by further experimentation and observation.

This pleasant property increasingly disappears as one enters the domains of complex industrial processes, environmental control systems, management systems, and socio-economic systems. The work described here aims to clarify the relationship between the smallness of the available observation sets for such systems and the degree of usefulness of the models obtained for them.

Until recently, the class of models which could be used in scientific investigations was restricted by a very practical consideration. The behaviour of the model had to be understood, and that understanding could only be obtained from the theory of the model. The model was constrained to be sufficiently simple for theoretical investigation to be possible.

The availability of the computer has changed this situation radically. It is now possible to investigate the behaviour of a model by simulation, with hardly any theoretical understanding of it. Consequently this constraint on the complexity of useful models has been removed, or at least greatly relaxed. It is now possible to postulate a complicated model structure, to observe its simulated behaviour, and to adjust the details of the model until its simulated behaviour resembles the behaviour of the system being investigated.

When is such a model useful? When does it give any understanding of how the system really works? When can it be used as a reliable guide to how the system will behave in the future? The purpose of the work reported in this thesis is to throw some light on these questions. A further aim is to investigate how rival models of the same system behaviour should be assessed. Most of the thesis is ostensibly concerned with the details of rival model assessment, but it is clear that the ability to distinguish between competing models is intimately connected with the ability to say how good an isolated model is.

Why should a simulation model of the type described above not be useful or reliable? If it reproduces the observed system behaviour, is that not sufficient evidence to indicate the quality of the model? In fact, is it not clear that the better the reproduction of the observed behaviour, the better the model? Our answer is no. The basic reason is the possibility of "overfitting" the model, since its complexity is relatively unconstrained, and it is being checked against a small set of data.

Consider the following simple example. Suppose that two measurements are taken of some variable at two different times, and that we have no other information about it. It is desired to predict the value of the variable at some third time. If a linear variation with time of the variable is proposed, on the basis of the two observations, then it is clear that the only reasonable assessment of confidence in the prediction of the model is nil. The predicted value is no more likely (in an intuitive sense) than any other value. However, if a third measurement is taken which agrees with the prediction of the model, confidence in the model immediately increases. It is now possible to say that values predicted by the model are better predictions, in some sense, than mere guesses. If further measurements are taken, and these also agree with the predictions of the model, then confidence in the model increases very quickly. It never amounts to certainty, of course, but after only ten observations, say, one would have little doubt that the next prediction would be correct (which does not imply that it would be).

The confidence which one is willing to ascribe to this model clearly depends on the difference between the number of observations which it "explains" and the number of observations required to construct the model. If all of the available observations are used to construct the model, then we have no confidence in its predictions. This situation can also be described by saying that if the number of arbitrary decisions that have been made about the model, in order to make it fit the observations, is the same as the number of observations, then we have no confidence in the model.

This point was made succinctly by Poincaré, when he dismissed Jeans' classical explanation of the ultraviolet catastrophe and the specific heat of solids (1):

> "It is obvious that by giving suitable dimensions to the communicating tubes between his reservoirs and giving suitable values to the leaks, Jeans can account for any experimental results whatever. But this is not the role of physical theories. They should not introduce as many arbitrary constants as there are phenomena to be explained; they should establish connections between different experimental facts, and above all they should allow predictions to be made."

On the other hand, the accuracy with which the model reproduces the observed behaviour is clearly significant. If only a slight increase in complexity results in a large increase in accuracy, then in some sense fewer "arbitrary constants" have been added to it than the additional "number of phenomena" which it now explains. What is required for model assessment is some "trade-off" between the complexity of a model and its accuracy. A prerequisite for this is a measure of complexity which is applicable to a wide class of models. A major innovation introduced in this work is the casting of models in such a form, that poorness of fit of model behaviour to the observed behaviour appears as a component of model complexity. The required trade-off is thus achieved by assessing model complexity in a suitable manner.

A more orthodox approach to the problem of model assessment would be to examine the assessment of models chosen from a small class, and to postulate some statistical framework. It may then be possible to formulate the assessment problem as a statistical decision problem. This type of approach has indeed been investigated, even for dynamical models of the type encountered in control studies (2) (3) (4) (5). We do not follow such an approach for the following reasons.

Any method arrived at from statistical considerations will be appropriate only for a narrow class of models (such as linear difference-equation models, for example), set in a particular (statistical) environment (such as "observations corrupted by white, Gaussian, additive noise"). Such a method will not be useful if two very different models are being compared - for example, if the system being investigated is the behaviour of competing firms in some market, it may be desired to compare a model based on Forrester's "Industrial Dynamics" techniques (6) with a model which uses game theory (7) to explain firms' actions and the market's responses.

Realistic simulation models often contain nonlinear elements. When such models are also dynamical, it is usually extremely difficult to describe the evolution of the probability distributions of relevant variables (8). Furthermore, when investigating environmental and socio-economic systems, the most interesting and important behaviours often occur under transient conditions. When modelling these, it may not be appropriate to assume stationariness of relevant processes. Finally,when few observations of a system are available, and there is little a priori knowledge about it, the statistical specification of the system's environment may itself be very uncertain. In this case little is lost by not assuming it to be known; in fact, misleading conclusions may be avoided.

These considerations indicate that it may be more fruitful to investigate the assessment of models of complex, poorly understood systems by making as few assumptions as possible and examining the general situation, rather than by a painstaking and difficult analysis of each model structure, as it arises.

1.2 Overview of Approach and Results.

We develop a characterisation of modelling which has three "components": the system to be modelled, a model of this system, and a criterion of quality of the model.

The system to be modelled is taken to be defined by a pair of sets of observations of its input and output. Since measurements are always obtained with limited resolution and accuracy, each observation is assumed to be rational. Each set of observations is assumed to be finite. The system therefore looks like a set of discrete-state, discrete-time measurements. However, it will become evident that this does not constrain the models of such a system to be of the same category. It merely reflects the realities of data collection. A system will be defined in more detail in sec. 1.3.

A model of the system is any algorithm which maps certain subsets of the observations onto the output observations. This definition is broad enough to admit algorithms which would not normally be of much interest, such as those whose interpretation implies a reversed direction of time, or even a lack of any time ordering. It also allows algorithms which compute functions defined only on the particular observations obtained. These are useless for deducing how the system may behave in a new situation (presumably the goal of the modelling exercise), but models of this type will serve as a reference, with respect to which the success of the modelling exercise will be assessed. Any restriction to models of a particular type is accomplished by specifying which subsets of the observations lie in the domain of the algorithm, and which elements of the output observations are to be the corresponding images.

For example, deterministic difference equation models need only map successive blocks of input observations to successive outputs, whereas stochastic predicting models of the Wiener - Kolmogorov or Kalman types must map successive blocks of input and past output observations to successive outputs.

The term "algorithm" may be interpreted as "computer program". Thus we think of models as programs for computing the output observations, and these programs may use the specified subsets of the observations to help them in this task. This viewpoint would be excessively arbitrary, if it were not for the power of Church's Thesis (9), which states that any procedure which satisfies the intuitive notion of an "algorithm" can be expressed in any one of the equivalent formalisations of the theory of algorithms, and hence can be expressed as a computer program.

When the model is written as a computer program in some programming lanaguage, the criterion of quality is taken to be the shortness of that program, as measured by the number of characters in the program. The length of the program is a measure of the number of arbitrary decisions which have been made (relative to the programming language) in constructing the model. Furthermore, a model is required to compute the output observations exactly (to the accuracy with which the observations were originally made). In order to do this, the model must generate internally those terms which would conventionally be thought of as "fitting errors". Since the programming language has a finite number of terminals, the length of the model increases when these terms increase. The criterion of quality thus incorporates a particular trade-off between complexity and approximation.

The above characterisation of modelling is explained in more detail in Chapter 3. Support for it is given in section 2.2. The essence of this support is that the length of the shortest program required to compute a sequence displays properties analogous to the properties of the entropy associated with a probability space. In particular, a long sequence, which requires a maximally long program to compute it, passes every effective test for randomness (asymptotically, with probability 1). This suggests that the amount by which it is possible to "compress" the program (model) required to compute a set of observations (system) represents the amount of information which it has been possible to extract from the observations. If the only model which has been found is one that merely reads out the observations from a look-up table, then no "compression" has been achieved, and such a model conveys no information about the observations.

A consequence of our characterisation is that no algorithm can exist for finding the best model (according to the above criterion of quality) of an arbitrary system.

The choice of programming language to be used, for assessing the quality of a model, can be viewed as the specification of "what is to be taken for granted". It should therefore be made in the light of the modeller's *a priori* knowledge about the system, and of the purposes of the modelling exercise. In Chapter 4 this connection is examined more closely. It is shown that, if the observation sets are large enough, then the results of model assessment are independent of the choice of programming language. This can be interpreted to mean that the modeller's *a priori* beliefs become less significant as the set of observations available to him grows.

Nevertheless, the assessment of models of small observation sets is dependent on the modeller's specification of his *a priori* beliefs. Consequently such an assessment cannot be taken to be definitive. However, this is mitigated by the fact that the modeller does not need to choose between mutually exclusive sets of *a priori* beliefs: he can stipulate programming lanaguages which imply a greater or smaller state of knowledge.

Several different models, even when written in the same language, will rarely use exactly the same features of that language. It is therefore questionable whether a comparison of their lengths gives a measure of their complexity relative to the *same* set of assumptions. Chapters 5 and 6 resolve this difficulty. Chapter 5 develops a formal equivalent of "a program makes use of such-and-such facilities of a