System Theory: From Classical State Space to Variable Selection and Model Identification

Diego Liberati
Italian National Research Council, Italy

INTRODUCTION

System Theory is a powerful paradigm to deal with abstract models of real processes in such a way to be accurate enough to capture the salient underlying dynamics while keeping the mathematical tools easy enough to be manageable. Its typical approach is to describe reality via a reduced subset of ordinary differential equations (ODE) linking the variables. A classical application is the circuits theory, linking the intensive (voltage) and extensive (current) variables across and through each simplified element by means of equilibrium laws at nodes and around elementary circuits. When such relationships are linear (like in ideal capacitors, resistances, and inductors, just to stay in the circuit field), a full battery of theorems does help in understanding the general properties of the ODE system. Positive systems, quite often used in compartmental processes like reservoirs in nature and pharmacologic concentration in medical therapy, enjoy most of the properties of the linear systems, with the nonlinear constraint of non negativity. More general nonlinear systems are less easily treatable unless a simple form of nonlinearity is taken into account like the ideal characteristic of a diode in circuit theory. When the physics of the process is quite known, like in the mentioned examples, it is quite easy to identify a small number of variables whose set would fully describe the dynamics of the process, once their interrelations are properly modeled: this is the classical way to approach such a problem.

Nowadays, on the other side, new fields are growing up, like bioinformatics, where, instead, many data are collected over several possibly correlated variables whose joint dynamics would follow a law not a priori known nor easily understandable on the basis of the state-of-the-art knowledge. Given the opportunity to have so much data not easy to correlate by the human reader, but probably hiding interesting properties, one of the typical goals one has in mind is to face the problem on the basis of a hopefully reduced meaningful subset of the measured variables. The complexity of the problem makes it thus worthwhile to resort to automatic classification procedures in order to preprocess the collected data. Then, the original question does arise of reconstructing the synthetic mathematical model, capturing the most important relations between variables, in order to infer their hidden relationships, like in systems biology.

BACKGROUND

The introduced tasks of selecting salient variables and identifying their relationships from data may be sequentially accomplished with various degrees of success in a variety of ways. Principal components order the variables from the most salient to the least one, but only under a linear framework. Partial least squares do allow extension to nonlinear models, provided that one has prior information on the structure of the involved nonlinearity; in fact, the regression equation needs to be written before identifying its parameters. Clustering may operate even in an unsupervised way without the a priori correct classification of a training set (Boley, 1998). Neural networks are known to learn the embedded rules with the indirect possibility (Taha & Ghosh, 1999) to make rules explicit or to underline the salient variables. Decision trees (Quinlan, 1994) are a popular framework providing a satisfactory answer to the recalled needs.

Four main general purpose approaches will be briefly discussed in the present article. In order to reduce the dimensionality of the problem, thus simplifying both the computation and the subsequent understanding of the solution, the critical problems of selecting the most salient variables must be solved. This step may already be sensitive, pointing to the very core of the information to look at. A very simple approach is to resort to cascading a divisive partitioning of data orthogonal to the principal directions—PDDP—(Boley, 1998) already proven to be successful in the context of

analyzing micro-arrays data (Garatti, Bittanti, Liberati, & Maffezzoli, 2007).

A more sophisticated possible approach is to resort to a rule induction method, like the one described in Muselli and Liberati (2000). Such a strategy also offers the advantage of extracting underlying rules, implying conjunctions and/or disjunctions between the identified salient variables. Thus, a first idea of their even nonlinear relations is provided as a first step to design a representative model whose variables will be the selected ones. Such an approach has been shown (Muselli and Liberati, 2002) to be not less powerful over several benchmarks than the popular decision tree developed by Quinlan (1994)

An alternative in this sense can be represented by Adaptive Bayesian Networks (Yarmus, 2003) whose advantage is also to be available on a commercial wide spread data base tool like Oracle.

Finally, a possible approach to blindly building a simple linear approximating model is to resort to piece-wise affine (PWA) identification (Ferrari-Trecate, Muselli, Liberati, & Morari, 2003).

The joint use of (some of) such approaches briefly described in the present contribution, starting from data without known priors about their relationships, will thus allows reduction of dimensionality without significant loss in information, then to infer logical relationships, and, finally, to identify a simple input-output model of the involved process that also could be used for controlling purposes.

VARIABLE SELECTION VIA UNSUPERVISED CLUSTERING

In this article, we will firstly resort to a quite recently developed unsupervised clustering approach, the PDDP algorithm, proposed in Boley (1998). According to the analysis provided in Savaresi and Boley (2004), PDDP is able to provide a significant improvement of the performances of a classical *k-means* approach (Hand, Mannila, & Smyth, 2001; MacQueen, 1967), when PDDP is used to initialize the *k-means* clustering procedure. The approach taken herein may be summarized in the following three steps:

1. A principal component analysis defines a hierarchy in the transformed orthogonal variables according the principal directions of the data set (Hand et al., 2001; O'Connel, 1974).
2. The unsupervised clustering is performed by cascading a non-iterative technique—the principal direction divisive partitioning (PDDP) (Boley, 1998) based upon singular value decomposition (Golub & van Loan, 1996)— and the iterative centroid-based divisive algorithm *k-means* (MacQueen, 1967). Such a cascade, with the clusters obtained via PDDP used to initialize *k-means* centroids, is shown to achieve best performances in terms of both quality of the partition and computational effort (Savaresi & Boley, 2004).
3. By analyzing the obtained results, the number of variables needed for the clustering may be reduced by pruning all the original variables that are not needed in order to define the final partitioning hyper-plane so that the classification eventually is based on a few variables only.

VARIABLE SELECTION VIA MINIMUM DESCRIPTION LENGTH

Based on information theory, the minimum description length (MDL) principle (Barron, Rissanen, & Yu, 1998) states that the best theory to infer from training data is the one that minimizes the length (i.e., the complexity) of the theory itself and the length of the data encoded with respect to it (Friedman, Geiger, & Goldszmidt, 1997). This approach can be applied to address the problem of variable selection by considering each single variable as the simplest predictive model of the whole system. As described in (Kononenko, 1995), each variable can be ranked according to its description length that reflects the strength of its correlation with the system dynamics. In this context, the MDL measure is given by (Yarmus, 2003) weighting the encoding length (with one submodel for each observed value of the variable) with the number of bits needed to describe the data, based on the probability distribution of the target value associated to each submodel. However, once all features have been ordered by rank, no a priori criterion is available to choose the cut-off point beyond which features can be discarded. To circumvent this drawback, one can adopt an iterative approach that starts with building a model on the set of the n-top ranked variables. Then, a new variable is