Annual Reviews in Control 34 (2010) 1–12

Contents lists available at ScienceDirect

Annual Reviews in Control

journal homepage: www.elsevier.com/locate/arcontrol

Perspectives on system identification☆,☆☆

Lennart Ljung

Division of Automatic Control, Linköpings Universitet, SE-581 83 Linköping, Sweden

ARTICLE INFO

Article history:
Received 28 October 2009
Accepted 10 December 2009
Available online 10 April 2010

Keywords:
System identification
Mathematical models
Estimation
Non-linear models
Statistical methods

ABSTRACT

System identification is the art and science of building mathematical models of dynamic systems from observed input–output data. It can be seen as the interface between the real world of applications and the mathematical world of control theory and model abstractions. As such, it is an ubiquitous necessity for successful applications. System identification is a very large topic, with different techniques that depend on the character of the models to be estimated: linear, nonlinear, hybrid, nonparametric, etc. At the same time, the area can be characterized by a small number of leading principles, e.g. to look for sustainable descriptions by proper decisions in the triangle of model complexity, information contents in the data, and effective validation. The area has many facets and there are many approaches and methods. A tutorial or a survey in a few pages is not quite possible. Instead, this presentation aims at giving an overview of the "science" side, i.e. basic principles and results and at pointing to open problem areas in the practical, "art", side of how to approach and solve a real problem.

1. Introduction

Constructing models from observed data is a fundamental element in science. Several methodologies and nomenclatures have been developed in different application areas. In the control area, the techniques are known under the term *system identification*. The area is indeed huge, and requires bookshelves to be adequately covered. Any attempt to give a survey or tutorial in a few pages is certainly futile.

I will instead of a survey or tutorial provide a subjective view of the state of the art of system identification—what are the current interests, the gaps in our knowledge, and the promising directions.

Due to the many "subcultures" in the general problem area it is difficult to see a consistent and well-built structure. My picture is rather one of quite a large number of satellites of specific topics and perspectives encircling a stable core. The core consists of relatively few fundamental results of statistical nature around the concepts of *information*, *estimation* (*learning*) and *validation* (*generalization*). Like planets, the satellites offer different reflections of the radiation from the core.

Here, the core will be described in rather general terms, and a subjective selection of the encircling satellites will be visited.

2. The core

The core of estimating models is statistical theory. It evolves around the following concepts:

Model. This is a relationship between observed quantities. In loose terms, a model allows for prediction of properties or behaviors of the object. Typically the relationship is a mathematical expression, but it could also be a table or a graph. We shall denote a model generically by $\mathcal{M}$.
True description. Even though in most cases it is not realistic to achieve a "true" description of the object to be modeled, it is sometimes convenient to assume such a description as an abstraction. It is of the same character as a model, but typically much more complex. We shall denote it by S.
Model class. This is a set, or collection, of models. It will generically be denoted by M. It could be a set that can be parameterized by a finite-dimensional parameter, like "all linear state-space models of order n", but it does not have to, like "all surfaces that are piecewise continuous".
Complexity. This is a measure of "size" or "flexibility" of a model class. We shall use the symbol C for complexity measures. This could be the dimension of a vector that parameterizes the set in a smooth way, but it could also be something like "the maximum norm of the Hessian of all surfaces in the set."
Information. This concerns both information provided by the observed data and prior information about the object to be modeled, like a model class.
Estimation. This is the process of selecting a model guided by the information. This includes both finding a suitable model

☆ An earlier version of this article was presented as a plenary paper at the 17th IFAC World Congress, Seoul, Korea, July 6–11, 2008.
☆☆ This work was supported by the Swedish Research Council under the Linnaeus Center CADICS and the Swedish Foundation for Strategic Research via the center MOVIII.

E-mail address: ljung@isy.liu.se.

1367-5788/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.arcontrol.2009.12.001

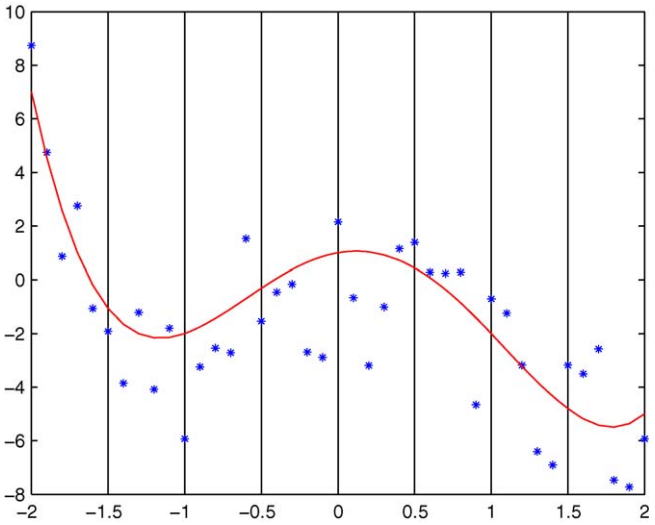
class and the particular model parameters in that class. The data used for selecting the model is called *estimation data* (or *training data*) and will be denoted by Z_e^N (with N marking the size of the data set). It has become more and more fashionable to call this process *learning*, also among statisticians.

Validation. This is the process of ensuring that the model is useful not only for the estimation data, but also for other data sets of interest. Data sets for this purpose are called *validation data*, to be denoted by Z_v. Another term for this process is *generalization*.

Model fit. This is a (scalar) measure of how well a particular model $\mathcal{M}$ is able to "explain" or "fit to" a particular data set Z. It will be denoted by $F(\mathcal{M}, Z)$.

To have a concrete picture of a template estimation problem, it could be useful to think of elementary *curve-fitting*.

Example 1. A template problem—curve-fitting



Consider an unknown function $g_0(x)$. For a sequence of x-values (regressors) $\{x_1, x_2, \ldots, x_N\}$ (that may or may not be chosen by the user) we observe the corresponding function values with some noise:

$$y(t) = g_0(x_t) + e(t) \tag{1}$$

The problem is to construct an estimate

$$\hat{g}_N(x) \tag{2}$$

from

$$Z^N = \{y(1), x_1, y(2), x_2, \ldots, y(N), x_N\} \tag{3}$$

This is a well-known basic problem that many people have encountered already in high-school. In most applications, x is a vector of dimension, say, n. This means that g defines a surface in $\mathbb{R}^{n+1}$ if y is scalar. If $y(k)$ itself is a p-dimensional vector, it is in this perspective convenient to view the problem as p separate surface-fitting problems, one for each component of y.

Two typical approaches are the following ones:

Parametric: Postulate a parameterized model set M, of say nth order polynomials $g(x,\theta)$, parametrized by the $n+1$ coefficients θ, and then adjust θ to minimize the least squares fit between $y(k)$ and $g(x_t,\theta)$. A complexity measure C could be the order n.

Nonparametric: Form, at each x, a weighted average of the neighboring $y(k)$. Then a complexity measure C could be the size of the neighborhoods. (The smaller the neighborhoods, the more complex/flexible curve.)

The border line between these approaches is not necessarily distinct.

2.1. Estimation

All data sets contain both useful and irrelevant information ("Signal and noise"). In order not to get fooled by the irrelevant information it is necessary to meet the data with a prejudice of some sort. A typical prejudice is of the form "Nature is Simple". The conceptual process for estimation then becomes

$$\hat{\mathcal{M}} = \arg\min_{\mathcal{M}\in M}[F(\mathcal{M}, Z_e^N) + h(C(\mathcal{M}), N)] \tag{4}$$

where F is the chosen measure of fit, and $h(C(\mathcal{M}), N)$ is a penalty based on the complexity of the model $\mathcal{M}$ or the corresponding model set M and the number of data, N. That is, the model is formed taking two aspects into account:

(1) The model should show good agreement with the estimation data.
(2) The model should not be too complex.

Since the "information" (at least the irrelevant part of it) typically is described by random variables, the model $\hat{\mathcal{M}}$ will also become a random variable.

The method (4) has the flavor of a parametric fit to data. However, with a conceptual interpretation it can also describe nonparametric modeling, like when a model is formed by kernel smoothing of the observed data.

The complexity penalty could simply be that the search for a model is constrained to model sets of adequate simplicity, but it could also be more explicit as in the curve-fitting problem:

$$V_N(\theta, Z_e^N) = \sum (y(t) - g(\theta, x_t))^2 \tag{5a}$$

$$\hat{\theta}_N = \arg\min_{\theta} V_N(\theta, Z_e^N) + \delta||\theta||^2 \tag{5b}$$

Such model complexity penalty terms as in (5b) are known as *regularization* terms.

2.2. Fit to validation data

It is not too difficult to find a model that describes estimation data well. With a flexible model structure, it is always possible to find something that is well adjusted to data. The real test is when the estimated model is confronted with a new set of data—validation data. The average fit to validation will be worse than the fit to estimation data. There are several analytical results that quantify this deterioration of fit. They all have the following conceptual form: Let a model $\hat{\mathcal{M}}$ be estimated from an estimation data set Z_e^N in a model set M. Then

$$\bar{F}(\hat{\mathcal{M}}, Z_v) = F(\hat{\mathcal{M}}, Z_e^N) + f(C(M), N) \tag{6}$$

Here, the left-hand side denotes the expected fit to validation data, while the first term on the right is the model's actual fit to estimation data ("the empirical risk"). The fit is typically measured as the mean square error as in (5a). Hence, to assess the quality of the model one has to adjust the fit seen on the estimation data with a quantity that depends on the complexity of the model set used. The more flexible the model set, the more adjustment is necessary. Note that $\hat{\mathcal{M}}$ is a random variable, so the statement (6) is a probabilistic one.