Perspectives on System Identification ⋆

Lennart Ljung *

* *Division of Automatic Control, Linköpings universitet, SE-581 83 Linköping, Sweden (e-mail: ljung@isy.liu.se)*

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 in the solar system, 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 $\mathbb{m}$.

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 $\mathcal{S}$.

Model Class. This is a set, or collection, of models. It will generically be denoted by $\mathcal{M}$. 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 $\mathcal{C}$ 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. 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.

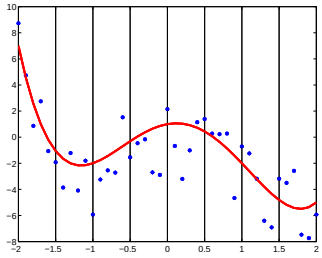
Validation. This is the process of ensuring that the model is useful not only for the estimation data, but also for other data sets of interest. Data sets for this purpose are called *validation data*, to be denoted by Z_v. Another term for this process is *generalization*.

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

⋆ This work was supported by the Swedish Research Council and the Swedish Foundation for Strategic Research via the center MOVIII.

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 $\mathcal{M}$, of say $d-1$:th order polynomials $g(x, \theta)$, parametrized by the d coefficients θ (for a scalar x), and then adjust θ to minimize the least squares fit between $y(k)$ and $g(x_t, \theta)$. A complexity measure $\mathcal{C}$ could be the order n.

Nonparametric: Form, at each x, a weighted average of the neighboring $y(k)$. Then a complexity measure $\mathcal{C}$ 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

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

where $\mathcal{F}$ is the chosen measure of fit, and $h(\mathcal{C}(\mathfrak{m}), N)$ is a penalty based on the complexity of the model $\mathfrak{m}$ or the corresponding model set $\mathcal{M}$ and the number of data. 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.

These aspects are somewhat contradictory, and a good trade-off must be found, as we shall discuss later. Since the "information" (at least the irrelevant part of it) typically is described by random variables, the model $\hat{\mathfrak{m}}$ 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 non-parametric 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 $\hat{\mathfrak{m}}$ be estimated from an estimation data set Z_e^N in a model set $\mathcal{M}$. Then

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

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). The quantity f is a strictly positive function which increases with the complexity $\mathcal{C}$ and decreases with the number N of estimation data. Hence, to assess the quality of the model one has to adjust the fit seen on the estimation data with this positive quantity. The more flexible the model set, the more detoriation of the fit should be expected. Note that $\hat{\mathfrak{m}}$ is a random variable, so the statement (6) is a probabilistic one.

For the simple curve fitting problem (5) with $\delta = 0$, the expression (6) leads to the well known forms

$$E\bar{\mathcal{F}}(\hat{\mathfrak{m}}, Z_v) \approx \frac{1 + d/N}{1 - d/N}\mathcal{F}(\hat{\mathfrak{m}}, Z_e^N) \tag{7a}$$

$$\approx (1 + 2d/N)\mathcal{F}(\hat{\mathfrak{m}}, Z_e^N) \tag{7b}$$

$$\approx \frac{1}{(1 - d/N)^2}\mathcal{F}(\hat{\mathfrak{m}}, Z_e^N) \tag{7c}$$

where the left hand side is the expected fit when applied to validation data with expectation also over $\hat{\mathfrak{m}}$, and $\mathcal{F}(\hat{\mathfrak{m}}, Z_e^N)$ is the model's fit to estimation data. The first expression is Akaike's Final Prediction Error (FPE), the second one is Akaike's Information Criterion (AIC) when applied to the Gaussian case, (Akaike, 1974), and the third

one is the generalized cross-validation (GCV) criterion, (Craven and Wahba, 1979). Here $d = \dim \theta$ serves as the complexity measure of the model set.

Remark: These expressions are derived with expectations over Z_e^N on both sides, e.g. Ljung (1999), Ch 16. However, they are typically used to estimate the quantity on the LHS, and then the expectation on the RHS is replaced with the observed fit, "the empirical risk".

When the (regularized) criterion (5) with $\delta > 0$ is used, d in the above expression is replaced with the effective number of parameters $d^*(\delta)$, see e.g. Ljung (1999), where

$$d^*(\delta) = \sum_{k=1}^{d} \frac{\sigma_k}{\sigma_k + 2\delta} \tag{8a}$$

$$\sigma_k = \text{the singular values of } V_N''(\theta, Z_e^N) \tag{8b}$$

Similarly, when Vapnik's learning theory is applied to function estimation, the model's behavior on validation data is such that with probability at least $1-\delta$, see Suykens et al. (2002)

$$\bar{\mathcal{F}}(\hat{\mathfrak{m}}, Z_v) \leq \mathcal{F}(\hat{\mathfrak{m}}, Z_e^N) \left(1 - \sqrt{\frac{\mathcal{C}(\log N/\mathcal{C}) + 1 - \log\delta}{N}}\right)_+^{-1} \tag{9}$$

Here the complexity measure $\mathcal{C}$ is the Vapnik–Chervonenkis (VC)-dimension of the model set, which measures how well the functions of the class can shatter (separate) random points.

Yet another result in this vein is the following one: Let $\hat{\mathfrak{m}}$ be an estimated model in a model set $\mathcal{M}$ for data Z_e^N, and let $\mathfrak{m}^*$ be the model that has the best fit (to Z_v) in $\mathcal{M}$. Then with probability at least $1 - \delta$ it will obey

$$\bar{\mathcal{F}}(\hat{\mathfrak{m}}, Z_v) \leq \bar{\mathcal{F}}(\mathfrak{m}^*, Z_v) + 2\mathcal{R}_N + c\sqrt{\frac{\log 1/\delta}{N}} \tag{10}$$

where $\mathcal{R}_N$ is the *Rademacher average* which describes how well the model structure $\mathcal{M}$ is capable of reproducing random functions. See Bartlett et al. (2005) for precise formulations.

The intuitive, and rather obvious conclusion of all this is that *one should not be so impressed by a good fit to estimation data, if the model set has been quite flexible.*

The formal results reviewed in this section also give a good rational ground for the pragmatic form of the estimation criterion (4), and give several concrete suggestions for the function h.

2.3 Bias and Variance

If we assume that there is a true description $\mathcal{S}$, we can conceptually write the model error

$$\mathcal{S} - \hat{\mathfrak{m}}$$

(actually we can interpret $\mathcal{S}$ and $\hat{\mathfrak{m}}$ to be any (scalar) property of the object, like the static gain of a dynamical system.)

The mean square error (MSE) is

$$W = E(\mathcal{S} - \hat{\mathfrak{m}})^2 = (\mathcal{S} - \mathfrak{m}^*)^2 + E(\hat{\mathfrak{m}} - \mathfrak{m}^*)^2 = B + V \tag{11a}$$

$$\mathfrak{m}^* = E\hat{\mathfrak{m}} \tag{11b}$$

where the MSE splits into a bias (square) contribution B and a variance error V. This is a very elementary and well known relation, but still worth some contemplation. It is obvious that B is a function of the model class complexity (flexibility) $\mathcal{C}$ and that it decreases as $\mathcal{C}$ increases. Analogously V increases with $\mathcal{C}$. (There are elementary expressions for this, but intuitively it is sufficient to realize that the wider model class is used, the more susceptible the model will be for picking up random misinformation in data.)

Conceptually,

$$V = g(\mathcal{C})/N$$

where g increases with $\mathcal{C}$ and N is size of the estimation data set.

Our search for the pragmatically best model will thus mean a search for the model complexity $\mathcal{C}$ that minimizes W. This will typically occur for an $\mathcal{M}$ such that $B \neq 0$, even when we in principle could find a more flexible $\mathcal{M}$ with no bias: *"We should thus not strive for the truth, but for reasonable approximations."* Another consequence is that minimizing W leads to increased model complexity when we have more data, since V decreases with N. Yet another consequence is that minimization of (11) favors small model sets that contain good approximations of the true object. This means that it is beneficial to shrink the model set as much as possible using physical (or other) insights into the nature of the object. In the control literature, such model sets are called *Grey-box models.* It is another matter that it may be computationally expensive and cumbersome to use such model sets.

2.4 The Information Contents in Data

The value of information in observed data must be measured with respect to what we already know. For example, suppose that we know the logarithm of the probability density function $\ell_Y(x, \theta)$ of an observed random variable Y up to a parameter θ. Then the *Fisher Information matrix* for θ is

$$\mathcal{I} = E\ell_Y'(Y, \theta)(\ell_Y'(Y, \theta))^T \tag{12}$$

where prime denotes differentiation w.r.t. θ. The celebrated Cramér-Rao

inequality, (Cramér, 1946), says that no (unbiased) estimator $\hat{\theta}$ can have a smaller covariance matrix than the inverse of $\mathcal{I}$:

$$\operatorname{Cov}\hat{\theta} \geq \mathcal{I}^{-1} \tag{13}$$

For the curve fitting problem (5a) with Gaussian errors, the information matrix is

$$\mathcal{I} = \frac{1}{\lambda}\sum_{t=1}^{N} g'(x_t, \theta)(g'(x_t, \theta))^T \tag{14}$$

where λ is the variance of the errors e.

3. THE COMMUNITIES AROUND THE CORE

The art and technique of building mathematical models of (dynamic) systems is crucial to many application areas. Hence, many scientific communities are active in developing theory and algorithms. With a few exceptions, this has taken place in surprisingly separated and isolated

environments, with journals and conferences of their own. So, we see separate subcultures in the general problem area, and it would be highly desirable to encourage more active exchanges of ideas. In particular, I am sure that the System Identification community would benefit from an influx of new ideas from other cultures.

3.1 Statistics

Mathematical statistics and time series analysis (cf Section 3.2) is in many respects the "mother" field of System Identification, see e.g. Deistler (2002). Here many of the basic results of Section 2 were developed. Statistics is clearly a very broad field, and it is not meaningful to give terse summary of recent trends.

Among developments with relevance to System Identification are for example the *bootstrap*, see e.g. Efron and Tibshirani (1993), and the EM algorithm, (Dempster et al., 1977). Other results of relevance to order selection are new techniques for regularization (variants of (5b)), such as *Lars, Lasso, NN-garotte*, see e.g. Hastie et al. (2001).

3.2 Econometrics and Time Series Analysis

Econometrics is a science that has grown out of statistics for extracting information from economic data, taking into account both the special features of such data and the a priori information coming from economic theory. Econometrics has a long tradition of giving inspiration to time series and difference equation modeling and its roots coincide with developments in statistics. The work on time series dates back to Jevons (1884), Yule (1927), and Wold (1938). The classic paper Mann and Wald (1943) developed the asymptotic theory for the LS estimator for stochastic linear difference equations (AR systems). The results were extended to simultaneous (multivariate) systems, where LS is not consistent, in Koopmans et al. (1950), where also central identifiability issues were sorted out and Gaussian Maximum Likelihood estimates were proposed and analyzed. Important extensions to the ARMA(X) case have been proposed by Anderson (1971) Hannan (1970) later on. The problem of errors-in-variables modeling (when there are disturbances on both input and output measurements) also has its origins in econometrics, (Frisch, 1934).

More recently, important focus has been on describing volatility clustering, i.e. more careful modeling of conditional variances for modeling and forecasting of risk (GARCH models, (Engle, 1982)), as well as on describing non-stationary behavior of interesting variables in terms of a common stationary linear combination ("cointegration"), (Engle and Granger, 1987), which gives the long run equilibrium relation between these variables. These two subjects were in focus for the Sveriges Riksbanks Prize in Economic Sciences in memory of Alfred Nobel in 2003.

3.3 Statistical Learning Theory

The coining of the term *Statistical learning*, e.g. Vapnik (1998), Hastie et al. (2001), has moved the fields of statistics and *Machine learning* closer together. Much effort has been devoted to convex estimation formulations, such as *support vector machines*, Vapnik (1982). An important feature of these techniques is that the classification/estimation is formulated in high-dimensional feature spaces, which by RKHS (Reproducing Kernel Hilbert Space) theory is transformed to kernels in the data space. The *kernel trick* and Mercer conditions are terms that relate to this transformation, see e.g. Wahba (1999). Issues of *convexification* have lately played an essential role, e.g. Bartlett et al. (2006).

3.4 Machine Learning

The term *machine learning* was coined in Artificial Intelligence, see e.g. the classical book Nilsson (1965). The area has housed many approaches, like Kohonen's self-organizing and self-learning maps, (Kohonen, 1984), to Quinlan's tree-learning for binary data, (Quinlan, 1986), and the early work on perceptrons, (Rosenblatt, 1962), that later led to neural networks. More recent efforts, include Gaussian Process Regression (kriging), e.g. Rasmussen and Williams (2006), which in turn can be traced back to general nonlinear regression. Overall, the fields on machine learning and statistical learning appear to be converging.

3.5 Manifold Learning

Another "learning topic" is *manifold learning*, which really is the important area of dimension reduction of high-dimensional data to nonlinear manifolds. This is a nonlinear counterpart of multivariate data analysis, such as Principal Component Analysis (PCA). Some techniques, like *kernel PCA*, (Schölkopf et al., 1999), are such extensions. Other methods are based on developing proximity matrices, often with nonparametric techniques, such as isomaps and variance unfolding. A special such technique that has been frequently used is LLE (*Local Linear Embedding*), (Roweis and Saul, 2000). It can be described as a way to determine a coordinate system in the manifold that inherits neighborhoods and closeness properties of the original data. Manifold learning has evolved into a community of its own, essentially because of its importance for computer vision and object recognition.

3.6 Statistical Process Control and Chemometrics

The term *chemometrics* is primarily used in process industry and stands for statistical methods for extracting information from data sets that often consist of many measured variables. The techniques are various forms of Multivariate data analysis, such as PCA, but in Chemometrics the use of Partial Least Squares (PLS), (Wold et al., 1984), has been a predominant way of projecting data onto linear subspaces. For a recent survey, see MacGregor (2003). The PLS methods are conceptually related to *subspace methods* in System Identification. The term *(Multivariate) Statistical Process Control* refers to identifying important indicators for the process and keeping track of their changes.

3.7 Data Mining

Data Mining or Knowledge Discovery has become a buzzword for sorting through large databases and finding relevant information. Data mining has been applied to a

variety of problems like intelligence, business, finance and searching the web for medical and genetic information. It employs many different techniques, in terms of computer science, numerical analysis, and visualization. Wikipedia describes data mining as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data". It is clear that this definition puts the subject in orbit of our core, and several techniques of statistical and model predictive nature, like Neural Networks, Decision Trees, and Nearest Neighbor classification are also found in the toolbox of data mining. See Tan et al. (2006) for an introduction.

3.8 Artificial Neural Networks

Neural Networks are a development of the perceptron, (Rosenblatt, 1962) and have seen a significant development over the past two decades, e.g. Haykin (1999) and the area has transformed into a community of its own, e.g. "The IEEE Neural Networks Society". This is quite remarkable, since the topic is but a flexible way of parameterizing arbitrary hypersurfaces for regression and classification. The main reason for the interest is that these structures have proved to be very effective for solving a large number of nonlinear estimation problems.

3.9 Fitting ODE-coefficients, and Special Applications

There is also a subarea of contributions and books that do not have roots neither in statistics nor artificial intelligence. This is a perspective that has come from physical modeling, simulation and solving ordinary differential equations (ODEs), and is thus more of numerical analysis and engineering. In such a context, if the ODE contains unknown parameters, it is natural to employ numeric optimization to fit the solutions to observed data, without invoking a statistical framework. This perspective is taken, e.g., in Schittkowski (2002).

To this culture we can also count books that deal with particular applications, (even though they may use also statistics), e.g., the extensive literature on aircraft modeling, like Klein and Morelli (2006).

3.10 System Identification

System Identification is the term that has been coined by Zadeh (1956) for the model estimation problem for dynamic systems in the control community. Two main avenues can be seen for the development of the theory and methodology (Gevers, 2006): One is the *realization avenue*, that starts from the theory how to realize linear state space models from impulse responses, Ho and Kalman (1966), followed by Akaike (1976), leading to so-called subspace methods, e.g. Larimore (1983) and Van Overschee and DeMoor (1996). The other avenue is the *prediction-error approach*, more in line with statistical time-series analysis and econometrics. This approach and all its basic themes were outlined in the pioneering paper Åström and Bohlin (1965). It is also the main perspective in Ljung (1999).

A main feature of dynamical systems is that the future depends on the past. Thus a prediction of the output $y(t)$ at time t, either being constructed by *ad hoc* reasoning or carefully calculated in a stochastic framework, depends on all or some previous measured inputs and outputs, $Z^{t-1} = \{y(t-1), u(t-1), y(t-2), u(t-2), \ldots\}$. Let us denote the prediction by $\hat{y}(t|t-1) = g(Z^{t-1})$. In case the system is not fully known, this prediction will be parameterized by a parameter θ (which typically is finite-dimensional, but could also conceptually capture nonparametric structures) so the prediction is

$$\hat{y}(t|\theta) = g(Z^{t-1}, \theta) \tag{15}$$

The distinguishing features as well as the bulk of efforts in System Identification can, somewhat simplistically, be described as

(1) Invent parameterizations $\hat{y}(t|\theta)$, suitable to describe linear and nonlinear dynamic systems. For underlying state-space realizations, *realization theory* has been an important source of inspiration. Questions of how prior physical knowledge can best be incorporated form another central issue.
(2) Translate the core material of Section 2 to properties of estimated systems, as well as estimation procedures.
(3) Find effective ways to estimate θ numerically for the chosen parameterizations. The curve-fitting criterion (5) forms a beacon for these efforts in the prediction error approach, typically leading to nonlinear optimization by iterative search. The realization avenue has developed techniques based on SVD and QR factorizations.
(4) The typical intended use of the model in this context is for prediction or control. This means that models of the noise affecting the system often are essential.
(5) Experiment design now becomes the selection of input signal. The effects of the experiment design can be evaluated from the core material, but can be given concrete interpretations in terms of model quality for control design, e.g. Gevers (1993). Specific features for control applications are the problems and opportunities of using inputs, partly formed from output feedback, e.g. Hjalmarsson (2005). An important problem is to quantify the model error, and its contribution from the variance error and the bias error, cf. (11), "model error models", e.g. Goodwin et al. (1992).

4. SOME OPEN AREAS IN SYSTEM IDENTIFICATION

System Identification is quite a mature area that has had an interesting and productive development. Much has been done, but many problems remain. I shall in this section outline a few areas that I believe are worthy of more studies, and I would like to encourage young researchers to have a go at these problems. Some open areas from an industrial perspective follow in Section 6.

4.1 Issues in Identification of Nonlinear Models

A nonlinear dynamic model is one where $\hat{y}(t|\theta) = g(Z^t, \theta)$ is nonlinear in Z^N (but could be any function, including linear, of θ). Identification of nonlinear models is probably the most active area in System Identification today, Ljung and Vicino (2005). It is clear from Section 3 that there is a corresponding wide activity in neighboring communities,

and I think it is important for the control community to focus on issues that are central for dynamic systems:

- What are the useful parameterizations of $\hat{y}(t|\theta)$ for nonlinear models of dynamic systems? Ideas and suggestions have proliferated and given rise to a confusing amount of models. In Section 5 we make an attempt at a taxonomy of such models.
- Stability of predictions and simulations: Consider the following simple example. Let a nonlinear model by given by $\hat{y}(t|\theta) = g(y(t-1), u(t-1), \theta)$. The predictions are simple enough, since they just involve two measured inputs and outputs. The simulation of this model for a given input is more tricky:
$$y_s(t) = g(y_s(t-1), u(t-1), \theta), t = 1, 2, ...$$
This is a nonlinear dynamic system, and to establish for which values of θ it is stable, is in general very difficult. For control applications, it would be very helpful to find flexible classes of models, for which simulation stability could be tested with reasonable effort.
- How to identify a nonlinear system that operates in closed loop and is stabilized by an unknown regulator?
- Develop Model Error Models for linear or nonlinear models of nonlinear systems that can be used for robust control design.
- Find effective data-based nonlinearity tests for dynamical systems.

4.2 Convexification

By convexification we mean formulating the estimation method as a convex optimization problem (or, more generally, one that does not give rise to local minima). As seen in Section 3, convexification has played a major role recently in several communities, and the development of convex and semidefinite programming has been booming in recent years, e.g. Boyd and Vandenberghe (2004). These activities have not been particularly pronounced in the System Identification community, which has largely been sticking to a maximum likelihood (or related) framework.

The likelihood function for estimating θ in a parameterization $\hat{y}(t|\theta)$ is typically multi-modal, so the estimation problem becomes plagued by the existence of local minima. The estimation procedure becomes dependent on good initial starting values, and such are often easy to find for black-box linear models. For other models, both nonlinear black-box models and linear grey-box models, the problem is a serious one. More efforts are required to understand convexification and utilize modern semidefinite programming techniques. One solution is of course to look for how far we can get with linear parameterizations ($\hat{y}(t|\theta)$ linear in θ), for example LS Support Vector Machines and other radial bases (kernel) techniques, cf. Suykens et al. (2002).

While convexification has not been a focus of System Identification research, there are a few results of this character:

Subspace methods, (N4SID, MOESP, etc.) form a one-shot estimation procedure for black-box linear state space models, e.g. Van Overschee and DeMoor (1996). Pushing the interpretation slightly, these methods can be seen as a form of regularized Least Squares for ARX models.

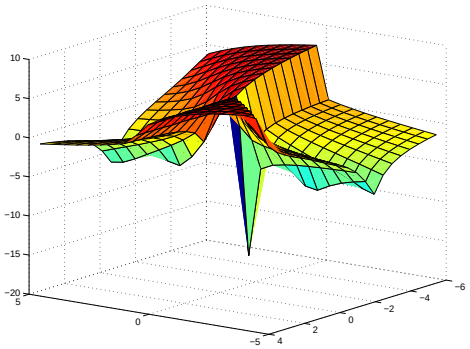


Fig. 1. The surface $\log EV_N(\theta)$. The "floor" is formed by the two parameters θ_1 and θ_2 and the "wall" is the value of V.

Reformulation via differential algebra. Consider the following example:

Example 2. (Constructing more measured signals). Let a model be described by (the Michaelis-Menten equation)
$$\dot{y} = \theta_1 \frac{y}{\theta_2 + y} - y + u \tag{16a}$$
$$y_m(t_k) = y(t_k) + e(t_k) \tag{16b}$$
The signals $u(t_k)$ and $y_m(t_k), k = 1, 2, \ldots, N$ are measured and e is white Gaussian noise. The maximum likelihood method for estimating θ is to minimize
$$V_N(\theta) = \sum_{k=1}^{N} (y_m(t_k) - \hat{y}(t_k|\theta))^2$$
where $\hat{y}(t|\theta)$ is the solution to (16a). The surface $\log EV_N(\theta)$ is shown in Figure 1 when the input u is an impulse. The surface does not depend on the variance of e. (That only shifts the level by a constant.) Clearly it is a difficult function to minimize: There are many local minima, the global minimum has a very narrow domain of attraction, and there are slopes that continue to roll off downhill as θ grows. This is true no matter how small the variance of the noise is. One might conclude that it is difficult to find the parameters of this model, and that information about them are well hidden in the data.

If we for the moment disregard the noise e, we can do as follows: Multiply (16a) with the numerator and rearrange the terms:
$$\dot{y}y + \theta_2\dot{y} = \theta_1 y - y^2 - \theta_2 y + uy + \theta_2 u$$
or
$$\dot{y}y + y^2 - uy = [\theta_1 \;\; \theta_2] \begin{bmatrix} y \\ u - \dot{y} - y \end{bmatrix} \tag{17}$$
Equation (17) is a linear regression that relates the unknown parameters and measured variables. We can thus find them by a simple least squares procedure. We have, in a sense, convexified the problem in Figure 1.

The manipulations leading to (17) are an example of Ritt's algorithm in Differential Algebra. In fact it can be shown, (Ljung and Glad, 1994), that any *globally identifiable model structure can be rearranged* (using Ritt's algorithm) *to a linear regression.* This is in a sense a general convex-

ification result for any identifiable estimation problem. A number of cautions must be mentioned, though:

- Although Ritt's algorithm is known to converge in a finite number of steps, the complexity of the calculations may be forbidding for larger problems.
- With noisy measurements, care must be exercised in differentiation, and also the linear regression may be subject to disturbances that can give biased estimates.

But the fact remains: the result shows that the complex, non-convex form of the likelihood function with many local minima is not inherent in the model structure.

Sum Of Squares Techniques in Linear Grey-Boxes. The non-convex nature of the criterion function is normally no serious problem for linear black-box models, since many consistent, non-optimal, non-iterative methods (e.g. Instrumental Variables, N4SID) can be used to get into the domain of attraction of the global minimum. The issue is much more pressing for physically parameterized grey-box models, even in the linear case. Monte-Carlo tests show, that, e.g. for third order linear grey-box models with six physical parameters, the success rate in reaching the global minimum from random initializations of the parameters, ranges from 0% to 18%. An attempt to convexify this problem is described in Parrilo and Ljung (2003): Let A, B and C be a black-box estimate of the linear system, and let $A_0(\theta), B_0(\theta), C_0(\theta)$, be the corresponding grey-box parameterizations which are assumed to be affine in θ. Then formulate the following polynomial optimization problem:

$$(\hat{\theta}, \hat{T}) = \arg\min_{\theta,T} h(\theta, T) \tag{18a}$$

$$h(\theta, T) = \|T \cdot A - A_0(\theta) \cdot T\|_F + \|T \cdot B - B_0(\theta)\|_F + \|C - C_0(\theta)T\|_F \tag{18b}$$

and solve it using the *sum of squares* technique, (Parrilo and Sturmfels, 2003). Here T corresponds to the unknown similarity transform that connects the two state-space bases. Also this approach suffers from complexity issues, and needs to be refined.

Manifold Learning. Manifold learning was mentioned in Section 3.5. The "re-coordinatization" of the LLE-technique can also be used to cover the nonlinearities of the mapping from the original regressor space to the measured output, so we can obtain a concatenated mapping

$$\mathcal{X} \to g(x) \to \mathcal{Z} \to h(z) \to \mathcal{Y}$$

where $\mathcal{X}$ is the original regressor space, $\mathcal{Z}$ is the manifold (typically, but not necessarily, of lower dimension than $\mathcal{X}$), and $\mathcal{Y}$ is the space of the observed outputs. The function g is the LLE mapping, and h may be chosen as a simple mapping (with convexity properties) relying upon g for the nonlinearities. Some attempts with this (for dynamic models) are reported in Ohlsson et al. (2008).

Model order reduction. Model reduction is closely related to System Identification, cf. Section 4.3. It is therefore interesting to follow convexification attempts for model order reduction problems, see Sou et al. (2008), and see if they have implications on system identification loss function formulations.

4.3 Model Approximation/Model Reduction

System identification is really system approximation. We attempt to find a model of acceptable accuracy from data, and the resulting model is by necessity an approximation of the true description. This means that the topics of model reduction and model approximation are closely related to identification. Now, model reduction is in itself a huge research area with a wide scope of application areas (e.g. `www.modelreduction.com`). It could rightly have been listed as one of the communities around the core, Section 3, but it lacks the data-estimation component. It is true that a model reduction perspective has been in focus for some work in system identification, but I am convinced that the identification community could learn a lot more by studying the model reduction research - especially for nonlinear systems.

Linear Systems – Linear Models. Model reduction for linear models is quite well understood. Balanced realizations, Moore (1981), show how the different states contribute to the input-output map and are a rational ground for reducing the state dimension by projecting the state-space to certain subspaces. As noted in the original contribution this is pretty much like Principal Component Analysis (PCA) in linear regression (and related to how the state space is selected in subspace methods, cf. Section 4.2). Linear model reduction can be a very useful tool in system identification (cf. the command `balred` in the System Identification Toolbox, Ljung (2007)), for example when concatenating single-output models to a bigger model. My impression is, though, that this possibility is much underutilized.

Nonlinear Systems – Linear Models. The situation becomes much more difficult and interesting when we want to approximate a nonlinear system with a linear model, (which is typically what happens in practice when you build linear models.) Certain issues are well understood, like what is the linear second-order equivalent to a nonlinear system, Ljung (2001), but the results can be surprising as seen from the following example (Enqvist, 2005):

Example 3. Consider the static and (slightly) nonlinear system

$$y(t) = u(t) + 0.01u^3(t) \tag{19}$$

For a certain (non-Gaussian and bounded) input, its linear second order equivalent is dynamic with a Bode plot as shown in Figure 2. It has a very high gain for low frequencies, and is very different from the Bode plot obtained by just ignoring the small nonlinear term. It is this linear model with the strange low frequency gain that an output error identification method will produce for data from (19).

Such investigations of nonlinear systems that are "perturbations" of linear ones are also carried out by Schoukens, Pintelon and coworkers, e.g. Schoukens et al. (2003)

Nonlinear Systems – Nonlinear Models. The most challenging problem is when we would like to approximate a nonlinear system with a simpler nonlinear model. For effective identification of nonlinear models, this is a topic which must be understood. There is a quite extensive

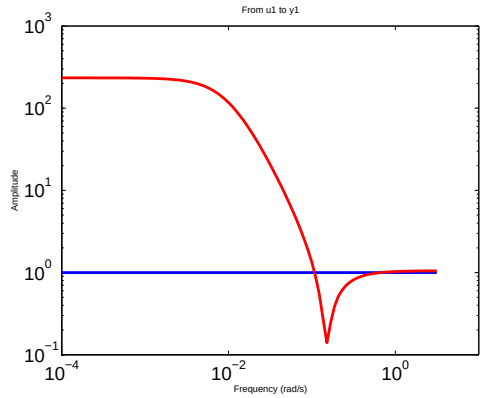


Fig. 2. Amplitude Bode plot of the second order linear equivalent of (19). Straight line: Bode plot of the linear system obtained by ignoring the small nonlinear term.

literature on this problem, but this is not the place to provide a survey of that. Let it suffice to note that among the approaches we see (1) linearization followed by reduction of the linear model, with its states fed back into the nonlinear model, (2) mimicking the balanced realization thinking in terms of contributions to observability and controllability, Scherpen and Gray (2002), and (3) various nonlinear Galerkin methods (truncations of function expansions).

There also exist some MATLAB packages for nonlinear model reduction, e.g. Sun and Hahn (2006).

4.4 Large Databases and Data Mining

We have just got used to the GigaByte and TeraByte world and the PetaByte world is around the corner. The impact on identification and model building cannot be overemphasized. It is now (or soon will be) possible to have a major industrial plant's entire process data recordings over decades available in one database. That in itself constitutes an (implicit) model of the plant, but it is a formidable task to condense it to useful formats. The tools of data mining (Section 3.7) need to be adapted and adopted to the model building and prediction problems of the System Identification community. Some first steps have been taken by Just-in-Time-models, Cybenko (1996) and the Model-on-Demand concept, Roll et al. (2005).

Bayesian Networks. Bayesian (or belief) networks are probabilistic graphical models that describe dependences between variables and they are used in a wide variety of applications, see e.g. Jensen (2001). They have not really been used in System Identification applications, but could be a very useful tool to infer and describe signal flows in a dynamical system, for example to find suitable model structures. By data mining in process databases, relevant dependencies between variables of interest could be found.

Sensor Networks. (Wireless) Sensor networks, e.g. Chong and Kumar (2003), is a rapidly evolving technology to collect information with (many) spatially distributed, autonomous devices. This has an interesting potential for industrial monitoring purposes and adds to the richness of information for model development.

5. NONLINEAR MODEL STRUCTURES: A PALETTE OF GREY SHADES

As mentioned in the previous section, identification of nonlinear models is a very active area. The goal can be said to find model structures that obey the conclusion of Section 2.3: to find descriptions that are flexible enough to cover many relevant nonlinear phenomena, at the same time as they allow inclusion of physical insight in order not to be too flexible. This has lead to a large, and sometimes confusing amount of approaches, and it is not easy to give a coherent description of the current status. Part of the problem is the negative definition: it has been commented that this area is as huge as "non-elephant zoology" (quote attributed to mathematician/physicist Stan Ulam). In this section we give a brief account of the dominating concepts. It is customary in estimation, as remarked in Section 2, to distinguish between grey-box models that are parameterizations based on physical insights, and black-box models, that are just flexible function surfaces. To bring some kind of order into nonlinear model structures we need to invoke a whole palette of grey shades from white to black.

5.1 White Models

White box models are the results of diligent and extensive physical modeling from first principles. This approach consists of writing down all known relationships between relevant variables and using software support to organize them suitably. Similarly, libraries of standard components and subsystems are frequently used.

Physical Modeling and DAEs Modern object-oriented modeling tools, like MODELICA, (Fritzson, 2003), do not necessarily deliver the resulting model in state space form, but as a collection of differential algebraic equations (DAE):

$$F_k(\xi(t), \dot{\xi}(t), z(t), e(t)), \quad k = 1, \ldots, K \tag{20}$$

Here z are measured signals, being inputs and outputs, but not necessarily distinguished as such, e are unmeasured disturbance signals, possibly modeled as stochastic processes, and ξ are so called internal variables that are used to describe the dynamic relationships.

5.2 Off-white Models

Models with lightest shade of grey are obtained when white-box models (20) contain some parameters that have unknown or uncertain numerical values.

The nonlinear identification problem is to estimate such parameters from the measured $z(t)$. In general, this is a difficult problem, that has not yet been treated in full generality. A good reference for a deterministic setting is Schittkowski (2002).

State-space Models If the model equations can be transformed into state space form

$$\dot{x}(t) = f(x(t), u(t), \theta) \tag{21a}$$
$$y(t) = h(x(t), u(t), \theta) + e(t) \tag{21b}$$

where e is white noise, a formal treatment is possible: For each parameter θ this defines a simulated (predicted) output $\hat{y}(t|\theta)$ which is a parameterized function

$$\hat{y}(t|\theta) = g(Z_e^{t-1}, \theta)$$

in somewhat implicit form. Minimizing a criterion like (5a) will then actually be the Maximum Likelihood method. This really requires e to be white measurement noise. Some more sophisticated noise modeling is possible, usually involving *ad hoc* nonlinear observers.

The approach is conceptually simple, but could be very demanding in practice, since the minimization problem will take substantial effort and the criterion may have several local minima.

A recent approach using the EM method, (Dempster et al., 1977) for the case where f and h in (21) are affine in θ is described in Schön et al. (2006). Particle filter techniques to deal with Maximum Likelihood methods to identify nonlinear systems are described in Andrieu et al. (2004).

5.3 Smoke-Grey Models

Semi-physical Modeling: By *semi-physical modeling* we mean finding nonlinear transformations of the measured data, so that the transformed data stand a better chance to describe the system in a linear relationship. The basic rule for this process (to ensure its leisurely aspect) is that only high-school physics should be required and the work must take no more than 10 minutes.

To give a trivial example, consider a process where water is heated by an immersion heater. The input is the voltage applied to the heater, and the output is the temperature of the water. Any attempt to build a linear model from voltage to temperature will fail. A moment's reflection (obeying the rules of semi-physical modeling) tells us that it is the power of the heater that is the driving stimulus for the temperature: thus let the squared voltage be the input to a linear model generating water temperature at the output. Despite the trivial nature of this example, it is good to keep as a template for data preprocessing. Many identification attempts have failed due to lack of adequate semi-physical modeling. See, e.g., Ljung (1999), Examples 5.1 and pages 533 - 536 for more examples of this kind.

5.4 Steel-Grey Models

Composite Local Models: Nonlinear systems are often handled by linearization around a working point.

The idea behind *composite local models* is to deal with the nonlinearities by developing local models, which are good approximations in different neighborhoods, and then compose a global model from these. Often, the local models are linear, so a common name for composite models is also *local linear models*. See, e.g. Johansen and Foss (1995), and Murray-Smith and Johansen (1997).

Let the partitioning into neighborhoods be based on a measured working point variable denoted by $\rho(t)$ (sometimes called *regime variable*). Let the regime variable be partitioned into d values ρ_k, $k = 1, \ldots, d$, and let the neighborhoods around ρ_k be defined by weighting functions $w_k(\rho)$, $k = 1, \ldots, d$. The partitioning into neighborhoods may not be known *a priori* and then a vector η may be formed from (unknown) parameters that describe the partitioning, which may be overlapping or selecting. Then the weigthing functions will depend on this parameter: $w_k(\rho, \eta)$. This means that the predicted output will be

$$\hat{y}(t|\theta, \eta) = \sum_{k=1}^{d} w_k(\rho(t), \eta)\hat{y}^{(k)}(t|\theta^{(k)})$$

where $\rho(t)$ is the known current value of the regime variable. The prediction $\hat{y}^{(k)}(t|\theta^{(k)})$ is the local model corresponding to ρ_k. This prediction depends on some parameters that are associated with the k :th local model, which we denote by $\theta^{(k)}$. (The vector θ will contain the parameters of all local models.) If this model is linear in the parameters, $\hat{y}^{(k)}(t) = \varphi^T(t)\theta^{(k)}$, we obtain

$$\hat{y}(t|\theta, \eta) = \sum_{k=1}^{d} w_k(\rho(t), \eta)\varphi^T(t)\theta^{(k)} \tag{22}$$

which for fixed η is a linear regression, since the regime variable $\rho(t)$ and the regression vector $\varphi(t)$ are measured and known.

LPV Models: So-called *Linear Parameter Varying (LPV)* models are closely related to composite local models. In state space form they are described by:

$$\dot{x}(t) = A(\rho(t))x(t) + B(\rho(t))u(t)$$
$$y(t) = C(\rho(t))x(t) + D(\rho(t))u(t)$$

where the *exogenous* or regime parameter $\rho(t)$ is measured during the operation of the system. Identification of such models has been the subject of recent interest. See, e.g., Lee and Poolla (1999) and Bamieh and Giarré (2002).

5.5 Slate-Grey Models

Hybrid Models: The model (22) is also an example of a *hybrid* model. It is piecewise linear (or affine), and switches between different modes as the "state" $\varphi(t)$ varies over the partition. The regime variable ρ is then a known function of φ. If the partition is given, so that η is known, the estimation problem is simple: It is a linear regression. However, if the partition has to be estimated too, the problem is considerably more difficult, due to the discrete/logical nature of the influence of η. Methods based on mixed integer and linear (or quadratic) programming are described in Roll et al. (2004). See also Bemporad et al. (2005) for another approach.

Block-oriented Models. A very useful idea is to build up structures from simple building blocks. This could correspond both to physical insights and as a means for generating flexible structures.

There are two basic building blocks for block-oriented models: linear dynamic system and nonlinear static transformation. These can be combined in a number of ways. Some combinations are known under well-known names, see Figure 3. Recently, such variations of structures have been found to be useful in several contexts, see Hsu et al. (2006) and Schoukens et al. (2003).

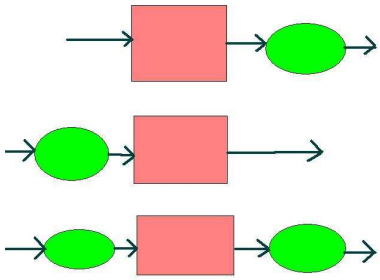


Fig. 3. Typical block oriented models, where squares are linear dynamic systems and ovals are static nonlinearities. Above: A Wiener model. Middle: A Hammerstein model, Below: A Hammerstein-Wiener model.

Remark: The actual shade of grey in slate-grey models may be in the eye of the beholder: For example block-oriented connections may correspond to physical phenomena. The Wiener model is a linear system followed by a nonlinear sensor and the Hammerstein model has a nonlinear actuator. Both these cases are common in practice, and then a block oriented model is rather smoke-grey. But one may also note that the Wiener model, if allowed to have multiple linear outputs, becomes a universal approximator to a wide class of nonlinear systems, cf. Boyd and Chua (1985), so it could as well be viewed as a black-box model. The same is true for hybrid models.

5.6 Black Models

Basis Function Expansion: In a black-box setting, the idea is to parameterize the function $g(x, \theta)$ in a flexible way, so that it can well approximate any feasible true functions $g_0(x)$. A typical choice is to use a function expansion

$$g(x, \theta) = \sum_{k=1}^{m} \alpha_k g_k(x) \tag{23a}$$

with some basis functions g_k.

It turns out that a powerful choice of basis functions is to let them be generated from one and the same "mother function" $\kappa(x)$ and scale and translate it according to

$$g_k(x) = \kappa(\beta_k(x - \gamma_k)) \tag{23b}$$

(here written as if x is a scalar.) The basis functions are thus characterized by the scale (dilation) parameters β_k and the location (translation) parameters γ_k. Typical choices of $\kappa(\cdot)$ are the sigmoid or the Gaussian bell.

When x is a vector, the argument can be interpreted as a scalar product with a vector β_k which then also determines suitable projections in the regressor space. Another possibility is to interpret the scaling as ellipsoidal symmetric.

The resulting structure (23) is very flexible and very much used. Depending on how the particular options are chosen, this contains, for example, radial basis neural networks, one-hidden-layer sigmoidal neural networks, neuro-fuzzy models, wavenets, least-squares support vector machines etc. See e.g Ljung (1999), Chapter 5.

6. SOME CENTRAL PROBLEMS FOR INDUSTRIAL USE

System Identification is an area of control where the gap between theory and practice is not very pronounced. More or less sophisticated identification methods are in daily use in practical applications in companies. Identification software of different kinds have a wide circulation in the industrial world. Still, there are a number of issues for industrial use that have not yet been addressed in a satisfactory way:

6.1 Automatically Polishing Data and Finding Informative Portions

Sometimes a carefully designed identification experiment is carried out in industry in order to build a model. However, more often one has to rely upon process records from normal production. In fact, the problem is often that one has too much data stored from previous production (cf. Section 4.4). Reasons why these are not used include that they contain missing data and outliers, that they are not informative (nothing has happened for a long time) or other deficiencies. What industry demands are automatic procedures for scanning large data records for segments that are informative with respect to a specified issue, polishing the data ("peak shaving") and marking and possibly reconstructing missing data. That would be a very useful step toward effective data mining for industrial data.

6.2 An Efficient Integration of Modeling and Parameter Estimation

Models and simulation are playing a more and more important role in industrial product and process development. Modeling and simulation tools like SIMULINK, DYMOLA, NI-MATRIXX, MODELICA, etc are ubiquitous for engineering work. Mostly, these models are derived from first principles, physical insight and sometimes they are black-box models obtained by system identification. The System Identification framework certainly offers grey-box techniques for mixing physical insights with information from measured data. But for these to be more used, they should be integrated in the engineer's daily modeling tool. An environment based on modern modeling from first principles, allowing Differential Algebraic Equations and model libraries, such as MODELICA, should be integrated with efficient parameter fitting to observed signals, and serious statistical model validation techniques. A study of how DAE modeling that includes stochastic disturbances can be adapted to system identification is given in Gerdin (2006).

6.3 Taking Care of Structural Information

For an engineer in industry, often certain structural information about the plant are known, like "the input flow rate cannot affect the temperature at stage 3" or "these two process components are identical and connected in cascade" etc. Modern control theory encourages us to take a multivariate view of the process and treat multiple inputs and multiple outputs simultaneously. Simple process

insights of the sort just mentioned may not be so easy to incorporate in this view. True, we can develop grey-box models to account for structural information, but it may seem as overkill if the purpose is just to describe simple, logical information flow constraints. There is a need to take care of this problem and provide simpler tools for handling structural information. A study of this kind is Wahlberg et al. (2008).

6.4 Simple Models for Control Tuning and Performance Monitoring

A major concern in industrial applications is monitoring. This includes both failure detection and predictive maintenance. A specific aspect is to monitor the performance of the control loops: Is it time to retune the controller or are there indications of valve stiction? Such questions can be answered by estimating simple models. Likewise, retuning of PID controllers can be made in terms of low order models that capture the essential dynamics in the relevant frequency band(s).

7. CONCLUDING REMARKS

I have tried to sketch a broad and subjective picture of where System Identification stands, by itself and among its neighbors. A main message has been that much more interaction between the different "communities around the core" would be very valuable. I have pointed to some open problem areas both regarding theory and industrial practice where progress would mean big steps forward in terms of understanding and usefulness of our field.

Let me end with a personal IFAC reflection. When I was IFAC Vice President and chairman of the Technical Board (1987–1993) there was a discussion if it was not time to abolish the long running (since 1967) Symposium series on System Identification, since the topic had lost its luster. In the end we did not do so. I hope I have convinced the readers that that was a wise decision.

8. ACKNOWLEDGMENTS

In the preparation of this paper, I have benefited from the discussions with my friends in our VUB micro-symposium group: Michel Gevers, Håkan Hjalmarsson, Paul van den Hof, Bart De Moor, Rik Pintelon and Johan Schoukens. Michel also provided most important, detailed comments on the manuscipts. Manfred Deistler gave me valuable input on econometrics, and my former students Martin Enqvist, Jacob Roll and Thomas Schön provided important feedback on the manuscript. Ulla Salaneck invented the palette and gave me valuable comments on the text.

REFERENCES

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19: 716–723, 1974.

H. Akaike. Canonical correlation. In R. Mehra and D. Lainiotis, editors, *Advances in System Identification*. Academic Press, New York, 1976.

T. W. Anderson. *The Statistical Analysis of Time Series*. Wiley, New York, 1971.

C. Andrieu, A. Doucet, S. S. Singh, and V. B. Tadić. Particle methods for change detection, system identification and control. *Proceeding of IEEE*, 92(3):423–438, 2004.

K. J. Åström and T. Bohlin. Numerical identification of linear dynamic systems from normal operating records. In *IFAC Symposium on Self-Adaptive Systems*, Teddington, England, 1965.

B. Bamieh and L. Giarré. Identification of linear parameter varying models. *Int. Journal of Robust and Nonlinear Control*, 12:841–853, 2002.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4): 1497–1537, 2005.

P. L. Bartlett, M. Jordan, and J. D. McAuliffe. Convexity, classification and risk bounds. *J. American Statist. Assoc.*, 101(473):138–156, 2006.

A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino. A bounded-error approach to piecewise affine system identification. *IEEE Transactions on Automatic Control*, 50 (10):1567–1580, October 2005.

S. Boyd and L. O. Chua. Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Transactions on Circuits and Systems*, CAS-32(11):1150–1161, 1985.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, England, 2004.

C. Y. Chong and S. P. Kumar. Sensor networks. evolution, opportunities and challenges. *Proc. IEEE*, August 2003.

H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, N.J., 1946.

P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31:377–403, 1979.

G. Cybenko. Just-in-time learning and estimation. In S. Bittanti and G. Picci, editors, *Identification, Adaptation, Learning*, NATO ASI Series, pages 423 –434, Berlin, 1996. Springer.

M. Deistler. System identification and time series analysis: Past, present and future. In *Stochastic Theory and Control: Festschrift for Tyrone Duncan*, pages 97 – 108. Springer, Kansas, USA, 2002.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithms. *J. Royal Statistical Society, ser. B*, 39(1):1–38, 1977.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of variance of United Kingdom inflation. *Econometrica*, 50:987–1008, 1982.

R. F. Engle and C. W. J. Granger. Co-integration and error-correction: Representation, estimation and testing. *Econometrica*, 55:251–276, 1987.

M. Enqvist. *Linear Models of Nonlinear Systems*. PhD thesis, Linköping University, Sweden, Dec 2005. Linköping Studies in Science and Technology. Dissertation No 985.

R. Frisch. Statistical confluence analysis by means of complete regression systems. Technical Report 5, Economics Institute, University of Oslo, Oslo, Norway, 1934.

P. Fritzson. *Principles of Object-oriented Modeling and Simulation with Modelica*. Wiley-IEEE Press, 2003.

M. Gerdin. *Identification and Estimation for Models Described by Differential-Algebraic Equations*. Linköping studies in science and technology. dissertations. no. 1046, Linköping university, SE-581 83 Linköping, Sweden, November 2006.

M. Gevers. A personal view of the development of system identification. *IEEE Control Systems Magazine*, 26(6): 93–105, 2006.

M. Gevers. Towards a joint design of identification and control? In H L Trentelman and J C Willems, editors, *Essays on control: Perspectives in the theory and its applications*, ECC '93 Groningen, 1993.

G. C. Goodwin, M. Gevers, and B. Ninness. Quantifying the error in estimated transfer functions with application to model order selection. *IEEE Trans. Automatic Control*, 37(7):913–929, 1992.

E. J. Hannan. *Multiple Time Series*. Wiley, New York, 1970.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

S. Haykin. *Neural Networks*. Prentice Hall, Upper Saddle River, NJ, 2nd edition, 1999.

H. Hjalmarsson. From experiment design to closed loop control. *Automatica*, 41(3):393–438, 2005.

B. L. Ho and R. E. Kalman. Effective constuction of linear state-variable models from input/output functions. *Regelungstechnik*, 14(12):545–548, 1966.

K. Hsu, T. Vincent, and K. Poolla. A kernel based approach to structured nonlinear system identification part I: Algorithms, part II: Convergence and consistency. In *Proc. IFAC Symposium on System Identification*, Newcastle, Australia, March 2006.

F. V. Jensen. *Bayesian Networks and Decision Graphs*. Springer, 2001.

W. S. Jevons. *Investigations in Currency and Finance*. MacMillan, London, 1884.

T. A. Johansen and B. A. Foss. Identification of nonlinear-system structure and parameters using regime decomposition. *Automatica*, 31(2):321–326, 1995.

V. Klein and E. A. Morelli. *Aircraft System Identification: Theory and Practice*. AIAA Education Series, Renston, VA, 2006.

T. Kohonen. *Self-Organization and Associative Memory*. Springer, Berlin, 1984.

T. C. Koopmans, H. Rubin, and R. B.Leipnik. Measuring the equation systems of dynamic economics. In T. C. Koopmans, editor, *Statistical Inference in Dynamic Economic Models*, volume 10 of *Cowles Commission Monograph*, chapter II. John Wiley and Sons, New York, 1950.

W. E. Larimore. System identification, reduced order filtering and modelling via canonical variate analysis. In *Proc. 1983 American Control Conference*, San Francisco, 1983.

L. Lee and K. Poolla. Identification of linear parameter-varying systems using non-linear programming. *ASME Journal of Dynamic Systems, Measurement and Control*, 121:71–78, 1999.

L. Ljung. Estimating linear time invariant models of non-linear time-varying systems. *European Journal of Control*, 7(2-3):203–219, Sept 2001. Semi-plenary presentation at the European Control Conference, Sept 2001.

L. Ljung. *System Identification - Theory for the User*. Prentice-Hall, Upper Saddle River, N.J., 2nd edition, 1999.

L. Ljung. *System Identification Toolbox for use with* MATLAB. *Version 7*. The MathWorks, Inc, Natick, MA, 7th edition, 2007.

L. Ljung and T. Glad. On global identifiability of arbitrary model parameterizations. *Automatica*, 30(2):pp 265–276, Feb 1994.

L. Ljung and A. Vicino, editors. *IEEE Trans. Automatic Control: Special Issue on Identification*, volume AC-50, Oct 2005.

J. F. MacGregor. Data-based methods for process analysis, monitoring and control. In P. van den Hof, editor, *Proc. of the IFAC Conference on System Identification, SYSID'03*, Rotterdam, The Netherlands, August 2003.

H. B. Mann and A. Wald. On the statistical treatment of linear stochastic difference equations. *Econometrica*, 11:173–220, 1943.

B. C. Moore. Principal component analysis in linear systems: Controllability, observability and model reduction. *IEEE Trans. Automatic Control*, AC-26:17–32, Feb 1981.

R. Murray-Smith and T. A. Johansen, editors. *Multiple Model Approaches to Modeling and Control*. Taylor and Francis, London, 1997.

N. Nilsson. *Learning Machines*. McGraw-Hill, New York, 1965.

H. Ohlsson, J. Roll, and L. Ljung. Regression with manifold-valued data. In *IEEE Conference on Decision and Control*, Cancun, Mexico, 2008. Submitted.

P. A. Parrilo and L. Ljung. Initialization of physical parameter estimates. In P. van der Hof, B. Wahlberg, and S. Weiland, editors, *Proc. 13th IFAC Symposium on System Identification*, pages 1524 – 1529, Rotterdam, The Netherlands, Aug 2003.

P. A. Parrilo and B. Sturmfels. Minimizing polynomial functions. In S. Basu and L. González-Vega, editors, *Algorithmic and Quantitative Real Algebraic Geometry*, volume 60 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. 2003. Preprint available from `arXiv:math.OC/0103170`.

J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

C. E. Rasmussen and C.K. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

J. Roll, A. Bemporad, and L. Ljung. Identification of piecewise affine systems via mixed-integer programming. *Automatica*, 40(1):37–50, Jan 2004.

J. Roll, A. Nazin, and L. Ljung. Non-linear system identification via direct weight optimization. *Automatica*, 41 (3):475–490, Mar 2005.

F. Rosenblatt. *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Spartan Books, 1962.

S. T. Roweis and L. K. Saul. Nonlinear dimension reduction by locally linear embedding. *Science*, 290: 2323–2326, December 2000.

J. M. A. Scherpen and W. S. Gray. Nonlinear Hilbert adjoints: Properties and applications to Hankel singular values. *Nonlinear Analysis: Theory, Methods, Applications*, 51(2):883–901, 2002.

K. Schittkowski. *Numerical Data Fitting in Dynamical Systems*. Kluwer Academic Publishers, Dordrecht, 2002.

B. Schölkopf, A. Smola, and K. R. Muller. Kernel principal component analysis. In B. Scholkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, pages 327–352. MIT Press, Cambridge, MA, 1999.

T. B. Schön, A. Wills, and B. Ninness. Maximum likelihood nonlinear system estimation. In *Proceedings of the 14th IFAC Symposium on System Identification*, Newcastle, Australia, 2006.

J. Schoukens, J. G. Nemeth, P. Crama, Y. Rolain, and R. Pintelon. Fast approximate identification of nonlinear systems. *Automatica*, 39(7):1267–1274, 2003. July.

K. C. Sou, A. Megretski, and L. Daniel. A quasi-convex optimization approach to parameterized model order reduction. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 27(3):456–469, 2008.

S. Sun and J. Hahn. *Model Reduction Routines for MATLAB*. Chemical Engineering, Texas A&M University, College Station, TX, 2006. Download from `cheweb.tamu.edu/orgs/groups/Hahn/Model_Reduction`.

J. A. K. Suykens, T. van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

P. N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.

P. Van Overschee and B. DeMoor. *Subspace Identification of Linear Systems: Theory, Implementation, Applications*. Kluwer Academic Publishers, 1996.

V. N. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, 1982.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

G. Wahba. Support vector machines, reproducing kernel Hilbert spaces, and the randomized GACV. In B. Scholkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, pages 69–88. MIT Press, Cambridge, MA, 1999.

B. Wahlberg, H. Hjalmarsson, and J. Mårtensson. On identification of cascade systems. In *Proc. 17th IFAC World Congress*, Seoul, South Korea, July 2008.

H. Wold. *A Study in the Analysis of Stationary Time Series*. Almqvist and Wiksell, Stockholm, 1938.

S. Wold, A. Ruhe, H. Wold, and W.J. Dunn III. The collinearity problem in linear regression, the partial least squares(PLS) approach to generalized inverses. *SIAM J Sci. Stat. Computs.*, 5(3):735–743, 1984.

G.U. Yule. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Phil. Trans. Royal Soc. London*, A 226:267:98, 1927.

L. A. Zadeh. On the identification problem. *IRE Transactions on Circuit Theory*, 3:277–281, 1956.