System Identification: Problems and perspectives

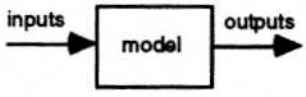
G. De Nicolao
Dipartimento di Informatica e Sistemistica, Università di Pavia,
Via Ferrata 1, 27100 Pavia, Italy,
denicolao@conpro.unipv.it

Abstract

The paper presents a tutorial overview of the main problems arising from (quantitative) system identification. The fundamental issues of identifiability, overparametrization and model comparison are addressed within a probabilistic framework. After having classified models according to their linearity with respect to the unknown parameters, it is explained why system identification is definitely easier for linear-in-parameters models. The paper also illustrates some specific features of dynamic system identification, namely the distinction between output error and equation error models, the need for persistently exciting input signals and the use of prefiltering.

Identification: From data to models

Mathematical models of natural and man-made systems play an essential role in today's science and technology. The applications of models range from simulation and prediction to control and diagnosis in heterogeneous fields such as all branches of engineering, economics, medicine, physiology, geophysics, and many others. It is therefore natural to pose the question where mathematical models come from. If we depict a model as a box containing the mathematical laws that link the inputs (causes) with the outputs (effects), the three main modelling approaches can associated with the "colour" of the box .



White box modelling: The model is derived directly from some first principles by taking into account the connection between the components of the system. Typical examples are found in mechanical and electrical systems where the physical laws ($F = ma$, for instance) can be used to predict the effects given the causes. Rather than white, the box should be termed "transparent", in the sense that we know the internal structure of the system.

Grey box modelling: Sometimes the model obtained by invoking the first principles is incomplete because the value of some parameter is missing. For instance, a planet is subject to the gravitation law but its mass is unknown. In this case, it is necessary to collect experimental data and proceed to a tuning of the unknown parameters until the outputs predicted by the model match the observed data. The internal structure of the box is only partially known (there are grey zones).

Black box modelling: When either the internal structure of the system is unknown or there are no first principles available, the only chance is to collect data and use them to guess the links between inputs and outputs. For instance, this is a common situation in economics and physiology. However, black box modelling is also useful to deal with very complex systems where the white box approach would be time consuming and expensive (an example: modelling the dynamics of an internal combustion engine in order to develop the idle-speed controller).

System identification is concerned with the development and analysis of methods for performing grey and black box modelling (Ljung, 1987), (Söderström and Stoica, 1989), (Haber and Unbehauen, 1990) (Juditsky et al., 1995), (Sjöberg et al., 1995). Differently from white-box modelling that is intimately related to the specific knowledge domain (mechanics, thermodynamics, electromagnetism, ...), system identification covers a number of methodological issues that arise whenever data are processed to obtain a quantitative model.

The present contribution is an attempt at giving a tutorial overview of the main problems arising from quantitative model identification. It is hoped that this could be a stimulus towards a profitable interaction between the quantitative and qualitative viewpoint.

Identification as hypersurface reconstruction

A mathematical model can be thought of as a mapping $f(\cdot)$ that expresses the dependent variables (the outputs y) as a function of the independent ones (the inputs u):

$$y = f(u) \tag{1}$$

To make a simple example, consider the fundamental law of

dynamics $F = ma$, that predicts the acceleration (the effect) as function of the applied force F (the cause) given the mass m. Then, $y = a$, $u = F$, and $f(u) = u/m$.

In general, both y and u can be vectors: $y = [y_1\ y_2\ \dots\ y_p]'$, $u = [u_1\ u_2\ \dots\ u_m]'$. For the sake of simplicity, hereafter it will be assumed that y is scalar ($p = 1$). Then, in the simplest case ($m = 1$), the map (1) corresponds to a curve in the (x,y)-plane a curve. If $m = 2$, then (1) represents a surface. For $m > 2$, (1) is an hypersurface in a suitable space.

When performing identification, only a finite number of noisy samples are available:

$$y(k) = f(u(k)) + v(k),\ k = 1,\ 2,\ \dots,\ N \qquad (2)$$

where the term $v(k)$ accounts for the (unavoidable) measurement errors. If we postulate the existence of a model (1) that explains the data, then the identification process is equivalent to reconstructing ("learning") the hypersurface $f(u)$ from the pairs $(u(k),y(k))$ ("examples", "training set") According to this viewpoint, there are clear connections with function approximation theory (Poggio and Girosi, 1990), learning theory, neural networks (Narendra and Parathasarathy, 1990) and, last but not least, statistics (Beck and Arnold, 1977) whenever the measurement errors are given a probabilistic description.

Linear vs. nonlinear models

At first sight there is no hope to reconstruct the hypersurface $f(u)$ from a finite set of pairs $(u(k),y(k))$ unless some further assumptions are introduced. In this respect it is common to assume that $f(u)$ belongs to a family of functions that share the same structure and differ for the values taken by suitable parameters. In other words, $f(u) = f(u,\theta)$, $\theta = [\theta_1\ \theta_2\ \dots\ \theta_p]'$. To make an example, one may assume that a good approximation for $f(u)$ is a third-order polynomial, i.e.

$$f(u,\theta) = \theta_1 + \theta_2 u + \theta_3 u^2 + \theta_4 u^3 \qquad (3)$$

In this case, the identification problem boils down to estimating the values of the four parameters θ_i.

As will be discussed later on, the identification problem is harder for models that are nonlinear *in parameters*. Note that this has little or nothing to do with the possible linearity of the model with respect to the input u. For instance, $f(u,\theta)$, in (3) is clearly nonlinear with respect to u, but linear with respect to the parameter vector θ. Conversely, the model

$$y = f(u,\theta) = exp(-\theta u) \qquad (4)$$

is nonlinear with respect to θ. Letting $\bar{y} = ln(y)$, model (4) can be reduced to a linear one, i.e.

$$\bar{y} = -\theta u \qquad (5)$$

This kind of trick, however, is not always possible. Moreover, there are statistical reasons that may suggest the use of the nonlinear model (see "The nonlinear case" section).

The probabilistic paradigm

If the data $y(k)$ were error-free, it would be relatively easy to estimate the parameter vector θ. In the most favourable case, it would suffice to take q measurements in order to uniquely determine (through the solution of a system of q equations) the q parameters $\theta_1, \theta_2, \dots, \theta_q$.

Given the unpredictable nature of the measurement error, it is rather natural to model the errors as random variables. They are usually assumed to be zero-mean and incorrelated. If the measurements have not the same precision, it is important to know their variances $\sigma^2_k = Var[v(k)]$ or at least their ratios. In conclusion, the probabilistic paradigm amounts to assuming that the data are generated according to

$$y(k) = f(u(k),\theta^\circ)+v(k),\ k = 1,\ 2,\ \dots,\ N$$

where θ° is the "true" parameter vector and $v(k)$ are errors with some known statistical properties.

The advantage of using a probabilistic formulation, is that the identification problem can be rigorously solved following the guidelines of statistical estimation theory. For instance, if the probability distribution of the errors is known (Gaussian, for instance), one can resort to the maximum likelihood estimator.

An important point of the probabilistic paradigm is that any parameter estimate $\hat{\theta}$, being a function of the measured data $y(k)$, is a random variable itself. This reflect the fact that repeating the identification procedure on a set of newly collected data (with different measurement errors) would lead to a different model. Due to such randomness, no identified model is 100% reliable so that it is indispensable to complement all estimated parameters with their confidence intervals. An estimator is good if the probability distribution function of the estimated parameters is centered around the true parameter values (i.e. $E[\hat{\theta}] = \theta^\circ$) and has small variance.

The linear case

Consider a linear-in-parameters model. Letting $Y = [y(1)\ y(2)\ \dots\ y(N)]'$ and $V = [v(1)\ v(2)\ \dots\ v(N)]'$ be the observations and errors vectors, it is always possible to write

$$Y = \Phi\theta^\circ + V \qquad (6)$$

where $\Phi = \Phi(u(1),u(2),\dots,u(N))$ is a suitable $N \times q$ matrix called sensitivity matrix. Hereafter it is assumed that $N > q$, i.e. there are more data than unknown parameters. It is also assumed that the errors are incorrelated and have all the same variance: $Var[v(k)] = \sigma^2 \neq 0,\ \forall k$.

Normal equations

Under the above assumptions, one can look for the so-called BLUE (best linear unbiased estimator), namely the estimator that has minimum variance among all linear estimators such that $E[\hat{\theta}] = \theta^{\circ}$. If the errors are Gaussianly distributed such an estimator coincides with the maximum likelihood one and, more importantly, has the minimum variance among *all* (linear and nonlinear) unbiased estimators.

Let $\varepsilon = [\varepsilon(1)\ \varepsilon(2)\ \dots\ \varepsilon(N)]' = Y - \Phi\theta$ denote the residuals vector. It turns out (Beck and Arnold, 1977) that the BLUE $\hat{\theta}$ is the vector θ that minimizes the sum of squared residuals

$$SSR(\theta) = (Y-\Phi\theta)'(Y-\Phi\theta) = \sum_{k=1}^{N} \varepsilon(k)^2$$

The values of θ that minimize $SSR(\theta)$ satisfy the so-called normal equations

$$\Phi'\Phi\hat{\theta} = \Phi'Y$$

If Φ has no linearly dependent columns, i.e. $rank(\Phi) = q$ (*identifiability condition*), the unique solution of the normal equations is the linear *LS* (*least squares*) *estimator*

$$\hat{\theta} = (\Phi'\Phi)^{-1}\Phi'Y$$

From a numerical point of view care must be used in the solution of the normal equations. However, this is not a major problem for the user as ready-made specific algorithms are available in commercial software packages.

It is also possible to assess the variance of the estimated parameters. Indeed, the variance of the j-th entry of $\hat{\theta}$ is given by the j-th diagonal entry of the matrix

$$Cov[\hat{\theta}] = (\Phi'\Phi)^{-1}\sigma^2$$

If σ^2 is not known, it can be estimated as

$$\hat{\sigma}^2 = \frac{SSR(\hat{\theta})}{N-q}$$

Identifiability

When the identifiability condition is not satisfied, $\Phi'\Phi$ is not invertible and the normal equations admit an infinite number of solutions so that the identification procedure fails to provide a unique solution. This is a symptom of overparametrization in the sense that some parameters (or linear combinations of them) are superfluous. In practice, rather than being singular, $\Phi'\Phi$ will be close to be singular. Nevertheless, overparametrization can still be detected from the condition number of $\Phi'\Phi$ (a measure of how far a matrix is from being singular) as well as from the SD's (standard deviations) of the estimated parameters (if the SD is more than twice the parameter estimate, there is a reasonable suspect that the parameter is not significantly different from zero)

There are two possible causes of overparametrization. First, it may be that an unnecessarily complex model has been chosen for describing the physical behaviour of the system at hand (example: a first-order electric circuit has been described by a second-order model). This calls for the comparison of different models in order to choose the "right" one (see the "Occam razor" below).

The second cause of overparametrization has to do with inadequate data collection. Even if the model is correct, it may be impossible to uniquely estimate all its parameters because the data do not bring sufficient information (example: in a linear model $y = \theta_1+\theta_2 u$, if $u(k) = \bar{u}$, $\forall k$, it will be impossible to estimate both θ_1 and θ_2). The only possible remedy is an accurate experiment design in order to thoroughly explore all significant regions of the hypersurface $f(u)$. In the dynamic case this has to do with the use of "persistently exciting" input signals (see "The dynamic case" section). When the experiment design cannot be changed, one will have to estimate a lower order model even if it is known that the "true" model is more complex.

Matrioshka models and the Occam razor

When performing black-box modelling, it is a common practice to identify several models which have different complexity and then compare their performances in order to select the "best" one.

To make an example, in order to model the curve $y = f(u)$, one could consider as candidate models all polynomials in u with order ranging from *0* to *3*. Then, there are four possible model structures:

$$M_1:\ y = \theta_1$$
$$M_2:\ y = \theta_1 + \theta_2 u$$
$$M_3:\ y = \theta_1 + \theta_2 u + \theta_3 u^2$$
$$M_4:\ y = \theta_1 + \theta_2 u + \theta_3 u^2 + \theta_4 u^3$$

It is apparent that M_k is a particular case of M_{k+1}, i.e. the models are nested as matrioshka dolls (the technical definition is "hierarchical models"). Note by passing that the problem of comparing matrioshka models can arise also within the context of grey box identification. An example could be the problem of assessing whether the kinetics of a certain drug is better described by a single- or two-compartment model.

Coming back to our example, after having computed the LS estimator in the four cases, the problem of finding the best model has to be addressed. Recalling that *SSR* minimization has been used as a criterion for estimating the parameter vector within a given model structure, one could be tempted to use the same criterion in order to compare model structures. However, this is a nonsense since it is easily seen that, letting SSR_k be the *SSR* of the LS estimate relative to the k-th model, it always holds that $SSR_{k+1} \le SSR_k$ (Beck and Arnold (1977), (Ljung, 1987), (Söderström and Stoica, 1989). In other words, an increase

of the model complexity inevitably leads to smaller residuals. In particular, it is well known that fitting N data by means of an $(N-1)$-th order polynomial yields null residuals.

At this point one could ask what is wrong with a model that interpolates the observed data or, compatibly with the desirable complexity, approximates them as closely as possible. The answer is that such a model would be good only in absence of measurement errors. When noise is present and "too many" parameters are estimated, the identified model uses the extra parameters to learn the noise in the data. This can be practically checked by testing the identified model on a set of "fresh" data (validation data) not used in the identification phase. A model that fits too closely the identification data will be unable to satisfactorily predict the validation data where the noise takes different values.

When the data are abundant, it is convenient to use only a part of them for identification leaving aside a validation set for model selection purposes. Then, after a set of models having different structure has been identified, one will select the model that minimizes the sum of squared residuals SSR^V computed on the validation data.

For a more formal analysis, assume now that there is a "true" model of order q, in the sense that the observed data are generated according to (6). Then, if an unnecessarily complex model of order $q+r$ has been identified (for instance, (6) corresponds to a quadratic function of u and a cubic polynomial has been identified, i.e. $q = 3$, $r = 1$), it can be proven that (Ljung, 1987), (Söderström and Stoica, 1989)

$$E[SSR^V] = \sigma^2(N+q+r) \tag{7}$$

Therefore, the presence of superfluous parameters deteriorates (on the average) the predictive performance of the identified model. Expression (7) can be regarded as the mathematical formulation of the *principle of parsimony* ("do not use additional parameters if they are not necessary"), which is a particular case of the so-called *Occam razor* ("entia non sunt multiplicanda praeter necessitatem").

In some cases there are no sufficient data to form an identification and a validation set. Then, a number of alternative criteria have been proposed for finding the best model within a set of matrioshka models. Among them, the most popular ones are based on the minimization of the following cost functions (Ljung, 1987), (Söderström and Stoica, 1989):

$$FPE = \frac{N+q}{N-q} SSR$$

$$AIC = \frac{2q}{N} + ln(SSR)$$

$$MDL = \frac{ln(N)\, q}{N} + ln(SSR)$$

FPE, which stands for Final Prediction Error, is an estimate of (7). *AIC* (Akaike Information Criterion) and *MDL* (Minimum Description Length) are based on information theoretic principles. Note that all the above criteria penalize the *SSR* but also include a penalty on the order q of the model. For instance $FPE \to \infty$ as $q \to N$. In general, the optimal model order according to the different criteria is not necessarily the same although it does not usually change too much. In particular, *FPE* and *AIC* are roughly equivalent (at least for large N), whereas *MDL* is more parsimonious in the sense that it leads to the choice of models with less parameters.

Other error models

So far, it has been assumed that the measurement errors have the same variance. If, on the contrary, $Var[v(k)] = \sigma^2_k$, just let $\Sigma_V = diag\{\sigma^2_k\}$ denote the covariance matrix of vector V. Then, it is possible to prove (Beck and Arnold, 1977) that the BLUE $\hat{\theta}$ is the minimizer of the weighted sum of squared residuals

$$WSSR(\theta) = \sum_{k=1}^{N} \frac{\varepsilon(k)^2}{\sigma^2_k}$$

The closed-form expression of $\hat{\theta}$ is the WLS (weighted least squares) estimator

$$\hat{\theta} = (\Phi'\Sigma_V^{-1}\Phi)^{-1}\Phi'\Sigma_V^{-1}Y$$

Then, by suitable adjustments, all the results relative to the (unweighted) LS case can be extended to the WLS case.

The nonlinear case

Under the probabilistic paradigm, it is always possible to write

$$Y = \Phi(\theta^\circ) + V \tag{8}$$

where Y, V, θ° have been defined before and $\Phi(\theta^\circ)$ is a suitable $N \times 1$ vector of functions (dependence of Φ upon u is omitted for the sake of notation). Hereafter it is assumed that $N > q$ and the errors $v(k)$ are independently and identically distributed Gaussian variables. All assumptions about $v(k)$ can be easily relaxed except for Gaussianity. Nevertheless, the following identification procedure is likely to provide satisfactory results also in the non-Gaussian case if the size of the errors is not too large.

The maximum likelihood estimator

Let $\varepsilon = [\varepsilon(1)\ \varepsilon(2)\ \ldots\ \varepsilon(N)]' = Y - \Phi(\theta)$ denote the residuals vector. Then, the ML (maximum likelihood) estimator $\hat{\theta}$ can be shown (Beck and Arnold, 1977) to be the vector θ that minimizes

$$SSR(\theta) = (Y-\Phi(\theta))'(Y-\Phi(\theta)) = \sum_{k=1}^{N} \varepsilon(k)^2 \tag{9}$$

Although the analogy with the linear case is apparent, there is a major difference in that *SSR* is no more a quadratic

function of θ. Hence, in general, there exists no closed-form formula for the ML estimate which, rather, must be searched through the numerical solution of the "nonlinear least squares" problem (9).

As already mentioned, some nonlinear models can be made linear by suitably transforming the output variable, see e.g. (4) and (5). In so doing however, also the errors are transformed so that minimizing the SSR for (5) will not yield the same estimate obtained by minimizing the SSR for (4) (obviously, the differences tend to vanish if the errors are small). Nevertheless, the estimate obtained from the linear model (5) can prove very useful as initialization of an iterative algorithm that calculates the ML estimate for (4).

Skiing in the fog

Nonlinear optimization is usually performed by means of iterative schemes of the type

$$\theta^{k+1} = \theta^k + \Delta(\theta^k) \qquad (10)$$

where θ^k denotes the approximation of the parameter vector at the k-th step of the algorithm. In (10), the correction term $\Delta(\theta^k)$ depends on $SSR(\theta^k)$ and possibly also on $dSSR(\theta)/d\theta$ and $d^2SSR(\theta)/d\theta^2$ evaluated at $\theta = \theta^k$. Some classic iterative algorithms (Dennis and Schnabel, 1983), (Fletcher, 1987) are the gradient (not very efficient), Gauss-Newton, and Newton-Raphson (more efficient but computationally expensive). Other possible algorithms include simulated annealing, pattern search methods, and genetic algorithms.

Differently from the linear case where the numerical issues are not a major concern, nonlinear optimization is made nontrivial by the possible presence of multiple local minima where the estimate can get trapped. A good way to appreciate the difficulty of the minimization problem (9) is to consider a model with two parameters. Then, $SSR(\theta_1,\theta_2)$ is a surface whose absolute minimum must be searched for. The iterative algorithm is similar to a person skiing on that surface who aims at reaching the lowest point in the valley. Since at each step the available information is of local type, this is like skiing in the fog.

In view of the possibility of ending trapped in a local minimum, it is clear that the result of the algorithm will be affected by the starting point. Sometimes, the algorithm may even fail to converge if the initialization is not sufficiently close to a local minimum. In order to maximize the probability of finding the absolute minimum it may be convenient to repeat the execution of the algorithm with different initializations but this obviously increases the computational effort. It is worth stressing that the availability of a good initial guess of $\hat{\theta}$ can play an essential role for the successful solution of the problem.

Confidence intervals, identifiability, model comparison.

Once, the ML estimate $\hat{\theta}$ has been computed, a number of issues including confidence intervals, identifiability and model comparison can be addressed by linearization (Beck and Arnold, 1977). Indeed, let $\tilde{Y} = Y - \Phi(\hat{\theta})$, $\tilde{\theta} = \theta - \hat{\theta}$, and $\tilde{\Phi} = d\Phi(\theta)/d\theta$ evaluated at $\theta = \theta^k$. Then, in a neighbourhood of $\theta = \hat{\theta}$,

$$\tilde{Y} \cong \tilde{\Phi}\tilde{\theta} + V$$

so obtaining, at least locally, a linear model to which the results of the previous section can be applied.

About neural networks

In the last decade there has been a growing interest for identification methods based on neural networks (Poggio and Girosi, 1990), (Narendra and parathasarathy, 1990), (Sjöberg et al., 1955). As a matter of fact, they are just particular classes of models which, depending on the type of neural network, can fall into the linear or the nonlinear case.

It is worth pointing out that the neural network community uses a particular jargon. Below, the main terms are reported together with their "translation" in the system identification terminology:

network	↔	model
weights	↔	parameters
train	↔	identify
examples, training set	↔	observations
overtraining	↔	overparametrization

Radial basis function neural networks

The output of an RBF (radial basis function) neural network (Poggio and Girosi, 1990) is just a linear combination of functions with radial symmetry centered in points $\bar{u}^k \in R^m$ called centers

$$y = f(u) = \sum_{k=1}^{q} \theta_k h(\|u - \bar{u}^k\|)$$

A typical choice for $h(\cdot)$ is a Gaussian function:

$$h(r) = e^{-\frac{r^2}{2c^2}}$$

Once the radial function $h(\cdot)$ and its parameters (e.g. c in the above Gaussian function) have been selected, the model parameters are the centers $\bar{u}_k$ and the amplitudes θ_k. Note that, if the Gaussian functions are interpreted as membership functions, there is some analogy with fuzzy models.

In practice it is rather common to assign the location of the centers by means of some heuristic algorithm. Once the centers have been fixed, the model is linear in the parameters θ_k, so that all the considerations made for the linear case can be applied.

Multilayer perceptrons

The output y of a single perceptron (Haykin, 1994) is given by

$$y = f(u) = h\left(\sum_{i=0}^{d} w_i u_i\right) = h(w'u)$$

where $w = [w_0\ w_1\ w_2\ \dots\ w_d]'$ is the weights vector and the so-called "activation function" $h(z)$ is a sigmoidal type nonlinearity, e.g.

$$h(z) = (1+exp(-\beta z))^{-1}$$

Once $h(z)$ has been selected, the free parameters are given by the weights w_i. A multilayer network of perceptron is obtained by connecting the outputs of the perceptron belonging to a given layer with the inputs of the perceptrons of the subsequent layer. The inputs of the perceptrons of the first layer are the model inputs $u_1\ u_2\ \dots\ u_m$, whereas the model outputs $y_1\ y_2\ \dots\ y_p$ are the outputs of the perceptrons of the last layer.

Differently from RBF neural networks, multilayer perceptrons are nonlinear in parameters. In fact, it is well known that their training can be difficult and time consuming due to the presence of local minima.

Identification of dynamic models

A dynamic model is characterized by the fact that the present output does not depend only on the present input but also on its past history. Such models are typically described by means of ordinary differential or difference equations and pose the most challenging identification problems.

In principle, the identification of a dynamic model can always be reduced to the schemes already analyzed in the previous sections. For instance, consider the first-order differential equation

$$\dot{x} = \theta x\ ,\quad x = x_0 \tag{11}$$

where θ is an unknown parameter and y_0 is the (known) initial condition (this problem could typically stem from grey box modelling). The solution is

$$x(t) = x_0\ exp(-\theta t)$$

so that, letting $y = x$, $u = t$, $f(u,\theta) = x_0\ exp(-\theta u)$, the model has been written in the form (1). The observations could be the values $y(k) = x(t_k)$ observed at the times $u(k) = t_k$. Although (11) is a linear differential equation, the model to be identified is nonlinear in θ. In this case it may be useful to refer to the linear model

$$ln(y) = ln(x_0) - \theta u$$

in order to obtain initial guesses for $\hat{\theta}$ (see "The nonlinear case" section).

Although the identification of dynamic models can be reduced to the general framework, they have some specific features that deserve to be discussed separately. To keep the exposition at an acceptable level of complexity it is assumed that both the input u and the output y are scalar signals which are uniformly sampled, so that reference is made to the sampled values $u_k = u(t_k)$, $y_k = y(t_k)$, $k = 1,\dots,N$, where T is the sampling period. Then, a fairly general dynamic model is represented by the difference equation:

$$y_k = f(y_{k-1},\dots,y_{k-n_y},u_{k-1},\dots,u_{k-n_u}) \tag{12}$$

where, as usual, $f(\cdot)$ is a hypersurface to be reconstructed.

Output error vs. equation error models

The most natural way to allow for the presence of the noise v_k is to assume that at each discrete-time instant k a noisy measurement

$$z_k = y_k + v_k \tag{13}$$

is available. Since the error has been added to the output, this is called *OE (output error) model* (Ljung, 1987), (Söderström and Stoica, 1989). Note that y_k is updated according to (12) where no noise term is present.

Alternatively, one can add the noise term within the difference equation (12) yielding

$$y_k = f(y_{k-1},\dots,y_{k-n_y},u_{k-1},\dots,u_{k-n_u}) + v_k \tag{14}$$

which is an *EE (equation error) model* (Ljung, 1987), (Söderström and Stoica, 1989). Now, the value y_k depends (through $y_{k-1},\dots,y_{k-n_y}$) also on the past values of the noise. In many cases this model is less natural but, as seen below, it may be far easier to identify.

As a particular case, $f(\cdot)$ may have a linear structure, i.e.

$$f = a_1 y_{k-1}+\dots+a_{n_y}y_{k-n_y} + b_1 u_{k-1}+\dots+b_{n_u}u_{k-n_u} \tag{15}$$

Then, the model is characterized by the parameter vector

$$\theta = [a_1\ a_2\ \dots\ a_{n_y}\ b_1\ b_2\ \dots\ b_{n_u}]'$$

In the linear case, the EE model is better known as ARX (AutoRegressive eXogenous) model. By analogy, nonlinear EE models are also known as NARX (nonlinear ARX) models.

Simulation vs. prediction

For a given OE model, the residuals are obtained by simulating the model using the inputs u_k and computing the difference between the measures z_k and the calculated output y_k^c. More precisely,

$$y_k^c = f(y_{k-1}^c,\ldots,y_{k-n_y}^c,u_{k-1},\ldots,u_{k-n_u})$$
$$\varepsilon_k = z_k - y_k^c$$

Then, the *SSR* is defined in the usual way and identification can be performed using standard algorithms. In general, even when f is linear as in (15), it turns out that the residuals are not a linear function of θ. Consequently, one has to cope with nonlinear estimation and all related problems (convergence, initialization, local minima, ...). A notable exception is when f in (15) does not depend on past values of y_k but only on $u_{k-1},\ldots,u_{k-n_u}$ Then, the model is linear-in-parameters and can be identified by linear least squares. In such a case, the coefficients $b_1\ b_2 \ldots\ b_{n_u}$ coincide with the impulse response of the system, so that this is called a *FIR (finite impulse response)* model. The main drawback is that a large number n_u of coefficients may be necessary to describe systems with slowly decaying impulse responses and this can cause overparametrization problems.

Conversely, for a given EE model, the residuals are the difference between y_k and the predicted output $\hat{y}_k$ calculated using past values of y_k and u_k. More precisely

$$\hat{y}_k = f(y_{k-1},\ldots,y_{k-n_y},u_{k-1},\ldots,u_{k-n_u}) \qquad (16.a)$$
$$\varepsilon_k = y_k - \hat{y}_k \qquad (16.b)$$

If an ARX model is considered,

$$\varepsilon_k = y_k - a_1 y_{k-1} - \ldots - a_{n_y} y_{k-n_y} - b_1 u_{k-1} - \ldots - b_{n_u} u_{k-n_u}$$

so that the residuals linearly depend on the vector θ. Consequently, ARX models fall within the linear case and can be easily identified by linear least squares. Remarkably, from (16) it is seen that also NARX models are linear-in-parameters provided that f is a linear function of θ (Chen et al., 1990) To make an example, the NARX model

$$y_k = \theta_1 y_{k-1} + \theta_2 y_{k-1}^2 + \theta_3 u_{k-1}^2 + v_k$$

is linear in $\theta_1, \theta_2, \theta_3$.

In conclusion, ARX (and also some NARX) models enjoy the advantage that their parameters can be directly computed by solving the normal equations. Their drawback is that they aim at minimizing the prediction error rather than the simulation one. In general it is easier to predict than simulate. In fact, a simulator calculates the output using only the past inputs, whereas the one step-ahead predictor can take advantage of the knowledge of the past values of the output. It may well happen that a model provides good one-step-ahead prediction but is largely unsatisfactory for what concerns simulation. For some applications such as the design of control systems, it may suffice to have a good predictor but this is not always the case. It is worth noting that the difference between the parameters estimated via the OE and EE approaches vanishes if the measurement errors are close to zero.

Persistent excitation

Assume that the observed data have been generated by a "true" ARX model having the same structure as the model to be identified. Even in this idealized case, identifiability is not guaranteed unless the input u_k is properly chosen. To make a limit example, no parameter can be identified if $u_k = 0$, $\forall k$, since (assuming zero initial conditions) this implies that $y_k = v_k$, i.e. the output is pure noise.

A formal analysis of the identifiability condition for ARX models leads to the definition of *order of persistent excitation* of a signal (Ljung, 1987), (Söderström and Stoica, 1989), which, roughly speaking, is equal to the number of spectral lines in the Fourier spectrum of the signal. For periodic signals n_p is not greater than their period, whereas, if u_k is a sequence of independent random variables, then $n_p = \infty$. A necessary condition for identifiability is that u_k has order of persistent excitation n_p at least equal to the number of b_k parameters in the ARX model.

More in general, the moral is that it is not possible to identify complex systems unless they are properly excited by their inputs which should contain as many harmonics as possible.

Prefiltering

As already mentioned, if the errors v_k are "small", the identification of the EE ARX model (14), (15) is roughly equivalent to identifying the linear OE model (13), (15) but is much more efficient. As a matter of fact, the possibility of approximating an OE model passing through a EE one is not restricted to the case of negligible errors.

Very often, one is interested with the dynamic behaviour of the system at "low frequencies" and the input u_k has a low-pass spectrum. Therefore, although the noise v_k is not negligible, it is likely that at low frequencies it will be dominated by u_k. Then, if both u_k and y_k are low-pass filtered, the resulting filtered signals will be practically independent of the values taken by v_k. Consequently, an ARX model identified using the filtered input and output will provide a good approximation (at low frequencies) of the dynamics of a linear OE error model identified from u_k and y_k (Ljung, 1987), (Söderström and Stoica, 1989).

Other issues

There are a number of topics concerning the identification of dynamic models that have not been treated for the sake of conciseness. Among them, one could mention alternative model structures such as the ARMAX (AutoRegressive Moving Average eXogenous) models, and adaptive identification algorithms (Ljung and Söderström, 1983) which perform on-line adaptation of the model parameters in order to track variations of the system dynamics.

References

Beck, J.V. and K.J. Arnold (1977). *Parameter Estimation in Engineering and Science*. John Wiley and Sons, New York.

Chen, S., S.A. Billings, C.F.N. Cowan and P.M. Grant (1990). Practical identification of NARMAX models using radial basis functions. *Int. J. Control*, 52, 1327-1350.

Dennis, J.E. and R.B. Schnabel (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, N.J..

Fletcher, R. (1987). *Practical Mehtods of Optimization*. 2nd ed., John Wiley and Sons, New York.

Haber, R. and H. Unbehauen (1990). Structure identification of nonlinear dynamic systems - A survey on input/output approaches. *Automatica*, 26, 651-677.

Haykin, S.(1994). *Neural Networks - A comprehensive foundation*. Macmillan, New York.

Juditsky, A., H. Hjalmarsson, A. Benveniste, B. Delyon, L. Ljung, J. Sjöberg and Q. Zhang (1995). Nonlinear black-box models in system identification: Mathematical Foundations. *Automatica*, 31, 1725-1750.

Ljung, L. (1987). *System Identification - Theory for the User*. Prentice-Hall, Englewood Cliffs, N.J..

Ljung, L. and T. Söderström (1983). *Theory and Practice of Recursive Identification*. MIT Press, Cambridge, Mass..

Narendra, K.S. and K. Parathasarathy. Identification and control of dynamical systems using neural networks. *IEEE Trans. Neural Networks*, 1, 4-27, 1990.

Poggio, T. and F. Girosi (1990). Networks for approximation and learning. *IEEE Proc.*, 78, 1481-1497.

Sjöberg, J., Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P.Y. Glorennec, H. Hjalmarsson, and A. Juditsky (1995). Nonlinear black-box modeling in system identification: A unified overview. *Automatica*, 31, 1691-1724.

Söderström, T. and P. Stoica (1989). *System Identification*. Prentice Hall, Cambridge.