PART I

INTRODUCTION TO IDENTIFICATION AND MODELS FOR LINEAR DETERMINISTIC SYSTEMS

1 Introduction

This chapter introduces the basic concepts of identification and provides an overview of the subject. The notion of model, its definition and applications are discussed with suitable examples. The main objective of this chapter is to provide foundations of identification, introduce certain basic terminology, offer a historical overview and describe a systematic procedure for building empirical models from data.

1.1 MOTIVATION

Analysis of process characteristics and inter-variable relationships is of paramount importance in prediction, control, monitoring, design and innovation of process systems. A key step in these analyses is the development of a (mathematical) description of the process under study, known as the *model*. Two contrasting approaches are generally followed for model development: (i) a theoretical (first-principles) approach that is based on fundamental laws of matter and energy, and (ii) an **empirical** approach that is based on analysis of observations (experimental or operating data). The latter approach is a highly practical alternative to the former and widely followed since most processes are too complex to be understood at a fundamental level. Observations potentially carry a wealth of *information* that remains otherwise obscure in a first-principles approach. *The subject of **System Identification** is concerned with the means and techniques for studying a process system through observed / experimental data, primarily for developing a suitable (mathematical) description of that system*.

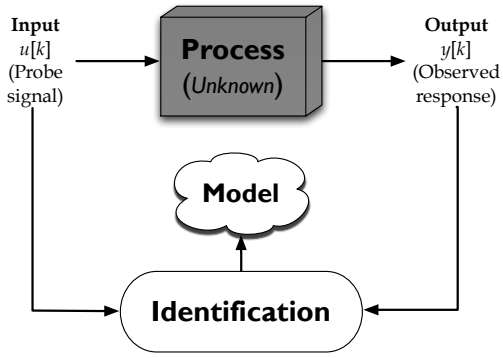

FIGURE 1.1 (**SEE COLOR INSERT**) Identification is the task of using input-output data to build a model: a mathematical abstraction of the process.

Figure 1.1 schematically portrays a typical exercise in System Identification. The prime objective is to develop a *model* from input-output[1] data. The resulting model is said to be *empirical* in contrast to being a *first-principles* model. A simple analogy to identification is that of taking a vehicle (process) on a *test drive*.

[1]Alternative terminologies for input and output signals are also prevalent; for example, *cause* and *effect*, *probe signal* and *response*, *independent* and *dependent*, *explanatory* and *predicted* variables, and so on.

Example 1.1: Vehicle Test Drive

In the test drive of a vehicle, the user develops a mental model of the vehicle's performance by examining its response to changes in various inputs such as fuel supply, pressure on the brake pedal, rotations of steering vehicle and so on. The end use of the model is in decision making; additionally, the exercise can provide insights into how to operate the vehicle in a safe and efficient manner.

Another analogy that is useful to relate to is the interview process.

Example 1.2: Interview

In a recruitment interview, the interviewer attempts to "identify" the candidate by asking questions related to the skills required for the advertised position. The end use of this model is once again in decision making.

The identified model in Figure 1.1 consists of two components (i) a *mathematical description* of the cause-effect relationships, usually known as the *deterministic* model and (ii) a statistical-plus-mathematical description of the uncertainties, known as the *stochastic* model. The latter component comes into play due to the presence of observation errors, process uncertainties and modeling errors (unaccounted dynamics) (further explained in §1.3). On the other hand, the main object of interest to the user is the *deterministic* (input-output) component of the model because it captures the dynamics of the physical process. *An important fact is that the accuracy and precision of the estimated deterministic model depends on the assumptions constituting the stochastic model*. This fact is often undermined or even ignored in several modeling exercises partly because it is not immediately obvious to a beginner in identification as to why such a connection exists. Only a careful study and practice of the subject establishes this point clearly.

The concepts and techniques of identification have largely originated from the domains of statistics, engineering and econometrics. The task of identification, however, appears in almost every walk of life. Shortly we shall come across examples that are illustrative of the versatility of this subject. In order to set up and solve these diverse problems, it is necessary to understand the principles governing identification concepts and techniques, which is the purpose of this text.

What factors motivate the need for identification? The two primary motivating factors are (i) the need for models in process analysis and automation and (ii) the practical limitations of the first-principles approach in developing models. These factors are discussed in that order below.

Incentives in model development

The benefits of developing models are enormous. Figure 1.2 depicts the applications of models to four major branches of process systems engineering, namely, design, estimation (prediction), control and monitoring.

Central to all these applications is the use of models in simulations and predictions, which constitute the key incentives of model development. Simulations, by their computational nature, offer a cost and time-effective, safe alternative to experiments that are usually time-consuming, expensive and marked with safety issues. Significant advances in the fields of computational science and technology have rendered simulations as powerful ways of understanding processes, bringing about innovations and testing new designs and strategies. The enormous benefits gained through the simulation route usually come at the cost of inaccuracies in simulations. In order to minimize these losses, high accuracy requirements are imposed on a model whose end use is in simulation.

On the other hand, *predictions* are on-line exercises where the model is used to forecast the process response for the prevailing operating conditions over a finite time-step horizon. Conceptually predictions are not much different from simulations; however, technically simulation is reserved for

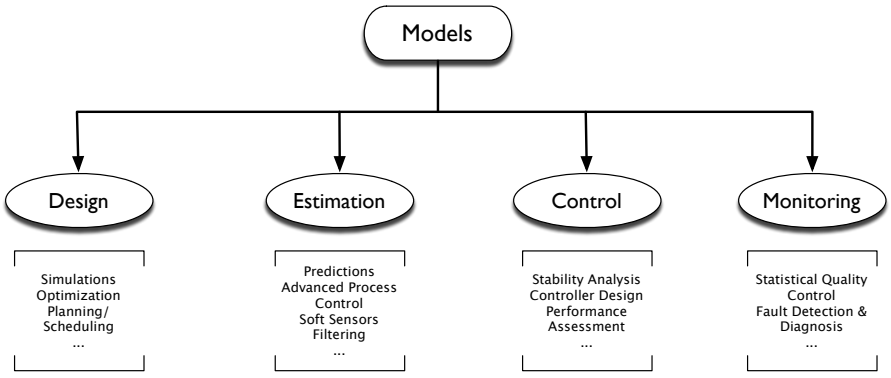

FIGURE 1.2 Applications of models in process systems engineering.

predictions under special conditions - the case of *infinite-step prediction horizon*. Chapter 18 offers technical definitions of predictions, how to compute them from models and explains this distinction in detail. Predictions are vital to design, control, fault detection, testing new schemes, etc. However, the accuracy requirements of the models are lower than in simulation-based applications.

Specific examples below highlight the role of identification in process operations and its *multi-disciplinary* facets.

Example 1.3: Reactor Control

Reactions between two or more chemical species are often accompanied by the release of heat.

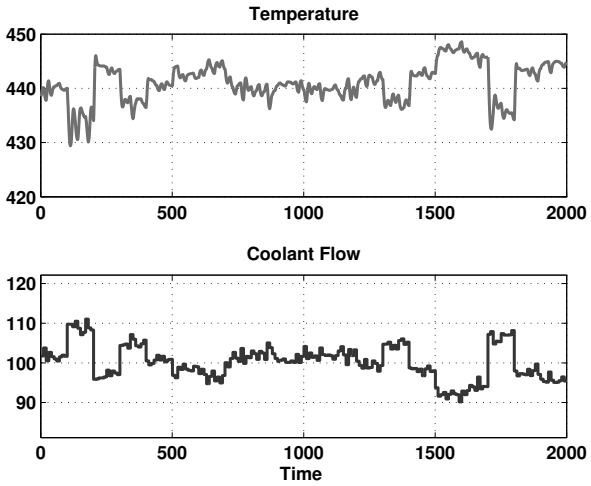

FIGURE 1.3 Temperature response of a reactor to changes in coolant flow rate is used in identification of the reactor system (source: DAISY (Moor et al., 1997)).

For stable and an efficient operation of a reactor, engineers are always confronted with the design of a suitable controller that adjusts the flow of the coolant in the jacket to regulate the reactor temperature. An engineer conducts experimental studies on the reactor by varying the coolant flow and measuring the temperature response, a typical snapshot of which is displayed in Figure 1.3 (data source Moor et al. (1997)). This data is used to identify a model between the manipulated variable and the controlled variable.

Identification is frequently used to build models for monitoring processes.

Example 1.4: Process Monitoring

Monitoring an industrial plant consists of first building a model of the plant under normal operating conditions followed by a projection of the fresh data onto this model online. Industrial processes can be too complex for the development of a first-principles model. The natural recourse is to identify a model relating process variables from routine normal operating data. This is the case of multivariable identification. The challenges are the multivariable nature, non-linearities and importantly the closed-loop conditions.

Models can be used to understand relationships between variables so as to be able to detect errors in fresh data and correct those errors.

Example 1.5: Data Reconciliation

Measurements from sensors are always not reliable in the sense that they are not necessarily consistent with the physics of the process. In a mixing process for example, the flow and temperature measurements may not satisfy the mass and energy balances. Data reconciliation consists of first detecting the presence of significant errors (above tolerable noise levels) and then rectifying the erroneous data. The first step requires a model that is consistent with the physical laws. For several processes, the form of the model equations is not known well enough. Identification comes to the rescue of setting up the simultaneous problem of model estimation and gross error detection from data.

The next example is concerned with the development of a model of an aircraft.

Example 1.6: Flight Control

One of the key problems in flight control is the attitude control, i.e., the control of the aircraft's orientation. Attitude control is carried out by exerting forces on the aircraft in the appropriate direction thereby generating the appropriate moment about its center of gravity. A model that predicts the orientation for a given direction of force is central to the design of this controller. A typical procedure is to apply pilot-generated or computer-generated frequency sweeps (inputs) and record dynamic responses using appropriate sensors. The challenges are the presence of highly coupled and unstable dynamics.

The problem of identification assumes an interesting form in the task of inferential sensing.

Example 1.7: Soft Sensing

There exist several physical variables for which sensing hardware does not exist or it is that they cannot be measured on-line. Examples include composition of product stream in a distillation column or a gasifier, molecular weight of a polymer, fineness of cement and so on. In situations such as these, an attractive alternative is to "soft" sense these variables (primary variables) using measurements of other variables (explanatory variables). The identification problem is that of building a model between the scarce offline (laboratory) measurements of the primary variable and the frequently available on-line measurements of explanatory variables.

System Identification principles are useful in building an approximate simplified empirical model of a first-principles high-order non-linear model.

Example 1.8: Model Approximation

High-order rigorous first-principles models allow process engineers to safely test new designs, control strategies and monitoring schemes. However, it may be often required to work with

a simplified model of a process, e.g., in controller design. These ideas find routine place in process control and aircraft system design. System Identification principles guide the user in designing the appropriate input signal, setting realistic simulation parameters and finally in building the desired model.

The following example presents an identification exercise in the context of weather forecasting.

Example 1.9: Prediction of Relative Humidity

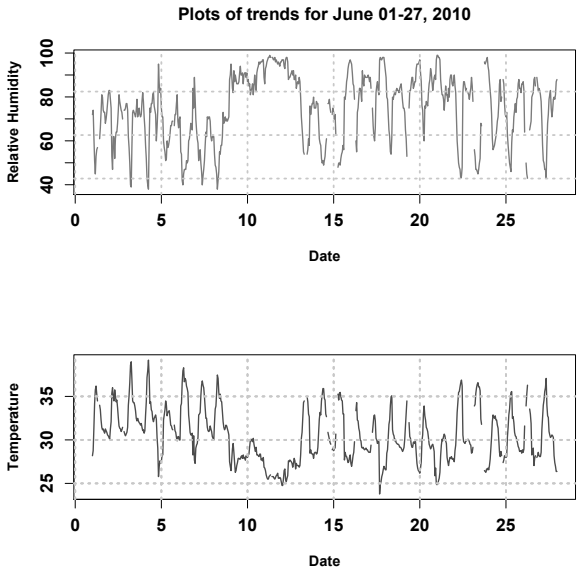

FIGURE 1.4 Plot of relative humidity and temperature at IIT Madras, Chennai, India from June 01 to 27, 2010. Data is missing at random.

In a statistical approach to weather forecasting, the relative humidity (RH) of air can be predicted using historical data. Additionally, past temperature measurements can be used to improve predictions. In the former case, we build what is known as a time-series model, whereas in the latter case we build a model typical of identification. The challenge here is that the user does not have the privilege to provide test inputs, i.e., changes in temperature. Besides, the input itself is a measurement (corrupted with observation errors) rather than a known quantity. Figure 1.4 shows a plot of RH and temperature recorded by an automated weather station (AWS) located at IIT Madras, Chennai, India for the period June 01-27, 2010. The missing segments in the plot correspond to missing data segments.

Modern applications of identification find increasing use in energy systems and biomedical engineering, a rapidly evolving field of engineering in medicine and biology.

Example 1.10: Optimization

A fuel cell system generates electrical energy by electrochemically combusting hydrogen with oxygen. For an efficient operation of a fuel cell system, the operating conditions (e.g., stack temperature and pressure) have to be set at their optimum. A dynamic model of the fuel cell system allows the designer to determine the optimum operating conditions.

Example 1.11: Insulin Delivery Control

A classical biomedical application of engineering principles is the development of an automated insulin delivery system for diabetic patients. To ensure efficient control, the biomedical engineer will first require a mathematical model relating the insulin delivery rate to the glucose levels in the body. Data from several patients are collated to identify a suitable model. One of the challenges is the presence of feedback naturally present in the metabolic pathways.

In recent times, models have been used to understand the causal (directional) relationships between variables.

Example 1.12: Network Reconstruction

In neurosciences, one of the prime interests is to uncover how certain active parts of the brain are connected to each other. Neuroscientists can build the underlying network connectivity by first collecting measurements from select regions (e.g., electroencephalogram (EEG) data) and then building *causal* models between such measurements. The coefficients of the model can be further examined in time- or in frequency-domain to detect the presence of *directed connectivities* between measurements in those regions.

In general, the modeling objectives and the type of model naturally vary with each layer. In process design, for instance, the usual aim is to identify the steady-state relationships between measurements, whereas in control and monitoring the aim is to identify the dynamic behavior of the plant. On the other hand, it may be necessary to develop a rigorous simulator of the process where conducting experiments on the actual plant is often neither safe nor cost- or time-effective. The end use of the model clearly governs the accuracy requirements of the model. As an example, a model used in the feedback control of a process is acceptable even if it is a low-order approximation of the process whereas the model used for optimizing or simulating the same process is acceptable only if it is able to deliver excellent predictions.

From the foregoing discussion, it is evident that the role and impact of identification in process operations and making innovative changes to processes is immense. While this being true, the quality of identified models depend on the level of sophistication in instrumentation and data acquisition. Thus, while identification helps in making innovative changes to systems, these enhancements, in turn, facilitate better ground for identification.

We discuss the second motivating factor for identification, namely, the limitations faced by the first-principles approach in developing models.

Benefits of empirical approaches

Identification offers a powerful and pragmatic alternative to **first-principles modeling**, which uses basic laws of material, momentum and energy balances combined with some constitutive (empirical) relationships. In the early days of control and automation, models were largely developed using the first-principles approach. These models primarily related continuous-time variables. A good understanding of the physics of the process is critical to the development of first-principles models. Most modern processes of interest are complex to the extent that precludes a fundamental approach. The natural recourse has been towards data-driven approaches since they assume minimal prior knowledge and largely depend on input-output observations for developing models. With advances in measurement technology and the dawn of the digital era in the early 1960s, data-based approaches gained large momentum. Rapid developments in estimation theory, advances in computational sciences combined with the benefits of digital technology over analog counterparts strongly nourished data-driven approaches.

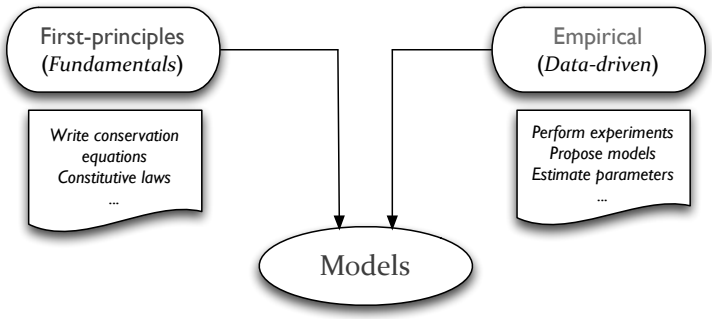

FIGURE 1.5 Two approaches to modeling: first-principles vs. experimental (empirical).

Figure 1.5 offers a simplistic comparison of the first-principles and empirical approaches to modeling. In addition to being natural alternatives, empirical approaches offer several practical benefits, salient among them being (i) the ability to build models with minimal process knowledge and (ii) the flexibility in selecting the model structure and (iii) the convenience of implementing them in a soft form, i.e., in the form of a computer-interpretable code. Not surprisingly, identification is today an integral part of modern industrial control and automation schemes.

There exist several distinguishing and contrasting features of first-principles and empirical models. These are discussed in §3.2.1. At this point of discussion, two important points merit attention. The first point is concerned with the transparency (physical meaning) of the models. First-principles models by their way of construction are *transparent* whereas the empirical models are generally *opaque* with respect to the physics of the process because they are developed largely using mathematical methods rather than physical laws. For this reason, models obtained from identification are termed as *black-box* models. The opacity of empirical models can be reduced by incorporating a priori process knowledge, for example, by imposing known structural constraints on the model, resulting in what are known as *grey-box* models. These ideas are relatively modern and usually treated as advanced topics in identification (§23.7.2 presents some preliminary ideas). The second point is related to the extrapolation capabilities of these models. Data-driven models have, in general, good predictive abilities only over the operating regime spanned by the data whereas first-principles models are superior in this respect.

As with any matured field, a brief study of history of its evolution and developments is beneficial in appreciating the depths and breadths of the subject, treading previously unexplored paths, avoiding accidental re-invention of wheels and in several other ways. With this motive, a historical account of System Identification follows.

1.2 HISTORICAL DEVELOPMENTS

The idea of identifying models from data is perhaps as old as the art of learning through experimentation. In today's era of automation, engineers increasingly turn to data-driven approaches for building approximate dynamic models. Empirical relationships are commonplace in thermodynamics and transport phenomena. There is hardly an application today where data-driven knowledge is not used. The formalization of concepts and techniques of identification as we learn today has evolved primarily through a variety of contributions from engineers and statisticians. Much of the presentation below is inspired from the surveys of Åström and Eykhoff (1971), Gevers (2006), and Ninness (2009a).

Modern System Identification has its roots in the eighteenth and nineteenth century developments of mathematics and probability theory (read Ninness (2009a) for a good account of developments). Some milestone results in this direction include the conditional probability due to Bayes (Bayes,

1763), the method of least squares (LS) due to Gauss and Legendre (Gauss, 1809), emergence of the illustrious Fourier transforms (Fourier, 1822; Proakis and Manolakis, 2005), Schuster's periodogram measure (Schuster, 1897). Of these the least squares method and the ideas therein have possibly had the maximum impact on data-based modeling and parameter estimation. It is expected to do so for many decades to come. Gauss's conception of the LS method itself originated from his proposition to determine planetary orbits from astronomical data instead of using physical laws such as Kepler's laws of motion. This was a classical example of system identification. The LS method is, in fact, regarded as a precursor to and the core engine of several present-age sophisticated estimation methods including the celebrated Kalman filter for state estimation (read Sorenson (1970) for a technical treatment).

Early decades of the twentieth century witnessed a surge of developments largely tilted towards modeling of stochastic processes. Fisher's concept of likelihood and the celebrated maximum likelihood estimation (MLE) method (Fisher, 1912, 1922), contributions by Yule and Walker in modeling auto-regressive processes (Walker, 1931; Yule, 1927) and more importantly, the landmark works of Khinchin, Kolmogorov, Wiener, Cramer and the likes laid solid foundations for modeling random processes in time and frequency (spectral) domains (see Priestley (1981) for a good historical account). Applying these concepts to real data gave birth to the field of *time-series analysis* (TSA, also known as *statistical signal processing*), whose main focus was *prediction of phenomena driven by unknown or unmeasurable causes*. The auto-regressive moving average (ARMA) model was theoretically shown to be capable of representing a wide range of linear stationary processes with the support of the *spectral factorization theorem* and the conception of *white-noise* stochastic process. Gradually TSA matured into a solid field with widespread applications in engineering, sciences and econometrics. A significantly distinct body of work started to emerge with the inclusion of known (deterministic) *eXogenous causes* into ARMA models, culminating into methods for identification such as ARX and ARMAX. These models are essentially *parametrized* difference equation descriptions with the input and the white-noise signal as the forcing functions for the deterministic and stochastic subsystems, respectively (read Box, Jenkins and Reinsel (2008)). The stochastic parts of these models, as remarked earlier, accounted for the effects of unmeasured disturbances and measurement noise. An interesting account of the connections between time-series analysis and system identification is chronicled in Deistler (2002) .

In a world of parallel developments, the field of control and systems theory was being enriched with contributions from leading engineers (particularly during the post Second World War era) resulting in formalization of systems concepts and systemic methods for tuning classical PID (proportional-integral-derivative) algorithms. Control engineers began to use simplified continuous-time *impulse-* and *step-response models* for controller design. The step response was found to be particularly convenient because of its *plant-friendliness* (step changes are easier to introduce than impulses) and the ease with which the three key pieces of information required for control, namely, *time-delay, time-constant and steady-state gain*, could be extracted. Numerous empirical methods to optimally estimate these parameters from the step response were developed. These models remain the backbone of several methods for tuning lower-level industrial control loops even today. Step response models have also been used successfully in several model predictive control (MPC) schemes (reference), known under the name dynamic matrix control (DMC). However, step-type signals are weak in frequency content and not suited to modeling of high-order systems. Consequently they also have limited predictive abilities. Adding to these limitations was the rising need for methods that can use *observed* input-output data rather than merely *experimental* data.

With the dawn of digital era in the 1950s, computer-based control schemes began to be foreseen as platforms for powerful next-generation industrial control strategies. Advances in measurement and instrumentation technology helped in materializing these ideas in practice. The much acclaimed sampling theorem due to Shannon (1948) and Whittaker (1935) opened up doors to the world of digital analysis of continuous-time systems. Engineers were able to introduce custom-designed inputs

with richer excitation instead of a simple step-type signal and also store large volumes of experimental / operating data. Emergence of novel efficient computational algorithms for signal processing tools introduced a paradigm shift in digital signal processing. A landmark invention was the Fast Fourier Transform (FFT) algorithm for spectral analysis of signals due to Cooley and Tukey (1965). The digital systems revolution was aptly supported by a rapid formalization of discrete-time systems theory. Data-driven discrete-time models began to emerge as natural and preferred alternatives. A paradigm shift in control and estimation theory occurred with the arrival of the celebrated Kalman filter (Kalman, 1960; Kalman and Bucy, 1961) which not only provided optimal algorithms for state estimation but also propelled the theory of optimal control. Combined with the new methods of identification, the notion of *approximate* as opposed to *true* models gained increasing emphasis within the control community. This is because optimal controller design based on simplified reduced-order models was computationally lighter. State-space descriptions gradually grew to be the natural choices for joint identification and signal estimation, as well as for multivariable control.

By mid-1960s, two distinct classes of methods precipitated: one class drawn on the techniques of time-series analysis (predominantly using parametrized input-output representations) and another built on the state-space descriptions. The seminal papers by Åström and Bohlin (1965) and Ho and Kalman (1966) can be treated as the foundational works for the two streams of identification methods. With the topic gaining the status of a field of its own, formalizations of definitions and estimations began to emerge. Zadeh (1962) gave a formal definition of identification as "*the determination on the basis of input and output, of a system within a specified class of systems, to which the system under test is equivalent.*" The work by Åström and Bohlin (1965) solved the parameter estimation methods for ARMAX models in the MLE framework, which was then extended to the general family of Box-Jenkins models (Box, Jenkins and Reinsel, 2008). On the other hand, the work of Ho and Kalman (1966) established a method for determination of state-space models from impulse response coefficients.

The field of System Identification experienced a transformation in theory and practice starting in the mid-1970s with the introduction of *prediction-error* (PE) identification methods due to Ljung (1976a,b, 1978). A significant change of approach to empirical modeling came about due to a marked shift in the problem formulation, wherein the restrictive search for *true* model structures was replaced by a broader and practical search for the best *approximate* models. The PE minimization (PEM)[2] approaches, which were devised for estimating *parametrized* models (both input-output and state-space), were shown to contain several well-known methods including the likes of LS and ML estimation. In fact, it is believed to have drawn inspiration from the ML estimation method. Expectantly, the PE methods resulted in almost always non-linear optimization problems (with unique solutions only for special model structures *and* quadratic objective functions).

Ljung (1976a,b) showed that the PEM estimates *asymptotically* (large sample conditions) converge to the true parameters under some mild assumptions on the data generating process (also see Caines (1976)). Establishment of the asymptotic Gaussian distributional properties of these methods under fairly general conditions (Ljung, 1999; Ljung and Caines, 1979) provided the necessary support for their practical use. The study of how error characteristics in parameter estimates translate to errors in transfer functions (frequency response functions) (Ljung, 1985b; Wahlberg and Ljung, 1981) solidified the position of prediction-error methods in the field of control. Expressions for bias (accuracy) and variance (precision) highlighted the *tunable* elements of identification and their impact on the final model quality. A number of other properties and implementation aspects of prediction-error methods were subsequently established, refined and reviewed in the years to follow. Several excellent articles, textbooks and innumerable conferences on identification have since been organized, especially on linear time-invariant (LTI) dynamic systems. Over the years, significant developments occurred in the areas of non-linear, time-varying and other branches of identification

[2]The acronym PEM is also used to abbreviate other related phrases in System Identification, e.g., *prediction-error models*.

as well. The PE minimization methods are today at the heart of many state-of-the-art identification algorithms (for parametrized models). Various other branches of identification oriented towards specific end use of models (e.g., control relevant identification) have also evolved over the years.

A limitation of the prediction-error approach is that it is not equipped to handle the joint problem of identification and signal (state) estimation that arises when the observed variable is different from the physical output of interest, but is related to it. Such situations are not uncommon in process systems. The state-space descriptions are apt choices for these classes of problems. An added advantage of state-space models is that they can represent multi-input, multi-output (MIMO) systems with elegance and simplicity. Two broad classes of state-space models exist (as with the input-output models), namely, the *non-parametrized* (unstructured) class that make minimal assumptions on the process and *parametrized* class that results as a consequence of assuming a certain structure on the state-space matrices. Estimating non-parametrized state-space models with PEM methods involve difficulties that are mitigated only through parametrizations. However, parametrization of multivariable models usually requires prior insights into process characteristics; further, a single scheme of parametrization does not suit all applications. Motivated by these considerations and equipped with the tools of linear algebra, statistics and optimization, algorithms for numerical (non-parametric) state-space identification were sought. The work of Ho and Kalman (1966) marked the beginning of such methods, wherein an elegant method for estimating a state-space model from the impulse-response coefficients using what are known as Hankel matrices was presented. Subsequently, two significant works by Akaike (1974b) and Kung (1978) laid the foundations for a set of novel methods, collectively known as the *subspace identification* (SSID) algorithms for estimation of non-parametric state-space models from measured data (Overschee and Moor, 1996; Qin, 2006).

The principal benefits of subspace methods are that they are *non-iterative* (unlike PEM), and based on efficient numerical methods such as singular-value decomposition (SVD) and QR factorization methods. Further, they contain an implicit implementation of the *numerical* Kalman filter. The classical Kalman filter assumes that a model is readily available whereas the numerical filter presents the optimal estimates *directly* from data through a series of orthogonal projections. Thus, vis-a-vis PEM approaches, subspace methods have an edge with respect to *parametrization*, convergence, numerical efficiency and also model-order reduction (Overschee and Moor, 1996). A highlight of the subspace methods is that it can "automatically" estimate the order of the state-space model unlike the classical non-parametric input-output identification methods. Thus, it requires minimal intervention of the user. Notwithstanding the numerous benefits, subspace methods often come in for two standard criticisms, at least in the early years of development: first, that they generate sub-optimal estimates (unlike PEM) and second, application to closed-loop identification is marked with challenges. While the gravity of the first point has been considerably reduced by a formal interpretation of subspace methods in the PEM framework, the second limitation has been overcome with considerable success in a number of works (Huang, Ding and Qin, 2005; Katayama and Tanaka, 2007; Larimore, 1996; Ljung and McKelvey, 1996; Verhaegen, 1993)

To summarize, there exist two broad classes of algorithms, namely, the *prediction-error* and the *subspace* identification algorithms for developing empirical dynamic models from input-output data. The PE methods find wide usage in estimation of parametric models and can be used in both open- and closed-loop conditions. They possess good large sample properties and result in Gaussian distributed estimates. Subspace identification on the other hand is suited for estimation of non-parametrized state-space models under open-loop and closed-loop conditions. While the PEM estimates are theoretically guaranteed to be optimal, subspace methods do not necessarily guarantee optimality. Since the latter deal with non-parametric models, it has been a recent practice to run the data through SSID methods prior to estimating a parametric model using the PE algorithms.

In closing, special mention should be made of a relatively less used, but powerful, method known as the *instrumental variable* (IV) method. The IV technique is primarily devised to produce *consistent* (convergent) and *efficient* (minimum error) estimates in situations where the LS method fails to

do so. It replaces the regressors in a LS method with what are known as *instruments*. The idea of the IV method was originally devised for identification problems in econometrics where the causes (explanatory variables) are known with error unlike in the classical identification of most engineering problems where the inputs are deterministic (known) variables (read Angrist and Krueger (2001) and Stock and Trebbi (2003) for an analysis of the historical origins). Gradually they were adopted to the engineering arena (Soderstrom and Stoica, 1994), sometimes grouped under the banner of *correlation methods* (Ljung, 1999).

Remarks: The historical account above is by no means exhaustive and has been intentionally restricted to the mainstay of this text, which is the sub-field of dynamic, linear time-invariant systems. There exist other branches of identification that have attracted considerable attention over the last two decades, such as black-box identification of continuous-time / non-linear / time-varying / multiscale systems and grey-box identification, to name a few. An interested reader is referred to articles of the likes of Bohlin (2006), Kerschen et al. (2006), Rao and Unbehauen (2006), Sjöberg et al. (1995), and Tangirala, Mukhopadhyay and Tiwari (2013) and the classical book of Ljung (1999), in addition to the previously referenced survey papers for a broad overview of developments in this field.

Through the foregoing sections it is hoped that the reader has garnered a bird's eye view of System Identification, recognized the key motivating factors and gained an overview of the prime developments and conceptions in this field. In the descriptions above, a number of terms such as parametrization, bias, variance, consistency, etc. have been used. Technical definitions and explanation of these terms are provided later at appropriate points in the text.

From a layman's perspective, identification appears synonymous to the fit of curves to data (curve-fitting). This is indeed true. However, it is also important to recognize that identification is *much more* than merely curve fitting. Identification exercises are challenged primarily by experimental factors (input design), process complexities, and measurement uncertainties and require a sound knowledge of models and their estimation. These challenges and limitations can be dealt deftly only through a holistic, careful and a step-by-step understanding of the theory. *A systematic approach to identification will overcome or in the least minimize the challenges in empirical model-building by a careful design of experiments for identification, judicious choice of models and an appropriate deployment of estimation algorithms.* A fact to remember is that the task of *identification is almost always an iterative exercise*. To rightly interpret the results at each iteration and to incorporate the end-of-iteration learning back into the identification exercise requires a good understanding of how different elements of identification impact the model that is being developed. A methodical approach will not only result in a high-fidelity model but will also guide us on how we should conduct our experiments or make improvements at different stages to improve model quality. The ability to attribute the success or failure of a model at the end of an identification exercise to the right factors can only rest with a learned practitioner.

Aims and objectives of this text

The grand objective of this text is to present a *self-contained learning* material on developing **dynamic empirical discrete-time models**, predominantly, *linear system identification*, aimed at beginners, instructors/researchers and practitioners of this subject. Specifically the objective is to explain and illustrate the following:

i. systematic procedure for identification
ii. different deterministic-plus-stochastic descriptions used in identification of LTI systems
iii. technical concepts such as parametrization, bias, variance, consistency, etc.
iv. classical and modern estimators, namely, method of moments, LS, MLE and Bayesian methods
v. estimation of signal properties such as auto- and cross-correlation, spectral density, coherence

vi. prediction-error and subspace-identification methods
vii. estimation of time-delays and non-parametric models
viii. practical aspects of model development: *how to assess the goodness of models, how to refine an estimated model using statistical and practical guidelines,* etc.
ix. concepts of design of inputs and experiments for identification

The emphasis is on the interpretations and practical aspects of identification with some affordable sacrifice of rigor. The underlying principles are explained with suitable demonstrations and illustrative examples while attention is also paid to the development of key theoretical expressions. It is hoped that a healthy balance of theory and practice is achieved in this process.

An overview of selected advanced topics is provided towards the end of this text. These concepts, although more challenging and complex than their linear counterparts, are still based on the principles of linear system identification.

1.3 SYSTEM IDENTIFICATION

Identification is the exercise of developing a mathematical relationship (model) between the causes (inputs) and the effects (outputs) of a system (process) based on observed or measured data. Stated otherwise, identification establishes a mathematical map between the input and output spaces as determined by the data.

Terminology and notation

The terms *input* and *output* in identification have generic meanings in identification. *Outputs* constitute all those signals that are measured and which one wishes to predict. They are also known as *responses* or *predicted variables*. *Inputs* collectively refer to all variables that are considered to influence the outputs. The input set consists of both that which *can be manipulated* by the user (*probe signals*, and that which *cannot be adjusted but can be measured*. The former subset of signals are frequently referred to as *inputs* and the latter subset as *measured disturbances*. This is the terminology that is followed in the present text.

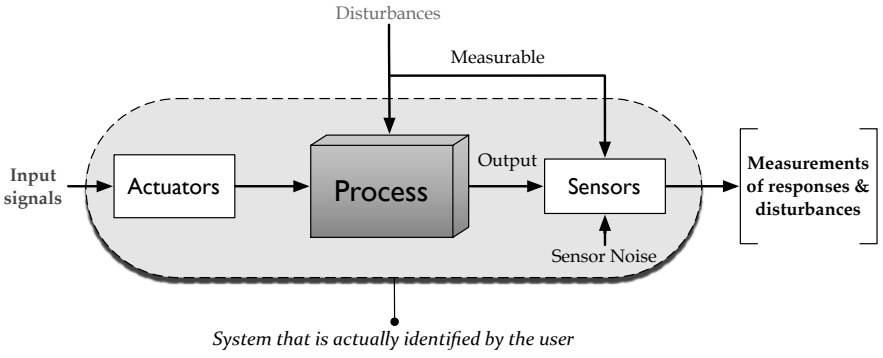

FIGURE 1.6 (**SEE COLOR INSERT**) The system being identified consists of the true process and additional elements.

Figure 1.6 depicts the participating elements in an identification exercise, namely, the *actuators, process* (which we wish to identify), *sensors* and *disturbances* in addition to the *probe signal* and the *response*. Disturbances themselves are categorized into two classes - those which can be measured (e.g., temperature fluctuations in a hot and cold fluid mixing process) and those which are unmeasured (e.g., wind disturbances experienced by an aircraft).

An important observation merits attention - *the discrete-time model built by the user is that of the system appearing in the shaded region and not that of the process alone*. In other words, the identified model not only explains the (continuous-time) process dynamics, but also the actuators, sensors and as already stated, effects of the disturbances. Extracting the model of the continuous-time process from the identified discrete-time model is certainly a non-trivial task (see also remark below).

The following notation is used in the text. **Discrete-time** physical input (adjustable) is denoted by $u[k]$, *measured* response by $y[k]$, unmeasured and measured disturbances by $d[k]$ and $d_m[k]$ respectively, where the index k stands for the k^{th} observation (sampling) instant. States and deterministic variables as the case may be are represented by $x[k]$. Continuous-time signals and functions are denoted by parentheses, for example, $u(t)$ and $y(t)$, where t is a continuous-time quantity. We shall for most portions of the text, turn off the *measured* disturbances; to include them in the analysis whenever necessary is fairly straightforward (at least for linear systems). The *effect* of the unmeasured disturbance $d[k]$ is represented by $v[k]$; it is *assumed to additively* corrupt $y_t[k]$, the true response of the process to the input. Vectors shall be indicated by boldfaced notation, e.g., $\mathbf{u}[k]$ for a vector of inputs. Matrices are denoted by (boldfaced) uppercase characters, as in $\mathbf{A}$ or $\mathbf{\Gamma}$.

1.3.1 THREE FACTS OF IDENTIFICATION

There are three universal facts of identification concerning the **accuracy** and **precision** of identified models, which also provide the guiding paths for identification:

1. *It is generally not possible to build an accurate model from finite-sample data*. Any model estimated from data contaminated with errors can never be accurate. A more important factor is the possibility of a model-process mismatch (any process is typically more complex than a presumed mathematical model). Mis-specification of the model structure (see §18.5.1 for a formal definition) usually lead to systematic errors in model estimates and predictions. Technically, such estimates (of models, parameters or signals) and predictions are termed as *biased*. The general effort in estimation is to produce *unbiased* estimates, but this can be achieved only with the "correct" specification of model structure, proper choice of estimation method and in many situations only with infinite (very large) observations. When the latter is achieved, the model is said to be *asymptotically unbiased*. Technical definitions of (finite-sample, statistical) bias and asymptotic bias are provided in §13.3.
2. *It is generally not possible to estimate a precise model from finite-sample data*. This stems from the fact that a single record of data is only one of the several possible data records for the same experiment. *Repeating* an experiment (while holding all the controllable factors at the *fixed values*) produces numerically a different set of readings. The cause for this variability in data across different runs of experiments is the randomness in disturbances and measurement noise. In practice, one usually works only with a *single* experimental finite-length record, the estimated model is only one among many models that could have been estimated from other possible realizations. Thus, the variability in data manifests as *impreciseness* in estimates (of models, parameters or signals). The statistical measure of variability is the **variance**. Every estimation method strives to drive the variance in estimates to zero. The fact is that *it is never possible to precisely estimate any parameter (model or signal) from finite-length data*. However, it may be achievable under *asymptotic* (large sample) conditions. This is a highly desirable property of every estimator. See §13.4 and 13.10 for technical definitions and explanations.
3. *The accuracy and precision of the optimally identified model, among other factors, is critically dependent on the (i)* ***input type*** *(excitation and shape, with the latter holding for non-linear systems) and the (ii)* ***signal-to-noise ratio*** *achieved in the experiment*. A generalized term capturing both of these aspects is **information**. The quality of the final model depends on how *informative* the data is. *Fisher's information* metric (Fisher, 1925) plays a fundamental role in this respect

(see §13.2).

To understand the role of information intuitively, recall Example 1.2 pertaining to an interview process. The success of selecting the "best" candidate depends on the "information" obtained by the interviewer through a set of suitable questions (number and nature) as well as the relative uncertainties for that candidate.

Therefore, an important task in identification is *input design*. The challenge is that a proper design of input requires sufficient knowledge of the system and the level of uncertainties, which is the purpose of input design itself! Identification is therefore inevitably an iterative exercise. Section 2.1 illustrates the role of input in identification on a simple example. Theoretical details are discussed in §22.3.

Whenever one speaks of accuracy and precision, technically one requires to have a reference point. typically the truth (Section 13.3 and 13.4 presents technical definitions of these terms). In the statements above as well, the description of a "true" system is tacitly assumed to be available. Practically, however, the complexity of the system that is being identified is beyond the reach of a mathematical description. Then, the reference point is the "best" approximation that can be realized for the chosen model structure and the given experimental data, especially, inputs. Nevertheless, for theoretical analysis, i.e., for academic purposes, one often assumes a "true" system primarily to study accuracy and precision properties (of a model structure) resulting from an identification exercise with a chosen estimation algorithm and the input design for a given system.

Given an input-output data set, there exist many approximate models that can *numerically* explain the input-output data with reasonable accuracy and precision. The main tools for realizing the "best" model are built on concepts of linear algebra, estimation (optimization), probability theory and stochastic signal processing. A point of practical importance is that the optimal estimation algorithm delivers the best model conditioned on a user-specified set of parameters (e.g., model class, order, time-delay, etc.). *The user is required to further optimize these free parameters, which can be done efficiently only through a formal study of identification*.

Remarks: The identified model is usually discrete-time in nature since it is based on *sampled data*. When the data is generated through sampling a continuous-time process[3] with the aid of sampler and hold devices (see Figure 1.8 and the associated discussion), there are certain merits in identifying the inherent continuous-time process. Justifiably, there has been a renewed interest in recent times to directly identify the continuous-time process from input-output data. Historically, methods for developing lower-order continuous-time models from step response data existed even before the emergence of modern identification (recall §1.2), but were overshadowed by the surge of discrete-time identification methods. The last two decades have seen the active development of efficient methods for directly identifying or indirectly recovering continuous-time models from sampled data. The theory related to these topics is outside the scope of this text. An interested reader is referred to the rich literature available on this topic (Garnier and Wang, 2008; Rao and Unbehauen, 2006)

We now turn our attention to the model, which is the *centerpiece* of identification. In the following section, only a brief overview of the models and certain important classifications is provided. Chapter 3 treats the topic of modeling in greater detail and presents the necessary technical details and terminology.

1.3.2 NOTION OF A MODEL

A model of a process is broadly defined as that entity which can emulate the characteristics of that process for a given set of operating conditions and parameters.

---

[3]A large class of processes naturally occur in discrete-time for which this discussion does not apply. Examples include population growth, average rainfall, etc.

It is the *lens* through which the observer analyzes the process. The usages of the term model in real-life phrases such as a *model* apartment, a *model* design also have a similar connotation: a substitute or a prototype for the original process. A few examples below offer glimpses of the different forms of models that are encountered in engineering and other scientific applications.

Example 1.13: Spring-Mass System

The steady-state displacement of a spring x under the application of a force F, under some mild assumptions is nicely modeled by Hooke's law

$$F = -kx$$

where k is the spring constant. The model is an example of a static model, relating instantaneous quantities.

Example 1.14: Liquid Level System

A liquid level system, commonly used as a buffer in many industrial processes, consists of a cylindrical tank with in and out flows. The transient relationship between the liquid level $h(t)$ and the inlet flow rate $F_i(t)$ can be modeled as

$$A_c \frac{dh(t)}{dt} = F_i(t) - C_v \sqrt{h(t)}$$

where A_c is the cross-sectional area of the tank and C_v is the valve coefficient of the valve at the outlet. This is an example of a dynamic, non-linear, ordinary differential equation model.

Example 1.15: RLC Circuit

The dynamic behavior of the output voltage $y(t)$ of an RLC circuit is observed at T_s time interval for a change in input current $u(t)$. The resulting discrete-time system is described by the set of equations

$$\begin{bmatrix} x_1[k+1] \\ x_2[k+1] \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1[k] \\ x_2[k] \end{bmatrix} + \begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix} u[k]$$
$$y[k] = \begin{bmatrix} 0 & R \end{bmatrix} \begin{bmatrix} x_1[k] \\ x_2[k] \end{bmatrix}$$

where R is the resistance and the elements $\{a_{ij}\}$, $\{b_{ij}\}$ depend on the values of inductance L, capacitance C and the sampling interval T_s. The discrete-time quantities are denoted by the integer index k.

A model such as this one is known as a discrete-time linear state-space model (of second order).

Example 1.16: Plug-Flow Reactor

A plug-flow reactor is a tubular reactor which is continuously fed with a reactant at one end of the reactor to produce a desired product at the other end. For a reactor carrying out first-order reaction $A \longrightarrow B$ under isothermal conditions, the unsteady-state description of the composition of the reactant $c_A(t,z)$ and product $c_B(t,z)$ can be modeled as

$$\frac{\partial c_A}{\partial t} = -v\frac{\partial c_A}{\partial z} - k_0 c_A \exp\left(-\frac{E_A}{RT}\right)$$
$$\frac{\partial c_B}{\partial t} = -v\frac{\partial c_B}{\partial z} + k_0 c_A \exp\left(-\frac{E_A}{RT}\right)$$

where v is the fluid velocity, T is the temperature of the reactor and the partial derivatives are along time t and length of the reactor $z \in [0, L]$, respectively.

This is an example of a distributed parameter model, which is used to describe a process that evolves in time as well as along another dimension (in this case, space).

Example 1.17: Spring-Mass Dampener

The spring-mass-dampener systems are representative of the shock-absorber systems and hydraulic systems that have second-order dynamics. When sinusoidal forcing functions are applied to these systems, the response at steady-state is obtained by means of the complex frequency response function

$$G(j\omega) = \frac{K\omega_n^2}{(\omega_n^2 - \omega^2) + 2j\zeta\omega_n\omega}$$

where ω_n and ω are the natural frequency of the system and the input, respectively, while ζ is the damping factor of the system. A model such as this is an example of a frequency-domain description of a system.

As discussed in the previous sections, a realistic identification exercise always results in a composite model consisting of *deterministic* (which explains effects of known inputs) and *stochastic* (describing impact of unknown causes) terms. The example below pertains to a widely used model class for identification.

Example 1.18: Auto-Regressive Exogenous (ARX) Model

A second-order **empirical** auto-regressive model with exogenous input has the following mathematical form:

$$y[k] = -a_1 y[k-1] - a_2 y[k-2] + b_1 u[k-1] + e[k] \tag{1.1}$$

where $y[k]$ is the measured output, $u[k]$ is the **physical** (exogenous) input and $e[k]$ is a purely **random** signal that is absolutely unpredictable. The latter can also be thought of as a **fictitious** input, whose statistical properties are partly known, but never known in amplitude. When written in transfer function form (introduced in Chapter 4), it is easy to see the measurement as actually a sum of two effects,

$$y[k] = \frac{b_1 q^{-1}}{1 + a_1 q^{-1} + a_2 q^{-2}} u[k] + \frac{1}{1 + a_1 q^{-1} + a_2 q^{-2}} e[k] \tag{1.2}$$

where q^{-1} is the backward shift operator such that $q^{-1}u[k] = u[k-1]$ (see Chapter 4).

The model in the foregoing example is a *parametric* model (characterized by parameters a_1, a_2 and b_1), different types of which are presented in Chapter 17.

Chapter 3 presents a rigorous definition of the model with a detailed presentation of the different classes of models. At this point we discuss two very broad classes of models, namely, *qualitative* (descriptive) and *quantitative* (numerical) models and two sub-classes in the latter category.

1.3.3 QUANTITATIVE VS. QUALITATIVE MODELS

Qualitative models, as the name suggests, merely describe the response of a system on a categorical basis with little or no involvement of numerical values. For example, when the heat input to a fluid heating system is increased, the temperature of the fluid increases; or when the product draw in a distillation column is increased, the purity of the top product decreases and so on. Quantitative

models, on the other hand, describe the relationship between quantified changes in input and output in terms of mathematical expressions.

Among these two classes, quantitative models prove to be, clearly, more useful in process control, optimization, monitoring as compared to qualitative models whose use is mostly restricted to making qualitative analysis of and improvements to process operations. The major advantage of the quantitative models is that they allow us to draw inferences and make decisions based on quantified criteria. Besides they can be easily programmed by means of a computer, which greatly reduces the need for human intervention in process operations.

Quantitative models can be further categorized into different pairs of classes depending on the nature of the processes they describe, the assumptions made about the underlying phenomena and the approach taken to develop them. We have already encountered two contrasting classes, namely, the first-principles vs. empirical models in Section 1.1. Two additional classifications that are frequently encountered in the identification literature are discussed below. A detailed description of different types of models and the basis for classification is provided in Chapter 3.

1.3.3.1 Deterministic vs. Stochastic Models

In Section 1.3, we made an important observation. The measured response not only contains effects of the known inputs, but also the effects of disturbances and measurement errors. The model that explains the effects of inputs is usually termed as a deterministic model, while the model that explains the effects of disturbances and sensor noise is termed as a stochastic model. In general, deterministic models accurately relate variables that are free of error, i.e., those which are known accurately while stochastic models describe the uncertain characteristics of the process using probability theory and time-series models. In an identification exercise, the developed model is a composite one, containing both the deterministic and stochastic components. Chapters 2 and 3 discuss the related aspects in detail.

1.3.3.2 Non-Parametric vs. Parametric Models

This forms a very important classification in the modeling arena. The term non-parametric should not be misunderstood as to be devoid of any unknowns or parameters. In fact, the term *parametric* refers to the *parametrization* of the model. As a simple example, consider the step response model of a system, which is simply the set of step response coefficients at the sampling instants (from start to steady state). This is a non-parametric model. However, if the system is assumed to have first-order characteristics, then the response can be characterized by three *parameters*, namely, gain, time-constant and time-delay. When the response coefficients are directly estimated, it is termed as non-parametric identification. Instead if the parameters are estimated, it is the case of *parametric* identification. Clearly, in the latter, the number of unknowns, are fewer than the former. However, this advantage of parametric models only sets in when prior knowledge is available.

Parametric models possess a specific structure and order and are characterized by fewer parameters while non-parametric models do not possess any specific structure or order but are characterized by a large number of unknowns. Difference equation descriptions are common examples of the former class while convolution models (impulse response models) are examples of the latter. From an identification viewpoint, non-parametric models can be estimated with minimal a priori knowledge while the estimation of parametric models demands some a priori knowledge on the user's part. This prior knowledge can be acquired by first estimating a non-parametric model. A more detailed presentation of these aspects is contained in Chapter 3.

Until this point we have studied the basic concepts of identification and obtained a preview of the different types of models. Now we address the most important topic, which is the identification procedure itself.

Essentials of Identification

Model development is the primary goal in identification. To ensure that the identification method delivers models with a desired accuracy and precision level, it is imperative that the user has a clear understanding of:

i. What are the possible model structures for a deterministic process? (Chapters 3 to 5)
ii. How does one represent uncertainties and randomness in processes / data? (Chapters 7 to 11)
iii. Which is the most appropriate model structure to begin with? (Chapter 17)
iv. What are the qualities of a good estimator? (Chapter 13)
v. What methods are available for estimating optimal models? (Chapters 14, 15, 19, 20, 21 and 23)
vi. How does one assess the goodness of estimated models and construct confidence regions? (Chapters 13 and 22)
vii. How to compute predictions given a model structure? (Chapter 18)
viii. What are the optimal inputs for a given identification problem and how to design them? (Chapter 22)

Answers to the above questions are provided by a formal study of the subject indeed as indicated by the respective chapters in the text. However, it is equally important to implement identification in a step-wise manner. The systematic procedure presented in the following section describes these steps starting with data acquisition. This procedure is revisited in greater detail later in Chapter 22.

1.4 SYSTEMATIC IDENTIFICATION

The procedure for identification can be divided into five salient steps, namely,

S1. Data Generation and Acquisition
S2. Data Pre-Processing
S3. Data Visualization
S4. Model Development
S5. Model Assessment and Validation

A flowchart delineating the salient steps in identification is shown in Figure 1.7. The flowchart highlights two important facets of identification. Firstly, that identification is an iterative exercise and secondly, that prior process knowledge can be factored into every stage of identification. In fact, it is recommended to incorporate as much prior knowledge as possible. The different steps are discussed below in detail.

1.4.1 DATA GENERATION AND ACQUISITION

This first step can be said to be the most influencing step in identification, rightly so since the success of every stage of identification and the confidence on the final model depends on the *information quality* (and quantity) of data. It is the *food* for identification.

Figure 1.8 portrays the schematic of a typical *sampled-data system*. Discrete-time input designed by the user is converted to an approximate continuous-signal by a hold device (also termed as the *digital-to-analog* converter), which then excites the process with the help of an actuator (also known as the final control element). The resulting response is then observed with the aid of sensing device that consists of a sampler and quantizer. The sensor is often known as the *analog-to-digital* (A/D) converter. Sampled data is passed on to a data storage device.

There are three key decision variables in this operation, namely, the *type of discrete-time input*, the *type of hold device* and the *sampling rate*. Input design is an identification-specific issue and has

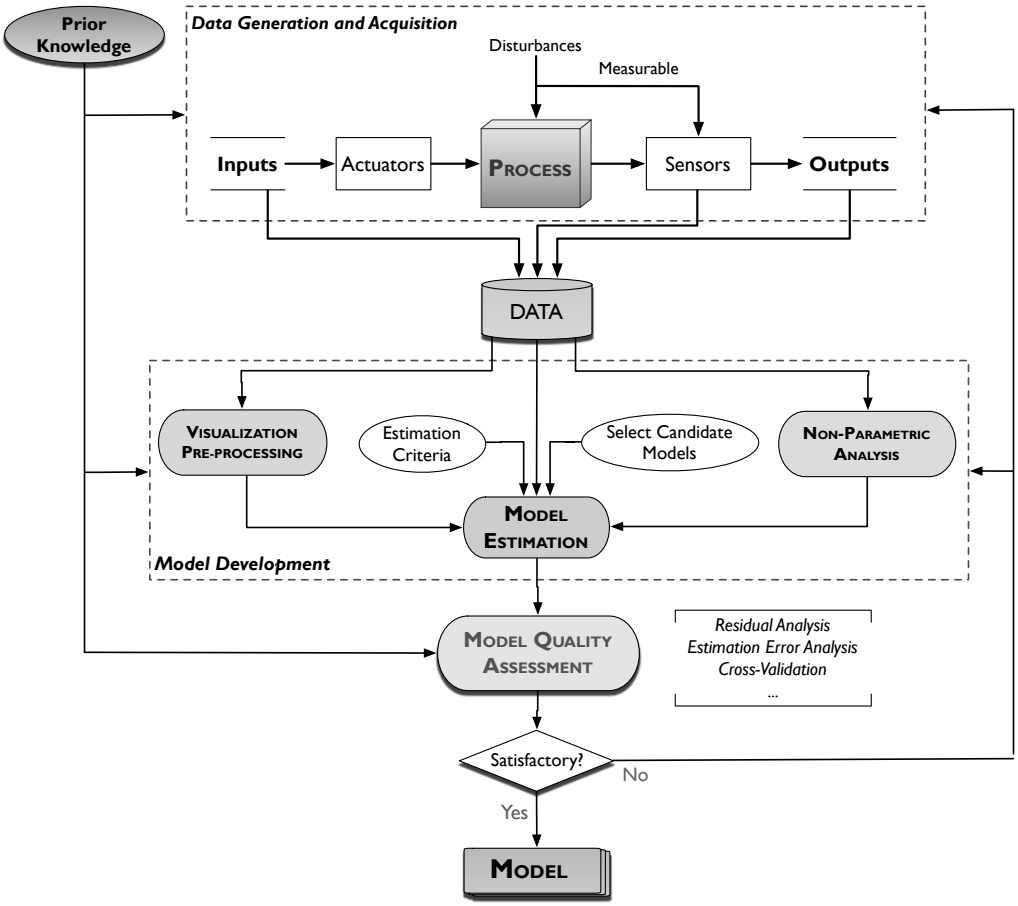

FIGURE 1.7 (**SEE COLOR INSERT**) A generic iterative procedure for System Identification.

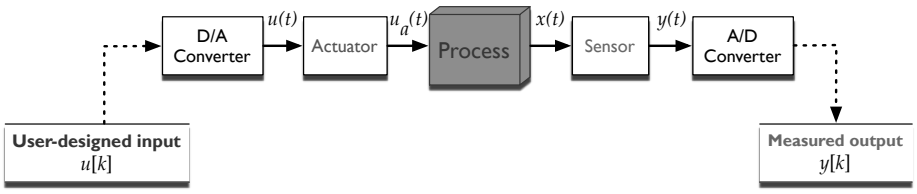

FIGURE 1.8 A standard sampled-data system.

been reasonably well studied but with plenty of open-ended problems. The basic question of interest is - *what kind of excitation is best for a given identification problem?* Once again, this is analogous to framing the right set of questions to be asked in an interview process or in an examination. Just as the number, type and complexity level of questions depends on the purpose of the examination, the number of candidates that take part and that have to be selected, the nature of the input is also tied to the end use of the model, i.e., whether it is eventually used for control, fault detection or in simulation and the accuracy plus precision requirements of the model.

The basic tenet is that the excitation in the input should be such that its effect in the measured output is larger than those caused by sensor noise / unmeasured disturbances. In closed-loop situations, the input moves are decided by the controller; in such situations, the user has no direct way of adjusting the input excitation. However, indirect access is available through the set-point or a dither signal (an external signal) that is introduced at the input or at the set-point. The central result in design of experiments for identification is strongly tied to the notion of *identifiability*, or more appropriately the concept of "informative" data. The basic idea is illustrated later in §2.1. Formalization of related concepts, as presented in §22.2 and §22.3, not only facilitates solutions to the input design problem, but also allows us to obtain a preliminary idea of the model complexity that can be built for a given data set. The concept of informative data stems from a central concept in identification, known as the identifiability, whose basic idea is illustrated in §2.1. A formal discussion of the same appears in §18.6.

Obtaining informative data is critically dependent on the sampling rate. The issue of choosing a suitable sampling rate has been at the roots of several fields, particularly in the field of communications and signal processing where it all began six decades ago (Lüke, 1999; Unser, 2000). As in the case of the input design problem, the major results in sampling theory are best formulated and presented in frequency-domain. The celebrated sampling theorem due to Whittaker, Shannon and Nyquist is a standing example of this fact. Despite the remarkable progress in the subject (of sampling), certain open issues still remain. One such open-ended problem is in the context of choosing a suitable sampling rate for multiscale systems, which are characterized by phenomena occurring at time-scales differing by at least an order (e.g., a fuel cell system that generates energy through electrochemical reactions).

Finally, the hold device plays an important role in identification because it alters the spectral characteristics of the designed discrete-time input as a consequence of approximating the continuous-time signal from the digital signal. Mathematically, this approximation can be carried out in infinitely different ways. A widely used method is the *zero-order hold* (ZOH) approximation. It holds the signal constant between two sampling instants. Essentially the interpolating curve is a zeroth order polynomial. ZOH devices are popular due to their ease of implementation. All higher-order devices are *non-causal* and therefore not realizable (causal versions exist, but also introduce additional complexities - they find use only in very specialized applications).

Chapter 6 explains the basics of sampling and hold operations, and reviews *discretization* - which is the procedure to derive the discrete-time equivalent of a continuous-time system.

1.4.2 DATA PRE-PROCESSING

The acquired data in its raw form is usually not ready to be used for model development. Often the data has to be subjected to quality checks and a pre-processing step before presenting it to the model estimation algorithm. There are many factors that affect data quality (in addition to noise), two among which are discussed below.

i. *Outliers*: Outliers are data which do not conform to other parts of the data largely due to sensor malfunctions and/or abrupt and brief process excursions. Detecting and handling outliers in data can be very complicated and challenging primarily due to the fact that there is no universal

definition of an outlier. Notwithstanding this fact, a few reasonably good statistical methods are available for this purpose.

ii. *Missing data*: The issue of missing data is prevalent in several applications. Intermittent sensor malfunctioning, power disruptions, non-uniform sampling and data transfer losses are some of the common reasons for missing data. In Figure 1.4 we observed missing observations (at random) of relative humidity and temperature. A few popular methods for handling missing data include the ML-based expectation maximization and the multiple imputation methods.
Section 22.4.2 gives an overview of some well-known methods for handling missing data and outliers.

Pre-processing of data may also be motivated by the assumptions, limitations and requirements of model development. For instance, data may contain drifts, trends and other non-stationarities, whereas most identification methods assume stationarity of data, a condition requiring the statistical properties of data to remain invariant with time. To bring such measurements into the realm of stationary signals, two approaches are commonly implemented: (i) explicit fitting of polynomials and working with residuals and (ii) difference the series and work with the differenced series. Sections 7.5.5, 9.7 and 22.4.1 discuss these approaches.

Pre-filtering data is an elegant way of encompassing methods for handling a variety of data characteristics such as drifts and noise. Further, it can be used to obtain preferentially accurate fits in select frequency ranges and to establish theoretical equivalences of different parametric model structures (see §17.5).

Data pre-processing can consume a significant amount of the overall time and effort in an identification exercise. The situation can be alleviated considerably by choosing a reliable instrumentation and data acquisition system, and a careful experimental design.

1.4.3 DATA VISUALIZATION

Visualizing data is a key step in information extraction and signal analysis. The value of information obtained from visual inspection of data at each stage of identification is immense. Prior to the pre-processing stage, visual examination assists in manually identifying presence of drifts, outliers and other peculiarities. It also provides an opportunity for the user to qualitatively verify the quality of data from an identification viewpoint (e.g., sufficient excitation). A careful examination can also provide preliminary information on the delay, dynamics and gain of the process. The information obtained at this stage can be used at the model quality assessment stage. For instance, if the user observes an underdamped behavior as a salient characteristic, any model that does not capture this behavior can be rejected.

Powerful methods exist for visualizing multi-dimensional or multivariable data. The user can exploit the effectiveness of these methods in selecting the right candidate models. Post model development, a visual comparison of the predictions vs. the observed values should be strongly preferred to a single index such as correlation or similarity factor for assessing the quality of the model.

Finally, visualization of data in a transform domain (e.g., Fourier domain) can prove very beneficial. Examination of the input and output in frequency domain can throw light on the spectral content and presence of periodicities. It is useful in obtaining a first-hand feel of the level of (input) excitation and the filtering nature of the system. Time-frequency analysis tools such as wavelet transforms (Mallat, 1999; Tangirala, Mukhopadhyay and Tiwari, 2013) can further provide valuable information on the time-varying characteristics of the process.

1.4.4 MODEL DEVELOPMENT

Development of a model is the central goal of identification, as we have already learned. As explained in §1.1, the general objective is to build a *deterministic-plus-stochastic* model. This part of

identification involves two steps: (i) specifying a model structure and order, and (ii) estimating the parameters of that model by solving the associated optimization problem.

Choice of candidate models

Choosing the candidate models is perhaps the most challenging and time consuming part of identification. The approach is usually iterative and governed by the following guidelines. Only a brief overview is provided. The illustrative example of §2.4 highlights a few of these aspects, while a detailed discussion with case studies is presented in Chapter 22.

1. *Accuracy (bias) and precision (variance) requirements*: For a given data set, the bias and variance of the estimated model are determined by its structure and the estimation algorithm. An important aspect is the *interplay* between the stochastic and deterministic parts of the model.
2. *Prediction accuracy and horizon*: The primary considerations here are the range (of time steps) over which prediction is sought and the accuracy (over a frequency range). Unless otherwise explicitly demanded by the application, the one-step ahead prediction is of interest.
3. *End-use of the model*: In addition to the predictive abilities, the end-application might impose other requirements. For instance, if the intended use is in control, the model should be as simple (low-order) as possible and should not have underestimated the delay. Further, the characteristics of the process over its bandwidth should have been well captured. The branch of control-relevant identification is an offshoot of these ideas.
4. *Estimation aspects*: The ease of estimation can be an important factor of consideration. Models that yield *linear-in-parameter predictors* are typically preferred to those that are non-linear due to convenience of computation and the existence of a unique solution. The price that is paid is the prediction ability of the model. A trade-off is therefore sought. These facts are theoretically elucidated and illustrated in Chapters 18, 17, 21 and 22.
5. *Prior knowledge*: The choice of model may be motivated by some prior information known on the type of models such as linear / non-linear, low- / high-order, etc. or on the structure of the parametric model. As mentioned earlier, this facet of identification falls under the purview of grey-box modeling (see §1.4.6 for more discussion and §23.7.2 for illustrations).

Notwithstanding the influence of the above factors, in almost all situations, the user begins with an initial guess of the structure and iteratively makes refinements using the results from the assessment of model quality.

Remarks: One of the biggest benefits of empirical modeling is the flexibility in selecting the model structure. As much as it equips the user with tremendous freedom, it also brings with it the risks of overfitting (see Example 2.4 for an illustration and also §2.4). Moreover, a "correct" model is beyond the reach of any modeling exercise. So to speak, no process can be accurately described by any mathematical description.

Therefore, *the goal of identification is not necessarily to develop a correct model, but is rather to build a good and useful working model.*

Estimation method

Once a candidate model is selected, what remains is its estimation, which is essentially an optimization problem. Section 1.2 gave an overview of the methods available for estimation, namely, the prediction-error minimization, instrumental variable and the subspace identification methods. A typical estimation criterion typically has a form of minimization of a function of the prediction errors. The function is usually based on a distance metric. For example, least squares methods minimize the Euclidean distance between the predicted and observed values (squared 2-norm of the prediction errors). Other factors such as quality of parameter estimates, number of parameters can be factored into the objective function. Methods based on the maximum likelihood principle, on the other hand, construct the objective function from probabilistic considerations.

In choosing an estimation algorithm, the prime factors for consideration are the goodness of estimates and the ease of computation. Usually these are conflicting factors. However, without sacrificing the key properties (e.g. precision) of an estimator, wherever possible a computationally simpler algorithm may be chosen. The final decision should be driven by a judicious, practical and informed choice of algorithms.

1.4.5 MODEL ASSESSMENT AND VALIDATION

The model quality assessment and validation step is an integral part of any model development exercise, be it identification or a first-principles approach. The focal points of analysis are

1. *How effectively has the model explained the output variations in the training data?* The overall goal is certainly to achieve as "small" a prediction error (minimum bias) as possible, but not at the cost of low precision (variance in model or parameter estimates). With this objective in mind, three tests are performed (Chapter 22 explains these points in detail with illustrations)
   - *i*. *Statistical analysis of prediction errors or residuals*: The key requirement is that *there is* **no residual information** *left for the model to capture*.
   - *ii*. *Error analysis of estimates*: The (standard) errors should be small *relative* to the estimated values.
   - *iii*. *Analysis of model fit*: Metrics such as adjusted R^2 and its variants are typically used for determining the degree of fit (of predictions).
2. *How accurately does the model predict the response on a test (fresh) data set?* This is the test of **cross-validation**. The primary purpose of this test is to determine whether the model has been trained to capture the *global* characteristics of the process as evidenced in the training data set or has specialized to the *local* features of the training data. When the latter occurs, its predictive capabilities for a fresh data set deteriorates and model is deemed to have been overfit. The performance itself is evaluated based on certain metrics of fit and prediction horizons.

The outcomes of the above diagnostic tests provide the necessary feedback for refining the decisions made in previous stages. If a model does not meet any of the aforementioned requirements, it inevitably calls for improvements at one or more of the previous steps. When the user has confidence in data quality, efforts should be directed towards refining the structure and/or order of the model. Naturally, a review of decisions in other stages may also be required. To achieve best results, the user should have a sound knowledge of the impact of choices and decisions in the preceding steps on the final quality of the model.

1.4.6 PRIOR PROCESS KNOWLEDGE

As remarked above quite often some a priori knowledge concerning the (i) structure of the model, (ii) order of the model, (iii) values of a subset of parameters and (iv) bounds on parameter values may be available. For instance, in developing a model for a two-tank (in series) system it is known that the process has second-order overdamped characteristics. It is certainly beneficial to incorporate this additional knowledge into model estimation in anticipation that (i) the resulting model will reflect the physics of the process in a more transparent manner than a classical black-box model and (ii) the resulting model will have better prediction capabilities than an unconstrained model.

Naturally, the degree of transparency increases with the available a priori information. The most transparent case is that of the known mathematical form of the model usually through a first-principles analysis of the process and only the parameters remain to be determined. This may be termed for lack of a better term, a *white-box* model.

Grey-box modeling arises prominently in development of state-space models with known structure. For a given input-output system, there exist infinite state-space realizations that produce the same input-output relationship. However, a particular structure may be desirable from an estimation

viewpoint or by the application of interest. Constrained optimization of parametric models provides the necessary framework for estimation of grey-box models.

Quite often it is the case that the prior knowledge is not known with certainty. For instance, the analyst knows that the parameters of interest fall within a range or that impulse response has a particular shape. Bayesian approaches offer a powerful framework for handling prior knowledge mixed with uncertainty. These methods are now being widely applied in several fields because they directly produce the uncertainty regions for the estimates unlike classical point estimators such as LS and MLE. A major limitation of the Bayesian estimators until recently was the high computational cost involved in implementing them. However, with the recently introduced Markov chain Monte Carlo (MCMC) simulations, the barrier has been significantly reduced. The mathematical paraphernalia of Bayesian approaches can be quite complex and is usually covered in advanced texts. In this text, nevertheless, some preliminary material is presented in §15.2 and §20.2.3.

Remarks: Black-box models are usually in for criticism because they do not explicitly take into account the physics of the process. Notwithstanding the technical correctness of these criticisms, a point in counter is that a careful analysis of the data and a systematic identification can produce good *working* black-box models that have predictive capabilities similar to or sometimes better than that of a first-principles model. Undoubtedly, the success of building these models is crucially anchored to the data quality.

1.4.7 SUGGESTIONS FOR OBTAINING A GOOD MODEL

It is appropriate to conclude this section by prescribing the ingredients for obtaining a good working model:

- *Good quality data*: Design and conduct experiments to generate informative data. The data should be rich enough to distinguish between two competing candidate models.
- *Data visualization*: Often this step is undermined. Visual analysis of data can extract valuable information with such ease that would otherwise require elaborate mathematical analysis. The information obtained by a qualitative analysis at this stage is useful in making decisions in model selection as well as in model validation.
- *Suitable model structure and domain*: Model structure should be based on end use, predictability and physical insights. The right choice of domain (e.g., frequency) for data pre-processing and modeling can significantly enhance the quality of the model.
- *Simplicity of the model*: Simple models can be good approximations of complicated systems. A complex model that produces marginal improvement in predictions at the price of large errors in parameter estimates should be discarded in favor of a simpler model.
- *Time-scale for modeling*: Identifying the appropriate time-scale for the phenomenon of interest is crucial. Slow sampling leads to loss of observability whereas excessively fast sampling produces large amounts of noise relative to the signal and can also push the system to the verge of instability (see Chapter 6).
- *Model validation*: Right interpretation of model validation and model quality assessment is essential to building a good model. The model diagnostic checks reveal considerable information on the model sufficiency and can contain clues to directions for model refinement.
- *Finally, there is no substitute for thinking, insight and intuition.*

Clearly, identification is both a science and an art since it is brought about by a careful blend of knowledge (of both the subject and process), design (of the experiment), experience and intuition (of the user).

The next section concludes this chapter with a description of the organization and flow of learning material in this text.

1.5 FLOW OF LEARNING MATERIAL

The subject of identification is essentially a *confluence* of the four broad subjects of theory of random processes, estimation theory, signal processing and systems theory.

Figure 1.9 schematically illustrates this fact. A mastery of the subject of identification invariably demands a strong foundation in these respective areas. From a pedagogical perspective, a contextual and a need-based learning is pragmatic. With this perspective, the first three parts of this text is devoted to an exposition of the requisite fundamentals in the founding areas.

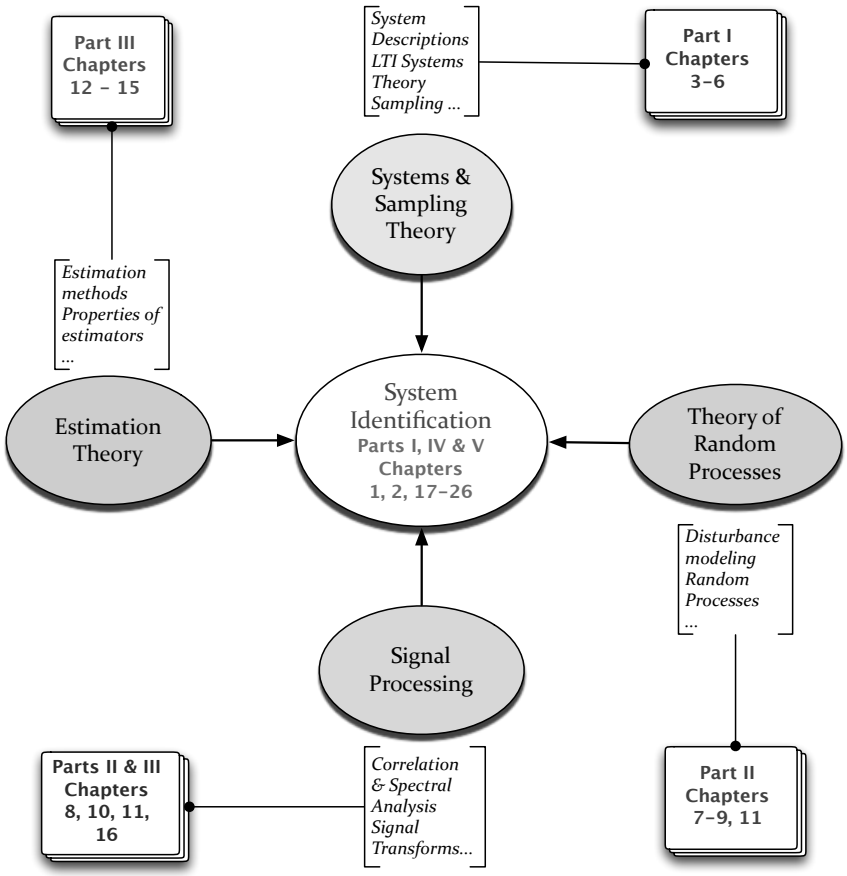

FIGURE 1.9 System Identification involves application of concepts from four broad fields in engineering, mathematics and statistics.

ORGANIZATION OF THIS BOOK

The book is divided into five parts. Part I provides the introductory material on identification, mathematical models and the various descriptions of deterministic linear time-invariant systems. In Part II, the focus is entirely on the theory of random processes, models for linear stationary processes and spectral (frequency-domain) representations. Part III provides the crucial foundations on estimation theory, methods for estimating statistical properties, techniques for (model) parameter estimation and most importantly on metrics for assessing the quality of estimates. The subject matter of Part IV is the culmination of the concepts and principles presented in the preceding parts, in the context of identification. Expectantly, it is the largest among all the parts and constitutes the core of this text. The topics in Part IV can be fully understood with a good comprehension of the preceding portions. Part V offers glimpses of advanced concepts in identification.

The content of each part is described below in greater detail.

Part I

Chapter 1 introduced the basic principles of identification and notions of modeling with an exposition of a systematic procedure for identification. Chapter 2 is a curtain raiser for the subject. It takes the reader through the inner world of identification by illustrating the procedure outlined in Section 1.4 on a simple case study. The main aim of Chapter 2 is to allow the user to absorb the tenets and ideas of identification (of LTI systems) with a qualitative understanding of the theory. A general introduction to mathematical descriptions and model classification is the subject of Chapter 3. Through this chapter, the reader obtains an overview of the library of models that one can choose from for a given process. Chapters 4 and 5 are exclusively devoted to the models of discrete-time LTI systems, which are the primary representations of interest. In addition to providing a standard review, the pros and cons of each of these descriptions from an identification standpoint are discussed. The connections between continuous-time and discretized systems under a sampling-and-hold operation are presented in Chapter 6. The material therein also reviews basic sampling theory in the context of signals and systems. A beginner should not be surprised to find a recurring need to revisit Chapters 4 and 6 during the course of reading this text.

Part II

In Section 1.3 we observed that the response of a system contains effects of both deterministic and stochastic quantities, thus the need for building a composite model. Chapter 7 introduces the reader to the founding concepts of random processes. A brief review of the theory of random variables and the founding concepts of stationarity and ergodicity constitutes the chapter. Predictability of random processes is tested by means of correlation functions, which are treated in Chapter 8. The chapter also introduces the building block for time-series modeling and analysis, the *white-noise process*, which is an ideal unpredictable random process. Time-domain models for linear stationary processes, namely, the auto-regressive moving average (ARMA) representations and ARIMA models are treated in Chapter 9. Theoretical correlation function properties of these models and Yule-Walker equations are the focal topics. Chapter 10 presents a review of Fourier transforms and the concepts of spectrum and spectral density, for deterministic signals. It offers the requisite foundations for the frequency-domain analysis and spectral representations of random processes presented in Chapter 11, which provide a complementary (to that of time-domain) and an excellent viewpoint of random processes. Frequency-domain equivalents of correlation functions, namely, the cross-power spectrum and coherence are described in this chapter. Further, the milestone result of *spectral factorization theorem* is reviewed. In each chapter, the concepts are illustrated through suitable examples in MATLAB®.

Part III

Parts I and II provide the theoretical foundations. In Part III, the focus shifts to the field of estimation, which connects the world of theory to the field of practice with a blend of optimization and statistics. Therefore this part serves as the most appropriate transition from the previous parts to the remainder portions of this text. The subject of estimation is a very exciting and challenging one. The reader should spend considerable time in carefully understanding the subtle and practical aspects of estimation in the constituent chapters. Chapter 12 presents an introduction to the field of estimation and briefly discusses the three estimation problems, namely, *prediction, filtering and smoothing*. Any estimation problem is only half-complete with the computation of an estimate. The remainder, an important one, constitutes the accuracy and precision analysis of the estimate (or the estimator). Chapter 13 deals with the topic of the "goodness" of the estimators. Key properties such as bias, variance and consistency are discussed in great detail with suitable examples. Expressions for these properties throw light on how experiments should be conducted for obtaining *efficient* (minimum variance) estimates.

There exist an innumerable set of algorithms for parameter estimation. However, it suffices to focus on four widely used classes of algorithms, namely, the method of moments, least squares methods, maximum likelihood methods and Bayes estimation algorithms. Chapter 14 describes the first two classes while Chapter 15 presents the remaining two classes. Part III concludes with Chapter 16, which is concerned with estimation of statistical properties (e.g., mean, correlation) and spectral densities. The estimators of signal properties presented in this chapter are vital to estimation of model parameters. As with previous parts, the concepts in each chapter are adequately illustrated by way of examples in MATLAB.

Part IV

Part IV forms the nucleus of this text. It constitutes the application of concepts and methods in the previous parts to the problem of identification. To begin with, the reader is introduced to *non-parametric* and *parametric* models for LTI systems in Chapter 17. The general class of parametric descriptions, i.e., the prediction-error model structures form the central subject of discussion in the chapter. A knowledge of prediction theory is vital to the practice of identification. Chapter 18 enunciates the basics of prediction theory, where concepts of one-step and infinite-step ahead predictions are elucidated. Predictor-based model descriptions are also introduced in this chapter. The central concept of *identifiability* and related notions such as *equality of models* are explained in the concluding part of this chapter. Conditions for (model) identifiability of linear time-invariant black box model structures are discussed. The notion of *system identifiability* and the requirements to guarantee the same are also presented.

Chapter 20 treats the identification of non-parametric models, namely, the (impulse, step and frequency response models. Classical and modern methods for estimation of IR coefficients are included. Standard estimators of frequency response functions are discussed.

Chapter 21 is concerned with the identification of parametric model structures that are described in Chapter 17 using the PEM methods, correlation approaches and IV methods. Techniques for estimating specific model structures (e.g., ARX, ARMAX, OE) are presented. Frequency-domain interpretations of quadratic PEM methods are also discussed. This chapter constitutes an integral part of the text.

Chapter 22 discusses the statistical and practical aspects of identification, beginning with the theory governing input design for identification of LTI systems. The ideas of informative data and *persistent excitation* are formally discussed. This is followed by a treatment of the input design in the classical sense. Popularly used class of input signals, with special attention to pseudo-random binary signals (PRBS), are discussed. Methods for handling outliers and missing data, namely, the robust identification and the EM algorithm are explained in detail with suitable examples. Subsequently, the time-delay estimation problem is discussed in good detail where an efficient frequency-domain method based on Hilbert transform relation along with the classical correlation and model-based methods are presented. In the second part of this chapter, model order determination and other options such as pre-filtering are reviewed. Finally, statistical tests for model quality assessment and validation are described with MATLAB-based illustrations.

Identification of state-space models is taken up in Chapter 23. A majority of this chapter is devoted to the subspace identification, the ideas and algorithms therein. A basic understanding of the Kalman filter is integral to comprehending subspace algorithms. A tutorial review of the same is therefore included. The core of this chapter is the general class of subspace identification algorithms with specializations such as N4SID, MOESP and CVA. This topic is usually counted among advanced topics in identification due to the underlying mathematical complexities. In view of this fact, the presentation begins with the simple deterministic case and gradually builds up to the deterministic-plus-stochastic scenario. Parametrized state-space models and their identification are described subsequently. The chapter concludes with an outline of grey-box identification, mainly by way of illustrations on a few simulated processes.

Finally, Chapter 24 presents a set of simulation and industrial case studies to illustrate the practical applications of the concepts learned in this text. These can also serve as motivating examples in a classroom session.

Part V

The final part consisting of Chapters 25 and 26 offers glimpses of the advanced topics in identification. The objective is to provide an overview of these topics, highlight the main challenges and discuss the principles of a few common methods. Therefore, the presentation is kept crisp and illustrative. Chapter 25 contains a collection of topics related to linear time-varying, non-linear and closed-loop identification of single-input, single-output (SISO) systems. Illustrative examples are presented for each of these classes. The book concludes in Chapter 26 with a cursory introduction to multivariable identification. A frequency-domain method for time-delay estimation in multivariable systems, as an extension of the technique described for SISO systems in Chapter 22, using partial coherence is introduced. Finally, a method for multivariable identification, when both inputs are outputs are known with errors, using the well-known principal component analysis (PCA) is put forth.

1.6 SOFTWARE

The primary software platform for the illustration of examples in this text is MATLAB® (Release 2014a), a powerful commercial computation, visualization and simulation software package developed by The MathWorks, Inc. The System Identification Toolbox (Ljung, 2014), a companion software exclusively designed for this subject, contains all the necessary tools for the analysis, simulation and identification of process systems from data. Simulations of continuous-time and complex process systems are facilitated by MATLAB and its companion software SIMULINK. The best way to learn MATLAB and SIMULINK is by going through some of the popular tutorials first and reproducing the results therein.

EXERCISES

E1.1. Define the term "System Identification."

E1.2. Explain the need for identification in process automation.

E1.3. Identify two key differences between the first-principles and empirical approaches to model development.

E1.4. What are the differences between deterministic and stochastic effects in a measured process response?

E1.5. A liquid buffer system is a commonly encountered unit in industrial processes. It consists of a vessel with an inflow F_{in}, an outflow F_{out} due to *gravity* and a liquid buffer maintained at a height h. Assuming a cylindrical geometry for the vessel, answer the following:

a. Write a first-principles model for the liquid level dynamics using the conservation of mass principle. Assume incompressible flow (constant density) and that the outflow $\propto$ to the square root of liquid level.
b. What kind of a model do you obtain in part (a)? (e.g., non-linear / linear)
c. Do you need any experimental data for developing your first-principles model? If yes, what kind of data do you require?

E1.6. Suppose you are completely unaware of the form of the first-principles model and you intend to empirically obtain the dynamic, *discrete-time* model (that relates the input-output observations) of interest in **E1.5.**. What would be your approach? Describe in detail the following

a. How would you choose a suitable sampling interval?

b. What input would you consider appropriate?
c. What would be a suitable model to begin with? Would a linear *dynamic* model be a good starting point?
d. Can you list some ideas of how to estimate the model that you postulate?
e. How do you propose to handle the measurement errors (assume no unmeasured disturbances are present)?
f. List the major challenges (if any) in this exercise.

E1.7. Give two examples of quantitative and qualitative models for a (i) mechanical system, (ii) electrical system, (iii) aerospace system and (iv) a chemical engineering system.

E1.8. Describe three uses of model in the context of different applications.

E1.9. Give two examples of identification problems in any field (engineering, medicine, humanities, etc.) of your choice.

E1.10. Explain the salient steps in identification.

E1.11. What role does model validation play in the development of a good model?