Appendix W3.7
System Identification

W3.7.1 A Perspective on System Identification

In order to design controls for a dynamic system, it is necessary to have a model that will adequately describe the system's dynamics. The information available to the designer for this purpose is typically of three kinds.

1. **Physical model**: First, there is the knowledge of physics, chemistry, biology, and the other sciences which have over the years developed equations of motion to explain the dynamic response of rigid and flexible bodies, electric circuits and motors, fluids, chemical reactions, and many other constituents of systems to be controlled. The model based on this knowledge is referred to as a "physical" model. There are many advantages to this approach, including ease of controller development and testing. One disadvantage of this approach is that a fairly high-fidelity physical model must be developed.

2. **Black box model**: It is often the case that for extremely complex physical phenomena the laws of science are not adequate to give a satisfactory description of the dynamic plant that we wish to control. Examples include the force on a moving airplane caused by a control surface mounted on a wing, and the heat of combustion of a fossil fuel of uncertain composition. In these circumstances, the designer turns to data taken from experiments directly conducted to excite the plant and measure its response. The second approach uses an empirical or heuristic model referred to as the "black box" model. In this approach, the control engineer injects open-loop commands into the system and records the sensor response. The process of constructing models from experimental data is called **system identification**. Standard system identification techniques (for example, linear least-squares) are used to identify a dynamic input/output model. The advantage of this technique is that the control engineer does not need to have a deep understanding of how the system physically behaves, but instead can design a controller solely based on the derived model. There are several major disadvantages to this approach. First, the control engineer must have access to working hardware. Another serious disadvantage of this approach is that it does not provide insight or physical understanding of how specific hardware modifications will affect the control—usually hardware modifications require the control engineer to repeat the full cycle of system identification, control design, and validation. The advantage of this approach is that we use logic and data to model

inputs and outputs, and the detailed knowledge of the physics is not required.

3. **Grey box model**: The third approach is the use of the combination of physical and empirical models referred to as "grey box" modeling.

In identifying models for control, our motivation is very different from that of modeling as practiced in the sciences. In science, one seeks to develop models of nature as it is; in control, one seeks to develop models of the plant dynamics that will be adequate for the design of a controller that will cause the actual dynamics to be stable and to give good performance. The initial design of a control system typically considers a small signal analysis and is based on models that are linear and time-invariant (LTI). This is referred to as a "control relevant" model. Having accepted that the model is to be linear, we still must choose between several alternative descriptions of linear systems. If we examine the design methods described in the earlier chapters, we find that the required plant models may be grouped in two categories: parametric and nonparametric. For design via root locus or pole assignment, we require a parametric description such as a transfer function or a state-variable description from which we can obtain the poles and zeros of the plant. These equivalent models are completely described by the numbers that specify the coefficients of the polynomials, the elements of the state-description matrices, or the numbers that specify the poles and zeros. In either case, we call these numbers the *parameters* of the model, and the category of such models is a **parametric description** of the plant model.

Parametric model

In contrast to parametric models, the frequency-response methods of Nyquist, Bode, and Nichols require the curves of amplitude and phase of the transfer function $G(j\omega) = Y(j\omega)/U(j\omega)$ as functions of ω. Clearly, if we happen to have a parametric description of the system, we can compute the transfer function and the corresponding frequency response. However, if we are given the frequency response or its inverse transform, the impulse response, without parameters (perhaps obtained from experimental data), we have all we need to design a lead, lag, notch, or other compensation to achieve a desired bandwidth, phase margin, or other frequency response performance objective without ever knowing what the parameters are. We call the functional curves of $G(j\omega)$ a **nonparametric** model because, in principle, there is no finite set of numbers that describes it exactly.

Nonparametric model

Because of the large data records necessary to obtain effective models and the complexity of many of the algorithms used, the use of computer aids is essential in identification. Developments such as Matlab's System Identification Toolbox are enormous aids to the practical use of the system identification techniques. For detailed discussion on system identification, the reader is referred to Franklin, Powell, and Workman (1998).

W3.7.2 Obtaining Models from Experimental Data

There are several reasons for using experimental data to obtain a model of the dynamic system to be controlled. In the first place, the best theoretical model built from equations of motion is still only an approximation of reality. Sometimes, as in the case of a very rigid spacecraft, the theoretical model is extremely good. Other times, as with many chemical processes such as papermaking or metalworking, the theoretical model is very approximate. In every case, before the final control design is done, it is important and prudent to verify the theoretical model with experimental data. Second, in situations for which the theoretical model is especially complicated or the physics of the process is poorly understood, the only reliable information on which to base the control design is the experimental data. Finally, the system is sometimes subject to online changes that occur when the environment of the system changes. Examples include when an aircraft changes altitude or speed, a paper machine is given a different composition of fiber, or a nonlinear system moves to a new operating point. On these occasions, we need to "retune" the controller by changing the control parameters. This requires a model for the new conditions, and experimental data are often the most effective, if not the only, information available for the new model.

Our sources of experimental data

There are four kinds of experimental data for generating a model:

1. **Transient response**, such as comes from an impulse or a step;
2. **Frequency-response data**, which result from exciting the system with sinusoidal inputs at many frequencies;
3. **Stochastic steady-state information**, as might come from flying an aircraft through turbulent weather or from some other natural source of randomness; and
4. **Pseudorandom-noise data**, as may be generated in a digital computer.

Each class of experimental data has its properties, advantages, and disadvantages.

Transient response

Transient-response data are quick and relatively easy to obtain. They are also often representative of the natural signals to which the system is subjected. Thus, a model derived from such data can be reliable for designing the control system. On the other hand, in order for the signal-to-noise ratio to be sufficiently high, the transient response must be highly noticeable. Consequently, the method is rarely suitable for normal operations, so the data must be collected as part of special tests. A second disadvantage is the data do not come in a form suitable for standard control systems designs, and some parts of the model, such as poles and zeros, must be computed from the

data.[1] This computation can be simple in special cases or complex in the general case.

Frequency response

Frequency-response data (see Chapter 6) are simple to obtain, but substantially more time consuming than transient-response information. This is especially so if the time constants of the process are large, as often occurs in chemical processing industries. As with the transient-response data, it is important to have a good signal-to-noise ratio, so obtaining frequency-response data can be very expensive. On the other hand, as we will see in Chapter 6, frequency-response data are exactly in the right form for frequency-response design methods; so once the data have been obtained, the control design can proceed immediately.

Stochastic steady-state

Normal operating records from a natural stochastic environment at first appear to be an attractive basis for modeling systems, since such records are by definition nondisruptive and inexpensive to obtain. Unfortunately, the quality of such data is inconsistent, tending to be worse just when the control is best, because then the upsets are minimal and the signals are smooth. At such times, some or even most of the system dynamics are hardly excited. Because they contribute little to the system output, they will not be found in the model constructed to explain the signals. The result is a model that represents only part of the system and is sometimes unsuitable for control. In some instances, as occurs when trying to model the dynamics of the electroencephalogram (brain waves) of a sleeping or anesthetized person to locate the frequency and intensity of alpha waves, normal records are the only possibility. Usually they are the last choice for control purposes.

Pseudorandom noise (PRBS)

Finally, the pseudorandom signals that can be constructed using digital logic have much appeal. Especially interesting for model making is the pseudorandom binary signal (PRBS). The PRBS takes on the value $+A$ or $-A$ according to the output (1 or 0) of a feedback shift register. The feedback to the register is a binary sum of various states of the register that have been selected to make the output period (which must repeat itself in finite time) as long as possible. For example, with a register of 20 bits, $2^{20} - 1$ (over a million) steps are produced before the pattern repeats. Analysis beyond the scope of this text has revealed that the resulting signal is almost like a broadband random signal. Yet this signal is entirely under the control of the engineer who can set the level (A) and the length (bits in the register) of the signal. The data obtained from tests with a PRBS must be analyzed by computer and both special-purpose hardware and programs for general-purpose computers have been developed to perform this analysis.

[1] Ziegler and Nichols (1943), building on the earlier work of Callender et al. (1936), use the step response directly in designing the controls for certain classes of processes. See Chapter 4 for details.

W3.7.3 Models from Transient-Response Data

To obtain a model from transient data, we assume a step response is available. If the transient is a simple combination of elementary transients, then a reasonable low-order model can be estimated using hand calculations. For example, consider the step response shown in Fig. W3.4. The response is monotonic and smooth. If we assume it is given by a sum of exponentials, we can write

$$y(t) = y(\infty) + Ae^{-\alpha t} + Be^{-\beta t} + Ce^{-\gamma t} + \cdots . \tag{W3.2}$$

Subtracting off the final value and assuming that $-\alpha$ is the slowest pole, we write

$$\begin{aligned} y - y(\infty) &\cong Ae^{-\alpha t}, \\ \log_{10}[y - y(\infty)] &\cong \log_{10} A - \alpha t \log_{10} e, \\ &\cong \log_{10} A - 0.4343\alpha t. \end{aligned} \tag{W3.3}$$

This is the equation of a line whose slope determines α and intercept determines A. If we fit a line to the plot of $\log_{10}[y - y(\infty)]$ (or $\log_{10}[y(\infty) - y]$ if A is negative), then we can estimate A and α. Once these are estimated, we plot $y - [y(\infty) + Ae^{-\alpha t}]$, which as a curve approximates $Be^{-\beta t}$ and on the log plot is equivalent to $\log_{10} B - 0.4345\beta t$. We repeat the process, each time removing the slowest remaining term, until the data stop is accurate. Then we plot the final model step response and compare it with data so we can assess the quality of the computed model. It is possible to get a good fit to the step response and yet be far off from the true time constants (poles) of the system. However, the method gives a good approximation for control of processes whose step responses look like Fig. W3.4.

EXAMPLE W3.5 *Determining the Model from Time-Response Data*

Find the transfer function that generates the data given in Table W3.1 and plotted in Fig. W3.5.

Solution. Table W3.1 shows, and Fig. W3.5 implies, that the final value of the data is $y(\infty) = 1$. We know that A is negative because $y(\infty)$ is greater than $y(t)$. Therefore, the first step in the process is to plot $\log_{10}[y(\infty) - y]$, which is shown in Fig. W3.6. From the line (fitted by eye), the values are

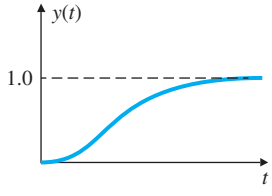


Figure W3.4
A step response characteristic of many chemical processes

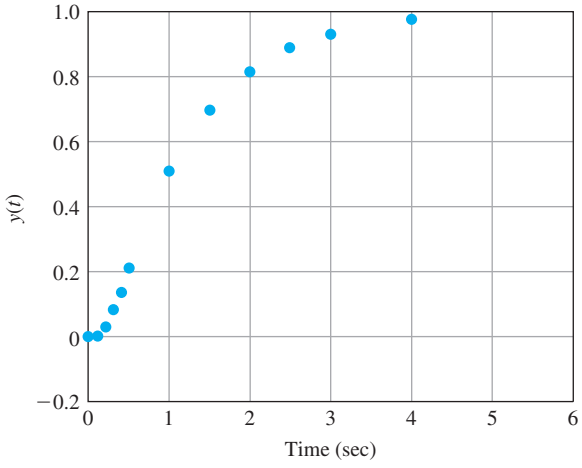
TABLE W3.1

Step Response Data

t	*y(t)*	*t*	*y(t)*
0.1	0.000	1.0	0.510
0.1	0.005	1.5	0.700
0.2	0.034	2.0	0.817
0.3	0.085	2.5	0.890
0.4	0.140	3.0	0.932
0.5	0.215	4.0	0.975
		∞	1.000

Based on Sinha, N. K. and B. Kuszta, Modeling and Identification of Dynamic Systems. NewYork: Van Nostrand, 1983.

Figure W3.5
Step response data in Table W3.1



$$\log_{10}|A| = 0.125,$$

$$0.4343\alpha = \frac{1.602 - 1.167}{\Delta t} = \frac{0.435}{1} \Rightarrow \alpha \cong 1.$$

Thus

$$A = -1.33,$$

$$\alpha = 1.0.$$

If we now subtract $1 + Ae^{\alpha t}$ from the data and plot the log of the result, we find the plot of Fig. W3.7. Here we estimate

$$\log_{10} B = -0.48,$$

$$0.4343\beta = \frac{-0.48 - (-1.7)}{0.5} = 2.5,$$

Figure W3.6
$\log_{10}[y(\infty) - y]$ versus t

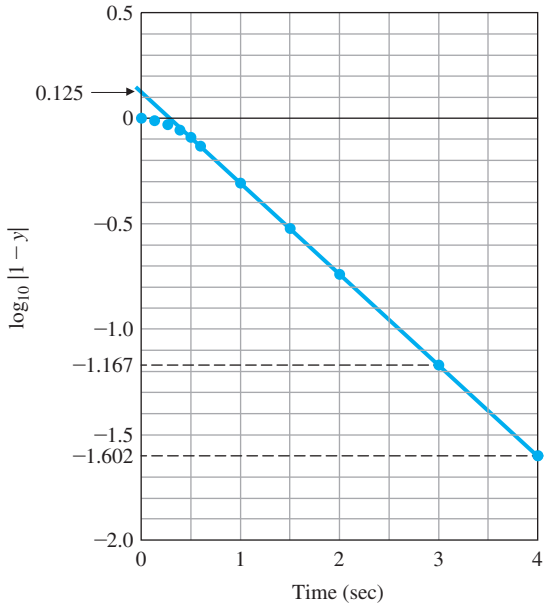
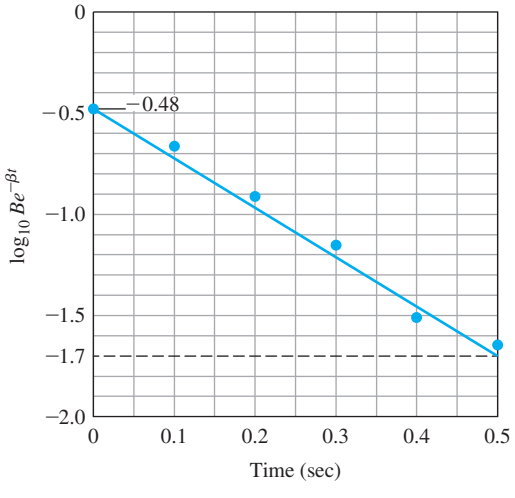


Figure W3.7
$\log_{10}[y - (1 + Ae^{-\alpha t})]$ versus t



$$\beta \cong 5.8,$$

$$B = 0.33.$$

Combining these results, we arrive at the y estimate

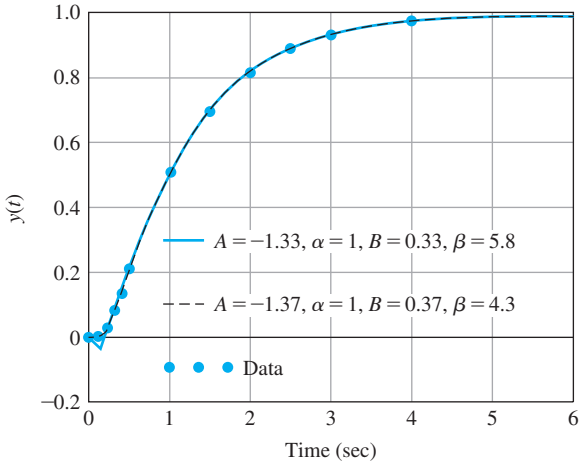
$$\hat{y}(t) \cong 1 - 1.33e^{-t} + 0.33e^{-5.8t}. \tag{W3.4}$$

Equation (W3.4) is plotted as the colored line in Fig. W3.8 and shows a reasonable fit to the data, although some error is noticeable near $t = 0$.

From $\hat{y}(t)$, we compute

$$\hat{Y}(s) = \frac{1}{s} - \frac{1.33}{s+1} + \frac{0.33}{s+5.8}$$

Figure W3.8
Model fits to the experimental data



$$= \frac{(s+1)(s+5.8) - 1.33s(s+5.8) + 0.33s(s+1)}{s(s+1)(s+5.8)}$$

$$= \frac{-0.58s + 5.8}{s(s+1)(s+5.8)}.$$

The resulting transfer function is

$$G(s) = \frac{-0.58(s-10)}{(s+1)(s+5.8)}.$$

Notice this method has given us a system with a zero in the RHP, even though the data showed no values of y that were negative. Very small differences in the estimated value for A, all of which approximately fit the data, can cause values of β to range from 4 to 6. This illustrates the sensitivity of pole locations to the quality of the data and emphasizes the need for a good signal-to-noise ratio.

By using a computer to perform the plotting, we are better able to iterate the four parameters to achieve the best overall fit. The data presentation in Figs. W3.6 and W3.7 can be obtained directly by using a semilog plot. This eliminates having to calculate $\log_{10}$ and the exponential expression to find the values of the parameters. The equations of the lines to be fit to the data are $y(t) = Ae^{\alpha t}$ and $y(t) = Be^{\beta t}$, which are straight lines on a semilog plot. The parameters A and α, or B and β, are iteratively selected so the straight line comes as close as possible to passing through the data. This process produces the improved fit shown by the dashed black line in Fig. W3.8. The revised parameters, $A = -1.37$, $B = 0.37$, and $\beta = 4.3$ result in the transfer function

$$G(s) = \frac{-0.22s + 4.3}{(s+1)(s+4.3)}.$$

The RHP zero is still present, but it is now located at $s \cong +20$ and has no noticeable effect on the time response.

This set of data was fitted quite well by a second-order model. In many cases, a higher-order model is required to explain the data and the modes may not be as well separated.

If the transient response has oscillatory modes, then these can sometimes be estimated by comparing them with the standard plots of Fig. 3.18. The period will give the frequency ω_d, and the decay from one period to the next will afford an estimate of the damping ratio. If the response has a mixture of modes not well separated in frequency, then more sophisticated methods need to be used. One such is **least-squares system identification**, in which a numerical optimization routine selects the best combination of system parameters so as to minimize the fit error. The fit error is defined to be a scalar **cost function**

Least-squares system identification

$$J = \sum_i (y_{data} - y_{model})^2, \; i = 1, 2, 3, \cdots, \text{ for each data point,}$$

so fit errors at all data points are taken into account in determining the best value for the system parameters.

W3.7.3.1 Models from Other Data

As mentioned early in Section 3.1.2, we can also generate a model using frequency-response data, which are obtained by exciting the system with a set of sinusoids and plotting $G(j\omega)$. In Chapter 6, we show how such plots can be used directly for design. Alternatively, we can use the frequency response to estimate the poles and zeros of a transfer function using straight-line asymptotes on a logarithmic plot.

The construction of dynamic models from normal stochastic operating records or from the response to a PRBS can be based either on the concept of cross-correlation or on the least-squares fit of a discrete equivalent model, both topics in the field of **system identification**. They require substantial presentation and background that are beyond the scope of this text. An introduction to system identification can be found in Chapter 8 of Franklin et al. (1998), and a comprehensive treatment is given in Ljüng (1999). Based largely on the work of Professor Ljüng, the Matlab Toolbox on Identification provides substantial software to perform system identification and to verify the quality of the proposed models.

W3.7.4 Obtaining a Pole-Zero Model from Frequency-Response Data

As we pointed out earlier, it is relatively easy to obtain the frequency-response of a system experimentally. Sometimes it is desirable to obtain an approximate model, in terms of a transfer function, directly from the frequency response. The derivation of such a model can be done

to various degrees of accuracy. The method described in this section is usually adequate and is widely used in practice.

There are two ways to obtain a model from frequency-response data. In the first case, we can introduce a sinusoidal input, measure the gain (logarithm of the amplitude ratio of output to input) and the phase difference between output and input, and accept the curves plotted from this data as the model. Using the methods given in previous sections, we can derive the design directly from this information. In the second case, we wish to use the frequency data to verify a mathematical model obtained by other means. To do so, we need to extract an approximate transfer function from the plots, again by fitting straight lines to the data, estimating break points (that is, finding the poles and zeros), and using Fig. 6.3 to estimate the damping ratios of complex factors from the frequency overshoot. The next example illustrates the second case.

EXAMPLE W3.6 *Transfer Function from Measured Frequency Response*

Determine a transfer function from the frequency response plotted in Fig. W3.9, where frequency f is plotted in hertz.

Solution. Drawing an asymptote to the final slope of -2 (or -40 db per decade), we assume a break point at the frequency where the phase is $-90°$. This occurs at $f_1 \cong 1.66$ Hz ($\omega_1 = 2\pi f_1 = 10.4$ rad/sec). We need to know the damping ratio in order to subtract out this second-order pole. For this, the phase curve may be of more help. Since the phase around the break-point frequency is symmetric, we draw a line at the slope of the phase curve at f_1 to find that the phase asymptote intersects the 0° line at $f_0 \cong 0.71$ Hz (or 4.46 rad/sec). This corresponds to $f_1/f_0 \cong 2.34$, which in time corresponds to $\zeta \cong 0.5$, as seen on the normalized response curves in Fig. 6.3b. The magnitude curve with the second-order factor taken out shows an asymptotic amplitude gain of about 6.0 db, or a factor of $10^{6.0/20} = 2.0$. As this is a gain rise, it occurs because of a lead compensation of the form

$$\frac{s/a + 1}{s/b + 1},$$

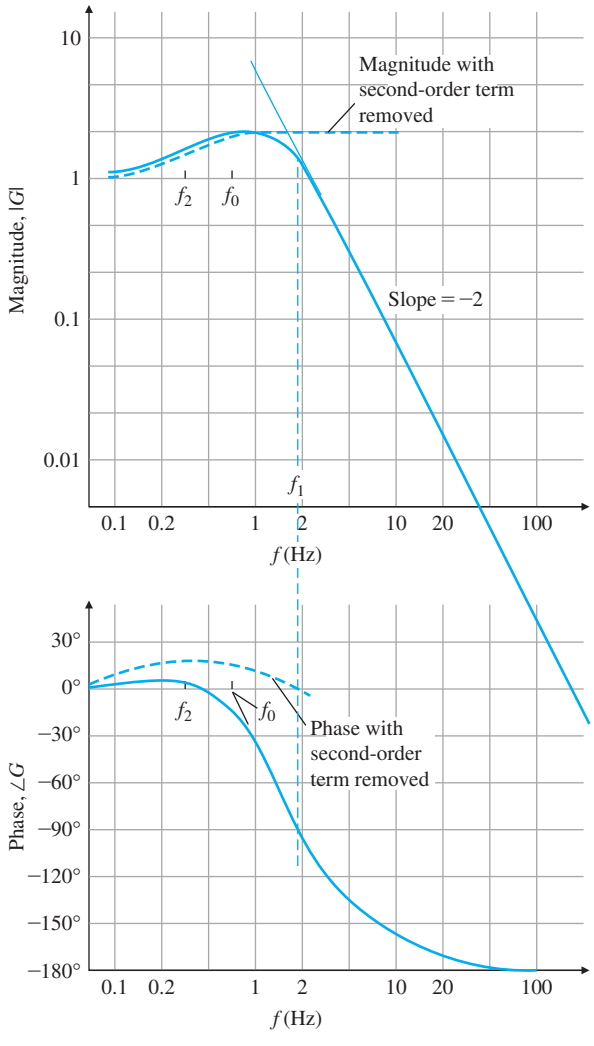
where $b/a = 2.0$. If we remove the second-order terms in the phase curve, we obtain a phase curve with a maximum phase of about 20°, which also corresponds to a frequency separation of about 2. To locate the center of the lead compensation, we must estimate the point of maximum phase based on the lead term alone, which occurs at the geometric mean of the two break-point frequencies. The lead center seems to occur at $f_2 \cong 0.3$ Hz (or $\omega_2 = 1.88$ rad/sec).

Thus, we have the relations

$$ab(1.88)^2 = 3.55,$$

$$\frac{b}{a} = 2,$$

Figure W3.9
Experimental frequency response



from which we can solve

$$2a^2 = 3.55,$$
$$a = 1.33,$$
$$b = 2.66.$$

Model from measured response

Our final model is given by

$$\hat{G}(s) = \frac{(s/1.33) + 1}{[(s/2.66) + 1][(s/10.4)^2 + (s/10.4) + 1]}. \tag{W3.5}$$

The actual data were plotted from

$$G(s) = \frac{(s/2) + 1}{[(s/4) + 1][(s/10)^2 + (s/10) + 1]}.$$

As can be seen, we found the second-order term quite easily, but the location of the lead compensation is off in center frequency by a factor of $4/2.66 \cong 1.5$. However, the subtraction of the second-order term from the composite curve was not done with great accuracy, rather, by reading the curves. Again, as with the transient response, we conclude that by a bit of approximate plotting we can obtain a crude model (usually within a factor of 1.4 (± 3 db) in amplitude and $\pm 10°$ in phase) that can be used for control design.

Refinements on these techniques with computer aids are rather obvious, and an interactive program for removing standard first- and second-order terms and accurately plotting the residual function would greatly improve the speed and accuracy of the process. It is also common to have computer tools that can find the parameters of an assumed model structure by minimizing the sum of squares of the difference between the model's frequency response and the experimental frequency response.

Further Reading for System Identification:

[1] L. Ljung, *Perspectives on System Identification, Annual Reviews in Control*, 34, pp. 1–12, Elsevier, 2010.

[2] L. Ljung, *System Identification: Theory for the User*, 2nd Ed., Prentice-Hall, 1999.

[3] G. F. Franklin, J. D. Powell, M. L. Workman, *Digital Control of Dynamic Systems*, 3rd Ed. Ellis-Kagle Press, 1998.

[4] M. B. Tischler and R. K. Remple, *Aircraft and Rotorcraft System Identification: Engineering Methods with Flight-Test Examples,* AIAA, 2006.

[5] R. Pintelon and J. Schoukens, *System Identification: A Frequency Domain Approach*, 2nd ed., Wiley-IEEE Press, 2012.

[6] System Identification Toolbox, The Mathworks.